

ABSTRACT

Title of Dissertation: Learning from Less Data:
 Perception and Synthesis

Divya Kothandaraman
Doctor of Philosophy, 2024

Dissertation Directed by: Prof. Dinesh Manocha
 Department of Computer Science

Machine learning techniques have transformed various fields, particularly in computer vision. However, they typically require vast amounts of labeled data for training, which can be costly and impractical. This dependency on data highlights the importance of research into data efficiency. In this thesis, we present advancements in data-efficient deep learning within the contexts of visual perception and visual generation tasks.

In the first half of the thesis, we focus on data efficiency in visual perception. Specifically, we tackle the challenge of semantic segmentation in autonomous driving, assuming limited access to labeled data in both the target and related domains. we propose self-supervised learning solutions to enhance segmentation performance in unstructured and adverse weather conditions, ultimately extending to a more generalized approach that is on par with methods using immense amounts of labeled data, achieving up to 30% improvements over prior work. Next, we address data efficiency for autonomous aerial vehicles, specifically in video action recognition. Here, we integrate concepts from signal processing into neural networks, achieving both data and

computational efficiency. Additionally, we propose differentiable learning methods for these representations, resulting in 8-38% improvements over previous work.

In the second half of the thesis, we will delve into data efficiency in visual generation. First, we focus on efficient generation of aerial-view images, utilizing pretrained models to create aerial perspectives from input scenes in a zero-shot manner. By incorporating techniques from classical computer vision and information theory, our work enables the generation of aerial images from complex, real-world inputs without requiring any 3D or paired data during training or testing. The approaches are on par with concurrent methods that use vast amounts of 3D data for training. Next, we focus on zero-shot personalized image and video generation, aiming to create content based on custom concepts. we propose methods that leverage prompting to generate images and videos at the intersection of various manifolds corresponding to these concepts and pretrained models, with applications in subject-driven action transfer and multi-concept video customization. These solutions are among the first in this area, showing significant improvements over baselines and related work. These approaches are also data and compute efficient, relying solely on pretrained models without the need for additional training data. Finally, we introduce a fundamental prompting solution inspired by techniques from finance and economics, demonstrating how insights from different fields can effectively address similar mathematical challenges.

Learning from Less Data: Perception and Synthesis

by

Divya Kothandaraman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:

Professor Dinesh Manocha, Chair/Advisor
Professor Maria Cameron, Dean's Representative
Professor Ming C Lin
Professor Jia-Bin Huang
Professor Tianyi Zhou

© Copyright by
Divya Kothandaraman
2024

Acknowledgments

This dissertation represents the collective efforts of many individuals and institutions over the past 4-5 years. I wish to express my heartfelt gratitude to all my collaborators, friends, family, administrative staff, and everyone else who has contributed to its successful completion.

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dinesh Manocha, for his unwavering support throughout my journey. His empowering and enabling guidance has been instrumental in helping me grow to my full potential and enhancing my capacity to work excellently. I am truly thankful for that. My first interaction with him was even before I had formally accepted my PhD offer, and his belief in me and encouragement were incredibly comforting. He has not just been a great advisor and mentor, but also a great friend. His approachable nature and willingness to listen have made this challenging journey much more manageable. Whether it was discussing research challenges, offering career advice, or simply sharing a moment of humor, his presence has always been reassuring and inspiring. I feel fortunate to have had the opportunity to work with someone who genuinely cares about my personal and professional growth.

Prof. Ming Lin has been an invaluable collaborator and mentor. Our research discussions have been incredibly useful and stimulating, significantly shaping my research and contributing to the success of my projects. Beyond her professional guidance, she has also been a great friend, always available for a chat about anything. Her unwavering support has made me feel very comfortable and confident throughout this journey.

I would like to extend my heartfelt thanks to my committee members, Prof. Dinesh

Manocha, Prof. Ming Lin, Prof. Tianyi Zhou, Prof. Jia-Bin Huang, and Prof. Maria Cameron for their invaluable feedback. Their insightful comments and suggestions have contributed to the improvement and refinement of this dissertation.

My internships at Adobe Research and Google DeepMind were instrumental in helping me reach my fullest potential. They provided much-needed industry exposure, honed my skills, and enhanced my understanding of practical applications and wide-ranging possibilities. These experiences also deepened my sense of curiosity and purpose, emphasizing how my research can add value from an application perspective. I am deeply grateful to Balaji Srinivasan and Mohammad Babaeizadeh for the incredible opportunities at Adobe and Google DeepMind.

My PhD work would not have been possible without the invaluable contributions of my collaborators: Tianyi Zhou, Rohan Chandra, Tianrui Guan, Xijun Wang, Ruiqi Xian, James Mullen, Mohammad Babaeizadeh, Ruben Villegas, Kihyuk Sohn, Taewon Kang, Kuldeep Kulkarni, Sumit Shekhar, and Pooja Guhan. I would like to extend my heartfelt thanks to them for the engaging discussions and enjoyable research projects we shared.

I truly enjoyed my time on the UMD campus; it has been incredibly inspiring and welcoming. I am a proud Terp and immensely grateful for the opportunity. I would also like to extend my gratitude to the funding sources, including ARL, ArtIAMAS, and UMD CS, for their generous support. I would like to extend my gratitude to Tom Hurst, Migo Gui, and the staff and faculty of the Computer Science Department at UMD and UMIACS for their exceptional support and responsiveness in addressing any bureaucratic or compute-related questions.

This thesis would not have been possible without the unwavering support of my friends, including Uttaran Bhattacharya, Pooja Guhan, Niall Williams, Xijun Wang, Vishnu Dorbala, Geonsun Lee, Ruiqi Xian, Laura Zheng, James Mullen, Trisha Mittal, Rohan Chandra, Senthil

Arul, Adarsh Jagan, Kasun Weerakoon, Mohamed Elnoor, Tianrui Guan, Archana Swaminathan, Namitha Padmanabhan, Sharath Girish, Saksham Suri, Pulkit Kumar, Shishira Maiya, Soumik Mukhopadhyay, Mara Levy, Matt Gwilliam, Kamal Gupta, Neha Kalibhat, Gowthami Somepalli, Ishani Mondal, Navita Goyal, Arpit Bansal, Siddharth Singh, Khushboo Khandelwal, Tripti Rajput, Abhilasha Sancheti, Anand Murali, Ishan Bansal, Shashank Gupta, Kartikey Pentapati, Anshul Shah, Tarun Kalluri, Sahana Ramnath, Aravind Shankaranarayanan, and Akhil Bhimraju.

I am deeply grateful for their constant encouragement and friendship. Their support, both academic and personal, has been a cornerstone of my journey. Each moment spent with them has been invaluable. They have celebrated my successes and provided comfort during challenging times. These cherished memories and shared experiences have not only made my PhD journey enjoyable but also incredibly fulfilling. My sincere apologies to anyone I may have inadvertently missed.

I would like to extend my deepest gratitude to my family for their unwavering support and love. Their encouragement and belief in my abilities have been a constant source of strength throughout this journey. I am also deeply thankful to the Almighty for guiding me at every step and helping me overcome obstacles. I am very grateful for His blessings and deeply thankful for the strength and resilience it has given me throughout this journey.

Table of Contents

Acknowledgements	ii
Table of Contents	v
List of Tables	ix
List of Figures	xiii
I Introduction and Overview	1
II Learning from Little to No Supervised Data: Domain Adaptation for Autonomous Driving	8
Chapter 1: BoMuDA: Boundless Multi-Source Domain Adaptation for Semantic Segmentation in Unstructured Traffic Environments	11
1.1 Prior Work	14
1.2 BoMuDANet	16
1.3 A Self-Training Algorithm for Multi-Source DA	17
1.4 Boundless Domain Adaptation	21
1.5 Experiments and Results	23
1.6 Results	24
1.7 Ablation Studies and Additional Experiments	29
1.8 User Study	29
1.9 Conclusion, limitations and future work	32
Chapter 2: SS-SFDA: Self-Supervised Source-Free Domain Adaptation for Road Segmentation in Hazardous Weather Conditions	35
2.0.1 Main Contributions	37
2.1 Related work	38
2.2 Method	40
2.2.1 Pre-Training Baselines Using Self-Attention	41
2.2.2 SS-SFDA	42
2.3 Few-Image Fine-Tuning via Model Distillation	45
2.4 Experiments and Results	46
2.4.1 Analysing Pre-Trained Self-Attention-based AutoEncoder	47

2.4.2	Results on Synthetic Datasets: Rain and Fog	47
2.4.3	Results on Real datasets: Foggy Zurich and Dark Zurich	51
2.4.4	Heterogeneous Real Datasets: Raincouver and Berkeley Deep Drive	55
2.4.5	Training Time and Convergence	56
2.4.6	Comparisons with Prior Work	56
2.5	Conclusions, Limitations and Future Work	57
Chapter 3:	SALAD: Source-free Active Label-Agnostic Domain Adaptation	58
3.1	Related work	61
3.2	Method	62
3.2.1	SALAD Description	63
3.2.2	Guided Attention Transfer Network (GATN)	64
3.2.3	Training Losses	66
3.2.4	Active Learning Strategy H_{AL}	68
3.2.5	SALAD Training Overview	69
3.3	Experiments and Results	70
3.3.1	Image Classification	71
3.3.2	Autonomous Driving: Synthetic to Real Segmentation on CityScapes	75
3.3.3	Document Layout Detection: DSSE	77
3.4	Conclusions, Limitations and Future Work	78

III Learning from Little Supervised Data: Aerial Video Action Recognition **79**

Chapter 4:	FAR: Fourier Aerial Video Recognition	82
4.0.1	Main contributions	84
4.1	Related Work	85
4.2	Method	87
4.2.1	Fourier Object Disentanglement	88
4.2.2	Space-Time Fourier Attention	90
4.2.3	Mathematical Analysis	93
4.3	FAR: Activity Recognition in UAVs	94
4.3.1	Randomly Initialized Uniform Sampling	95
4.4	Experiments and Results	95
4.4.1	Datasets	95
4.4.2	Implementation Details	96
4.4.3	Main Results: UAV Human RGB	97
4.4.4	Results: More UAV Datasets	102
4.5	Conclusions, Limitations and Future Work	102
Chapter 5:	DiffFAR: Differentiable Fourier Aerial Video Recognition	104
5.0.1	Main Contributions	105
5.1	Related Work	107
5.2	DiffFAR	108

5.2.1	Background: Differentiability	110
5.2.2	Differentiable Static-Dynamic Frequency Mask	111
5.2.3	Differentiable Disentanglement	113
5.2.4	Frame Sampling	114
5.3	Experiments	115
5.3.1	Network and Training Details	115
5.3.2	Results: UAV Human	117
5.3.3	Results: NEC Drone	120
5.4	Conclusion, Limitations and Future Work	121

IV Generating Aerial-view Images using a Pretrained Model and an Exemplar Front-view Image 122

Chapter 6:	Aerial Diffusion: Text Guided Ground-to-Aerial View Translation from a Single Image using Diffusion Models	125
6.1	Related work	128
6.2	Aerial Diffusion	131
6.2.1	Postulates	131
6.2.2	Method	134
6.3	Experiments and Results	139
6.3.1	Implementation details	139
6.3.2	Qualitative evaluation	140
6.3.3	Ablations and comparisons	142
6.4	Conclusion, Limitations and Future Work	145
Chapter 7:	HawkI: Homography and Mutual Information Guidance for 3D-free Single Image to Aerial View	146
7.1	Related work	149
7.2	Method	151
7.2.1	Test-time optimization	152
7.2.2	Mutual Information Guided Inference	154
7.3	Experiments and Results	158
7.3.1	Comparisons against text + exemplar image based methods	160
7.3.2	Comparisons against 3D based novel view synthesis (NVS) methods	162
7.3.3	Downstream application.	163
7.3.4	Ground-truth comparison	163
7.3.5	Ablation studies	164
7.3.6	Comparison with other metrics for guidance.	165
7.3.7	Text controlled view synthesis: Other views	166
7.3.8	3D-free HawkI + 3D priors?	166
7.4	Conclusions, Limitations and Future Work	166

V Generating Personalized Images using Pretrained Text to Image Diffusion Models and Exemplar Control Images **178**

Chapter 8:	ImPoster: Text and Frequency Guidance for Subject Driven Action Personalization using Diffusion Models	182
8.1	Related work	186
8.2	Method	188
8.2.1	Training	189
8.2.2	Inference	189
8.3	Experiments and Results	192
8.4	Conclusions	197
8.5	Limitations and Future Work	201
Chapter 9:	Text Prompting for Multi-Concept Video Customization by Autoregressive Generation	204
9.1	Related work	207
9.1.1	Text to Video Models	207
9.1.2	Personalization of generative models	208
9.2	Method	209
9.2.1	Problem Formulation	209
9.2.2	Proposed Method	209
9.3	Experiments	213
9.3.1	Setup	213
9.3.2	Qualitative Results	218
9.3.3	Quantitative results	224
9.4	Limitations and Future Work	226
Chapter 10:	Prompt Mixing in Diffusion Models using the Black Scholes Algorithm	228
10.1	Related work	231
10.2	Background	232
10.2.1	The Black Scholes Model	232
10.2.2	Relation to Diffusion Models	233
10.3	Prompt Mixing using Black Scholes Model	234
10.3.1	Problem Formulation	234
10.3.2	Method	236
10.4	Experiments and Results	237
10.4.1	Experimental Settings	239
10.4.2	Analysis and Comparisons	241
10.5	Conclusions, Limitations, and Future Work	243

VI Conclusions and Future Directions **249**

List of Tables

1.1	Main Results: We evaluate BoMuDANet on IDD using CityScapes (CS), Berkeley Deep Drive (BDD), SynScapes (SC), and GTA as sources. Higher mIoU and mAcc indicates direction of better performance. Bold indicates best while blue indicates second-best. Experiments I and II differ with respect to the sources, while experiment III differs with respect to the baseline used. Conclusion: Our unsupervised multi-source self-training algorithm outperforms the single-source baselines by 3.3% – 54.15%.	22
1.2	Pseudo labeling strategy for boundless DA: We show that pseudo labeling can provide semantic information about categories that do not belong to any source domain, for instance, auto-rickshaws (Auto) found in the IDD dataset (in bold). Moreover, pseudo labeling is simple in that the generated proxy labels do no need to be re-trained as there is no marked improvement in mIoU. We also show that thresholding outperforms KL divergence via an ablation study.	25
1.3	Comparison with SOTA: We compare with the SOTA in both unstructured (IDD) as well as structured (CS) traffic. Higher ($\uparrow$) mIoU and mAcc indicates direction of better performance. Bold indicates best while blue indicates second-best. mIoU(S) and mIoU(P) denote the performance on shared/known and private/unknown classes, respectively. Conclusion: Our model is SOTA on IDD by 5.17% – 42.9% and on CS in the multi-source setting by 12.70% – 90.13%, with a reduction in model size by upto to $5.2\times$	27
2.1	List of datasets: The second column categorizes the datasets as synthetic or real and the third column describes the images contained in the dataset.	45
2.2	Effect of self-attention: We show for various backbone architectures that self-attention improves the accuracy. DeepLabv2 (E-1) and DeepLabv2 (E-2) denote the removal of 1 and 2 layers from DeepLab respectively. We select DRN-D-38 with self-attention (bolded) as the baseline for all further experiments.	47
2.3	Benchmarking and comparing the self-attention based pre-trained model: We train our self-attention model on various datasets in a supervised manner. These supervised numbers help us conduct a relative study of the performance of SS-SFDA, which is unsupervised in the target domain. Additionally, we observe that our pre-trained model is comparable to the state-of-the-art (Experiment IV). Best results are in bold fonts, second best results are in blue).	48

2.4	Results on synthetic rain and synthetic fog [1]: We evaluate our model under three settings - A: (Baseline) Testing the clear weather CityScapes model, B: SS-SFDA without curriculum learning, C: SS-SFDA. We show that the CityScapes model generalizes well to light synthetic rain and fog, and its performance degrades as the intensity of rain and fog increases, which is restored by SS-SFDA. Comparison to supervised results: On synthetic rain, SS-SFDA achieves 90.07%-90.8% and 92.53%-97.6% of supervised mIoU and recall, respectively. On synthetic fog, SS-SFDA achieves 90.8%-96.78% and 91.12%-98.82% of supervised mIoU and recall respectively.	49
2.5	Curriculum learning improves performance by improving the accuracy of pseudo labels: The first and second columns correspond to the datasets our model is trained and tested, respectively. Evaluating on 30m fog reveals that a model finetuned to 75m fog performs better than the counterpart 150m fog model, which in turn performs better than that the CityScapes model. Similar conclusions can be drawn for other intensities of rain/fog.	52
2.6	Results on the real fog dataset, Foggy Zurich. We observe that SS-SFDA improves mIoU by 101.84% over the corresponding baseline. We notice that training the model on light fog followed by medium fog imparts the model with improvised self-training abilities, the mIoU improves by 22.3%. Additionally, the ablations on each minibatch indicate that the two step training procedure in SS-SFDA is indeed helpful.	52
2.10	Comparisons against the state-of-the-art: We adopt the state-of-the-art methods in self-training, domain adaptation, and SFDA classification to road segmentation, and present the results on Rain 100mm. We pick the best performing source-free model and conduct experiments on BDD. We observe an improvement of 10.26%.	53
2.7	Results on the night driving dataset, Dark Zurich. We observe that SS-SFDA improves mIoU by 31.47. We observe that our two-stage training routine within the curriculum learning algorithm helps the network perform well.	54
2.8	Results on Raincouver: SS-SFDA achieves 72.9% of supervised mIoU. We demonstrate the effectiveness of FewIm-FT over varying values of k in Experiment II. $k = 10$ results in an improvement of 21.8% over SS-SFDA, which is 88.82% of supervised mIoU. Additionally, we show the benefit of model distillation in Experiment III.	54
2.9	Results on BDD: SS-SFDA improves baseline mIoU by 7.12%. Experiments on FewIm-FT demonstrate an improvement of 10.30% over SS-SFDA for $k=10$, which is 93.11% of supervised mIoU.	54
3.1	Problem Settings: We highlight various domain adaptation settings. Src. Data, Src. Model, Lab. Tar., and Un. Tar. refer to abundant labeled source data, Source Model, Scarce Labeled Target Data and Unlabeled Abundant Target Data, respectively.	61
3.2	Results on digits classification in the shared label space setting	71
3.3	Comparing SALAD with prior art on SFDA on digits datasets. Budget reflects the percentage of total number of target samples used for active learning. We demonstrate that our SF-ADA approach outperforms prior art on SFDA by a large margin using a very small proportion of annotated target samples.	72

3.4	Results on adaptation from SVHN to MNIST, for label space shift. We consider two scenarios: case 1 - source data does not contain digits ‘3’ and ‘9’, case 2 - source data does not contain digits ‘7’, ‘5’, ‘4’, ‘1’. SALAD, with H_{AL} and GATN, achieves 99.4% and 99.8% of the accuracy in the scenario with no label shift. The accuracy at a budget of 300 images, without label shift, is 97.16%	73
3.5	Ablation Experiments for variations in transfer loss parameters ($\lambda_{Tr,L}$, $\lambda_{Tr,UL}$) and active learning strategies (A_G , A_E) from SVHN to MNIST.	73
3.6	GTA5 to CityScapes Adaptation: We show that SALAD imparts a relative improvement of 31.5%, 46.03%, 52.5% and 62.1% over the baseline source model with budgets of 50, 100, 200, and 500 images, with $3 \times -25 \times$ improvement on specific classes like “Bike”, “Train”, “MBike”, “Sidewalk”, etc.	75
3.7	Comparisons with the state-of-the-art SFDA approaches for adaptation from GTA to CityScapes. We show that SALAD achieves state-of-the-art mIoU with a small budget of 50 images, the mIoU obtained by naive finetuning [2] on 50 images is 39.5.	76
3.8	Ablation experiments for adaptation from GTA5 to CityScapes. In table (a), we study ablations on transfer loss for GATN, Table (b) shows ablations on GATN components at a budget of 50 images. In table (c), we ablate on the active learning heuristic, H_{AL} at a budget of 500 images.	76
3.9	Adaptation for document layout detection from PubLayNet to DSSE: Fine-tuning with SALAD improves performance by 31.3% over fine-tuning without SALAD.	76
4.1	Results on UAV Human RGB. Table (a): FAR can be embedded within any 3D action recognition backbone to achieve state-of-the-art performance. Pretraining with Kinetics boosts performance, and large input sizes work better since FA and FO are designed to capture global, as well as local knowledge. FAR imparts improvements of 2.20%-38.69% over 3D action recognition backbones across training configurations. Table (b) - Ablation experiments: We demonstrate that each component of FAR imparts substantial improvement in top-1 accuracy by upto 8%	97
4.2	Comparisons with state-of-the-art self-attention based transformer methods on UAV Human. We initialize all our models with Kinetics pre-trained weights. We observe higher accuracy and computational benefits (up to 3x) with FAR.	99
4.3	Results on more UAV datasets. We demonstrate that FAR improves the state-of-the-art accuracy by 8.02%-17.61% across popular UAV datasets.	102
5.1	Symbols used in the chapter	109
5.2	We evaluate our method on UAV Human. Across two 3D CNN backbones (I3D, X3D) and datasets, we demonstrate relative improvements of 14.28% - 38.05% in the Top-1 accuracy over the corresponding baselines.	116
5.3	We demonstrate the effectiveness of each component of FD with ablation experiments on UAV Human. All our experiments use RIUS for training as well as testing, and I3D backbone, and a spatial and temporal resolution of 540×960 and 8 frames respectively.	117

5.4	We compare our method against prior methods on UAV Human. We demonstrate state-of-the-art performance by at least 5.72%. Our method outperforms transformer architectures while using far fewer computations (upto 61.8 ×) and network parameters (upto 31.94 ×). Our method imposes minimal memory overhead over the backbone neural network.	117
5.5	We present ablation experiments on the frame sampling method. None of our models use data augmentation.	119
5.6	We present experiments on NEC Drone. We demonstrate improvements of 13% - 22% over prior work.	120
10.1	We follow prior work [3, 4] and analyze using CLIP Scores. The Black Scholes algorithm for prompt mixing in diffusion models achieves superior results, compared to other prompt-mixing techniques, as also evidenced by the qualitative results. .	239

List of Figures

1.1	We present a novel unsupervised deep learning-based approach called BoMuDANet for visual scene understanding in unconstrained and unstructured traffic environments. In this example, we demonstrate the benefits of BoMuDANet on images taken from the challenging IDD dataset. BoMuDANet accurately segments out dirt roads as terrains as well as a building, while preserving its shape. In contrast, the single source baselines (GTA/CityScapes) do not identify dirt and unstructured roads well, misclassify parts of sky as building, and fail to capture the shape of the unstructured building. BoMuDANet benefits from its ability to selectively distil information from various sources by iterative self-training, in addition to exploiting a chosen best source via domain adaptation.	12
1.2	Extension to SOTA in domain adaptive semantic segmentation. Our approach is the first method to <i>simultaneously</i> perform unsupervised multi-source boundless DA segmentation and can handle unstructured traffic environments.	15
1.3	Overview of BoMuDANet: The input consists of N sources $(s_1, s_2, \dots, s_N)$, from which the best-source is selected by the self-training algorithm. The self-training algorithm proceeds in an unsupervised fashion to generate the final set of pseudo-labels that are used to recognize out-of-distribution objects. The final output consists of the segmentation map of an image in the target domain.	15
1.4	In this example, we demonstrate the benefits of BoMuDANet on an image from the IDD dataset, depicting a mixture of challenging driving conditions. The top image in the second column shows the prediction of our model, and the bottom image shows the ground-truth. We observe that BoMuDANet accurately segments the autorickshaws (open-set object, a new type of vehicle - the third column zooms into the region containing the autorickshaw in the prediction and ground-truth), in addition to handling dense traffic and dirt roads.	24
1.5	Visual Results: BoMuDANet works well in various unconstrained environments including unmarked lanes, dirt roads, heavy traffic, and boundless category objects (auto-rickshaws) and results in higher accuracy. Each color represents a different object as shown in the color scheme.	26

1.6	User Study Results: In the first user study, we observe that users are more likely to report our result as the better segmentation compared to state-of-the-art. In the second study, we observe that majority of the users confirm that the IDD traffic scenes are complex and that it would be hard to drive in such environments whereas almost all participants find the driving environments in the CityScapes dataset to be easy.	34
2.1	We highlight the results generated by SS-SFDA on night [5] and fog benchmarks [6], compared to the baseline source model pre-trained on clear weather CityScapes. The purple regions (right) denote the segmented road pixels. The overall accuracy of our self-supervised algorithm in terms of mIoU is (88 – 96%) of supervised methods.	36
2.2	Our Approach: In stage 1, our model is pre-trained on a clear weather source dataset. In stage 2, our model is initialized with the pre-trained model from stage 1 and trained using our self-supervised algorithm, SS-SFDA, on the unlabeled adverse weather dataset. For heterogeneous weather datasets, we perform additional refinement steps based on model distillation (stage 3).	39
2.3	Qualitative results. Our model generates results that closely resemble the ground-truth (GT) compared to the baseline CityScapes pre-trained model. Purple indicates the segmented road region.	53
3.1	We present a generic method, SALAD, for the task of adapting from a pre-trained source network to a target domain with a small budget for acquiring annotation, under a possible shift in label space. SALAD consists of two complementary components: Guided Attention Transfer Network (GATN) and an active learning strategy H_{AL}	63
3.2	Overview of the computations for guided spatial attention and channel attention representations.	65
4.1	Fourier Object Disentanglement (FO) and Space-Time Fourier Attention (FA): FO empowers the network to intrinsically separate out the moving human agent from the background, without the need for any annotated object detection bounding boxes. This enables our network to explicitly focus on the low resolution human agent performing action, and not just learn from background cues. FO inherently characterizes salient and non-salient, and static and dynamic regions of the scene via the amplitudes of the feature maps it computes. FA elegantly exploits the mathematical properties of the Fourier transform to imbibe the properties of self-attention and capture contextual knowledge and long-range space-time dependencies at a much lower computational complexity. . . .	87

4.2	Qualitative results on UAV Human RGB. We show the effect of our Fourier Object Disentanglement (FO) method. In each sample, the images, in order, correspond to a frame from the video, feature representation before disentanglement and the feature representation after disentanglement respectively. Notice the effectiveness of FO in scenes with light noise (Row 1 Image 2, Row 2 Image 3), dim light (Row 1 Image 2), dynamic camera and dynamic background (Row 1 Image 1). Regions of the scene corresponding to moving human actor (or salient dynamic) are amplified most (solid yellow). Static background is completely suppressed (solid purple). Static salient regions are slightly amplified (e.g. lower body of human actor in Row 2 Image 3 - yellow), and dynamic backgrounds are suppressed to a great extent (pale yellow in Row 1 Image 1). We show more results in the supplementary material.	98
4.3	FAR converges much faster than the state-of-the-art action recognition method X3D-M [7]. In the left curve, we show the top-1 train accuracy as a function of the networks' training iterations. In the right figure, we plot the training loss curve. We demonstrate that FAR imparts convergence benefits over prior work, under the same hyperparameter and GPU configurations.	98
5.1	Human activity recognition in videos recorded using dynamic cameras at oblique and aerial angles typically consist of human actors that occupy less than 10% of the spatial resolution of the video. Background motion, noise, and low-resolution activities add to the complexities faced by video recognition systems. Our method achieves upto 41.86% and 80.75% top-1 accuracies on UAV Human and NEC Drone, respectively.	104
5.2	We present a differentiable frequency-based method for aerial video recognition. Our differentiable static-dynamic frequency mask provides a prior for disentangled regions relevant to action recognition. This mask is used to guide the learning of disentangled features within the layers of the neural network using an identity function. Furthermore, we propose a frame sampling strategy that chooses the best frame within each uniform video segment at test time, using the static-dynamic frequency mask and temporal difference.	109
6.1	Given a single ground-view image and the corresponding text description as input, Aerial Diffusion generates corresponding aerial-view image. Our method does not require any supervision from aerial-view data, pairs of ground-aerial view, depth maps, semantic maps, multi-views, etc. It is one of the first approaches to achieve ground-to-aerial view translation in an unsupervised manner.	126
6.2	Results. We apply Aerial Diffusion on diverse images such as animals/ birds, natural scenes, human actions, indoor settings, etc and show that our method is able to generate high-quality high-fidelity aerial images.	129

6.3	Aerial Diffusion. In step 1, we apply a homography transformation to the ground-view image I_S . This creates a gap between I_{Sh} and txt, which in turn reduces the bias of the model towards I_{Sh} in step 2. In step 2, we use the e_{src} to optimize to e_{opt} and finetune the diffusion model to reconstruct I_{Sh} for the e_{opt} . Using the e_{src} to find e_{opt} reduces the bias of the model towards I_{Sh} due to the disparity between I_{Sh} and the e_{src} . In stage 3, we manipulate the text embedding by using an alternating strategy to find the optimal solution in a higher-dimensional non-linear space to generate a high-fidelity aerial image I_T	130
6.4	Results. We apply Aerial Diffusion on diverse images such as animals/ birds, natural scenes, human actions, indoor settings, etc and show that our method is able to generate high-quality high-fidelity aerial images.	133
6.5	Effect of α. Low values of α generate images that are less aerial, high values of α generate low-fidelity images. A trade-off between the viewpoint and fidelity generates high-fidelity aerial images. The transformation, with α , is not smooth, reinforcing Postulate 2.	137
6.6	Based on the text description, Aerial Diffusion can generate aerial views with scene entities slightly different from the ground-view. The hallucination of background and unseen parts of the scene can also be controlled by text.	139
6.7	Ablations and Comparisons. Ablations: we prove the effectiveness of the homography, our alternating strategy over linear interpolation and finetuning with e_{src} instead of e_{tgt} . SOTA Comparisons: IMAGIC [8] (CVPR 2023) is unable to generate aerial views due to high bias towards the input image, domain gap and restricting the solution search to the linear interpolation manifold. Comparisons with other embedding space manipulation strategies that utilize both e_1 and e_2 reveal that our Alternating strategy is better.	140
6.8	Failure cases. A. The identity of the dog and cat is interchanged. B. The fidelity of the background (bed) is lost.	141
7.1	HawkI generates aerial-view images from a text description and an exemplar input image. It builds on a text to 2D image stable diffusion model and does not require any additional 3D or multi-view information at fine-tuning or inference.	147
7.2	Overview. HawkI generates aerial-view images, using a text description and a single image I_S as supervisory signals. It builds on a pretrained text-to-image diffusion model, and does not use any 3D or multi-view information. It performs test-time finetuning to optimize the text embedding and the diffusion model to reconstruct the input image and its inverse perspective mapping in close vicinity. Such a mechanism enables the incorporation of image specific knowledge within the model, while retaining its imaginative capabilities (or variance). At inference, HawkI uses mutual information guidance to maximize the information between the probability distributions of the generated image and I_S , to generate a high-fidelity aerial-view image.	151
7.3	Compared to state-of-the-art text + exemplar image based methods, HawkI is able to generate images that are “more aerial”, while being consistent with the input image. The top three images are from the HawkI-Syn dataset, the bottom three images are from the HawkI-Real dataset.	157

7.4	HawkI achieves the best viewpoint-fidelity trade-off amongst prior work on text + exemplar image based aerial-view synthesis, on various quantitative metrics indicate of text-alignment (for viewpoint and a broad description of the scene) and image alignment (for fidelity w.r.t. input image).	158
7.5	(Left figure.) Ablation experiments show that Inverse Perspective Mapping helps in the generation of images that are aerial, mutual information guidance helps in preserving the contents w.r.t. input image. (Right figure.) We compare with latest related work on novel view synthesis: Zero-1-to-3[9] and Zero123++ [10]. Both methods use the pretrained text-to-2D-image stable diffusion model along with the 800k 3D objects dataset, Objaverse [11], for training. Our method uses just the pretrained text-to-2D-image stable diffusion model to generate better results for the task of aerial view synthesis, guided by text and a single input image.	161
7.6	HawkI can be extended to generate other text-controlled views as well.	165
7.7	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	167
7.8	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	168
7.9	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	169
7.10	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	170
7.11	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	171
7.12	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	172
7.13	We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.	173

7.14	We show a few examples for comparisons with other metrics for diffusion guidance such as L2 distance and Wasserstein distance. Our mutual information guidance method is better at preserving the fidelity w.r.t the input image, as also evidenced by higher SSCD scores. The SSCD score for Wasserstein guidance, L_2 guidance, and mutual information guidance are 0.3181, 0.3224 and 0.3345 respectively, averaged over all images in the HawkI-Real dataset.	174
7.15	We compare with latest related work on novel view synthesis: Zero-1-to-3 and Zero123++ on images from HawkI-Syn. Both of these methods use the pretrained stable diffusion model and the 3D objects dataset, Objaverse with 800k+ 3D objects, for training. Our method uses just the pretrained stable diffusion model for the task of aerial view synthesis from a single image.	174
7.16	We compare with latest related work on novel view synthesis: Zero-1-to-3 and Zero123++ on images from HawkI-Real. Both of these methods use the pretrained stable diffusion model and the 3D objects dataset, Objaverse with 800k+ 3D objects, for training. Our method uses just the pretrained stable diffusion model for the task of aerial view synthesis from a single image. 3D generation methods like Zero123++ are capable of generating different views with high fidelity by using pretrained stable diffusion models to finetune on large-scale 3D objects datasets. However, their generalization capabilities are limited. Our method is able to generate high quality aerial images for the given input images without any 3D data and using just the pretrained text-to-2D image stable diffusion model, however, there is scope for improving the fidelity of the generated aerial image w.r.t the input image. Moreover, our method controls the viewpoint via text and does not provide the provision to quantitatively control the camera angle. Both of these limitations of our method can be alleviated by exploring the combination of pretrained Zero123++ models (or other 3D models) and our method, as a part of future work.	175
7.17	Additional results on extending HawkI to generate other text-controlled views. . .	176
7.18	Can any data augmentation be used in place of Inverse Perspective Mapping (IPM)? One question that arises from the usage of Inverse Perspective Mapping is related to whether it actually provides pseudo weak guidance, in addition to increasing variance (or reducing bias) in the representation space that is being conditioned for aerial view generation. The latter can be achieved with any random data augmentation. To understand this, we use a 45 degrees rotated image in place of the image corresponding to the Inverse Perspective Mapping in the second stage of finetuning the text embedding and the diffusion UNet. We do not use mutual information guidance in any of our experiments. Our finding is that results with models that use Inverse Perspective Mapping are generally better in terms of the viewpoint being aerial, while preserving the fidelity with respect to the input image, than models that use the 45 degree rotated image. Thus, we conclude that rather than using any random data augmentation technique, it is beneficial to use IPM as it is capable of providing pseudo weak guidance to the model for aerial view synthesis. This finding also paves direction for future work on using carefully crafted homography priors for view synthesis corresponding to different camera angles and viewpoints.	177

8.1	<i>Given a single “source” image and a “single” driving image and the corresponding text descriptions, ImPoster generates an image of the source subject performing driving action. We show how ImPoster is able to make a monkey and an alien meditate and play violin.</i>	183
8.2	Given a single source-driving image pair, ImPoster generates an image of the source subject performing the action depicted in the driving image. ImPoster first finetunes the text-to-image diffusion model on the source-driving image pair. At inference, ImPoster begins by first denoising in the direction of the driving image manifold followed by moving towards the manifold corresponding to the desired target image. At every step of the inference, frequency guidance steers the generation of an image with source subject characteristics and driving action.	186
8.3	ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues. Please see the appendix for more results generated using ImPoster and comparisons with Baseline++.	193
8.4	ImPoster is able to successfully transfer the driving motion while retaining the characteristics of the source subject, and achieves a better trade-off between driving action (CLIP/Phase score) and source subject (SSCD/DINO) than prior work, as also evidenced by our qualitative results. Stepwise text prompting creates a prior for the driving action to enable the model generate the driving action. The image frequency guidance formulations help the model in preserving the characteristics of the source subject, while reinforcing the driving action.	195
8.5	Ablations. Without stepwise text prompting, there is no prior for the driving action, which inhibits the model from generating the driving action accurately. The frequency (amplitude and phase) guidance methods help in generating the characteristics of the source subject (here, monkey) accurately - notice that there are changes to the color of the monkey (column 6), missing details in the fingers (column 4), changes in the size of the monkey (column 5).	195
8.6	ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.	198
8.7	ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.	199
8.8	ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.	200
8.9	ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.	201

8.10	Failure cases. Our method is ineffective in a few cases - it is unable to make a rabbit perform yoga or dance, and a cat perform yoga or meditate. We believe that this is due to bias-variance trade-off issues, that remain unresolved even after the application of our stepwise text prompting and image frequency guidance strategies. Usage of stronger vision and language backbones in the text to image generation pipeline, as they are made open-source to the community, can help the model disentangle features of the image better to alleviate these bias-variance trade-off issues. Besides, further research on this problem can lead to the development of newer methods that can help improve these results.	202
9.1	We present a method for multi-concept customization of text-to-video models. Our method only relies on a pre-trained text to video model and concept specific data in the form of image(s) or video(s). In the illustration above, our model uses an image of a teddy bear and ocean background (top row) and an image of a dog and video of playing violin (bottom row) to generate the customized videos. . . .	205
9.2	Proposed method for multi-concept video customization. First, we add adapter layers to the transformer architecture in an autoregressive T2V model and finetune these additional layers on the given images or videos associated with the N custom concepts and their corresponding text prompts. The goal is to find the solution at the intersection of the video manifolds corresponding to various custom concepts. Then, we condition on m ($= 5$) prior frames and sequentially generate the custom concepts and their interactions in a controlled manner using a set of prompts $p_0 \dots p_N$. The prompts $p_0 \dots p_N$ are designed to represent the scene in a top-down manner, each prompt adding a custom concept and the associated interaction.	207
9.3	Two subject customization. While Phenaki contains some prior knowledge of the interactions between different kinds of subjects, the finetuned model is unable to generate the interactions between the custom subjects. Through controlled and sequential autoregressive generation of concepts and their interactions, our method is able to generate customized videos with two subjects interacting with each other.	214
9.4	Subject Action customization. Phenaki, due to lack of prior knowledge, is unable to generate certain actions. The finetuned model is able to generate the custom subject, but not the precise action, even after finetuning. Generating the motion corresponding to the action first, followed by generating the custom subject performing action (conditioned on the motion) enables the generation of the custom subject performing the custom action. Thus, we are not only able to ‘teach’ the action to the model, but are also able to customize the subject performing the action.	215
9.5	Subject BG customization. While subject-background customization is relatively easy for the Phenaki model, the finetuned model is unable to emphasize on the background due to bias issues. This is resolved by our model, which first generates the background and then generates video frames of the custom subject performing action, conditioned on the background.	216

9.6	Three-concept customization. In the first case, the baseline models mix the attributes of the various concepts. Our method, by sequential generation of concepts and interactions, is able to generate the complex video with all desired concepts and interactions in the scene. Similarly, in the second example, while the finetuned model is able to bring in all concepts into the video, it is unable to generate the interaction of playing drums. Conditioning on motion information helps our method do so.	221
9.7	Single concept customization and compositionality. Causality can be a useful tool for generating videos where compositionality is desired, an interesting direction for future work.	222
9.8	In the first figure, we show the videoCLIP and DINO scores for our method and finetuning baseline. In the second figure, we show results of the human evaluation.	222
10.1	We present more results of our method along with comparisons. Vanilla SD fails to capture clear characteristics of individual text prompts, omitting distinct features such as those related to avocado/raccoon, muffin, parrot, and oil painting style. Linear interpolation generates images not consistent with the prompts, due to issues with non-linear manifolds. CLIP Min. generates images biased towards one of the prompts. Alt. and Step prompt selection methods suffer from artifacts and are not very successful in blending the characteristics of objects corresponding to the individual text prompts - the avocado/raccoon, muffin/dog are not blended well. In the parrot/dog image, the characteristics of the parrot are missing. Alt. generates artifacts in the Times Square/ oil painting image, while Times Square is not characterized well in the image generated by Step. In contrast, the Black-Scholes model adeptly overcomes these limitations, generating realistic images that meticulously balance and preserve the unique characteristics of each individual text prompt. The images are from set 1, set 2, set 3 and set 4 respectively.	235
10.2	Image variations from our method using the Black Scholes model for a single prompt, starting from different random Gaussian noise initializations.	243
10.3	Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, it misses out on the characteristics of the dog, parrot and ice-cream. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min. generates images biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. For instance, in order, the issues are: missing characteristics of dog, missing characteristics of coffee mug and artifacts in teddy bear, artifacts and missing characteristics of parrot, missing characteristics of teddy bear and artifacts in teddy bear, missing characteristics of rabbit. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 1.	244

10.4	Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, it misses out on the characteristics of the dog, cat, corgi dog and basket/ teapot mixing. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min. is biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. For instance, in order, the issues are: artifacts in dog/banana and missing characteristics of dog, missing characteristics of cat, the corgi dog in the dog/apple image does not describe the dog as well as our Black Scholes method does, the dog is not well described in the fourth example, the teapot is not well described in the final image. Our method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 2.	245
10.5	Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, in the first and fifth image, the characteristics of the teddy bear are missing, the second, third and fourth image do not describe the cat well, Linear interpolation generates images that are not consistent with the text prompts. CLIP Min is again biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. Specifically, the first image does not describe the subjects well, the second and third images are reasonably generated but our Black Scholes method generates a clearer image of the penguin blended cat, and teddy bear blended cat, the results of Alt. for the fourth case has artifacts and Step does not describe the cat well, the teddy bear is not well described in the fifth image. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 3.	246

10.6	Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. The first two images are cartoon-ish, and Northern lights is not clearly depicted. The third image also misses a clear description of Northern lights. In the fourth image, some of the billboards in Times Square are green, but there is no sign of Northern lights. In the fifth image, the sunset is missing. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min results are biased towards one of the prompts for all images, except the second one. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. In the first image, the kangaroo is not generated well (almost flying in Alt's result, and the characteristics of the kangaroo are not well generated by Step). In the second image, the characteristics of the tulip garden are not well represented. In the third image, notice how our model generates the characteristics of the sunset and Northern lights better, while ensuring that they are blended well to form a realistic image. In the fourth image, Alt and Step miss out on the characteristics of Times Square and Northern lights respectively. In the fifth image, Alt and Step miss out on the characteristics of Times Square and the sunset respectively. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 4.	247
------	--	-----

Part I

Introduction and Overview

The success of deep neural networks can be largely attributed to the availability of vast amounts of data. For instance, models like AlexNet and VGGNet were trained on the ImageNet dataset, which contains over 1.2 million labeled images, while ResNet-50 also utilized this extensive dataset to achieve high performance. In natural language processing, BERT was pre-trained on a corpus of 3.3 billion words, and GPT-3, with its 175 billion parameters, was trained on hundreds of billions of words from the internet. Similarly, YOLOv3 for object detection was trained on the COCO dataset with over 330,000 images. Foundational models like Meta’s LLaMA and Google’s Gemini also rely on massive datasets to perform a wide range of tasks effectively

While deep neural networks thrive on abundant data, which allows them to learn complex patterns and make accurate predictions, collecting labeled or even unlabeled data for every possible domain and task is often impractical due to time, cost, and resource constraints [12, 13]. In such scenarios, pragmatic approaches that are data efficient [14, 15] become essential. These data efficient approaches leverage data and pretrained models from related domains to acquire generic prior knowledge pertaining to the target domain. This enables the model to perform well even with limited data, by transferring to efficiently learn the characteristics of new domains and enables models to generalize better and perform well even with limited data.

Moreover, integrating techniques from classical computer vision such as camera models, and signal processing techniques such as frequency domain analysis, within deep neural networks can significantly enhance [16, 17] data-efficient learning. Classical methods, such as edge detection, feature extraction, and signal filtering, provide robust and interpretable features that can complement the deep learning models. By combining these techniques, we can create hybrid models that are not only data-efficient but also more robust and interpretable.

our thesis explores data-efficient learning across various perception and synthesis tasks, addressing key challenges such as:

- **Leveraging Data from Related Domains:** How can data from related domains, similar to the domain and task of interest, be utilized to enhance performance in a new, unseen domain without requiring labeled data specific to that domain?
- **Incorporating Classical Techniques:** How can classical signal processing and computer vision techniques improve performance in supervised settings where available data is limited?
- **Utilizing Pretrained Foundation Models:** How can foundation models, pretrained on extensive datasets and providing insights into the general characteristics of world knowledge, be employed for zero-shot image generation, enabling the creation of images without additional training data?

1. Data Efficiency for Perception

Visual scene understanding plays a crucial role in the perception systems of autonomous ground vehicles. Current systems are used for tracking, prediction, object detection, classification, localization, and semantic segmentation [18, 19]. While existing systems have made significant progress, they often operate in highly controlled or structured environments—sparse traffic, well-marked roads, and minimal debris. However, real-world scenarios present challenges: unstructured environments with heterogeneous traffic, less adherence to traffic rules by human drivers, and poorly marked roads. Additionally, adverse weather conditions like snow, rain, or fog impact perception accuracy, affecting road segmentation [5, 6]. Autonomous vehicles struggle to distinguish drivable regions from non-drivable ones, increasing accident risks [20]. Collecting labeled data in

these domains is expensive, prompting the need for unsupervised methods that leverage knowledge from related annotated driving scenes. Domain adaptation (DA), a transfer learning technique, addresses this by using large-scale annotated data from a ‘source’ domain to improve performance on the scarce ‘target’ domain. The first part of our thesis focuses on domain adaptation for autonomous driving.

In chapters 1-3, we develop domain adaptation methods for *data-efficient semantic segmentation in autonomous driving*. Specifically, we propose self-supervised algorithms for boundless multi-source domain adaptation and source-free domain adaptation. The former leverages self-training and knowledge distillation from structured datasets (such as CityScapes [21] and Berkeley Deep Drive [12]) to adapt to unstructured road images. The latter adapts neural networks from clear weather datasets such as CityScapes [21] to adverse weather images in an unsupervised manner, without annotations for adverse weather. Remarkably, the proposed algorithms achieves performance similar to fully supervised methods while eliminating the need for ongoing access to the clear weather dataset. These works extend to source-free active domain adaptation, optimizing knowledge transfer from pretrained networks and intelligent sample annotation for unsupervised datasets.

Similar to visual perception for autonomous ground vehicles, visual perception for autonomous aerial vehicles presents key challenges. One such challenge is aerial view action recognition—a critical task for security, surveillance, and tracking. While deep learning techniques have been widely used for activity recognition [7], analyzing scenes captured by UAV cameras [22, 23] is significantly more challenging than ground-camera datasets [24, 25]. UAV videos are harder to capture and label, resulting in fewer and smaller labeled datasets compared to ground videos. For example, Kinetics-400 [24] contains 306,245 ground-camera videos, whereas the recent UAV-Human database [22] has only 22,476 videos. Additionally, in UAV videos, the human actor—the

object of interest—typically occupies fewer pixels or area than the background, providing less information than a front view capture. The second part of this thesis focuses on developing data-efficient algorithms for aerial view action recognition.

In chapters 4-5, we propose novel algorithms for *data efficient aerial video action recognition*—a challenging problem due to limited available data compared to traditional action recognition. In our initial work, we develop a frequency-domain method to disentangle moving human actors from the rest of the video scene and use frequency attention to encode contextual information. In the subsequent chapter, we extend this disentanglement approach to handle static regions of the scene, improving the learning of salient regions. Additionally, we introduce a differentiable frequency-domain method for further performance enhancements.

2. Data Efficiency for Image and Video Generation

The scarcity of aerial data and the challenges associated with capturing data from aerial cameras or UAVs pose difficulties in training large neural networks for aerial image and video analysis. Cross-view synthesis, as explored in previous work [26, 27], enables the generation of realistic aerial view images from readily available ground-view images. This controlled synthesis approach produces diverse and realistic aerial data consistent with ground-view imagery, which can be valuable for representation learning and data augmentation in aerial view tasks. Beyond its applications in AR/VR and autonomous driving, cross-view synthesis involves a significant translation task for the network. It must create an entirely different view of all scene elements and the background while maintaining consistency in semantics, colors, spatial relations, pose, and other details. To enhance efficiency, the third part of our thesis focuses on developing algorithms

that optimize inference time for single-image cross-view synthesis. This approach eliminates the need for paired data or additional modalities like depth and 3D information, making it both data-efficient and domain-agnostic.

In chapters 6-7, we develop algorithms for *data efficient aerial-view image synthesis*, from a pretrained model and an exemplar front-view image. Our initial work, Aerial Diffusion, employs classical computer vision techniques—specifically Inverse Perspective Mapping (IPM) and an innovative alternating sampling strategy—to generate aerial views from a single image without relying on 3D or multi-view supervision. In Chapter 7, HawkI, we address the limitations of Aerial Diffusion, including residual diagonal artifacts and its inability to handle complex scenes. To achieve this, we propose a multi-stage fine-tuning approach that leverages IPM and incorporates information theory principles, including mutual information guidance.

Single-image cross-view image synthesis falls within the broader problem domain of controllable image and video generation. Another critical task in this field relates to personalization [28]. Recent advances in visual generative Awe have highlighted the importance of personalization [28]. Leveraging pretrained models and guidance signals from a single given image to control the content in a generated image or video requires approaches that utilize prior knowledge from these models. The goal is to generate desired images with minimal data points, all while operating in an unsupervised manner. The fourth part of this thesis focuses on image and video personalization. In chapters 8 - 9, we focus on *data efficient image and video personalization*, using a pretrained model and exemplar images describing the concepts contained in the personalized images and videos. In Chapter 8, we develop techniques to transfer actions or poses from an exemplar driving image to a source subject. Leveraging a pretrained diffusion model, we combine text prompt engineering and image frequency guidance to achieve this transfer. In Chapter 9, we propose an

autoregressive video generation method which allows customization of various concepts within the video scene, including subjects, motion, and background. These contributions enhance the controllability and personalization of image and video content, addressing important challenges in generative AI.

Further, we delve into prompt mixing in Chapter 10 - an essential aspect of our image and video personalization algorithms aimed at generating new and unseen concepts without additional training data, using pretrained models. Specifically, our goal is to develop an algorithm that generates images intersecting multiple text prompts. To achieve this, we draw inspiration from the connection between diffusion models (rooted in non-equilibrium thermodynamics) and the Black-Scholes model [29] from finance. By analogizing the variables in both contexts, we derive an algorithm that leverages properties related to the dynamics of the Markovian model from the Black-Scholes algorithm. This enables us to make informed text conditioning decisions during diffusion denoising, forecasting predictions at each time step. Our data-efficient approach utilizes a pretrained model and operates autonomously, without requiring human intervention or hyperparameter tuning. The solution for prompt mixing to generate text-controllable new and unseen concepts in a data-efficient manner, using pretrained models and without any additional training data. By leveraging these connections, we aim to enhance the generalizability and effectiveness of prompt-based image and video generation in future work.

In concluding this thesis, we provide directions for future research. we hope that the contributions presented here will inspire further investigations into data-efficient approaches for AI/ML problems related to perception and synthesis.

Part II

Learning from Little to No Supervised Data: Domain Adaptation for Autonomous Driving

In this section of the thesis, we explore the feasibility of learning representations for an unseen domain in an unsupervised manner, leveraging either supervised data or pretrained models from related domains. We propose several solutions for enhancing perception in this novel context.

This section comprises of three chapters. In the first chapter (Chapter 1), we present an unsupervised adaptation approach for visual scene understanding in unstructured traffic environments. Our method is designed for unstructured real-world scenarios with dense and heterogeneous traffic consisting of cars, trucks, two-and three-wheelers, and pedestrians. We describe a new semantic segmentation technique based on unsupervised domain adaptation (DA), that can identify the class or category of each region in RGB images or videos. We also present a novel self-training algorithm for multi-source DA that improves the accuracy. Our overall approach is a deep learning-based technique and consists of an unsupervised neural network that achieves 87.18% accuracy on the challenging India Driving Dataset. Our method works well on roads that may not be well-marked or may include dirt, unidentifiable debris, potholes, etc. A key aspect of our approach is that it can also identify objects that are encountered by the model for the first time during the testing phase. We compare our method against the state-of-the-art methods and show an improvement of 5.17% – 42.9%. Furthermore, we also conduct user studies that qualitatively validate the improvements in visual scene understanding of unstructured driving environments.

In the second chapter (Chapter 2), we present a novel approach for unsupervised road segmentation in adverse weather conditions such as rain or fog. This includes a new algorithm for source-free domain adaptation (SFDA) using self-supervised learning. Moreover, our approach uses several techniques to address various challenges in SFDA and improve performance, including online generation of pseudo-labels and self-attention as well as use of curriculum learning, entropy

minimization and model distillation. We have evaluated the performance on 6 datasets corresponding to real and synthetic adverse weather conditions. Our method outperforms all prior works on unsupervised road segmentation and SFDA by at least 10.26%, and improves the training time by $18 - 180\times$. Moreover, our self-supervised algorithm exhibits similar accuracy performance in terms of mIOU score as compared to prior supervised methods.

In the third chapter (Chapter 3), we present a novel method, SALAD, for the challenging vision task of adapting a pre-trained “source” domain network to a “target” domain, with a small budget for annotation in the “target” domain and a shift in the label space. Further, the task assumes that the source data is not available for adaptation, due to privacy concerns or otherwise. We postulate that such systems need to jointly optimize the dual task of (i) selecting fixed number of samples from the target domain for annotation and (ii) transfer of knowledge from the pre-trained network to the target domain. To do this, SALAD consists of a novel Guided Attention Transfer Network (GATN) and an active learning function, H_{AL} . The GATN enables feature distillation from pre-trained network to the target network, complemented with the target samples mined by H_{AL} using transfer-ability and uncertainty criteria. SALAD has three key benefits: (i) it is task-agnostic, and can be applied across various visual tasks such as classification, segmentation and detection; (ii) it can handle shifts in output label space from the pre-trained source network to the target domain; (iii) it does not require access to source data for adaptation. We conduct extensive experiments across 3 visual tasks, *viz.* digits classification (MNIST, SVHN, VISDA), synthetic (GTA5) to real (CityScapes) image segmentation, and document layout detection (PubLayNet to DSSE). We show that our source-free approach, SALAD, results in an improvement of 0.5% – 31.3% (across datasets and tasks) over prior adaptation methods that assume access to large amounts of annotated source data for adaptation.

Chapter 1: BoMuDA: Boundless Multi-Source Domain Adaptation for Semantic Segmentation in Unstructured Traffic Environments

Visual scene understanding is a key component of perception systems in autonomous vehicles (AVs) that deals with tracking, prediction, object detection, classification, localization, and semantic segmentation [18, 19]. These systems are responsible for understanding or interpreting the environment for safe and efficient navigation and collision avoidance. There has been considerable work on developing systems that are now deployed in the current generation of AV technologies [30]. However, most of the scene understanding algorithms and systems have been developed for highly controlled or structured environments. This includes sparse or homogeneous traffic, well-structured roads with clear lane marking, absence of debris, potholes or unidentifiable objects etc. Many times, the AVs need to operate in unstructured scenarios that consist of heterogeneous traffic with cars, trucks, bicycles or pedestrians and there is less adherence to traffic rules or regulations. Furthermore, roads are not well-marked or may include dirt or unidentifiable debris. There are still many challenges in terms of developing robust perception systems for such unstructured environments.

Recent developments in deep learning [18, 31] have resulted in significant advances in visual scene understanding [32, 33] in structured traffic environments [34, 35]. However, they do not work well in unstructured or unconstrained environments like India Driving Dataset [36].

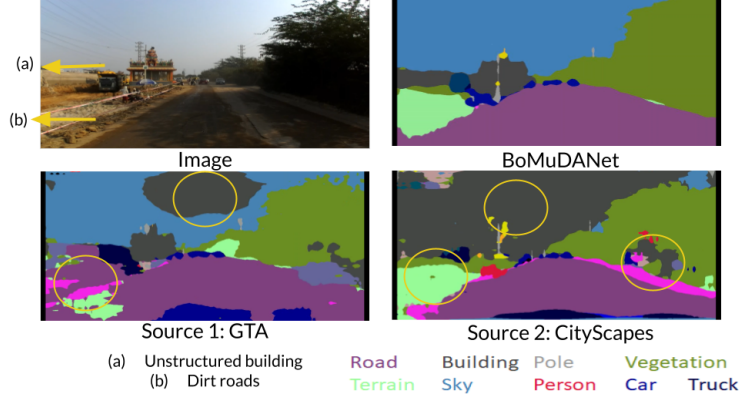


Figure 1.1: We present a novel unsupervised deep learning-based approach called BoMuDANet for visual scene understanding in unconstrained and unstructured traffic environments. In this example, we demonstrate the benefits of BoMuDANet on images taken from the challenging IDD dataset. BoMuDANet accurately segments out dirt roads as terrains as well as a building, while preserving its shape. In contrast, the single source baselines (GTA/CityScapes) do not identify dirt and unstructured roads well, misclassify parts of sky as building, and fail to capture the shape of the unstructured building. BoMuDANet benefits from its ability to selectively distill information from various sources by iterative self-training, in addition to exploiting a chosen best source via domain adaptation.

This is mainly due to lack of good datasets that consists of unstructured traffic scenes or outlier objects. For example, current traffic datasets used for training may not have vehicles such as auto-rickshaws [37]. An approach that can overcome these challenges is domain adaptation (DA) [38], a transfer learning technique that takes advantage of the availability of large scale annotated data in a different domain called the ‘source’ domain to perform a task on the ‘target’ domain, for which data is typically scarce. More specifically, DA can leverage many available large-scale structured traffic datasets such as CityScapes [35], Berkeley Deep Drive [34], GTA [39], and SynScapes[40] (source domains) to learn robust feature representations in unstructured environments.

Main Contributions: We present a new deep neural network called BoMuDANet for visual scene understanding in unstructured traffic environments. Our approach consists of a semantic segmentation technique based on unsupervised domain adaptation. BoMuDANet includes two

novel components:

1. Unconstrained traffic environments are highly heterogeneous (wide range of object classes) [41]; consequently, using only one source dataset (single source DA) is not sufficient in terms of providing the network with adequate information for optimal performance on the complex target domain. BoMuDANet benefits from its ability to selectively extract relevant knowledge across different and widely available structured environment datasets [34, 35]. Moreover, we perform multi-source DA [42] by alternating between adaptation from a selected source and knowledge distillation from the remaining sources. Based on the classical EM algorithm in statistical pattern recognition, BoMuDANet performs repeated rounds of training, alternating between adaptation and distillation to improve performance in each step.
2. Unconstrained traffic scenes may contain objects that are typically non-existent in current structured environment source domain datasets. Our approach in BoMuDANet uses a simple pseudo-labeling strategy for handling unknown objects encountered for the first time during the testing phase. The final probability predictions of the self-training algorithm are directly used to assign proxy labels to unknown object classes depending on their similarity to known objects classes in the training dataset. Our approach helps BoMuDANet detect new objects that are common in unstructured driving environments.

We have evaluated our approach extensively using the Indian Driving Dataset (IDD), CityScapes, Berkeley DeepDrive, GTA V, and the Synscapes datasets. In unstructured environments (IDD as the target dataset), we show that our unsupervised approach outperforms other unsupervised SOTA benchmarks by 5.17% – 42.9%. In structured environments (CityScapes as the target dataset), we show that our method outperforms other multi-source DA methods by 12.70% –

90.13%. We have performed extensive ablation experiments to highlight the benefits of our approach. Moreover, we also perform a user study to highlight the qualitative benefits of our approach. Overall, ours is the first unsupervised domain adaptation method for handling unstructured traffic environments.

1.1 Prior Work

Semantic Segmentation Semantic segmentation is a pixel-level task, which involves assigning a label to each pixel in an image. The advent of deep learning has resulted in a many segmentation techniques for autonomous driving [18, 31, 43, 44, 45, 46]. However, these methods suffer from three issues: (i) the networks need to be trained in a supervised manner, thus there is a demand for large volumes of annotated data; (ii) current labeled datasets are limited to structured environments; (ii) current learning methods do not scale well to unstructured environments.

Unsupervised Domain Adaptation Domain adaptive semantic segmentation has been explored under three different machine learning paradigms that differ based on the underlying learning approach. At one end of the spectrum, fully supervised [47] methods achieve higher accuracy on average, but are limited by the availability of annotated data. On the other end of the spectrum, unsupervised methods [48, 48, 49, 50, 51, 52, 53, 54, 55, 56] benefit from the lack of dependence on any training data, but are outperformed by fully supervised or semi-supervised methods. Semi-supervised approaches [57, 58, 59, 60] form a middle ground between the two paradigms. While there has been some work on using pseudo labels [61, 62] for training DA models, they do not scale well in the presence of multiple sources.

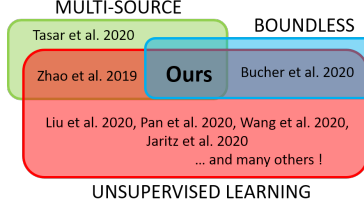


Figure 1.2: Extension to SOTA in domain adaptive semantic segmentation. Our approach is the first method to *simultaneously* perform unsupervised multi-source boundless DA segmentation and can handle unstructured traffic environments.

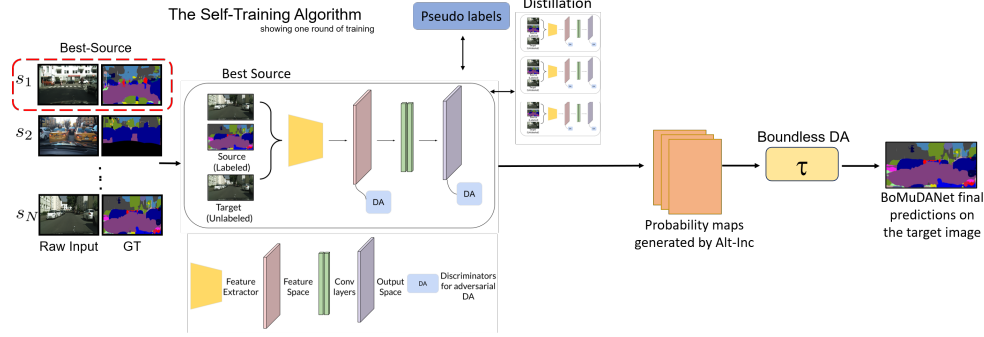


Figure 1.3: **Overview of BoMuDANet:** The input consists of N sources ($s_1, s_2, \dots, s_N$), from which the best-source is selected by the self-training algorithm. The self-training algorithm proceeds in an unsupervised fashion to generate the final set of pseudo-labels that are used to recognize out-of-distribution objects. The final output consists of the segmentation map of an image in the target domain.

Open-Set and Boundless DA If the set of labels in the source is equal to the set of labels in the target, then this type of DA is known as *closed-set DA*. On the other hand, if the target domain contains additional class labels that are not present in the source domain, then this type of DA is called *open-set DA*. In open-set DA, the additional class labels in the target domain that do not belong to the source domain are labeled as an “unknown” class [63]. While open-set DA has been proposed for object detection and classification [64, 65, 66], they don’t extend well for pixel-level tasks like semantic segmentation.

An extension to open-set DA is boundless DA, where the extra classes present in the target domain are explicitly labeled. Boundless DA has been recently studied by [61] for semantic segmentation, where the authors successfully classify open-set classes, but at the cost of degraded

accuracy on the closed-set categories.

Multi-Source Domain Adaptation While multi-source DA has been extensively studied in the context of other perception tasks like object recognition and classification [67, 68, 69, 70, 71], it has not been explored in detail for semantic segmentation [42, 72]. Prior approaches in multi-source DA suffer from heavy overhead in terms of the requirement of data from all sources at every point of training. In contrast, BoMuDA requires only the pre-trained models, along with data from a chosen source ‘best source’ domain. We present the first method for unsupervised multi-source boundless domain adaptive semantic segmentation. However, our approach can be generally applied towards domain adaptation in different perception tasks such as object recognition. This is a part of our future work.

1.2 BoMuDANet

In this section, we formally specify our problem, introduce the notation and present details of our neural network used for visual scene understanding in unstructured traffic environments.

Problem Setup and Notation Given an RGB image or video of unconstrained traffic selected from the target domain at test time, our goal is to identify the correct object class label of each pixel. In the training phase, we are provided with a set $\mathcal{S}$ of N source domains, in which each source domain is represented as S_i , where $i = 1, 2, \dots, N$, and one target domain T . The set of all categories in the target domain is denoted by $\mathcal{C}_T$, while the set of all categories in the i^{th} source domain is denoted by $\mathcal{C}_i$. In the boundless DA setting, the target domain may consist of open-set categories *i.e.* classes that are not present in any of the source domains. More formally,

$$\mathcal{C}_T \setminus \{\cup_i \mathcal{C}_i\} \neq \emptyset.$$

The output probability map (Each value in this map is the probability of the pixel, in the corresponding location in the input image, belonging to class C_i .) for an input image belonging to the i^{th} source domain is denoted as $P_i \in \mathbb{R}^{C_i \times h \times w}$, while the ground truth label for the same image is denoted by $y_i \in \mathbb{R}^{h \times w}$. In the unsupervised DA setting, the ground-truth labels for target domain images are not available.

Overview Our method, BoMuDANet, trains a deep neural network using a novel self-training algorithm. The input consists of images and their corresponding labels from multiple source domains and images from the target domain. The self-training algorithm generates probability maps corresponding to the target domain image.

1.3 A Self-Training Algorithm for Multi-Source DA

The main challenge with unsupervised multi-source domain adaptation is in setting up a cost function [73] for training the deep neural network (DNN). This is due to two reasons: (i) absence of target domain labels and (ii) variations between different source domains, and each source and target domain. In the proposed approach, we use the idea of “pseudo” labels to act in place of the missing target domain labels. The pseudo labels are the class predictions from the ‘best source’ single-source DA model, which is explained below in “Initialization”. These pseudo labels, along with the pre-trained single-source DA models from the remaining sources are used for training the deep neural network with an improvised cost function.

In concurrence with the notion of self-training, we observed that repeated re-training of the deep neural network with an enriched cost function each time leads to a significant boost in

accuracy. This is because the model weights are optimized after each round of training, which in turn optimizes the pseudo labels, leading to an increasingly accurate cost function to be used in the next round of training¹. These two optimizations occur in an alternating manner along with domain adaptation from best source and multi-source distillation, incrementally improving the accuracy.

The motivation behind the self-training algorithm comes from the Expectation-Maximization (EM) algorithm [74], a classical unsupervised learning algorithm in statistical pattern recognition. The EM algorithm consists of two alternating steps— the E step and the M step. The E step sets up a cost function from observed data, while the M step finds the model parameters that minimize the cost function. Our self-training algorithm mimics the principles behind the EM algorithm.

To setup the cost function for the first iteration, the self-training algorithm selects the best-performing source from the N source domains to generate pseudo-labels that serve as a proxy for the missing target domain labels. This best-performing source dataset is termed as the “Best-Source”. The inputs to the neural network are images and GT from the “Best-Source”, pre-trained single-source DA models for all the sources, images from the target domain and the pseudo labels. In summary, the self-training algorithm trains the network as follows:

1. *Initialize* $\leftarrow$ Best-Source model.

2. *Perform the following in an alternating manner:*

- Use pseudo labels from previous round of training to set up a cost function for BoMuDANet.
- Use this cost function along with the remaining $N-1$ pre-trained single-source DA

¹We provide evidence of the benefits of repeated re-training in the supplementary material.

models to train BoMuDANet in an end-to-end manner till convergence.

We now describe each step in detail.

Initializing the Best Source Model We begin by training single-source DA models using each source dataset, and the target dataset, using an adversarial DA paradigm [50]. The single-source domain discriminators (binary classifiers, see architecture below) characterize how indistinguishable the target domain is from each source domain. The output of the discriminators averaged over all target images characterizes the dissimilarity between each source domain and the target domain. The source domain with the least dissimilarity is selected as the “best source”. The deep neural network (DNN) used to train the best source-target pair is termed the “Best-Source” model.

Architecture: Consistent with adversarial domain adaptation [50], our network consists of a DNN for semantic segmentation, and domain discriminators. The backbone of the DNN consists of SOTA architectures such as VGG-16 [75], Dilated Residual Network [18], or DeepLab [31] (we experiment with different backbones). Domain discriminators are neural networks that aim to distinguish whether the predicted segmentation map is from the source or target.

Training: The inputs to this model consist of raw images and GT from the best source domain, raw images from the target domain T , and the pseudo labels. The model weights are initialized with parameters corresponding to the Best-Source baseline obtained in the initialization step. The cost function consists of a domain adaptation loss formulation to adapt from the best-source, a knowledge distillation loss formulation to selectively distil relevant information from the remaining sources, and a self-training step that utilizes the pseudo labels. We now describe the three loss functions that are used to train the network:

- **The supervised loss function** (L_{sup}): This is the standard cross entropy supervised loss function that is used to minimize the distance between the probability map outputs and the ground truth labels.

$$\mathcal{L}_{\text{sup}} = - \sum_{h,w} \sum_{c \in C_{bs}} y_{bs} \log(P), \quad (1.1)$$

where c denotes the object category, h, w denote the height and width of the input images, and $P \in \mathbb{R}^{C_{bs} \times h \times w}$ is the output of the model on source domain images.

- **The unsupervised loss function** ($\mathcal{L}_{\text{unsup}}$): For each target image, we use the trained model from the previous iteration of self-training to generate pseudo labels for self-training. The pseudo labels are generated using the probability map predictions, $P \in \mathbb{R}^C \times h \times w$. More formally,

$$y_{\text{pseudo}} = \arg \max_{c \in C} \text{Softmax}(P). \quad (1.2)$$

The pseudo-label y_{pseudo} is used in the unsupervised cross entropy loss function, L_{unsup} , as follows,

$$L_{\text{unsup}} = - \sum_{h,w} \sum_{c \in C} y_{\text{pseudo}} \log(P). \quad (1.3)$$

- **Multi-source distillation** (L_{distill}): From each of the single source DA networks, we generate their corresponding target domain probability maps $P_i, i \in [N]$. To selectively impart relevant knowledge from various sources, the target domain predictions of BoMuDANet are distilled using a weighted combination of KL divergence [76] loss terms corresponding

to each of the single-source DA predictions. KL divergence aligns probability distributions and a weighted combination of KL divergence from multiple sources aids in selective extraction of relevant knowledge. The weights (w_i) are determined by the dissimilarity between each source-target pair (see initialization step above).

$$L_{\text{distil}} = \sum_i w_i \times KL(P_{\text{bs}} || P_i). \quad (1.4)$$

The three loss functions are combined as follows:

$$L_{\text{overall}} = \lambda_{\text{sup}} L_{\text{sup}} + \lambda_{\text{unsup}} L_{\text{unsup}} + \lambda_{\text{distil}} L_{\text{distil}}, \quad (1.5)$$

where λ_{sup} , λ_{unsup} , λ_{distil} denote the hyperparameters for the respective loss terms. The domain discriminators are trained in an adversarial [50] fashion.

1.4 Boundless Domain Adaptation

We present a new method for performing Boundless DA *i.e.* to label categories that exist in the target dataset, but not in any of the source datasets (“open-set” or “private” or “unknown” categories). Categories that are common to both the source and the target domains are called “closed-set” or “shared” or “known” categories. The key assumption in our solution is that the open-set categories are physically similar to the closed-set categories. For instance, open-set categories such as auto-rickshaws are similar to vehicles like cars and vans. CityScapes [35] provides a definition for grouping semantically similar classes in autonomous driving environments. Classes that belong to the same high-level category will have feature maps that are semantically

Model	Experiment	mIoU	mAcc	Road	SW	Bldg	Wall	Fnc	Pole	Lt	Sign	Veg	Trn	Sky	Ped	Rdr	Car	Trk	Bus	Mb	Bike
I. CS, BDD, GTA to IDD (Baseline: [50])																					
Baselines	CS to IDD	24.43	65.23	82.46	22.55	25.93	13.22	9.30	15.26	1.92	19.02	75.16	20.41	29.54	31.37	8.12	49.81	8.53	10.41	10.29	6.55
	GTA to IDD	26.74	75.40	79.83	9.54	44.12	16.58	12.16	17.59	0.85	14.35	65.36	18.20	82.61	22.90	6.56	41.53	24.13	15.40	9.02	0.76
	BDD to IDD	35.75	85.65	93.33	27.17	59.77	13.18	15.56	21.03	3.65	29.93	80.52	33.21	93.64	30.62	5.59	53.03	38.34	32.24	6.46	6.27
BoMuDANet	Multi-source	37.66	86.50	94.02	31.89	61.79	15.51	16.89	20.61	2.73	35.43	81.75	36.52	94.16	32.12	4.67	54.74	42.64	38.61	5.42	8.51
II. SC, BDD, GTA to IDD (Baseline: [50])																					
Baselines	SC to IDD	31.55	83.04	92.46	21.25	52.59	4.61	7.87	17.02	2.73	12.60	77.52	4.43	92.38	31.54	23.32	66.59	4.09	18.35	27.27	11.25
	GTA to IDD	26.74	75.40	79.83	9.54	44.12	16.58	12.16	17.59	0.85	14.35	65.36	18.20	82.61	22.90	6.56	41.53	24.13	15.40	9.02	0.76
	BDD to IDD	35.75	85.65	93.33	27.17	59.77	13.18	15.56	21.03	3.65	29.93	80.52	33.21	93.64	30.62	5.59	53.03	38.34	32.24	6.46	6.27
BoMuDANet	Multi-source	36.93	86.30	93.82	30.53	61.13	13.34	16.43	21.21	3.57	34.90	81.64	34.54	94.19	31.70	4.64	53.48	40.77	35.54	5.68	7.64
III. CS, BDD, GTA to IDD (Baseline: [49])																					
Baselines	CS to IDD	38.53	86.68	93.67	27.08	64.62	25.89	17.80	23.39	4.18	31.29	83.06	29.83	94.22	32.28	11.18	61.68	39.86	33.32	12.08	8.23
	GTA to IDD	35.85	84.64	89.96	14.06	61.14	22.24	20.10	19.17	4.34	19.88	77.15	28.84	92.14	27.03	11.98	62.87	41.04	34.67	13.10	5.74
	BDD to IDD	38.29	86.74	93.80	33.33	62.57	14.94	15.35	23.66	3.80	31.95	81.72	34.47	94.26	33.00	8.71	57.11	42.87	39.16	9.41	9.22
BoMuDANet	Multi-source	39.23	87.18	93.18	29.97	63.46	24.18	20.97	19.18	4.56	25.64	81.99	35.39	94.19	30.06	11.23	62.01	46.65	39.30	13.39	10.87

Table 1.1: **Main Results:** We evaluate BoMuDANet on IDD using CityScapes (CS), Berkeley Deep Drive (BDD), SynScapes (SC), and GTA as sources. Higher mIoU and mAcc indicates direction of better performance. **Bold** indicates best while **blue** indicates second-best. Experiments I and II differ with respect to the sources, while experiment III differs with respect to the baseline used. **Conclusion:** Our unsupervised multi-source self-training algorithm outperforms the single-source baselines by 3.3% – 54.15%.

similar, and vice versa. This assumption is mild and is commonly made in many zero-shot learning strategies [77].

The underlying idea behind training our approach on open-set classes is to generate the corresponding pseudo-labels from the labels of the physically similar closed-set categories. More formally, let $y_{ST} \in \mathbb{R}^{h \times w}$ be the final labels obtained using Equation 1.2 from the self-training algorithm. Further, let $o \in O$ denote an open-set class from the set of open-set classes, O , and C_o denote the set of closed-set classes that are physically similar to o . We apply thresholding on y_{ST} such that pixels with softmax scores lower than a threshold τ for a physically-similar closed-set class are re-labeled as the open-set class. More formally, let $\hat{y}_{pseudo}$ denote the labels after thresholding, then $\hat{y}_{pseudo}$ is computed using,

$$\hat{y}_{pseudo} = T y_{ST} \quad (1.6)$$

where $T(\cdot)$ is a pixel-level thresholding operator. If l_{ab} denotes the class label of a pixel in the a^{th} row and b^{th} column with confidence score c_{ab} (maximum probability value over all classes,

as determined by the output probability map of the self-training algorithm), then the threshold operator at (a, b) is defined as,

$$T(a, b) = \begin{cases} l_{ab} \leftarrow o & c_{ab} \leq \tau \text{ and } l_{ab} \in C_o \\ l_{ab} & \text{otherwise} \end{cases}$$

An alternative to the thresholding operator is to use the KL divergence metric [76] to measure the similarity between open-set and closed-set object classes. We empirically observe, however, via an ablation study that thresholding in fact outperforms using the KL divergence metric.

1.5 Experiments and Results

Datasets and Evaluation Protocol We use five datasets - GTA5 [39], SynScapes (SC) [40], CityScapes (CS) [35], India Driving Dataset (IDD) [36] and Berkeley Deep Drive (BDD) [34]. GTA5 and SC contain synthetic simulated traffic videos while CS and BDD consist of real-world traffic in Europe and the USA, respectively. IDD consists of dense and unconstrained traffic conditions and heterogeneous road agents (e.g. autorickshaws) unobserved in any of the source domains. In addition to containing new objects, the pixel count (per class) in IDD is 5–10 $\times$ that of CS

We evaluate our models on the validation set images of the target domain, using the standard segmentation metrics [78]: Intersection over Union (IoU) and pixel-wise accuracy.

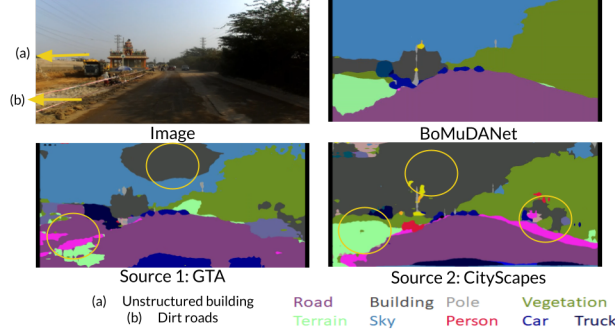


Figure 1.4: In this example, we demonstrate the benefits of BoMuDANet on an image from the IDD dataset, depicting a mixture of challenging driving conditions. The top image in the second column shows the prediction of our model, and the bottom image shows the ground-truth. We observe that BoMuDANet accurately segments the autorickshaws (open-set object, a new type of vehicle - the third column zooms into the region containing the autorickshaw in the prediction and ground-truth), in addition to handling dense traffic and dirt roads.

1.6 Results

Main Results on IDD: We present results of three sets of experiments using IDD as the target dataset. In each experiment, we compare BoMuDANet with single-source baseline models using the BDD, CS, SC, and GTA datasets, with the BDD dataset selected as the Best-Source. Note that we do not compare with a combination of single source datasets as combining multiple sources and treating them as a single source for DA has been shown to be ineffective [73].

We perform the first experiment with two real datasets (CS, BDD) and one synthetic dataset (GTA5), and show an improvement of $1.91 - 13.23$ (5.34% - 54.15%) mIoU points over the single-source baselines. In the second experiment, we replace the two real source datasets with two synthetic source datasets (SC, GTA) and one real dataset (BDD) with the BDD dataset as the Best Source, and show an improvement of 3.3% - 38.1% mIoU points over the single-source baselines. By comparing these two sets of experiments, we demonstrate that using multiple real-world datasets is more beneficial than using multiple synthetic datasets.

Experiment	mIoU	Car	Truck	Bus	Auto
CS, BDD, GTA to IDD, Baseline: [50]					
Pseudo labeling (PL)	35.68	51.16	33.89	28.99	9.39
PL + training	35.72	52.18	33.93	31.65	9.38
SC, BDD, GTA to IDD, Baseline: [50]					
Pseudo labeling (PL)	34.60	48.36	30.78	20.82	9.68
PL + training	34.40	49.14	30.44	22.59	9.48
CS, BDD, GTA to IDD, Baseline: [49]					
Pseudo labeling (PL)	37.27	58.76	36.58	22.15	11.78
PL + training	37.09	58.63	36.65	24.26	11.85
Ablation Experiments on IDD, Baseline: [50]					
KL Divergence	35.43	52.12	33.99	31.46	9.29

Table 1.2: **Pseudo labeling strategy for boundless DA:** We show that pseudo labeling can provide semantic information about categories that do not belong to any source domain, for instance, **auto-rickshaws** (Auto) found in the IDD dataset (in **bold**). Moreover, pseudo labeling is simple in that the generated proxy labels do not need to be re-trained as there is no marked improvement in mIoU. We also show that thresholding outperforms KL divergence via an ablation study.

In the third experiment, we replace the AdaptSegNet [50] backbone with a stronger SOTA backbone ADVENT [49], and use LS GAN for adversarial training instead of Vanilla GAN, and achieve a higher mIoU of 39.23. This suggests that the performance of our approach will increase as newer robust backbone architectures are proposed. We also validate this hypothesis by conducting experiments on structured environments as the target domain.

Results for the Boundless Case: We show the results for the proposed pseudo labeling strategy for boundless DA method in the table. Note that the thresholding operator, τ , is a tunable hyperparameter. A low value of τ will create a bias towards the private classes, a high value of τ will create a bias towards shared classes. A trade-off determines the optimal value of τ for best performance on both private and shared classes. Typically, tuning between 80% – 90% of max confidence score for the particular class works well.

The unsupervised loss function in Equation 1.3 can be used to retrain the network on $\hat{y}_{\text{pseudo}}$, along with y_{ST} which acts as a regularizer. The first row in each experiment shows the

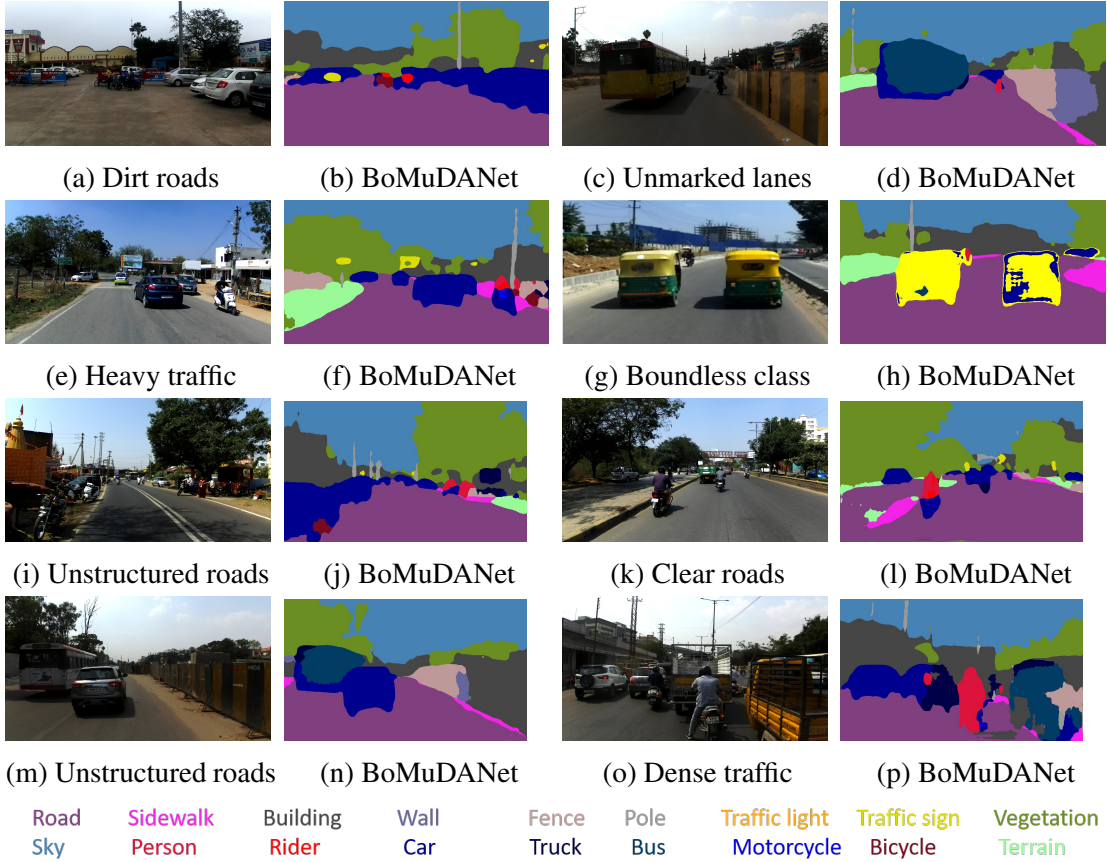


Figure 1.5: **Visual Results:** BoMuDANet works well in various unconstrained environments including unmarked lanes, dirt roads, heavy traffic, and boundless category objects (auto-rickshaws) and results in higher accuracy. Each color represents a different object as shown in the color scheme.

results obtained by proposed strategy of pseudo-labeling while the second row shows the results obtained by re-training the generated pseudo labels. However, we found that the performance of the retrained model is very similar to original model, mitigating the need for costly retraining and therefore contributing to the simplicity of the proposed pseudo labeling strategy.

Qualitative Results and Realtime Performance: As shown in the figure, our method works well in environments that have dirt roads, absence of clear lane markings, multiple road objects and unstructured traffic. BoMuDANet can recognize auto-rickshaws (boundless category object) reasonably well.

On IDD				
Method	Model	# Size(M)	mIoU(S)	mIoU(P)
[79]	ResNet-18	11.70	27.45	NA
[77]	ResNet 101	44.50	29.20	7.90
[61] (UDA)	ResNet 101	44.50	32.40	8.10
[61] (Apt.)	ResNet 101	44.50	32.70	8.60
[61] (Comb.)	ResNet 101	44.50	37.30	18.50
BoMuDANet	DRN-D-38	26.50	39.23	11.85

On CS				
Method	Model	Size(M)	mIoU(S)	# Sources
[80]	ResNet-101	44.50	39.40	Single
[50]	ResNet-101	44.50	42.40	Single
[49]	ResNet-101	44.50	43.10	Single
[49]	ResNet-101	44.50	43.80	Single
[48]	ResNet-101	44.50	46.30	Single
[81]	ResNet-101	44.50	49.90	Single
[72]	VGG-16	138.00	29.40	Multi
[42]	VGG-16	138.00	41.40	Multi
BoMuDANet	DRN-D-38	26.50	44.63	Multi
BoMuDANet	ResNet-101 [49]	44.50	49.59	Multi
BoMuDANet	ResNet-101 [81]	44.50	55.90	Multi

Table 1.3: **Comparison with SOTA:** We compare with the SOTA in both unstructured (IDD) as well as structured (CS) traffic. Higher ($\uparrow$) mIoU and mAcc indicates direction of better performance. **Bold** indicates best while **blue** indicates second-best. mIoU(S) and mIoU(P) denote the performance on shared/known and private/unknown classes, respectively. **Conclusion:** Our model is SOTA on IDD by 5.17% – 42.9% and on CS in the multi-source setting by 12.70% – 90.13%, with a reduction in model size by upto to $5.2\times$.

We have included a video demonstration of BoMuDANet in realtime in 6 diverse traffic videos containing both unconstrained (IDD) as well as structured (CityScapes) environments, along with comparisons. BoMuDANet operates at 2 fps on IDD and 21 fps on CityScapes with a model size of 26.5 million parameters. We refer the reader to the supplementary video.

Comparisons with SOTA

In Unstructured Environments (On IDD): , we compare our approach against other unsupervised segmentation methods. ZS3Net [77] does zero shot semantic segmentation, while [61] (UDA) and [61] (Apt.) builds upon ZS3Net for domain adaptation. [61] (Comb.) refers to the combined approach for boundless unsupervised domain adaptation (“BUDA”). It can be clearly observed that our method surpasses all past unsupervised segmentation methods by 5.17% - 34.34% on shared classes, with a much smaller architecture which is beneficial for practical autonomous

driving real-time applications.

Our hypothesis for the superiority of BoMuDANet over BUDA is that the latter comprises performance on closed-set classes in order to achieve improved performance on open-set classes [61, 77]. In contrast, our method classifies open-set categories *without* sacrificing accuracy on closed-set categories. The decreased performance of BUDA on “shared” classes could be due to decreased generalization capabilities of the model when trained on the new classes. We also outperform the semi-supervised method [82] by 42.9%, that uses ground-truth in 100 samples for supervision. [82] fails to acknowledge differences between various domains, which leads to a degradation in performance.

In Structured Environments (On CS): We additionally benchmark BoMuDANet in structured environments, using CS as the target domain and BDD, IDD and GTA as the source domains. Our method is SOTA in the multi-source setting by at least 12.70%–90.13% with a reduction in model size by upto $5.2\times$. Methods with ResNet-101 backbone have an inference time of 156.44 ms, and models with DRN-D-38 backbone have an inference time of 51.58 ms. On CS, BoMuDA outperforms the corresponding single-source DA baselines by 2.5% – 21.2% respectively. Furthermore, stronger backbones will help our model benefit accordingly. Further, comparison of our network (with corresponding backbones) against single-source baselines reveals that our model is the SOTA.

The core step in the approach of [42] is the use of the CycleGAN [83], which uses images and ground truth from all source domains at every training step. Our multi-source approach, in contrast, is more computationally efficient and requires data only from the “best source”. Pre-trained single-source adaptation weights can be directly used for the other datasets, thus

offsetting the need for images and GT from all source domains. The improvement in our approach comes from individually distilling relevant information from multiple domains as opposed to considering images from all source domains in every iteration.

1.7 Ablation Studies and Additional Experiments

We show the benefits of using multiple sources compared to a single source. The multi-source model outperforms the corresponding single source baselines by 3.3%–54.15%, demonstrating the efficiency of using multiple sources. In boundless DA, we replace the thresholding operator with the KL divergence loss to measure the similarity between the open-set classes and physically similar categories. We demonstrate the iteration wise performance of the self-training algorithm and a study of tuning the hyperparameters λ_{distil} and λ_{unsup} in the supplementary material. Additionally, selecting the best source at the pixel-level degrades accuracy by 18.05 % due to loss of contextual information. Finally, thresholding on pseudo labels [61] to reduce the number of false positives reduces the mIoU by 4.54 %.

1.8 User Study

We conducted a web-based user study to evaluate the visual scene understanding results generated by our approach in unstructured driving environments and evaluate their perceptual benefits compared to prior methods. We also test our hypothesis of classifying open-set objects. These open-set objects correspond to new objects that our model has never seen before in the training dataset. Finally, we perform an additional study where we highlight the complexity of our task by asking users to rate the level of difficulty of driving in the kinds of scenes evaluated

by our method.

Procedure The study consists of three sections and is approximately five minutes in duration. In the first section, we show users six comparisons (selected from the dataset randomly) between the results obtained by our unsupervised adaptive multi-source approach versus the state-of-the-art single-source baselines [50], along with the corresponding ground truth image. We do not reveal the identity of the methods and swap the order of the results to reduce the bias. We ask users to visually inspect both the results and report the result that best resembles the ground truth image. The aim of this study is to complement and support the quantitative results with human visual feedback.

In the second section, we show users results of our approach on two images containing an open-set and a closed-set object (selected from the dataset randomly) , respectively. We ask the participants to select which result in their view is better. We purposefully avoid defining what constitutes a better segmentation result to eliminate any bias. Higher probabilities of selecting the open-set object validates our hypothesis of classifying open-set objects.

In the third and final section, we show participants five pairs of images of structured and unstructured driving environments (selected from the dataset randomly) . Following the previous two sections, we do not label the images. Users were asked to rate the difficulty of driving for each image on a 5-point Likert scale (1– “Very Easy” to 5– “Very Hard”).

Participants 47 anonymous users participated in our study, recruited over the web. To remove bias related to users with driving experience in a particular traffic environment, we invited participants belonging to regions with traffic that is similar to the traffic found in both CityScapes and the IDD

datasets. We do not consider age or gender of the users in our study.

Evaluation: BoMuDANet vs. state-of-the-art Each participant picked the segmentation map (out of baseline, and BoMuDANet) that they felt best resembled the ground-truth. We summarize the percentage of responses for each of the 6 images in the figure, arranged in the order of increasing complexity. All participants agree that BoMuDA performs better than the baseline for image 1 and 2. While images 3 – 6 are hard to segment accurately, at least 60 – 90% of the participants agree that BoMuDA results in a better solution. This, coupled with the fact that participants did not apriori which segmentation maps were generated by the baseline single-source models and BoMuDANet indicates that our model is superior. One possible explanation of participants rating BoMuDANet segmentation maps higher is that BoMuDANet outperforms the single-source baseline by selectively extracting the best features from each source and is thus capable of segmenting dirt roads, persons, motorcycles in a better fashion.

Open-set analysis Define p as the probability that a participant will choose the open-set result (auto-rickshaw) when shown a pair of results generated by our model on an one open-set object and a regular object in the training dataset. We wish to test the hypothesis that either a participant cannot tell the difference between the two results ($p = 1/2$) or the open-set result is good enough such that participants choose the same ($p > 1/2$). Thus, we test the following right-tailed hypothesis test:

$$H_0 : p < \frac{1}{2} \quad \text{against} \quad H_1 : p \geq \frac{1}{2}.$$

Let N_1 be the number of participants who selected the open-set result as the better segmentation result. Then, $\hat{p} = N_1/N$. We test the above hypothesis with the test statistic $(\hat{p}-p)/\sqrt{(\hat{p}(1-\hat{p}))/N}$,

which is asymptotically normal with mean 0 and standard deviation 1, under the null hypothesis. We observe in our study that $N_1 = 30$ corresponding to $\hat{p} = 0.63$ with a p-value of 0.032884 ($\alpha = 0.05$ significance level). Our result is statistically significant indicating we can reject the null hypothesis (H_0) and accept the alternative (H_1).

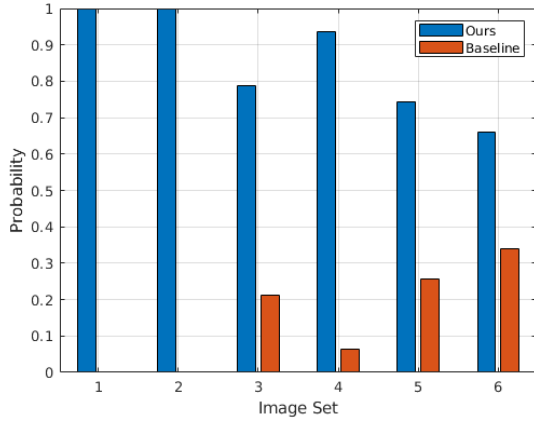
Complexity of the task Each participant was shown 5 pairs of images depicting scenes from CityScapes and IDD (unknown apriori) and asked to rate the difficulty of driving. Therefore, we have a total of 235 responses each, rating the difficulty of driving on a 5-point scale. We summarize the responses in the figure. About 93% of the participants reported that it they felt it was easy to drive in CityScapes-like environments, while 63% of the participants reported that driving in an environment similar to IDD is hard. In contrast, only 14% of the participants reported that driving in an IDD-like environment is easy and a meagre number of participants felt that navigating through the traffic depicted in CityScapes is hard. This is in line with prior studies [84], which state that driving in unstructured environments with heterogeneous traffic is difficult.

1.9 Conclusion, limitations and future work

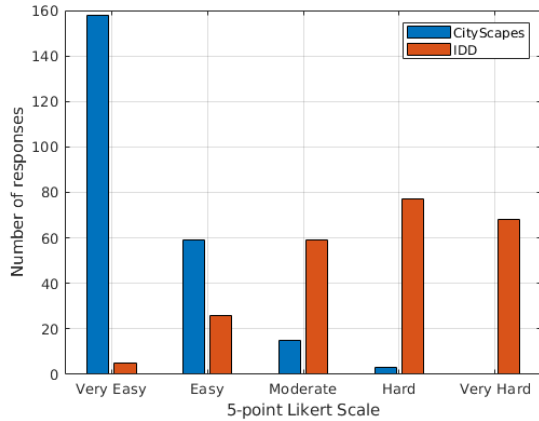
We present a novel learning methods for visual scene understanding in unstructured traffic environment. Our approach consists of a semantic segmentation technique that solves three key aspects of domain adaptation: unsupervised, multi-source and boundless, in unconstrained environments. We present a novel training routine that builds on the ideas of self-training and pseudo-labeling. The self-training routine is used to selectively distil information from various sources by iterative self-training, in addition to exploiting a chosen best source via domain adaptation. In addition, BoMuDANet can identify unknown objects encountered during the

testing phase via a simple pseudo labeling strategy. We highlight the benefits of our approach in terms of performing accurate segmentation and visual scene understanding in challenging datasets such as IDD. We highlight improved accuracy over prior methods and perform qualitative evaluation based on a user study.

Our approach has some limitations. Our current approach can only recognize new objects by taking advantage of the structural similarities between various objects in road environments. Currently, our model is unable to detect classes like animals and other classes that do not share any similarities with the ‘known’ classes. In addition, the existence of multiple ‘unknown’ objects that share similarities with the same set of ‘known’ classes can cause inter-class confusion. As part of future work, it would be useful to use the pseudo labeling strategy as a prior, and develop a training method that exploits zero-shot learning strategies. The core step in the self-training algorithm is the selection of a ‘best source’, which can vary from image to image. Our current formulation does not account for this factor during training. In addition, we do not account for variations within the target dataset, where images can have varying levels of similarities with the source datasets. Future work in this area can focus on an importance weighting scheme for a multi-source domain adaptation network that is more robust. We would also like to evaluate our approach in other challenging scenarios and integrate with planning and navigation.



(a) Study 1: We show the probability of users selecting the result generated by our multi-source approach (blue) on six randomly chosen images of unstructured driving environments from the IDD dataset compared to single-source baselines (red) [50]



(b) Study 3: We report the number of users corresponding to the 5-point Likert scale summed over all five sets of images.

Figure 1.6: User Study Results: In the first user study, we observe that users are more likely to report our result as the better segmentation compared to state-of-the-art. In the second study, we observe that majority of the users confirm that the IDD traffic scenes are complex and that it would be hard to drive in such environments whereas almost all participants find the driving environments in the CityScapes dataset to be easy.

Chapter 2: SS-SFDA: Self-Supervised Source-Free Domain Adaptation for Road Segmentation in Hazardous Weather Conditions

Research in autonomous driving continues to advance in terms of improving the perception capabilities of self-driving cars for greater safety. This includes detecting both static and dynamic obstacles such as pedestrians [85], tracking and predicting the trajectories of other vehicles [41, 86, 87, 88], and scene segmentation [45, 89]. Immense progress along these lines has led to the deployment of level 2 and almost level 3 autonomous vehicles (AVs) in urban traffic environments [90]. However, these advances in perception technology have been primarily designed to work well in safe and clear weather conditions. Driving in adverse weather and lighting conditions such as snow, rain, or fog is challenging not just for autonomous vehicles but even for humans. These conditions result in a degradation in accuracy of perception techniques including road segmentation [5, 6]. Consequently, AVs are unable to distinguish drivable regions of the road from the non-driveable region (which may be affected by snow, rain, or fog), thereby increasing the likelihood of road accidents [20]. In this chapter, we address the problem of road segmentation in adverse weather conditions.

The road segmentation [91, 92] problem corresponds to identifying the pixels in an RGB image or video that belong to the ‘road’ class. While general models designed for semantic segmentation in computer vision can be directly used for road segmentation, they suffer from

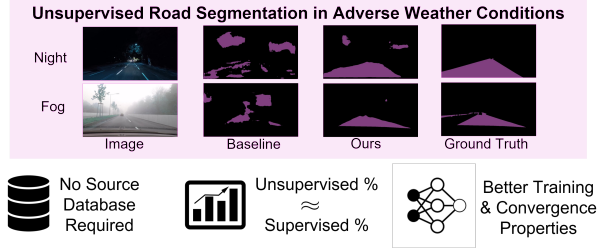


Figure 2.1: We highlight the results generated by SS-SFDA on night [5] and fog benchmarks [6], compared to the baseline source model pre-trained on clear weather CityScapes. The purple regions (right) denote the segmented road pixels. The overall accuracy of our self-supervised algorithm in terms of mIoU is (88 – 96%) of supervised methods.

the inability to capture semantic relationships between different objects due to the lack of unique labels for each class. The use of self-attention techniques [45, 93] can mitigate this issue by capturing long-range dependencies.

However, one major challenge in road segmentation in adverse weather is the lack of ground-truth annotations for road pixels. A common approach in deep learning for handling lack of training data is domain adaptation (DA) [49, 94]. However, DA-based methods assume access to source datasets (clear weather dataset in our context) at all times which can be prohibitive in terms of storage, memory, data corruption and privacy concerns. Recently, [95, 96] have proposed source-free domain adaptation (SFDA) in which deep neural networks (DNNs) do not require access to the source dataset during the adaptation stage; instead, DNNs are pre-trained on a source dataset (clear weather dataset) and the pre-trained model is directly used to adapt to the unlabeled target domain (adverse weather dataset). Current methods for SFDA are used for image classification [97, 98, 99, 100] and may not work well for semantic segmentation due to the inherent differences between the classification and segmentation tasks. Moreover, many current SFDA methods use GANs to produce a “copy” of the original source domain distribution. In addition to being computationally intensive due to the difficulty in training GANs, image generation for segmentation requires GANs to capture contextual information and

semantic relationships between multiple objects and the background, which can be complicated in road scenes. As a result, prior SFDA techniques have not been used for road segmentation.

2.0.1 Main Contributions

We present a new approach for road segmentation in adverse weather conditions. Our approach is based on a novel algorithm for SFDA using self-supervised learning. We initialize our model with an auto-encoder baseline network using self-attention to generate a pre-trained model on the clear weather source dataset. Using self-attention improves the overall model by capturing long-range dependencies within the image. Our novel contributions include:

1. We present a novel two-step self-supervised SFDA approach called SS-SFDA. In the first step, our method uses entropy minimization to enrich the noisy pseudo-labels generated by the pre-trained auto-encoder. In the second step, we use a novel self-training method that generates pseudo labels in an *online* manner, as opposed to iterative self-training use by prior methods. We use curriculum learning to implement these two steps. This results in the following benefits compared to prior GAN-based approaches:
 - SS-SFDA directly exploits the pre-trained model and trains via curriculum learning to progressively bridge the domain gap between the pre-trained source domain and target domain and achieve faster training times.
 - Our online self-training scheme overcomes the saturation issues.
2. For heterogeneous adverse weather datasets, we propose a method that extends SS-SFDA by leveraging a few labeled images from the target domain to improve the accuracy using model distillation.

We have evaluated our approach on 6 datasets corresponding to real and synthetic adverse weather conditions. Overall, our mIoU score is 88 – 96% of prior supervised methods. We also improve the training time over prior SFDA approaches by 18 – 180 $\times$. Finally, our improvement in terms of mIoU over the best SFDA approach is 10.26% on real adverse weather data.

2.1 Related work

Road Segmentation Research in deep learning for semantic segmentation [18, 31, 44, 45, 78, 89, 101] has paved the way for segmentation in urban traffic scenes like CityScapes [21]. These methods have been extended for supervised road segmentation [91, 92, 102, 103]. Our approach based on self-supervised learning is complimentary to these methods.

Domain Adaptation and Source Free Domain Adaptation Traditional domain adaptation [50, 94, 104, 105, 106, 107] methods have achieved remarkable success in adapting models from one domain to another for clear weather conditions. However, these methods need access to the source data. Many domain specific solutions have been proposed for adverse weather conditions, including specific solutions for driving in rain, fog, etc. [6, 20, 108, 109, 110, 111] In contrast, we propose a generic method that neither relies on specific details from each domain, nor requires access to source data during the adaptation stage.

In source-image free domain adaptation (SFDA), a deep neural network (DNN) pre-trained on a source dataset is required to directly make predictions on the target domain dataset in an unsupervised manner. This approach has been primarily used for image classification tasks. In SFDA, generative approaches [97, 98, 99, 100, 112] are used to either emulate the source data by using the feature representations of the pre-trained model or create a negative source

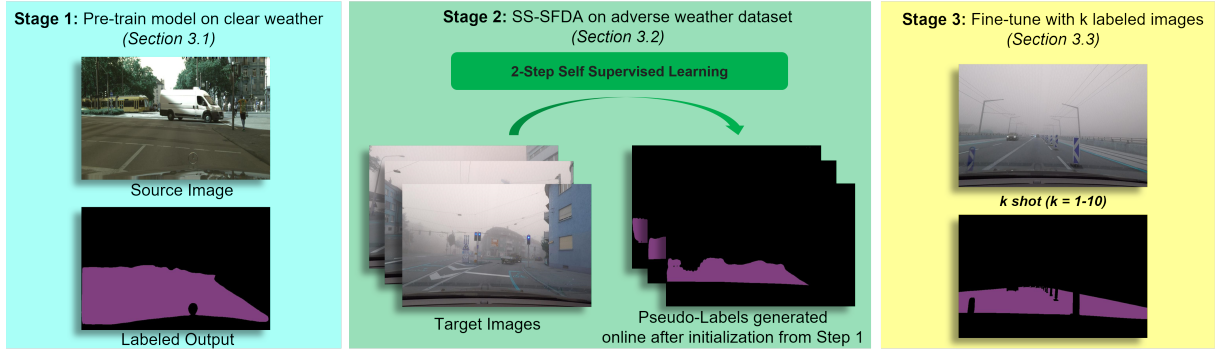


Figure 2.2: **Our Approach:** In stage 1, our model is pre-trained on a clear weather source dataset. In stage 2, our model is initialized with the pre-trained model from stage 1 and trained using our self-supervised algorithm, SS-SFDA, on the unlabeled adverse weather dataset. For heterogeneous weather datasets, we perform additional refinement steps based on model distillation (stage 3).

dataset during the pre-training stage. Non-generative approaches [95, 113, 114, 115] rely on computing adaptive class specific prototypes, and progressively learn on the target images. While generative approaches work well for classification tasks, they have not been effective for urban scene segmentation. This is because segmentation, being a pixel level task, requires the network to encode context, inter-class semantic relations, structure, and intricate boundaries [116, 117, 118]. In addition, training a generator for a complex segmentation task can lead to memory and computational overheads, and thereby adds to the difficulty of training GANs. While non-generative classification methods are easier to adapt for segmentation, the computation of prototypes results in similar issues due to the inherent differences between classification and segmentation tasks. Bateson et al. [119] explored SFDA in the context of medical segmentation and minimized entropy by the incorporation of class priors [119]. However, preservation of class priors does not extend well to urban road scenes because the number of road pixels in each image in the dataset can vary quite drastically. In contrast, we present a new method for SFDA using self-supervised learning that overcomes issues related to training and convergence that are frequent generative processes. Our method is designed to be complimentary to GAN-based SFDA.

Self-Supervised Learning Self-supervised learning has been used in semi-supervised learning [120, 121] and domain adaptation [122]. Most self-supervised learning methods are centered around the ideas of pseudo labeling [123, 124, 125, 126, 127], entropy minimization [49, 128], and curriculum learning [129]. Domain adaptation methods have access to source domain images. In contrast, we propose a novel self-training routine for SFDA and a completely unsupervised problem setting, where we have access to only a pre-trained model, and target domain images.

2.2 Method

In this section, we present our approach for source-image free domain adaptive (SFDA) road segmentation based on self-supervised learning. Our approach consists of three main components:

1. **Pre-training using the self-attention auto-encoder:** During this stage, we train the self-attention auto-encoder architecture on a clear weather dataset. This generates a model that encapsulates knowledge about road pixels.
2. **SS-SFDA : A Self-Supervised learning algorithm for SFDA:** During this stage, we initialize the model using the pre-trained model from the previous step and the target domain images. We use a combination of curriculum learning and entropy minimization to bridge the domain gap between the pseudo-labels and the target domain images. We first sort the target domain images in the increasing order of entropy, and create mini-batches of the dataset. The next task is to execute the following steps on each mini-batch to progressively self-train the model:
 - Optimize the model with an entropy minimization constraint to bridge the domain gap.

- Self-train the model by generating enriched pseudo-labels in an online manner.

3. **Few-Image Regularization:** For heterogeneous weather datasets, we use a very small number of labeled images (5 – 10) from the target domain to boost the performance of SS-SFDA via model distillation.

2.2.1 Pre-Training Baselines Using Self-Attention

The first step in SFDA is to pre-train a DNN on the source dataset for the task of road segmentation. In our case, the source dataset corresponds to traffic videos with clear weather conditions. While networks developed for semantic segmentation [31, 44, 89, 101] can be directly used for road segmentation, there is loss of context i.e. the model is unable to capture relationships between various semantic classes in the images like cars and roads, pedestrians and roads, sky and roads, etc. The loss of such context can lead to local ambiguities in classifying pixels [116, 117, 118].

Self-attention benefits from its capability to capture long-range dependencies between various regions of the image. Thus, using self-attention in road segmentation can allow neural networks to alleviate the degradation in performance due to loss of context. We use a simple autoencoder self-attention architecture that can be combined with any “off-the-shelf” segmentation network. We begin by taking an input RGB image $\mathcal{I} \in \mathbb{R}^{w \times h \times 3}$, which is passed through an encoder E to generate feature maps F_{en} ($F_{en} = E(I)$). Next, we apply self-attention SA [93] on these feature maps to obtain *attention* maps F_{sa} (same dimensions as F_{en}). These feature maps encapsulate the semantic relationships between various parts of the image. These feature maps F_{sa} are used to learn the final predictions $\mathcal{P}_{out} \in \mathbb{R}^{w' \times h' \times 1}$, which corresponds to the probability that each

pixel is classified as ‘road’. In the supervised setting where ground-truth labels $\mathcal{Y} \in \mathbb{Z}^{w' \times h' \times 1}$ are available, the network is optimized with a binary cross-entropy loss function,

$$L_{\text{CE}} = - \sum_{h,w} Y \log(P_{\text{out}}) + (1 - Y) \log(1 - P_{\text{out}}). \quad (2.1)$$

We use this pre-trained model for SFDA on the adverse weather datasets [1, 5, 6, 12, 130].

2.2.2 SS-SFDA

In this section, we describe our two-step self-supervised learning algorithm for unsupervised source-image free domain adaptive road segmentation in adverse weather conditions. We use the pre-trained model from the previous step. The pseudo-labels generated by the model are noisy leading to a domain gap between the pseudo-labels generated by the pre-trained model (on the source dataset) and the target domain images. Thus, directly self-training using the pseudo labels can hamper the performance of the model. To counter this, we propose an entropy minimization step which encourages the network to generate more accurate pseudo labels.

In addition, to bridge the domain gap between the pre-trained model and the target domain, we use curriculum learning [131, 132] in which the DNN is allowed to train on samples progressively in their increasing order of entropy of predictions. Given a probability map P denoting the probability that pixels are classified as road pixels, the entropy is computed as $-\sum P \times \log(P)$. This is because learning from samples with low entropy (low rain, for example) yields better pseudo labels on samples with higher entropy (high rain, for example) [109, 129, 129]. We create mini-batches of the dataset characterizing the difficulty of the images. For datasets which provide labels on the intensity (light rain vs heavy rain) of the weather condition, mini-batches

can be created directly. For other datasets, we sort the images in increasing order of the entropy and then split them into m ($m \sim 4-5$, determined by hyperparameter tuning) equal mini-batches. The model is self-trained on the mini-batches in a sequential manner. For the first mini-batch, the model is initialized with the pre-trained model. For subsequent mini-batches, our model is initialized with weights obtained by training the network on the previous mini-batch. For each mini-batch, the network is trained in two stages, as described below:

Step 1: Bridging the Domain Gap via Entropy Minimization The pre-trained model has a low entropy (i.e. high prediction probability or better generalization) on images that are similar (for example, similar geography, light rain, light fog) to source domain images [5] and vice versa. Thus, initializing the network with these pre-trained weights, followed by training by entropy minimization [59, 128] allows the network to generate enriched pseudo-labels. The inputs to the network are images from the target domain. Let the predictions of the network be denoted by $\mathcal{P}_{out} \in \mathbb{R}^{w' \times h' \times 1}$, corresponding to the probability that each pixel is classified as ‘road’. The cost function for entropy minimization is given by,

$$L_{EM} = - \sum_{\forall \text{pixels}} P^{h,w} \log P^{h,w}, \quad (2.2)$$

where $P^{h,w}$ is the probability that a pixel belongs to a class ‘road’ at a given location, and $-P^{h,w} \log P^{h,w}$ is the entropy.

Step 2: Online Self-Training Using Enriched Pseudo-Labels The network trained in Step 1 generates enhanced pseudo labels with high probability, and is thus a better representative of the target domain than the pre-trained source model. These enriched pseudo labels which can be

used to self-train the model further to improve performance. A traditional method of self-training using pseudo-labels is iterative [48] in which the network is trained to convergence (validation loss less than a given threshold) over multiple iterations. In each iteration, pseudo labels from the trained model in the previous iteration are used to set up the binary segmentation cost function. We show that iterative self training does not lead to any improvement in performance. This is because, the pre-training step, which is imperative for acquiring initial knowledge about road pixels since the problem is unsupervised in the target domain, causes the network to saturate quickly. Hence, we generate the pseudo labels in an online manner, *i.e.* the pseudo labels are generated from the network that is being trained. This allows the network to self-train from the improved pseudo labels as they are learnt.

The network in this stage is initialized with the weights obtained in Step 1. The inputs to the network are images from the target domain. Pseudo labels are generated in an online fashion from the network being trained as follows,

$$Y_{\text{pseudo}} = \begin{cases} 1 & \text{if } P^{h,w} \geq \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where $P^{h,w}$ is the probability that a pixel belongs to the class ‘road’ at a given location and τ is a threshold. The network is optimized with these pseudo-labels using a binary cross entropy loss term.

Dataset	Syn./Real	Weather
Rainy CityScapes [1]	Syn	Varying intensities (1mm - 200mm) of rain
Foggy CityScapes [1]	Syn	Varying intensities (750m - 30m) of fog
Foggy Zurich [6]	Real	Light and medium Fog
Dark Zurich [5]	Real	Twilight, Night
Raincouver [130]	Real	Rain, night
BDD [12]	Real	Snow, Fog, Rain, Night

Table 2.1: **List of datasets:** The second column categorizes the datasets as synthetic or real and the third column describes the images contained in the dataset.

2.3 Few-Image Fine-Tuning via Model Distillation

Some heterogeneous weather datasets like Raincouver [130] and Berkeley Deep Drive (BDD) [12] contain a mixture of adversities within the same image (for instance night+rain in Raincouver, see Table 2.1). Furthermore, these datasets are captured from different geographic conditions (i.e., source and target datasets may be from different regions). To make our model robust against such factors, we use ground truth labels for a few images (order of 5 – 10 images) from the target dataset in a final refinement step described below. In a nutshell, given a model trained on the unlabeled target dataset using SS-SFDA, and $k \leq 10$ labeled images from the target domain images, our goal is to learn enhanced feature maps for the target domain in the presence of adversarial factors such as mixtures of adversities and different geographical regions.

We empirically observe that directly fine-tuning the SS-SFDA model on the k images is sub-optimal due to overfitting. To prevent overfitting, we propose a model distillation [112] regularizer. Let the weights of the SS-SFDA model be denoted by $\omega_{\text{SS-SFDA}}$, and the weights of the model being currently trained be denoted by ω_{fewIm} . The cost function for model distillation is given by,

$$L_{\text{model-distil}} = C(\omega_{\text{SS-SFDA}}, \omega_{\text{fewIm}}),$$

where C represents a distance function such as MSE distance or L1 distance. In our benchmarks,

MSE distance works best.

The network is first initialized with weights of the SS-SFDA model. The model distillation term $L_{\text{model-distil}}$ with weight parameter $\lambda_{\text{model-distil}}$ is applied in conjunction with the binary cross entropy loss function (Equation 2.1) to constrain the probability predictions and ground-truth labels for k images. The $\lambda_{\text{model-distil}}$ term balances between extracting domain specific characteristics from the k images (such as mix of adverse weather, geographical features etc.) and prevents the weights of the model from diverging from the SS-SFDA weights (for better generalization). The overall equation follows as,

$$\mathcal{L}_{\text{overall}} = L_{\text{CE}} + \lambda_{\text{model-distil}} L_{\text{model-distil}} \quad (2.3)$$

2.4 Experiments and Results

We use the CityScapes dataset as the clear weather source domain. We conduct evaluation experiments on 6 datasets captured in adverse environmental conditions. We evaluate our model using four metrics: mean Intersection over Union (mIoU), Recall (or accuracy), Precision, and F1 score. All our models are trained using one NVIDIA GeForce GPU, and we implement the model using the PyTorch framework. We will make all code publicly available. The hyperparameters generalize across our experiments on all datasets. For the segmenatation model, we use the SGD optimizer with a learning rate of $2.5e - 4$, and momentum of 0.9 and weight decay of 0.0005. Dataset specific details are provided in the table below. Images are downsampled (by a factor of 2, where necessary) by bilinear sampling, and the corresponding ground-truth labels are downsampled by nearest neighbour downsampling.

Model	Acc.(%) (w/o. SA)	Acc.(%) (w. SA)
DeepLabv2 [31]	89.59	90.54 (+0.95)
DeepLabv2 (E-1) [31]	87.50	90.78 (+3.28)
DeepLabv2 (E-2) [31]	88.13	88.62 (+0.49)
DRN-D-105 [18]	83.92	85.32 (+1.40)
DRN-D-38 [18]	90.69	91.44 (+0.75)

Table 2.2: **Effect of self-attention:** We show for various backbone architectures that self-attention improves the accuracy. DeepLabv2 (E-1) and DeepLabv2 (E-2) denote the removal of 1 and 2 layers from DeepLab respectively. We select DRN-D-38 with self-attention (**bolded**) as the baseline for all further experiments.

2.4.1 Analysing Pre-Trained Self-Attention-based AutoEncoder

In this section, we analyse the performance of our self-attention auto-encoder model and benchmark its performance on various datasets in the supervised setting. We show that the usage of self-attention within various conventional semantic segmentation models to encode semantic relationships via capturing long-range dependencies improves performance over the corresponding baseline. We use the DRN-D-38 model [18] with self-attention in all further experiments. We benchmark the model on various weather datasets in the supervised setting. We use these supervised numbers in the following subsections to perform a comparative study of our self-supervised model for a summary). We also show that the self-attention autoencoder is comparable to the state-of-the-art on various datasets.

2.4.2 Results on Synthetic Datasets: Rain and Fog

We perform three evaluation experiments. Experiment A corresponds to testing the pre-trained CityScapes model on varying intensities of rain and fog. Experiment B corresponds to results obtained by training SS-SFDA without curriculum learning (*i.e.* by initializing with the CityScapes pre-trained model and training on higher intensities of rain and fog directly), and experiment C corresponds to results obtained by training SS-SFDA as proposed (*i.e.* with

Intensity	mIoU	Recall	Prec.	F1
I. Synthetic Rain [1]				
1mm	93.23	96.13	96.86	96.49
5mm	94.06	96.95	96.92	96.93
17mm	93.92	96.87	96.85	96.86
25mm	93.07	96.4	96.42	96.41
50mm	92.45	95.72	96.43	96.08
75mm	92.05	96.25	95.47	95.86
100mm	91.57	96.66	94.56	95.60
200mm	90.57	95.69	94.42	95.05
II. Synthetic Fog [1]				
750m	95.53	97.46	97.97	97.71
375m	94.74	97.08	97.51	97.30
150m	92.72	95.71	96.74	96.22
75m	91.64	96.12	95.15	95.63
50m	90.59	94.97	95.16	95.06
40m	90.21	95.68	94.03	94.85
30m	89.00	94.33	94.02	94.18
III. Real datasets				
Raincouver [130]	71.88	80.36	87.19	83.64
BDD [12]	89.19	92.86	95.76	94.29
IV. SOTA Comparisons				
Dataset	Method	mIoU	Recall	F1
CityScapes	FCN [78]	89.90	95.70	94.68
	CA [45]	88.93	94.05	94.14
	BoT [133]	76.55	85.82	86.72
	DeepLabv3 [101]	90.87	95.35	95.21
	s-FCN-loc [102]	91.04	96.11	95.36
	Zohourian [103]	86.34	96.76	92.44
	RBA [92]	96.00	98.13	98.00
	Ours	91.44	96.34	95.52
100mm Rain	SNE-Seg [91]	90.80	96.80	95.80
	Ours	91.57	96.66	95.60
100mm Rain	SNE-Seg [91]	90.50	93.00	94.47
	Ours	91.64	96.12	95.63

Table 2.3: **Benchmarking and comparing the self-attention based pre-trained model:** We train our self-attention model on various datasets in a supervised manner. These supervised numbers help us conduct a relative study of the performance of SS-SFDA, which is unsupervised in the target domain. Additionally, we observe that our pre-trained model is comparable to the state-of-the-art (Experiment IV). Best results are in **bold** fonts, second best results are in **blue**).

Intensity	Experiment	mIoU	Recall	Prec.	F1
I. Synthetic Rain					
1 mm	A	94.73	97.16	97.42	97.29
5 mm	A	94.09	97.01	96.90	96.95
17 mm	A	93.95	96.83	96.93	96.88
25 mm	A	93.01	96.30	96.46	96.38
50mm	A	89.82	93.71	95.58	94.64
75mm	A	87.15	90.92	95.45	93.13
75mm	B	88.08	93.94	93.38	93.66
100mm	A	82.97	86.88	94.84	90.69
100mm	B	86.67	92.88	92.84	92.86
200mm	A	65.98	68.94	93.90	79.50
200mm	B	80.25	86.96	91.22	89.04
200mm	C	81.58	88.55	91.20	89.85
II. Synthetic Fog					
750m	A	94.90	96.68	98.09	97.38
375m	A	92.71	94.60	97.88	96.21
150m	A	85.55	87.57	97.38	92.21
150m	B	89.74	94.59	94.59	94.59
75m	A	70.81	72.51	96.79	82.9
75m	B	80.67	83.01	96.62	89.30
75m	C	87.48	92.99	93.65	93.32
50m	A	57.29	58.92	95.39	72.85
50m	B	71.64	73.69	96.24	83.47
50m	C	85.47	90.75	93.63	92.16
40m	A	46.27	48.45	91.12	63.27
40m	B	52.62	53.94	95.57	68.96
40m	C	83.65	89.19	93.09	91.10
30m	A	35.04	38.05	81.59	51.90
30m	B	22.24	22.92	88.29	36.39
30m	C	80.82	85.96	93.11	89.39

Table 2.4: **Results on synthetic rain and synthetic fog [1]:** We evaluate our model under three settings - A: (Baseline) Testing the clear weather CityScapes model, B: SS-SFDA without curriculum learning, C: SS-SFDA. We show that the CityScapes model generalizes well to light synthetic rain and fog, and its performance degrades as the intensity of rain and fog increases, which is restored by SS-SFDA. **Comparison to supervised results:** On synthetic rain, SS-SFDA achieves 90.07%-90.8% and 92.53%-97.6% of supervised mIoU and recall, respectively. On synthetic fog, SS-SFDA achieves 90.8%-96.78% and 91.12%-98.82% of supervised mIoU and recall respectively.

curriculum learning).

Results on Synthetic Rain: The results of SS-SFDA on Synthetic Rainy CityScapes [1] are shown in the table. For low intensities of rain, the performance of the CityScapes model is more or less preserved. For higher intensities of rain, there is a degradation in performance. We observe that the decrease in performance is highest for 200mm rain, at 27.15%. For 75mm, 100mm, and 200mm, Experiment B imparts an improvement of 1.03%, 4.45%, and 21.63% respectively over the corresponding baselines. Experiment C leads to a cumulative improvement of 23.64% over the corresponding baseline. Furthermore, we demonstrate that the mIoU and recall of our SS-SFDA, which is completely unsupervised, is 90.07% – 90.80% and 92.53% – 97.60% of the counterpart supervised mIoU.

Results on Synthetic Fog: The results of our SS-SFDA on Synthetic Foggy CityScapes [1] are shown in the table. Similar to synthetic rain, the performance of the CityScapes model on light fog is more or less preserved. For higher intensities of fog, there is a degradation in performance. We notice that the degradation is very high for visibility distances less than 150m, and is the highest at 60.6% for 30m fog. Experiment B leads to an improvement of 4.89%, 13.9%, 25.04%, 13.7% on 150m, 75m, 50m, and 40m fog respectively. Direct application of the self-training algorithm on 30m fog (without curriculum learning) degrades performance since the generalization of CityScapes on 30m is very poor. Experiment C (Training SelfTr-Road Seg by curriculum learning i.e. progressively from 150m fog to 30m fog) cumulatively improves performance over the baselines (Experiment A) by 23.54%, 28.18%, 80.78% and 130.65% on 75m, 50m 40m, and 30m fog respectively. Furthermore, we demonstrate that the mIoU and recall of our SS-SFDA, which is completely unsupervised, is 90.8%-96.78% and 91.12%-98.82% of the supervised mIoU.

Benefits of Curriculum Learning: On a given dataset, the performance of our self-training algorithm depends heavily on the generalization capabilities of the pre-trained model used for initialization. Therefore, progressively initializing and training the model on increasing intensities of rain and fog will lead to the best accuracies. This implies that progressively training from low intensities to high intensities of rain and fog improves the quality of pseudo labels and prediction probabilities on high intensities of rain and fog. In simpler terms, for the 100mm rain dataset, a model trained on 75mm rain will work better than the CityScapes clear weather model. Similarly, for the 200mm rain dataset, a model trained on 100mm rain will work better than the 75mm rain model, which in turn will generalize better the CityScapes clear weather model. A similar intuition can be drawn for synthetic fog too. We validate this hypothesis in the table. The first and second columns correspond to the datasets that the model is trained and tested on respectively. We observe that curriculum learning progressively improves performance, which results in high quality pseudo labels with high confidence, a boon for self-training.

Generalization trends: We observe that our self-supervised SS-SFDA preserves the accuracy on the clear weather source dataset, CityScapes. The mIoU and accuracy of the fog model on clear weather CityScapes are at 96.9% and 99.00% of the supervised CityScapes model, respectively. The corresponding accuracy numbers for synthetic rain are 98.11% and 98.72%, respectively.

2.4.3 Results on Real datasets: Foggy Zurich and Dark Zurich

Analysis: Foggy Zurich The pre-trained CityScapes model fails to generalize to the real fog dataset due to the domain gap. In accordance with the curriculum learning strategy, the self-training algorithm is first applied on images with light fog, and then on images with medium

Training	Testing	mIoU	Recall	Prec.	F1
CS	100mm rain	82.97	86.88	94.84	90.69
75mm rain	100mm rain	86.31	92.63	92.68	92.65
CS	200mm rain	65.98	68.94	93.9	79.5
75 mm rain	200mm rain	78.97	85.86	90.77	88.25
100 mm	200mm rain	79.61	86.27	91.15	88.65
CS	75m fog	70.81	72.51	96.79	82.9
150m fog	75m fog	83.82	87.44	95.28	91.19
CS	50m fog	57.29	58.92	95.39	72.85
150m fog	50m fog	76.07	79.49	94.64	86.41
75m fog	50m fog	84.38	89.44	93.71	91.52
CS	40m fog	46.27	48.45	91.12	63.27
150m fog	40m fog	65.6	69.98	91.28	79.22
75m fog	40m fog	78.17	83.51	92.43	87.74
CS	30m fog	35.04	38.05	81.59	51.9
150m fog	30m fog	53.61	59.62	84.17	69.8
75m fog	30m fog	68.49	75.21	88.45	81.3

Table 2.5: **Curriculum learning improves performance by improving the accuracy of pseudo labels:** The first and second columns correspond to the datasets our model is trained and tested, respectively. Evaluating on 30m fog reveals that a model finetuned to 75m fog performs better than the counterpart 150m fog model, which in turn performs better than that the CityScapes model. Similar conclusions can be drawn for other intensities of rain/fog.

Experiment	mIoU	Recall	Prec.	F1
I. Baseline: CityScapes pre-trained model				
-	36.91	57.04	51.13	53.92
II. SS-SFDA on Light Fog; Init: CS model				
Step 1	59.1	68.27	81.48	74.29
Step 2	60.9	69.09	83.7	75.7
III. SS-SFDA on Medium Fog; Init: Light fog model;				
Step 1	73.34	76.22	95.1	84.62
Step 2	74.5	76.74	96.23	85.39

Table 2.6: **Results on the real fog dataset, Foggy Zurich.** We observe that SS-SFDA improves mIoU by 101.84% over the corresponding baseline. We notice that training the model on light fog followed by medium fog imparts the model with improvised self-training abilities, the mIoU improves by 22.3%. Additionally, the ablations on each minibatch indicate that the two step training procedure in SS-SFDA is indeed helpful.

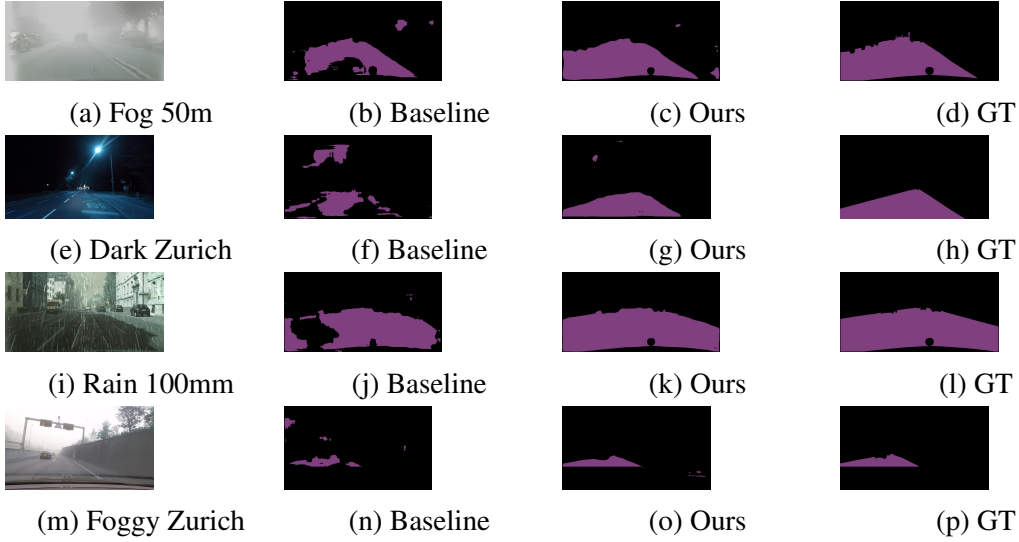


Figure 2.3: **Qualitative results.** Our model generates results that closely resemble the ground-truth (GT) compared to the baseline CityScapes pre-trained model. Purple indicates the segmented road region.

fog. We show results for each of the two stages of SS-SFDA to demonstrate how the two-step training procedure gradually improves performance. Training with light fog improves the mIoU by 64.99% over the corresponding clear weather baseline. Initializing the model trained on light fog, and fine-tuning on medium fog using SS-SFDA further improves the performance by 22.3%, thus resulting in a cumulative improvement of 101.84%. We do not show comparisons with supervised methods due to a lack of labeled datasets.

Experiment	Method	Source Data	mIoU	%(sup)
I. Synthetic Dataset: 100 mm rain				
Domain Adaptation	ADVENT [49]	✓	85.96	93.87%
Domain Adaptation	PrDA [95]	✓	86.12	94.04%
Self-training	CBST [126]		78.64	85.87%
SFDA	PrDA [95]		84.77	92.57%
SFDA	Ours		86.67	94.64%
II. Heterogeneous Real Dataset: Berkeley Deep Drive				
SFDA	PrDA [95]		75.32	82.43%
SFDA	Ours		83.05	93.11%

Table 2.10: **Comparisons against the state-of-the-art:** We adopt the state-of-the-art methods in self-training, domain adaptation, and SFDA classification to road segmentation, and present the results on Rain 100mm. We pick the best performing source-free model and conduct experiments on BDD. We observe an improvement of 10.26%.

Experiment	mIoU	Recall	Prec.	F1
I. Baseline: CityScapes pre-trained model				
-	54.9	63.89	79.6	70.88
II. SS-SFDA on Twilight Zurich; Init: CS model				
Step 1	69.11	74.75	90.15	81.74
Step 2	69.13	74.77	90.16	81.74
III. SS-SFDA on Night Zurich; Init: Twilight model				
Step 1	71.75	76.22	92.44	83.55
Step 2	72.18	76.51	92.72	83.84

Table 2.7: **Results on the night driving dataset, Dark Zurich.** We observe that SS-SFDA improves mIoU by 31.47. We observe that our two-stage training routine within the curriculum learning algorithm helps the network perform well.

Experiment	mIoU	Recall	Prec.	F1
I. SS-SFDA	52.42	93.96	54.25	68.78
II. Results on FewIm-FT				
k=1	54.71	90.7	57.96	70.74
k=2	56.74	90.45	60.36	72.4
k=10	63.85	79.39	76.54	77.93
III. Effect of model distillation (MD); k=10				
Without MD	57.08	70.99	74.45	72.67
With MD	63.85	79.39	76.54	77.93

Table 2.8: **Results on Raincouver:** SS-SFDA achieves 72.9% of supervised mIoU. We demonstrate the effectiveness of FewIm-FT over varying values of k in Experiment II. $k = 10$ results in an improvement of 21.8% over SS-SFDA, which is 88.82% of supervised mIoU. Additionally, we show the benefit of model distillation in Experiment III.

Experiment	mIoU	Recall	Prec.	F1
CS pre-trained model	70.28	74.49	92.54	82.54
SS-SFDA	75.29	83.3	88.7	85.92
FewIm-FT, k=10	83.05	93.06	88.53	90.74

Table 2.9: **Results on BDD:** SS-SFDA improves baseline mIoU by 7.12%. Experiments on FewIm-FT demonstrate an improvement of 10.30% over SS-SFDA for k=10, which is 93.11% of supervised mIoU.

Analysis-Dark Zurich: The CityScapes pre-trained model does not work well on Dark Zurich. SS-SFDA first trains on the twilight images, and then on the night images. Training with twilight images improves the mIoU by 24.91% over the corresponding clear weather baseline. Initializing the model with the model trained on twilight images, and fine-tuning on night images further improves performance by 4.41%, thus resulting in a cumulative improvement of 31.47%. We do not show comparisons to supervised accuracies due to the unavailability of training labels.

2.4.4 Heterogeneous Real Datasets: Raincouver and Berkeley Deep Drive

Raincouver [130] and Berkeley Deep Drive (BDD) [12] are complex datasets with images containing a mix of weather conditions in addition to scenes from different geographical regions. Raincouver consists of images captured under rain during the night. BDD consists of images in snow, fog, low light, glare, rain, etc. We show that SS-SFDA benefits from supervised finetuning with just 5 – 10 images. The models converge in 40 iterations, which takes 2 minutes to train on one NVIDIA GeForce GPU with 11GB memory.

Analysis-Raincouver: SS-SFDA results in an mIoU of 52.42, which is 72.9% of supervised mIoU. We show the effectiveness of the fine-tuning step. As k (number of supervised images used by the algorithm) increases, the performance of the model improves. Using just 1 image ($k = 1$) results in an improvement of 4.3% (over SS-SFDA). $k = 10$ achieves 88.82% of supervised mIoU. We demonstrate that model distillation improves mIoU by 11.86%. Hyperparameter tuning on the model distillation hyperparameter $\lambda_{\text{model-distil}}$ reveals that a value of 1.0 works best.

Analysis-Berkeley Deep Drive: The CityScapes model results in an mIoU of 70.28 (baseline). Training with SS-SFDA improves performance by 7.12% over the baseline, and is at 84.41%

of supervised IoU. In Experiment III, we pick k random images from the dataset and finetune the network using the fine-tuning step. In concurrence with our intuition, we observe that the performance improves as the number of images increases. For $k = 10$, we demonstrate an mIoU improvement of 10.30% over SS-SFDA, which is 93.11% of supervised accuracy.

2.4.5 Training Time and Convergence

The pre-training step helps our model converges in 1/6 epoch (the accuracies are similar over multiple random runs), thus bringing the training time down to 15 minutes on an NVIDIA GeForce GPU with 11GB memory. Generative SFDA models [97] take 30 epochs to converge, while prior work on self-training [126] and SFDA [95] applied (where feasible) to road segmentation converge in 3 epochs. Thus, we improve training time over prior SFDA approaches by $18 - 180\times$.

2.4.6 Comparisons with Prior Work

We adapt (where feasible) the state-of-the-art method in the following categories: self-supervised learning [126], source-free DA[95], and DA [49] to road segmentation for comparisons. Other methods like feature alignment by class-wise prototype learning [114], generative methods [97, 98, 99, 100, 112], and models that optimize source domain class priors [132] do not scale well to SFDA road segmentation. We performed evaluations using each of the methods on synthetic rain and observe that our model outperforms all prior methods. We finally train the best performing SOTA model on BDD for comparisons on heterogeneous real datasets which carry the highest level of difficulty. On BDD our method outperforms prior methods by 10.26%.

2.5 Conclusions, Limitations and Future Work

We propose a new method for road segmentation in adverse weather conditions using a novel self-supervised source-free domain adaptation approach. Through our evaluations on 6 real and synthetic datasets, we show that that our self-supervised model that has access to only a pre-trained clear weather model and unlabeled target images exhibits accuracy that is comparable to completely supervised models which have access to labels for all target domain images. In addition, we exhibit benefits in terms of faster training time and state-of-the-art performance.

There are a few limitations of our work. Currently, our approach is designed for binary segmentation and cannot perform multi-class segmentation. Extending the current method to multi-class segmentation would generalize the approach beyond road segmentation. Moreover, it would interesting to investigate if SFDA via self-supervised learning could be extended to other computer vision problems such as object recognition, image classification, scene understanding, etc.

Chapter 3: SALAD: Source-free Active Label-Agnostic Domain Adaptation

Deep learning solutions for visual applications such as semantic segmentation [49, 50], image classification, and document layout analysis [134, 135] require a large amount of annotated data. Two popular trends to deal with the lack of sufficient annotated data are Domain Adaptation (DA) and Active Learning (AL). In Active Learning (AL) [136, 137, 138, 139], the model mines and annotates samples within a fixed budget (*e.g.* 5% of the available corpus of unlabeled data [140]) to maximize the models performance. Typical active learning strategies include modelling diversity and uncertainty for efficient sampling [140, 141]. Domain adaptation [49] aims at transferring knowledge from a “source” domain to the “target” domain of interest.

An amalgamation of active learning and domain adaptation, Active Domain Adaptation (ADA) [137, 141, 142] has explored the use of annotated “source”-data from a related domain to adapt to the “target”-domain dataset, within a fixed budget of annotating the “target” data. The drawback of ADA is that it requires access to annotated source data, which might be prohibitive due to privacy issues or storage constraints [97, 99]. Recently, there have been explorations on source-free adaptation as well [115, 143]. However, these methods don’t generalize across label shifts and tasks, which can be enabled by annotating a small budget of samples from the target domain via active learning.

Main Contributions: In this work, we focus on the novel problem of Source-free Active Domain Adaptation, SF-ADA, where we have access to a pre-trained “source” network, but the source data is not available due to privacy concerns or otherwise. Further, an unlabeled target dataset, and a small budget for acquiring labels in the target domain is specified. Moreover, the target domain may also have a shift in the label space from the source domain. We propose **SALAD**, a generic novel framework for SF-ADA, which jointly optimizes sampling target data for annotation and source-free adaptation of the neural network to the target domain.

SALAD holistically addresses two key challenges of the problem setting via two complementary components: the Guided Attention Transfer Network (GATN) for source-free adaptation from the pre-trained network to the target domain and an active learning algorithm, H_{AL} , for mining samples from the target domain for annotation:

- *Source-free Adaptation:* GATN enables adaptation at the feature level from the pre-trained network to the target network (Figure 9.2). GATN uses a transformation network to modulate the features of the pre-trained network, in alignment with the target domain, followed by guided attention for selective distillation to the target network. The target network is guided in adaptation via samples selected through active learning.
- *Active Learning:* The effectiveness of GATN depends on the samples selected for annotation from the target dataset. While it is important to choose samples that are similar to the source distribution, we need to ensure that the chosen samples are informative to the network w.r.t. the target dataset. To do so, H_{AL} combines adaptability from the pre-trained network, as well as uncertainty w.r.t. the target network.

Our method does not attempt to emulate source data using generative approaches, which

is common in task-specific source-free domain adaptation [97, 144]. This makes our neural networks easy to train. SALAD has multiple benefits. (i) The architecture is task-agnostic, and can be applied across various visual tasks such as classification, detection and segmentation. (ii) Adaptation happens at the feature space (output of the network before the decoder). Thus, our architecture is label space-agnostic and can handle shifts in label space, where the source and target domains contain different number and types of classes (iii) The source data is not required for adaptation. Further, the pretrained source network is not required while testing and can be discarded after training.

We evaluate SALAD across three tasks. On classification datasets (MNIST, SVHN), we demonstrate that even without the source data, SALAD performs similar to or better than the prior active domain adaptation methods [137, 142] that use large amounts of annotated source data. Next, we evaluate on MNIST under 2 distinct cases of shift in output label space and show that SALAD is able to achieve 99.4% of the accuracy in case of no shift in the label space, thus establishing the effectiveness of our model in scenarios with label shift. Our experiments on the CityScapes dataset for semantic segmentation improves accuracy by 5.57% over fine-tuning. We also highlight the benefits of SALAD over other adaptation paradigms in Table 3.7. Finally, we conduct experiments on adaptation for document layout detection from PubLayNet to DSSE, where there is a shift in label space. SALAD imparts a relative improvement of 31.3% over fine-tuning of the target network on the small dataset.

Problem	Src. Data	Src. Model	Lab. Tar.	Un. Tar.
Semi Supervised DA (SSDA) [2]	Yes	Yes	Yes	Yes
Unsupervised DA (UDA) [50]	Yes	Yes	No	Yes
Source-Free DA (SFDA) [97]	No	Yes	No	Yes
Active DA (ADA) [142]	Yes	Yes	Yes	Yes
Source Free Active DA (SF-ADA)	No	Yes	Yes	Yes

Table 3.1: **Problem Settings:** We highlight various domain adaptation settings. Src. Data, Src. Model, Lab. Tar., and Un. Tar. refer to abundant labeled source data, Source Model, Scarce Labeled Target Data and Unlabeled Abundant Target Data, respectively.

3.1 Related work

Active Learning Active Learning (AL) aims to acquire a given small budget of labeled data while maximizing supervised training performance. Uncertainty-based methods select examples with the highest uncertainty [145, 146], using entropy [145], minimum classification margins [146], least confidence, etc. Diversity-based methods choose some points representative of the data, e.g. core-set selection [138, 147]. Recent approaches combine these two paradigms [136, 137, 148].

Domain Adaptation Domain adaptation aims to transfer the knowledge learned by a source domain model to an unlabeled target domain. Some of the existing works align feature spaces of the source and target domains by learning domain invariant feature representations by divergence-based measure minimization [149, 150], adversarial training [151, 152, 153], source or target domain data reconstruction [154, 155], image-to-image translation [156, 157] or normalization statistics [158, 159]. However, domain adaptation methods typically require access to annotated source data.

Active Domain Adaptation Active domain adaptation aims to adapt a model trained on source domain data to target domain by annotating a fixed budget of target domain samples. [160] introduced the task of ADA with applications to sentiment classification for textual data. They proposed a method employing a sampling strategy based on model uncertainty and a learned

domain separator. More recently, [142] studied ADA in the context of CNNs and proposed a method wherein samples are selected based on their uncertainty and targetedness, followed by adversarial domain adaptation. [59] proposed an algorithm that identifies uncertain and diverse instances for labeling followed by semi-supervised DA. [161] proposed a three-stage active adversarial training of neural networks using invariant feature space learning, uncertainty and diversity-based criteria for sample selection and re-training. [141] addressed the problem of lack of guarantee of good transfer-ability of features in domain adaptation. However, all of the above works use source domain data, which is prohibitive in terms of data privacy.

Source-free Domain Adaptation [112] introduced the paradigm of domain adaptation where source domain data is not available due to privacy issues and only a model pre-trained on the source domain data is available. Existing works employ generative approaches where the trained model is used to generate source samples using batch normalization [115] or energy-based methods [99] for classification task [95, 143, 162]. Others use a combination of distillation-based approach [163] or information maximization-based approach [144]. However, these methods do not consider using active learning to boost the performance, and typically do not generalize across tasks.

3.2 Method

We propose a novel method, SALAD , for the problem statement where we assume (i) a network, N_S , pre-trained (and frozen) on a source-domain, S (the source data is not available for adaptation) and (ii) an unlabeled target domain T from which we are allowed to annotate B images. The goal is to mine B images in total Tot_{AL} cycles, from the target domain for adapting

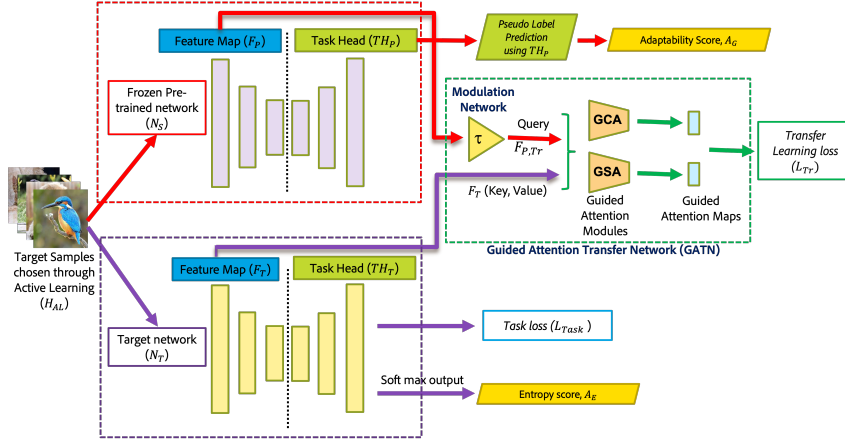


Figure 3.1: We present a generic method, SALAD, for the task of adapting from a pre-trained source network to a target domain with a small budget for acquiring annotation, under a possible shift in label space. SALAD consists of two complementary components: Guided Attention Transfer Network (GATN) and an active learning strategy H_{AL} .

network N_S to a target network, N_T which learns robust task-specific features for the target domain. For each cycle, $c = 1, 2, \dots, Tot_{AL}$ of active learning, the labeled set in target domain is denoted as $\mathcal{I}_{T,L,c}$ and unlabeled set as $\mathcal{I}_{T,UL,c}$. The proposed solution, SALAD, consists of two components: (i) Guided Attention Transfer Network (GATN) for adaptation from the pre-trained network (ii) active learning strategy, H_{AL} , for optimized acquisition of samples from the target domain.

3.2.1 SALAD Description

SALAD consists of two complementary components: an adaptation strategy, Guided Attention Transfer Network (GATN) and an active learning strategy H_{AL} . GATN combined with H_{AL} enables domain-specific learning for target network using chosen samples, where relevant domain-agnostic knowledge is transferred from the pre-trained source network via guided attention. GATN performs adaptation at the feature level.

The frozen pre-trained network, N_S is first split into a frozen feature encoder which generates

feature maps F_P and a frozen task head TH_P . Similarly, the trainable feature encoder of the target network, N_T is split into target feature map F_T and a trainable task head, TH_T . F_P and F_T are passed through the Guided Attention Transfer Network (GATN), which constrains the target network via a transfer learning loss L_{Tr} . The guided attention enables transfer of the domain-agnostic features from the pre-trained network, and discard domain-specific features. The GATN, as well as the pre-trained network, N_S , can be discarded after training, i.e. they are not required in the evaluation phase.

The active learning strategy, H_{AL} chooses samples using a combined metric of adaptability from source network and uncertainty in the target network, indicating the informativeness w.r.t. the target domain [137].

3.2.2 Guided Attention Transfer Network (GATN)

In this section, we describe the Guided Attention Transfer Network (GATN) in details. The GATN first transforms F_P to F_{P-tr} through a modulation network τ , which is a 4-layer fully convolutional network. Despite the transformation of the features from the pre-trained network, not all knowledge contained in F_{P-tr} is useful to the target domain. This is because τ is a CNN and has no filtering layers. It is important to transfer only domain agnostic features from F_{P-tr} . This is done by the guided attention networks. The transformed features F_{P-tr} and target network features F_T are then fed to two guided attention modules to compute attention across spatial and channel dimensions.

The guided attention modules consist of Guided Spatial Attention (GSA) and Guided Channel Attention (GCA) [45, 164] to compute attention across spatial and channel dimensions

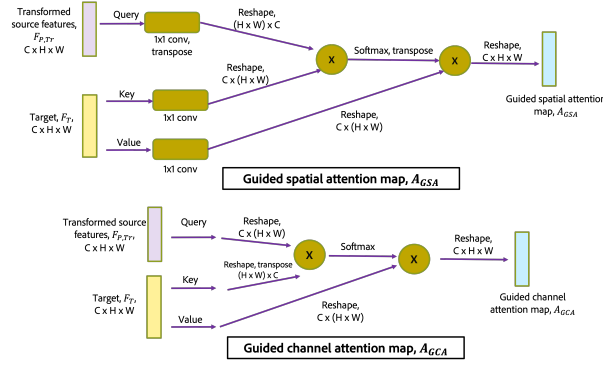


Figure 3.2: Overview of the computations for guided spatial attention and channel attention representations.

respectively. F_{P-tr} , along with F_T , are passed to spatial guided attention (GSA) and channel guided attention (GCA) modules. The attention modules compute alignment at the spatial and channel levels for the modulated source feature map F_{P-tr} and the target network feature map F_T . The target network contains finite domain-specific knowledge through the labeled target subset, that can be leveraged to guide adaptation. We build a guided attention module to do this, which builds on the math synonymous to self-attention [93]. We wish to reflect that the notion of guided attention is built on the idea described in co-attention [165]. Attention literature [45, 164, 166] describes the attention function as mapping a query and a set of key-value pairs to an output, where key, query, value, and output are all vectors. We denote F_{P-tr} as the query vectors and, key and value are assigned to target network feature maps F_T .

We incorporate the notion of guided spatial and channel attentions [164] using spatial and channel level feature vectors [45]. The goal of the Guided Spatial Attention (GSA) module (which generates attention representations A_{GSA}) is to highlight spatial regions of the transformed pre-trained network feature map, F_{P-tr} that align well with the target feature map, F_T . The goal of the Guided Channel Attention module (which generates attention representations A_{GCA}) is to highlight attributes (or channel level features at each spatial location) of transformed source

network features F_{P-tr} at each spatial location that align well with the target network feature maps F_T . Let the dimensions of the feature maps be $C \times H \times W$. In spatial guided attention, 1×1 convolution layers first transform the key, query, and value feature maps. This is followed by reshaping these feature maps to the shape $C \times (H \times W)$, which are used to compute attention weights, A_{GSA} . For channel attention A_{GCA} , reshaped feature maps of dimensions $C \times (H \times W)$ are used to compute attention maps without any convolution. Mathematically,

$$A_{GSA} = S(C_q(F_{P-tr}^\top) \odot C_k(F_T))^\top \odot C_v(F_T),$$

$$A_{GCA} = S(F_{P-tr} \odot F_T^\top)^\top \odot F_T. \quad (3.1)$$

where $\top$, S , and $\odot$ denote transpose operation, softmax and matrix multiplication operations, respectively. C_k, C_q, C_v denote trainable 1×1 convolutions followed by reshaping of the key, query, and value feature maps respectively.

3.2.3 Training Losses

We jointly train GATN (consisting of τ and the guided attention networks), and the target network, N_T with the following loss terms:

- *Transfer learning loss L_{Tr}* : L_{Tr} is computed as the attention weighted mean squared difference between transformed pre-trained feature maps F_{P-tr} and target network feature

maps F_T . Mathematically,

$$L_{Tr} = \sum A_{GSA} * [F_{P-tr} - F_T]^2 + \sum A_{GCA} * [F_{P-tr} - F_T]^2. \quad (3.2)$$

This loss is applied to all target images, labeled as well as unlabeled, and scaled by hyper-parameters $\lambda_{Tr,L}$ and $\lambda_{Tr,UL}$, empirically chosen to 0.1 or 0.01 depending on the task.

Hence, the total loss for labeled and unlabeled images is:

$$\lambda_{Tr,L} \sum_{\mathcal{I}_{T,L,c}} L_{Tr} + \lambda_{Tr,UL} \sum_{\mathcal{I}_{T,UL,c}} L_{Tr}$$

.

- *Task-specific loss L_{Task}* : To learn target domain-specific information, we compute the task specific loss for the target domain images, viz. multi-class cross entropy for classification and semantic segmentation, and focal loss for object detection. This is computed for both labeled and unlabeled target samples. For unlabeled samples, pseudo label is computed by thresholding the soft max output of N_S . Overall task-specific loss is:

$$\sum_{\mathcal{I}_{T,L,c}} L_{Task} + \lambda_{pseudo} \sum_{\mathcal{I}_{T,UL,c}} L_{Task}$$

.

- *Overall Loss*: The overall training loss is sum of the transfer and task losses computed above.

3.2.4 Active Learning Strategy H_{AL}

The training of GATN and the target network involves a task specific loss on the labeled subset of the target domain. Hence, the effectiveness of domain adaptation from the pre-trained network, and learning of domain specific features by the target network depend on the samples mined by the AL strategy H_{AL} . Thus, it is important to for H_{AL} to annotate samples (i) that facilitate adaptation from the pre-trained network as well as (ii) encode all aspects of target domain-specific knowledge.

- *Adaptability from the pre-trained network:* The prediction confidence of the pre-trained network, N_S , in computing the outputs for target domain samples implies the similarity of target samples to the source domain. Samples with high similarity facilitate knowledge transfer from the pre-trained network to the target network. We quantify ‘similarity’ by defining an ‘adaptability score’. For computing the adaptability score, we threshold the final softmax output of the pre-trained network, N_S , to compute the pseudo label map for the target samples. The task-specific loss is computed between the output of the pre-trained network and the pseudo-labels. The gradient score or adaptability score, A_G is the total l_2 norm of the gradient (without any gradient update - since the network is frozen) of the network N_S wrt the computed loss. Low gradient implies high confidence and hence high adaptability from N_S .
- *Uncertainty of the target network:* We need to choose target domain samples that provide the target network with exhaustive information about the domain. Hence it is important to choose target domain samples that the network is unsure about, which can be quantified

using the entropy of predictions from the target network. The softmax output of the target network provides the class-wise probability score map $\mathbf{p}$. This is used to compute the entropy score AE as $-\sum p \log p$. High entropy for a target sample indicates high uncertainty and hence should be selected for labeling.

A combination of the above two measures is maximized greedily to mine the samples for labeling:

$$H_{AL} = -\lambda_G \log A_G + \lambda_E \log A_E \quad (3.3)$$

where λ_G, λ_E , are binary variables (0/1) that toggle the metrics used for sampling. For the first batch of AL, we set $\lambda_G = 1$, and $\lambda_E = 0$, as the target network, N_T is not trained. For further epochs, we set $\lambda_G, \lambda_E = 1$.

3.2.5 SALAD Training Overview

We train the target network, N_T and GATN network together by mining samples from the target domain dataset in mini-batches using H_{AL} until the desired budget B is achieved. We use the SGD optimizer with a learning rate of $2.5e - 4$, and momentum of 0.9 and weight decay for 0.0005 for the training. The steps in the training routine are as follows:

- Initialize target network, N_T with parameters of the pre-trained source network, N_S
- For active learning cycles, c in range(0, Tot_{AL}):
 - Use the AL strategy H_{AL} , the pre-trained network N_S and the target network N_T to mine B/Tot_{AL} samples from the target dataset, and accumulate labeled target subset $\mathcal{I}_{T,L,c}$ and unlabeled target subset $\mathcal{I}_{T,UL,c}$.

- Create batches of images combining $\mathcal{I}_{T,L,c}$ and $\mathcal{I}_{T,UL,c}$. Optimize overall training loss for N_T and GATN for given number of epochs or till convergence.
- Sample for next active learning cycle using H_{AL} , using the current version of N_T and GATN.

3.3 Experiments and Results

We present results across the classification, detection and segmentation tasks, with standard evaluation metrics. Under classification settings, SALAD, even without access to annotated source data, performs similar to or better than (with a variance of 0.5% in accuracy) prior active domain adaptation methods [137, 142] that use large volume of annotated source data. Furthermore, we conduct experiments on MNIST under 2 distinct cases of shift in output label space, and show that SALAD can achieve 99.4% of the accuracy achieved when trained with all the labels, with only 5% sampling budget. Thus, SALAD can handle shifts in label space. Our experiments on CityScapes for semantic segmentation at various budgets reveal an improvement of 5.57% over fine-tuning (*i.e.* training the model without L_{tr}). Finally, we conduct experiments on document layout adaptation from PubLayNet to DSSE where there is a shift in output label space, SALAD imparts a improvement of 31.3% over fine-tuning (*i.e.* training the model without L_{tr}). We set number of cycles Tot_{AL} to be equal to 3 for MNIST, 1000 for SVHN, 50 for CityScapes.

Method	Source Data	B=100	B=200	B=300
Source only accuracy: 62.25				
O-ALDA [139]	Yes	79.10	81.40	82.70
CDAN [167] + Entropy [142]	Yes	93.10	94.60	95.00
CDAN [167] + BvSB [142]	Yes	94.20	95.00	95.90
CDAN [167] + Uniform [137]	Yes	90.00	94.00	94.50
CDAN [167] + BADGE [136]	Yes	92.90	94.90	96.50
SSDA MME [59]	Yes	93.00	95.00	95.50
AADA [142]	Yes	94.20	95.20	95.50
CLUE [137]	Yes	95.50	96.20	96.50
SALAD	No	91.64	95.96	97.16

(a) **Results on adapting from SVHN to MNIST:** With a budget of 300 images (0.5% of target data MNIST, last column of table), we show that, even without source data, SALAD outperforms prior work on active domain adaptation, using annotated source data. The fully supervised accuracy is 99.2%, we achieve accuracies comparable to fully supervised accuracies using just 0.5% of the dataset.

Method	1000	2000	4000	10000
Source only accuracy: 27.27				
FT+Uniform	68.0	76.2	80.0	84.7
FT+Entropy	68.0	75.1	81.2	87.8
FT+BADGE [136]	70.1	79.2	83.7	88.1
FT+Coreset [138]	70.0	78.8	82.8	88.2
FT+Margin [146]	71.0	78.0	83.2	88.4
FT+CLUE [137]	72.1	76.4	83.0	87.8
SALAD	74.2	82.2	86.6	88.6

(b) **Results on adapting from MNIST to SVHN:** We compare with prior methods on active learning, and demonstrate state-of-the-art performance. The fully supervised accuracy is 90.44%, we achieve accuracies comparable to fully supervised accuracies using just 1.8% of the dataset.

Table 3.2: **Results on digits classification in the shared label space setting**

3.3.1 Image Classification

We present our results on digits classification datasets under two settings: (i) shared label space, and (ii) shift in label space. In the shared label space setting, the label space that the pre-trained network was trained on and the label space of the target domains are the same. In the label space shift space setting, the target dataset contains labels not used for training the source network. We used ResNet-101 features for the experiment, consistent with the baselines. We set $\lambda_{Tr} = 0.01$, and λ_G to 1.0 and λ_E to 1.0 after first round of sampling. We set $\lambda_{pseudo} = 1.0$ for the dataset. We use feature heads from the penultimate layer of the ResNet-101 [168] classifier backbone.

Shared label-space setting, SVHN to MNIST: Concurrent with our intuition, the accuracy is better at higher budgets. When benchmarked at a budget of 300 samples, which is 0.5% of the total samples that MNIST contains, we observe that SALAD, even without any annotated source data, outperforms prior work on active domain adaptation which use large amounts of annotated

Method	Budget	Accuracy	Method	Budget	Accuracy
SDDA (WACV 2021) [99]	-	75.5	SDDA (WACV 2021) [99]	-	42.2
SDDA-P (WACV 2021) [99]	-	76.3	SDDA-P (WACV 2021) [99]	-	43.6
SALAD	0.16%	91.64	SALAD	0.18%	74.2

(a) SVHN to MNIST

(b) MNIST to SVHN

Table 3.3: **Comparing SALAD with prior art on SFDA on digits datasets.** Budget reflects the percentage of total number of target samples used for active learning. We demonstrate that our SF-ADA approach outperforms prior art on SFDA by a large margin using a very small proportion of annotated target samples.

source data (600000 images). Moreover, we observe that the accuracy of 97.16% with 300 images is 97.64% of fully supervised accuracy with 60k images.

Comparisons with SFDA methods SALAD outperforms prior art on SFDA by a large margin using a very small proportion of annotated target samples.

Comparisons with SFDA + AL methods. SALAD outperforms prior work on ADA. Since ADA (or AL + DA) methods are better than AL + SFDA methods, by transitivity, a holistic solution for SF-ADA, such as SALAD, is more beneficial than a naive combination of SFDA and AL.

Shared label-space setting, MNIST to SVHN: The complexity of SVHN is higher than that of MNIST, reflected in the source-only accuracy which stands at 27%. As per the prior work, we cap the net budget for mining samples via active learning at 10000 images, which is 1.8% of the total size of the dataset. We demonstrate state-of-the-art performance at varying budget of 1000, 2000, 4000 and 10000 images. Moreover, the accuracy at 10k images is 93.44% of fully supervised accuracy, which uses around 500k images.

Shift in label space, SVHN to MNIST: We present results on adapting from SVHN to MNIST under a shift in label space. In case 1, we train the source network on SVHN after removing samples corresponding to two classes (class 3 and class 9, randomly chosen). Similarly, in case 2, we remove 4 classes from the source dataset. Direct testing reveals that the accuracy for these

Class/Exp.	mean	0	1	2	3	4	5	6	7	8	9
Case 1: Remove the digits ‘3’ and ‘9’ from source SVHN											
Source only	56.88	69.20	86.80	79.10	0.00	53.80	95.70	41.00	78.20	63.00	0.00
B=100	88.29	97.80	98.90	94.60	83.80	94.30	96.60	84.90	91.00	93.80	48.20
B=200	96.27	98.50	98.60	98.10	94.30	96.70	97.80	97.90	93.60	92.80	94.40
B=300	96.61	99.10	98.70	98.10	95.00	97.50	97.90	98.00	91.30	95.40	97.10
Case 2: Remove the digits ‘7’, ‘5’, ‘4’, ‘1’ from source SVHN											
Source only	41.90	0.00	70.80	65.60	89.60	0.00	0.00	83.00	0.00	66.00	47.60
B=100	87.11	98.80	99.40	93.80	97.30	82.70	49.90	94.20	85.60	66.40	95.40
B=200	92.53	97.60	98.90	94.10	97.40	96.10	60.80	95.90	90.50	94.50	94.60
B=300	97.00	99.20	98.90	97.20	99.10	96.20	95.70	97.50	95.30	94.70	96.30

Table 3.4: **Results on adaptation from SVHN to MNIST, for label space shift.** We consider two scenarios: case 1 - source data does not contain digits ‘3’ and ‘9’, case 2 - source data does not contain digits ‘7’, ‘5’, ‘4’, ‘1’. SALAD, with H_{AL} and GATN, achieves 99.4% and 99.8% of the accuracy in the scenario with no label shift. The accuracy at a budget of 300 images, without label shift, is 97.16%

GATN Specification		AL Specification		Budget	Acc
$\lambda_{Tr,L}$	$\lambda_{Tr,UL}$	A_G	A_E		
Baseline source-free accuracy: 62.25%					
0	0	Yes	No	300	88.96
0	0.01	Yes	No	300	89.56
0.01	0.01	Yes	No	300	92.56
0.01	0.01	Yes	Yes	300	97.16

Table 3.5: Ablation Experiments for variations in transfer loss parameters ($\lambda_{Tr,L}$, $\lambda_{Tr,UL}$) and active learning strategies (A_G , A_E) from SVHN to MNIST.

classes is 0. Sampling with our AL strategy and gradually training with GATN gradually up to a budget of 300 images (again, 0.5% of target samples) restores accuracy to 96.61% and 97.00%, respectively, which is 99.4% and 99.8% of the accuracy achieved in the scenario with no label shift. The term A_E in H_{AL} ensures class balance, i.e., samples corresponding to classes not learnt by the pre-trained network are selected, along with classes with high prediction uncertainty. Thus, our method works well when there is a shift in label space.

Ablation experiments on H_{AL} and GATN. We present ablation experiments on adaptation from SVHN to MNIST. We set the net active learning budget at 300 images. Since the network has no prior knowledge about the target domain, the uncertainty metric can be applied only

from the second round of active sampling. In the first round of active sampling, we apply only the adaptability metric. In the first experiment, we study the impact of training the target network, without GATN, with all samples mined using just the adaptability score, the accuracy is 88.96%. This proves that even in the absence of adaptation from the pre-trained network, samples mined using the pre-trained networks' confidence score is advantageous since the target network is initialized with the weights of the pre-trained network. Next, we apply the distillation loss dictated by GATN only on the labeled subset and correspondingly set $L_{L,Tr} = 0.01$. We observe that knowledge transfer, in addition to sampling using the adaptability score improves performance by 0.6% points. This reinforces the quality of samples mined by the pre-trained network. Next, we apply the distillation loss to the unlabeled subset as well ($L_{Tr,UL}$), which leads to an absolute improvement of 3%. This is an indicator of selective distillation capabilities of GATN, where only useful features are distilled to improve performance. Finally, we experiment by using the uncertainty score, as well as the diversity score, to achieve an accuracy of 97.16%, an improvement of 34.91% points over the baseline and an improvement of 5.6% points over Experiment 3. Hence, the best adaptation using GATN is obtained when samples are mined intelligently (Experiment 4).

Ablation experiment on the modulation network τ . The accuracy on MNIST to SVHN, with 300 images, without the modulation network is 94.15% while the accuracy with the modulation network is 97.16%.

Experiment	mIoU	mAcc	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Neg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MBike	Bike
Source only	34.91	77.84	70.14	21.6	76.27	18.8	16.27	21.31	27.85	15.40	77.67	31.29	74.83	49.47	3.60	79.45	28.71	31.39	4.70	12.43	2.10
TENT [169] (Source-free)	38.9	-	87.3	39.0	79.8	24.3	19.6	21.2	25.1	16.6	83.8	34.7	77.7	57.9	17.8	85.0	24.9	20.8	2.0	16.6	4.5
S4T [170] (Source-free)	43.98	-	88.2	44.2	84.4	28.9	27.6	38.6	41.5	8.4	86.3	41.0	79.2	58.7	25.3	85.4	20.1	26.4	6.3	10.8	8.4
B=50	45.93	89.22	92.09	52.57	83.43	23.72	18.37	33.33	35.90	44.01	84.24	39.23	85.82	55.39	20.16	84.02	38.57	37.77	2.72	16.09	25.26
B=100	50.98	90.33	93.6	57.79	84.16	23.4	21.98	36.07	38.12	45.8	85.39	41.33	86.34	57.67	30.41	86.1	43.81	45.02	26.02	19.15	46.48
B=200	53.34	91.18	94.82	63.83	85.29	29.01	27.85	36.84	39.84	47.53	86.33	42.16	88.4	60.16	31.85	86.88	48.64	48.45	26.29	20.46	48.91
B=500	56.59	92.09	95.59	68.77	86.41	33.08	34.88	39.49	42.54	52.44	87.30	48.17	89.73	62.96	33.91	88.18	53.67	52.41	29.08	23.46	53.15

Table 3.6: **GTA5 to CityScapes Adaptation:** We show that SALAD imparts a relative improvement of 31.5%, 46.03%, 52.5% and 62.1% over the baseline source model with budgets of 50, 100, 200, and 500 images, with $3 \times -25 \times$ improvement on specific classes like “Bike”, “Train”, “MBike”, “Sidewalk”, etc.

3.3.2 Autonomous Driving: Synthetic to Real Segmentation on CityScapes

We conduct experiments on a dense pixel-level task, segmentation, where we adapt from GTA5 (25000 images) to CityScapes. To effectively transfer from GTA5 and address uncertainty of the target network while sampling, we set $\lambda_G = \lambda_E = 1$ from the second round of sampling. We set $\lambda_{T_r} = 0.01$. We use feature heads from the layer 3 of the underlying DeepLabv2 ResNet-101 backbone [101]. We present the experiments in the table. In the first row, we directly test the pre-trained GTA5 model, which gives an mIoU of 34.91. We next apply H_{AL} for batch active learning, and train the network using GATN after each round of sampling. A cumulative budget of 50 images, 100 images, 200 images, and 500 images leads to relative improvements (over the source only mIoU) of 31.5%, 46.03%, 52.5% and 62.1% respectively. Classes like ‘Sidewalk’, ‘Wall’, ‘Fence’, ‘Pole’, ‘Sign’, ‘Rider’, ‘Train’, ‘MBike’, ‘Bike’ have very low mIoU ($\sim 20\%$ or less) when directly transferred from the source model. We show that SALAD improves performance by $3 \times$ to $25 \times$. Class-wise comparisons with prior art on SFDA reveal that our AL heuristic H_{AL} strategically chooses classes like Bike and MBike with low confidence or high uncertainty to boost performance, while not compromising on performance w.r.t. transferable classes like road and sky.

Comparisons with SFDA methods: SALAD performs better with a budget of only 50 images

Method	mIoU
UBNA [171] (WACVW 2022)	36.1
UBNA+ [171] (WACVW 2022)	36.5
TENT [169] (ICLR 2021)	38.86
TENT + MS [169] (ICLR 2021)	36.89
SFDA (w/o IPSM) [163] (CVPR 2021)	41.35
SFDA [163] (CVPR 2021)	43.16
URMA [172] (CVPR 2021)	45.1
S4T [170] (ArXiv 2021)	43.98
S4T + MS [170] (ArXiv 2021)	44.83
SALAD	45.93

Table 3.7: **Comparisons with the state-of-the-art SFDA approaches for adaptation from GTA to CityScapes.** We show that SALAD achieves state-of-the-art mIoU with a small budget of 50 images, the mIoU obtained by naive finetuning [2] on 50 images is 39.5.

Budget B	$(\lambda_{Tr,L}, (0.1,0))$	$(\lambda_{Tr,UL}, (0.1,0.1))$	GSA	GCA	mIoU	H_{AL}	mIoU
50	44.62	45.93	No	No	39.50		
100	48.72	50.98	Yes	No	45.43		
200	50.49	53.34	No	Yes	45.45	A_G	50.29
500	53.46	56.59	Yes	Yes	45.93	$A_G + A_E$	56.59

(a)
(b)
(c)

Table 3.8: **Ablation experiments for adaptation from GTA5 to CityScapes.** In table (a), we study ablations on transfer loss for GATN, Table (b) shows ablations on GATN components at a budget of 50 images. In table (c), we ablate on the active learning heuristic, H_{AL} at a budget of 500 images.

(1.5%).

Ablation experiments We present ablation experiments. We study the effectiveness of GATN at various budgets. In the second column, we apply L_{Tr} to only the labeled subset, along with active learning using H_{AL} . The third column reflects mIoUs obtained by using our complete model, through training on labeled as well as unlabeled subsets along with active learning using H_{AL} . A comparison of the second and third columns indicates the benefits of selective transfer learning using GATN. Moreover, our model results in improvements of 16.27%, 16.92%, 13.24%

Experiment	mAP	Text	Caption	Figure	Table	List	Section
FT w/o SALAD	23.11	36.12	13.49	25.60	22.24	29.57	11.64
FT w. SALAD	30.36	44.59	11.61	35.57	24.80	37.48	28.18

Table 3.9: **Adaptation for document layout detection from PubLayNet to DSSE:** Fine-tuning with SALAD improves performance by 31.3% over fine-tuning without SALAD.

and 5.57% over the baseline numbers [2] obtained by naively fine-tuning the target network (with pre-trained weights initialization) without GATN, and by random sampling [2], at budgets of 50 images, 100 images, 200 images, and 500 images respectively. We also study the effectiveness of the different components of GATN, at a budget of 50 images. In the first experiment, we do not use either GCA or GSA [2]. Without GATN, our system will reduce to simple fine-tuning of the target network with the annotated target samples. GATN forms a bridge between the pre-trained network and the target network, and removing it would break the adaptation process. In the subsequent experiments, we show the impact of using channel and spatial features. We also demonstrate the effectiveness of using the fusion AL heuristic comprising of the adaptability score w.r.t. the pre-trained network as well as the uncertainty score w.r.t. the target network, as opposed to using just the adaptability score.

3.3.3 Document Layout Detection: DSSE

We adapt from the medical documents dataset PubLayNet to documents belonging to DSSE, a dataset containing magazines, receipts and posters. The documents in the two domains are quite different. Medical documents are written in a two-column format, with uniform text, figures and tables. In contrast, the target domain, DSSE, a new unseen dataset, is small (only 150 documents) and is extremely diverse. Moreover, PubLayNet has 5 classes, while DSSE has 6 classes. Hence, there is a *shift* in label space. Direct testing of PubLayNet on DSSE results in a mAP of 15.67. Since the dataset is very small, we do not apply H_{AL} , and instead directly use all 150 images for GATN. We use feature heads from the FPN of the underlying RetinaNet ResNet-101 backbone [173]. Fine-tuning without SALAD results in a mAP of 23.11, and fine-tuning

with SALAD improves performance by 31.3%, to 30.36.

3.4 Conclusions, Limitations and Future Work

We propose a generic source-free method, SALAD, for the task of adapting from a pretrained network to target domain, under a possible shift in label space, with the provision to annotate a small budget of samples in the target domain. SALAD consists of two complementary components: an active learning strategy H_{AL} , and GATN for effective adaptation and sampling. We evaluate the performance across 3 tasks and show improved or on par performance with methods using source data. One drawback of our method is that we use binary weights for the scores in H_{AL} , using learnable weights could be an interesting direction for future work. Moreover, we expect that SALAD can be extended to tasks and modalities where traditional domain adaptation has been useful. These include text classification, neural machine translation, sentiment analysis, cross lingual question answering, and domain stylization.

Part III

Learning from Little Supervised Data: Aerial Video Action Recognition

In the preceding section, we delved into autonomous driving; however, capturing data for autonomous aerial vehicles (AAVs) poses significant challenges. Despite limited supervised data availability, we investigate the feasibility of learning representations for AAVs. Specifically, we focus on the critical problem of aerial video action recognition.

This section comprises of two chapters. In the Chapter 4, we present an algorithm, Fourier Activity Recognition (FAR), for UAV video activity recognition. Our formulation uses a novel Fourier object disentanglement method to innately separate out the human agent (which is typically small) from the background. Our disentanglement technique operates in the frequency domain to characterize the extent of temporal change of spatial pixels, and exploits convolution-multiplication properties of Fourier transform to map this representation to the corresponding object-background entangled features obtained from the network. To encapsulate contextual information and long-range space-time dependencies, we present a novel Fourier Attention algorithm, which emulates the benefits of self-attention by modeling the weighted outer product in the frequency domain. Our Fourier attention formulation uses much fewer computations than self-attention. We have evaluated our approach on multiple UAV datasets including UAV Human RGB, UAV Human Night, Drone Action, and NEC Drone. We demonstrate a relative improvement of **8.02%** - **38.69%** in top-1 accuracy and up to **3** times faster over prior works.

In the Chapter 5, we present a learning algorithm, DifFAR, for human activity recognition in videos. Our approach is designed for UAV videos, which are mainly acquired from obliquely placed dynamic cameras that contain a human actor along with background motion. Typically, the human actors occupy less than one-tenth of the spatial resolution. DifFAR simultaneously harnesses the benefits of frequency domain representations, a classical analysis tool in signal processing, and data driven neural networks. We build a differentiable static-dynamic frequency

mask prior to model the salient static and dynamic pixels in the video, crucial for the underlying task of action recognition. We use this differentiable mask prior to enable the neural network to intrinsically learn disentangled feature representations via an identity loss function. Our formulation empowers the network to inherently compute disentangled salient features within its layers. Further, we propose a cost-function encapsulating temporal relevance and spatial content to sample the most important frame within uniformly spaced video segments. We conduct extensive experiments on the UAV Human dataset and the NEC Drone dataset and demonstrate relative improvements of 5.72% – 13.00% over the state-of-the-art and 14.28% – 38.05% over the corresponding baseline model.

Chapter 4: FAR: Fourier Aerial Video Recognition

Deep learning techniques have been widely used for activity recognition [7, 24, 174]. Video analysis of scenes captured using UAV cameras [22, 23] is much harder than activity recognition in ground-camera datasets [24, 25]. In these UAV videos, the object of interest, i.e. the human actor (any individual appearing in the video performing scripted or non-scripted actions), is typically much smaller in terms of number of pixels or the area than the corresponding background, and thus provides less knowledge than a front view capture. Moreover, it is harder to capture and label UAV videos. Overall, there are fewer and smaller labeled datasets of aerial videos, as compared to ground videos. For instance, ground-camera datasets like Kinetics-400 [24] contain 306,245 videos while the recent UAV-Human [22] database has 22,476 videos.

Given that the size of the human actor in UAV videos is much smaller than the corresponding background, a neural network trained on these datasets may learn to infer more from the background [175] than the human actor. While both background and context are important [176], the network must learn to first identify the human actor and the corresponding action, and then deduce relations of the human actor with the surroundings in a judicious manner. In the absence of annotated detection boxes that can demarcate the human actor, the network needs to be able to differentiate the moving human actor from the background in an intrinsic manner. One approach is to detect the object of interest via object detection [177]. However, action recognition models that heavily

rely on localization of the human actor require near to perfect object detection accuracy [178]. While it is practically not feasible to annotate all datasets for object detection, object detectors trained on ground camera datasets will not generalize well UAV videos due to domain gap issues [179, 180, 181]. Domain adaptation solutions do not lead to perfect generalization yet.

On the other hand, traditional optical flow [182] techniques require hundreds of optimization iterations each frame, and split the network into RGB and motion streams which increases computation and model parameters [183]. Low computation alternatives such as deep learning based optical flow [184, 185, 186], motion feature networks [187] and ActionFlowNet [188] are inferior in performance compared to optical flow. Techniques such as background subtraction [189] and motion segmentation [190] are not very promising either [191, 192]. Thus, the network needs to learn to automatically *disentangle* [193, 194] object feature representations from the corresponding entangled state containing both the object and background information.

In addition to object-background separation, it is important for the network to acquire knowledge [174] about the context, relationships between the object, and background and intra-pixels correspondences as well. Self-attention [93, 166] can model this information by capturing long-range dependencies within an image/ video. Prior work on attention based video activity recognition [174, 195, 196] has seen two classes of self attention networks by either directly applying self-attention on convolutional layers or using self-attention as the building block. Mathematically, the core step in the computation of self-attention is matrix multiplication, which makes it computationally expensive.

4.0.1 Main contributions

We present a novel method, FAR , for UAV video action recognition. The design of FAR in the frequency domain is motivated by the fact that frequency spectrums contain knowledge about a signals' characteristics that are not easily interpretable in the time domain. Our novel components include:

- We propose a novel Fourier Object Disentanglement method (FO) to bestow the network with the ability to *intrinsically* recognize the moving human actor from the background. FO operates in the frequency domain dictated by the spectrum of the Fourier transform corresponding to the temporal dimensions of the video. It characterizes the motion of the human actor based on the magnitude and rate of temporal change of feature maps that encode information about the spatial pixels of the video. The amplitudes at each spatial-temporal location of the feature maps are innately representative of dynamic salient, static salient, dynamic non-salient and static non-salient regions, in the same order of relevance. This also empowers the network to handle videos with moving background pixels and dynamic cameras.
- We present Fourier Attention (FA) to encapsulate context and long range space-time dependencies within a video. Fourier attention works in the frequency domain corresponding to the space-time dimensions of the video, and emulates the benefits of self-attention. The time complexity of FA is $O(n^2 \log n)$ as opposed to $O(n^3)$ for traditional self-attention, and the accuracy of Fourier-attention approximates that of self-attention.

Moreover, such a representation promotes global mixing. FAR has multiple benefits. (i) It

elegantly exploits the mathematical properties of Fourier transform to achieve the desired objectives of object background separation and context encoding by performing fewer computations than traditional methods. (ii) It is parameter-less, i.e., it does not have any learnable layers/ parameters. (iii) FAR can be embedded within any 3D action recognition network such as I3D [7, 24] to achieve state-of-the-art performance. (iv) FAR converges faster than the corresponding 3D action recognition backbone.

We experimentally demonstrate that FAR outperforms prior work by 8.02% – 38.69% performance across multiple UAV datasets including UAV Human RGB [22], UAV Human Night [22], Drone Action [23], and NEC Drone [197]. We compare with the state-of-the-art Fourier method, efficient attention method and self-attention based transformer methods and demonstrate accuracy, computation and memory benefits.

4.1 Related Work

Action Recognition: Action recognition is a well studied topic in computer vision. The emergence of large-scale ground-camera videos datasets [24, 25, 198] has led to development of deep learning techniques for action recognition. We refer the reader to [199] for a survey on action recognition. Broadly speaking, three classes of network architectures have been proposed for action recognition. The first [200, 201, 202, 203, 204] builds on the two-stream theory in cognition to model space and time separately. The second [7, 24, 174, 205, 206, 207] models space and time jointly via 3D CNNs. The third class includes transformer-based architectures [174, 208, 209, 210, 211]. These transformer-based solutions are built on self-attention [93, 166] and have high computational complexity. In the interest of optimizing GPU memory, frame sampling strategies [212, 213,

214, 215] for video action recognition have been proposed. The above solutions are focused on challenges pertaining to action recognition in ground-camera videos. However, UAV video action recognition is much more difficult.

UAV Action Recognition: UAV video databases [22, 23, 197, 216] have been used to develop solutions [217, 218, 219, 220] for UAV action recognition. However, these solutions are directly based off techniques designed for ground-camera datasets [24, 198], where the size of the object is comparable to the background. Moreover, for ground camera videos, an auxiliary guidance factor based on object detection [177] is a viable option. However, these assumptions do not hold true in UAV videos [221, 222]. **FFT and Deep Learning** FFT has been immensely used in traditional image [223, 224] and video processing [225, 226] applications. Fast Fourier Transform (FFT) has been recently used in deep learning methods. One of its first applications was to accelerate convolution operations [227]. Incorporating FFT between NN layers [228, 229, 230] instead of CNNs to transform the feature space to the frequency domain, and aid global mixing of knowledge, has been used to improve accuracy for image classification, detection and ground-camera action recognition. An interesting application of FFT includes image stylization [231] as a guiding factor for domain adaptation. Most recently, FFT was used to naively replace self-attention layers [166] for NLP applications.

Efficient attention Methods to improve memory efficiency of transformers include modifications in matrix multiplication [232], low rank approximations [233], kernel modifications [234] for linear time complexity [235, 236, 237, 238]. While most of these solutions are focused on NLP and image-based computer vision tasks, EA [232] demonstrates results on temporal action localization and STAR [239] performs skeleton action recognition. The former can be regarded as a localization task w.r.t. the temporal dimension while the latter uses pose information making

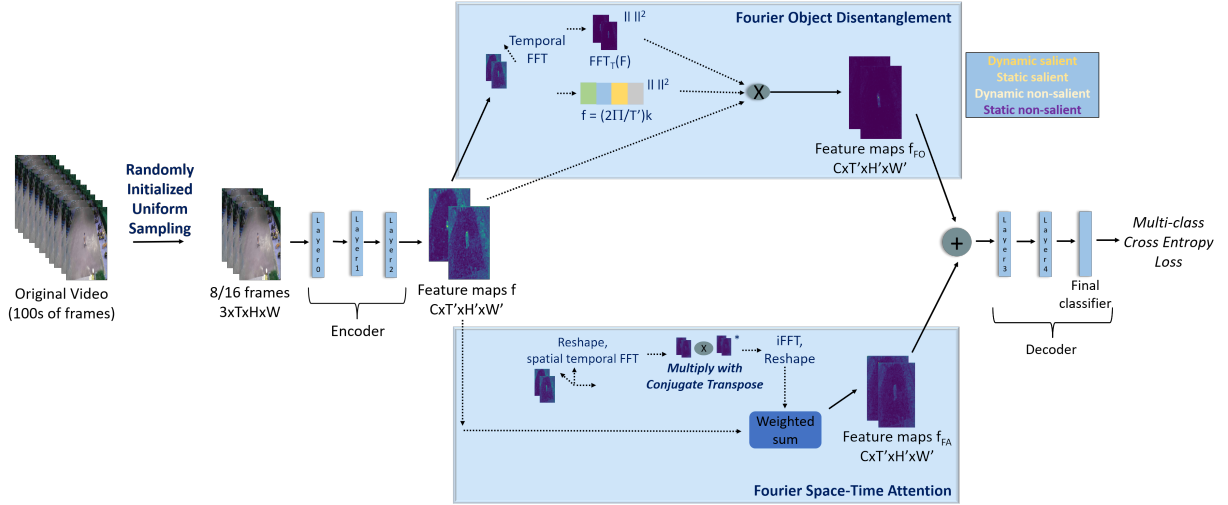


Figure 4.1: **Fourier Object Disentanglement (FO) and Space-Time Fourier Attention (FA):** FO empowers the network to intrinsically separate out the moving human agent from the background, without the need for any annotated object detection bounding boxes. This enables our network to explicitly focus on the low resolution human agent performing action, and not just learn from background cues. FO inherently characterizes salient and non-salient, and static and dynamic regions of the scene via the amplitudes of the feature maps it computes. FA elegantly exploits the mathematical properties of the Fourier transform to imbibe the properties of self-attention and capture contextual knowledge and long-range space-time dependencies at a much lower computational complexity.

the task of classification easier. None of these solutions are customized to UAV action recognition which brings forth different challenges.

4.2 Method

In this section, we describe our approach. We design two novel methods to decipher the human actor performing action, and encode context. Fourier Object Disentanglement (FO) disentangles the object from the background in an automatic manner. Fourier Space-Time Attention (FA) imbibes the properties of self-attention to capture long range space-time relationships at a lower computational cost. These modules can be embedded within any state-of-the-art 3D video recognition backbone such as I3D [24] or X3D [7] for improved action recognition. We now describe the methods in detail.

4.2.1 Fourier Object Disentanglement

We present a Fourier Object Disentanglement (FO) method to automatically separate the human actor from the background. Movement of the human actor in the scene can be characterised by temporal change of feature maps encoding spatial pixels (across space dimensions $H \times W$) in the video frames. The rate and magnitude of change of a signal can be quantified by amplitude of a signal at different frequencies. Thus, to identify the movement, we first transform the feature maps to a temporal frequency space. We perform this computation using 1D Fourier transform along the temporal dimension. Specifically, let $f(c, t, h, w) \in C \times T' \times H' \times W'$ denote the feature maps on which FO is applied, where C is the number of channels and T' and $(H' \times W')$ denote the temporal and spatial dimensions of the feature maps, respectively. The amplitude of the temporal Fourier transform at the frequency $-2\pi k/N$ is:

$$\mathcal{F}_{\mathcal{T}}(f)(k) = \sum_{n=0}^{n=T'} f(c, t, h, w) \times e^{-2\pi kn/N}, \quad (4.1)$$

which can be computed efficiently using the FFT algorithm [240]. $\mathcal{F}_{\mathcal{T}}(f)(k)$ mathematically represents the amplitude of the temporal signal at every spatial and channel location of the feature map f , at various frequencies. Intuitively, high frequency in the temporal dimension corresponds to the movement, and low frequency represents static regions of the scene. Therefore, regions corresponding to the moving human actor should have higher amplitude of Fourier transform at high frequencies. To infer the presence of the moving human actor at various spatial locations, we encapsulate the relationships between amplitudes and frequencies by multiplying the L2-norm-square of the amplitude at each frequency with the L2-norm-square of the frequency itself.

L2-norm ensures that frequencies and amplitudes are positive. L2-norm-square amplifies high amplitudes of the Fourier transform of the signal at high frequencies and suppress low amplitudes at low frequencies for disentangling dynamic regions of the scene. The frequencies, in order, are: $fr_k = [e^{-2\pi k/N}]$, $k = 1 \dots T'$.

Note that the frequencies are independent of the input video. Thus, the dynamic mask $\mathcal{M}_{FO}$ can be represented as

$$M_{FO} = \|F_T(f)(k)\|_2^2 \times \|fr_k\|_2^2, \quad (4.2)$$

where $|a|_2^2$ is L2 norm-square of a vector $|a|$. M_{FO} disentangles (or amplifies) parts of the scene corresponding to moving pixels. This may include moving background (and camera motion) in addition to moving human actor. Our next task is to use M_{FO} to demarcate moving object pixels from **moving background pixels**.

To further separate out only the moving actor, we capitalize on the activation maps f computed by the model. While not perfect, the activations at salient regions of the scene are higher than those at the non-salient regions. Hence, the final object disentangled representations can be represented as a dot product of M_{FO} and network features f , which amplifies dynamic, salient regions of the scene. Mathematically,

$$F_{FO} = f \odot M_{FO}. \quad (4.3)$$

According to this formulation, dynamic salient regions are amplified the most, and static non-salient regions are heavily suppressed. The amplitude at static salient regions and dynamic non-

salient regions is lower than the amplitude at dynamic salient regions. Due to the l_2 operation in the computation of M_{FO} and linear application of f , static salient regions have a higher amplitude than the dynamic non-salient regions. Thus, the ordering of amplitudes that is formed as: dynamic-salient $>$ static-salient $>$ dynamic-non-salient $>$ static-non-salient, in concordance with the relevance for decision making for action recognition. Thus, static as well as dynamic background regions have lower amplitudes than static and dynamic regions of the object executing action.

Time complexity: The time complexity of FO depends on the time complexity of 1D FFT, which is $n \log(n)$, for an n -element input vector. Consider the classical case [168] where the temporal and spatial dimensions at the mid level feature representations is half and one-fourth of the number of frames sampled and spatial dimensions of the image respectively. The number of FFTs that need to be computed is $C \times (H/8) \times (W/8)$ where C, H, W correspond to the number of channels at the mid-level, and spatial dimensions of the image. Therefore, the total time complexity is $C \times (H/8) \times (W/8) \times (T/2) \log(T/2)$.

4.2.2 Space-Time Fourier Attention

Consider a scene that depicts a human actor swimming in a swimming pool. Here, it is important to decipher the relationship between the human actor and the pool. While explicit modeling of correspondences between different pixels illustrating pose, orientation, and joint movements may not be necessary, it is crucial for the neural network to inherently capture this knowledge. Space-time self-attention for video action recognition [174, 195, 196] is capable for extracting this knowledge, but comes at the cost for expensive matrix multiplications.

We propose Fourier Space-Time Attention (FA) for acquiring knowledge about the long-range space-time relations within a video. Fourier attention approximates self-attention in an elegant fashion at a reduced computational cost. To understand the mechanics of Fourier attention, we first succinctly present self-attention [93]. The inputs to self-attention are key, query and value vectors which are representations obtained by 1×1 convolutions using a common input feature map. Vaswani et al. [166] describe the computation of self-attention as “a weighted sum of the values, where the weight (or sub-attention) assigned to each value is computed by a compatibility function of the query with the corresponding key.” Key, query and value are 1×1 convolution layers transforming the input feature maps. Mathematically, with x representing the input feature maps, and $\odot$ denoting matrix multiplication,

$$\text{Attention} = \text{Value}(x) \odot [\text{Query}(x)^T \odot \text{Key}(x)]^T \quad (4.4)$$

Our space-time Fourier attention method proceeds as follows. The first step is to obtain a representation equivalent to the key-query computation, termed *Fourier sub-attention*. Fourier sub-attention is motivated by autocorrelation, which is the correlation coefficient between different parts of the same signal. We define *Fourier sub-attention* as the element-wise product of the Fourier transform of feature maps with the conjugate transpose of the Fourier transform of these feature maps. To compute this space-time Fourier sub-attention, we reshape the video feature maps f to a 3D representation $C \times T' \times (HW)$, which are transformed to the frequency domain via 2D Fourier transform along the space and time axes as follows:

$$\mathcal{F}_{ST}(f)(m, n) = \sum_{h, w} f(c, t, h, w) e^{-2\pi m h / M} e^{-2\pi n w / N}. \quad (4.5)$$

computed efficiently using the FFT algorithm [240]. FFT is a representation of the signal as a whole at a wide spectrum of frequencies, and enables inherent and exhaustive global mixing between various spatial and temporal regions of the video. The space-time Fourier sub-attention $\mathcal{A}_{ST}$ in the Fourier domain is simply the element wise multiplication between $\mathcal{F}_{ST}$ and its complex conjugate $\mathcal{F}_{ST}^*$:

$$\mathcal{A}_{ST} = \mathcal{F}_{ST} \times \mathcal{F}_{ST}^* \quad (4.6)$$

Next, we compute the inverse FFT ($\mathcal{IF}$) of $\mathcal{A}_{ST}$ to obtain the correlations in the time domain, and reshape to $C \times T' \times H' \times W'$. These sub-attention “weights” are then used in a dot product (or element wise multiplication) with the input feature maps f to compute the final space-time Fourier attention maps f_{FA} . A scaling factor λ_{FA} , chosen empirically to be 0.01, scales these Fourier attention maps, which are then sum-fused with the input feature maps. Mathematically,

$$f_{FA} = F + \lambda_{FA} \times \mathcal{IF}(\mathcal{A}_{ST}), \quad (4.7)$$

Time complexity: Traditional self-attention [166] requires the model to perform two matrix multiplications. In the first matrix multiplication of self attention, we multiply the query matrix ($THW \times C$) with the key matrix ($C \times THW$). The time complexity is $O(C \times THW \times THW)$. In the second matrix multiplication, we multiply the value matrix ($C \times HWT$) with the attention matrix ($HWT \times HWT$). The complexity of this stage is $O(C \times HWT \times HWT)$. Hence, the overall time complexity of space-time self-attention [174] is $O(HWT \times HWT \times C)$.

In contrast, our Fourier attention solves the problem via one 2D FFT and one 2D iFFT. 2D FFT is computed on a matrix of dimensions $HW \times T$. The number of 2D FFTs that need

to be computed is equal to the number of channels (C). Hence, the complexity is $O(C \times HWT \log(HWT))$. The complexity of 2D FFT and 2D iFFT are the same. Therefore, the overall time complexity of Fourier attention is $O(C \times HWT \log(HWT))$. Clearly, Fourier attention is much more efficient than self attention. In terms of accuracy, space-time Fourier attention is comparable to space-time self-attention [174].

4.2.3 Mathematical Analysis

Lemma: Given an input matrix A , Fourier attention as well self-attention [166, 174] encapsulate long-range relationships for global mixing by computing outer products.

Proof: Without loss of generality, let $[a_{ij}]$ denote the elements of a square matrix A (with dimensions N) in $2D$. f, g, h represent 1×1 convolutions for key, query, value computations in self-attention. The self-attention matrix S_{mn} is:

$$S_{mn} = \sum_{l=1}^N h a_{ml} \sum_{k=1}^N [g a_{lk} \times f a_{kn}] \quad (4.8)$$

Fourier attention F_{mn} can be expanded using the above equations. f, g, h are 1×1 convolutions, and that the exponential terms span the entire spectrum of frequencies lets us define f, g, h for Fourier attention. Thus, the equation for Fourier attention can be simplified as:

$$F_{mn} = \sum_{b=1}^N \sum_{c=1}^N h_{mn}(b, c) a_{mn} \times \left\{ \sum_{j=1}^N \sum_{i=1}^N f_{mn}(b, c) a_{ij} \times g_{mn}(b, c) a_{ij} \right\}$$

In self-attention, f, g, h are learnable. In contrast, in Fourier attention, f, g, h are pre-defined by the Fourier spectrum. Nonetheless, they exhaustively cover the Fourier spectrum. Moreover, the

terms involved and the structure of computations (multiplications followed by summation) in the equations are similar, both promote global mixing and encapsulate long-range relationships.

4.3 FAR: Activity Recognition in UAVs

We present FAR, a network for video action recognition in UAVs. FAR samples 8-16 frames from the input video by using randomly initialized uniform sampling, described in Section 4.3.1. These frames are passed through the first few layers of the 3D backbone network (or encoder) to generate feature maps f . These features contain entangled object and background information along the space-time dimensions. The choice of this intermediate layer in the backbone network that extracts feature maps f is a careful trade-off between the spatial-temporal resolutions needed for FAR to work well and the amount of knowledge contained in the networks' layers. We describe this choice in detail in this section, as well as present ablation experiments to justify our choice.

The *Fourier Object Disentanglement* module, and the *Fourier Space-Time Attention* module act on f , in parallel, to generate f_{FO} and f_{FA} , respectively. f_{FO} and f_{FA} are sum fused and passed through the remaining layers of the neural network to generate the final action classification probability distribution, used in a multi-class cross entropy loss term with the ground-truth label for back-propagation.

Incorporating FO within the 3D backbone: Typically, to encapsulate temporal movement at each spatial location, we need to ensure that the spatial temporal dimensions of the feature map is not too small. Thus, it is useful to perform this operation using mid-level features (output from the middle layer of the network) that strike a fine balance between generic features that capture

context, and focused high level features (at output layer).

Incorporating FA within the 3D backbone: After FO, the network does not contain any background signal. Hence, Fourier attention needs to be applied either before FO or in parallel with FO. FO is applied on mid-level features. Applying FA at a high level is not very effective because the extracted features do not have sufficient information. Hence we apply FA on the mid-level features as well, in parallel with the Fourier object disentanglement module.

4.3.1 Randomly Initialized Uniform Sampling

It is computationally expensive to use all the frames in a video. In traditional uniform sampling, T frames are sampled at uniform intervals. The standard way of uniform sampling under-utilizes [213, 214] the knowledge that can be gained from the original video, which adds to the pre-existing issue of limited data. We use a variation of uniform sampling to improve the variance of the network and hence boost accuracy. First, we compute the step size as the ratio of total number of frames in the video and number of frames that we desire to sample. Next, we generate a random number between 0 and step size, and correspondingly designate the first frame to be sampled. This is followed by uniformly sampling video frames at step size intervals from the designated first frame.

4.4 Experiments and Results

4.4.1 Datasets

In this section, we briefly describe the UAV datasets used for evaluating FAR. **UAV Human RGB** [22] is the largest UAV-based human behavior understanding dataset. Split 1 contains

15172 and 5556 images for training and testing respectively captured under various adversities including illumination, time of day, weathers, etc. **UAV Human Night Camera** [22] contains videos similar to UAV Human RGB captured using a night-vision camera. The night vision camera captures videos in color mode in the daytime, and grey-scale mode in the nighttime. **Drone Action** [23] is an outdoor drone video dataset captured using a free flying drone. It has 240 HD RGB videos across 13 human actions. **NEC Drone** [197] is an indoor UAV video dataset with 16 human actions, performed by human subjects in an unconstrained manner.

4.4.2 Implementation Details

Backbone network architecture: We benchmark our models using two state-of-the-art video recognition backbone architectures (i) I3D [24] (CVPR 2017) (ii) X3D-M [7] (CVPR 2020). For both X3D and I3D, we extract mid-level features after the second layer.

Training details: Our models were trained using NVIDIA GeForce 1080 Ti GPUs, and NVIDIA RTX A5000 GPUs. Initial learning rates were $\{0.01, \text{ and } 0.001\}$ across datasets. We use the Stochastic Gradient Descent (SGD) optimizer with weight decay of 0.0005 and momentum of 0.9, and cosine/ poly annealing for learning rate decay. The final softmax predictions of all our models were constrained using multi-class cross entropy loss.

Evaluation: We report top-1 and top-5 accuracies.

Table 4.1: **Results on UAV Human RGB. Table (a):** FAR can be embedded within any 3D action recognition backbone to achieve state-of-the-art performance. Pretraining with Kinetics boosts performance, and large input sizes work better since FA and FO are designed to capture global, as well as local knowledge. FAR imparts improvements of 2.20%-38.69% over 3D action recognition backbones across training configurations. **Table (b) - Ablation experiments:** We demonstrate that each component of FAR imparts substantial improvement in top-1 accuracy by upto **8%**.

Backbone	FAR	Frames	Input Size	Init.	Top-1	Top-5
(i) I3D	No	8	540×960	Kinetics	21.06	40.81
(ii) I3D	Yes	8	540×960	Kinetics	29.21	50.27
(iii) X3D-M	No	16	224×224	None	27.0	44.2
(iv) X3D-M	Yes	16	224×224	None	27.6	44.1
(v) X3D-M	No	16	224×224	Kinetics	30.6	50.3
(vi) X3D-M	Yes	16	224×224	Kinetics	31.9	50.3
(vii) X3D-M	No	8	540×540	Kinetics	36.6	57.1
(viii) X3D-M	Yes	8	540×540	Kinetics	38.6	59.2

(a)

FO	FA	Sampling	Top-1
No	No	No	21.06
Yes	No	No	25.89
No	Yes	No	24.15
Yes	Yes	No	27.00
Yes	Yes	Yes	29.21

(b)

4.4.3 Main Results: UAV Human RGB

4.4.3.1 Benchmarking FAR

FAR can be embedded within any 3D action recognition backbone to achieve state-of-the-art performance. We show results on UAVHuman RGB at different frame rates, input sizes, backbone network architectures and pre-trained weights initialization. In experiment (i) and (ii), we use the I3D backbone, and initialize the network with pretrained Kinetics weights. Spatially, we downsample the input video by a factor of 2, and sample 8 frames per video. This configuration gives the network full access to the spatial portions of the video at every stage of training and testing. FAR imparts a relative improvement of **38.69%** and **23.18%** in top-1 and top-5 accuracy, respectively.

In the subsequent experiments, we use X3D-M, as the backbone. Many vision-based methods crop the original video into small patches of resolution 224×224 . We explore this in experiments (iii)-(vi) under two settings: without initializing with Kinetics pretrained weights, and by initializing with Kinetics pretrained weights. Concurrent with our intuition, initializing

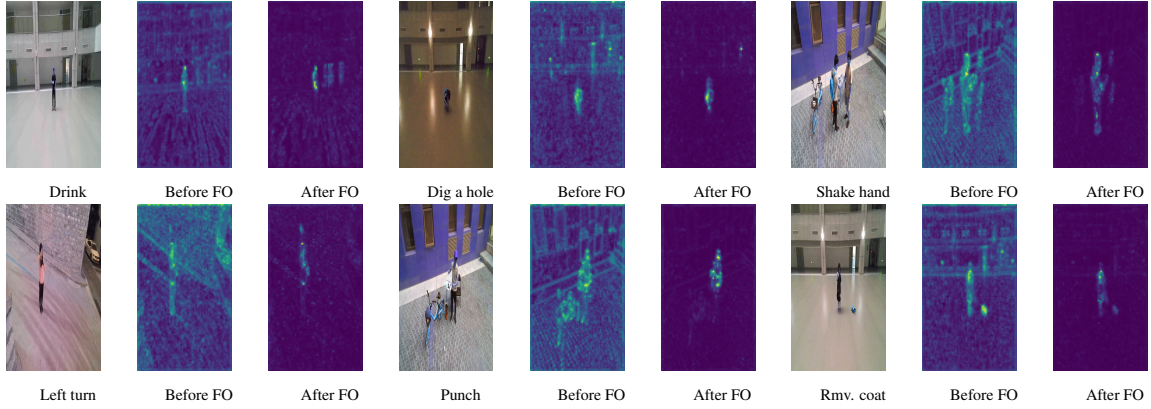


Figure 4.2: **Qualitative results on UAV Human RGB.** We show the effect of our Fourier Object Disentanglement (FO) method. In each sample, the images, in order, correspond to a frame from the video, feature representation before disentanglement and the feature representation after disentanglement respectively. Notice the effectiveness of FO in scenes with light noise (Row 1 Image 2, Row 2 Image 3), dim light (Row 1 Image 2), dynamic camera and dynamic background (Row 1 Image 1). Regions of the scene corresponding to moving human actor (or salient dynamic) are amplified most (solid yellow). Static background is completely suppressed (solid purple). Static salient regions are slightly amplified (e.g. lower body of human actor in Row 2 Image 3 - yellow), and dynamic backgrounds are suppressed to a great extent (pale yellow in Row 1 Image 1). We show more results in the supplementary material.

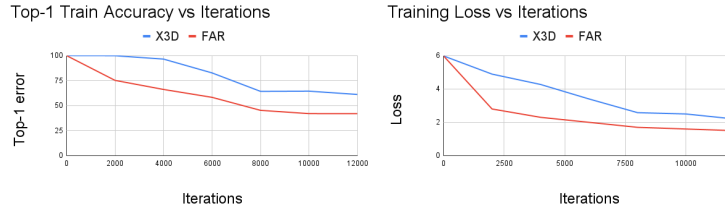


Figure 4.3: **FAR converges much faster than the state-of-the-art action recognition method X3D-M [7].** In the left curve, we show the top-1 train accuracy as a function of the networks’ training iterations. In the right figure, we plot the training loss curve. We demonstrate that FAR imparts convergence benefits over prior work, under the same hyperparameter and GPU configurations.

with Kinetics pretrained weights results in better performance than without initializing with Kinetics pretrained weights. In both cases, with small crop siz, FAR improves performance over the corresponding baselines by 2.2% – 4.24%. At a resolution of 224×224 , there is a slight decrease in Top-5 accuracy (0.1%).

Video action recognition is a global level task. Hence, it is important for the network to see a larger spatial region of the video to understand context, and get a better view of the human actor. Moreover, since the design of FAR is specifically inspired of challenges pertaining to object

Table 4.2: **Comparisons with state-of-the-art self-attention based transformer methods on UAV Human.** We initialize all our models with Kinetics pre-trained weights. We observe higher accuracy and computational benefits (up to **3x**) with FAR.

Method	Param M	Top-1 %	FLOPs GFlops/video	GPU GB/video	Inference Time (sec)/video
I. Baseline non-attention models					
X3D-M	3.8	36.6	14.39	3	0.08
I3D-M	12	21.06	346.55	10	0.1
II. Comparisons with the state-of-the-art Fourier-based method					
I3D + FNet [241] (2021)	12	24.39	346.56	10	0.1
I3D + FAR (Ours)	12	29.21	346.6	10	0.2
III. Comparison with the state-of-the-art efficient-attention method					
I3D + Efficient Attention [232] (WACV 2021)	12	21.13	462	13.3	0.12
I3D + Fourier Attention (Ours)	12	24.15	346.57	10	0.19
IV. Comparisons with the state-of-the-art self-attention based transformer methods					
ViT-B-TimeSformer (ICML 2021) [174]	121.4	33.9	2380	32	0.27
MVIT (ICCV 2021) [242]	36.6	24.3	70.8	9	0.16
X3D-M + FAR (Ours)	3.8	38.6	14.41	3	0.09

background separation, and context encoding, the margin of improvement imparted by FAR to the backbone architecture is larger when the crop size is higher. At a crop size of 540×540 , FAR improves top-1 and top-5 accuracies by 5.46% and 3.67% respectively, over the corresponding baselines.

4.4.3.2 Ablation Experiments

FAR ablations. We present ablation experiments on the components of FAR. We use the I3D backbone [24], and sample 8 frames per video. We initialize with Kinetics pretrained weights, and spatially downsample (and then feed in the entire frame) the video by a factor of 2. In the first four experiments, we uniformly sample 8 frames from frame 0. The first row is the baseline experiment with neither Fourier object disentanglement nor Fourier attention. We observe in the experiment corresponding to Row 2 that object disentanglement improves performance by 22.9% over the baseline. FO is a high pass filter (L2). The usage of a linear (L1) high pass filter results in an accuracy of 25.56%. Thus, we used the L2 high-pass filter as it results in a higher accuracy. Next, we determine the importance of context and long-range space-time relationships by using

only Fourier attention, and demonstrate an improvement of 14.67% in Row 3.

FO and FA complement each other - the former disentangles object from the background, while the latter decodes contextual information and inter-pixel, inter-frame relations. Using FO and FA in parallel, and sum-fusing the resultant feature maps cumulatively improves performance by 28.2% over the baseline. Finally, we incorporate the sampling scheme, vis-a-vis, randomly initialized uniform sampling along with FO and FA. This results in a final accuracy of 29.21%, 38.69% over the baseline.

FA ablations. We conduct ablation experiments on our proposed Fourier Attention (FA). We use the I3D backbone, and a video resolution of $540 \times 960 \times 8$. In all these experiments, FO is applied on level 2. In the first experiment, we extend FA to channels [45] in addition to space-time, at level 2, the accuracy is 26.48. In the second experiment, we apply channel FA at level 4 [45] while retaining space-time FA at level 2, the accuracy further degrades to 25.77. In contrast, the accuracy with space-time Fourier attention is 27.00. Thus, we find that global mixing at channel level does not contribute to improvement in performance.

Next, we explore the notion of multi-level FA, where FA is applied at multiple levels and not just level 2, the accuracy is 29.16. In contrast, FAR’s accuracy is 29.21. Our conclusion is that FA extracts knowledge prerequisite to learning long-range space-time relationships at level 2, to its maximum capacity, and applying it at more layers is redundant.

4.4.3.3 State-of-the-art comparisons

We report state-of-the-art comparisons. For all our experiments, we set the temporal and spatial resolutions at 8 frames and upto 540 (short side) respectively. We establish the baseline

accuracies using non-attention networks vis-a-vis, I3D [24] and X3D [7], in experiment I.

Comparisons with FNet. We report the accuracy using FNet, which is the state-of-the-art Fourier transform based self-attention method. FNet naively replaces every self-attention layer with the Fourier transform of the feature maps at that level. The motivation is to "mix" different parts of the feature representation and thus gain global information. Originally designed for NLP, it achieves 92 – 97% of the accuracy of BERT counterparts on the GLUE counterparts. However, when applied to video activity recognition on UAV Human RGB with the I3D backbone, we find that the accuracy is just 24.39%. In contrast, with the same backbone and hyperparameter settings, we demonstrate that FAR achieves 29.21%, an improvement of 19.76%.

Comparisons with efficient attention methods. We compare with the current state-of-the-art efficient attention method [232] in experiment III. For fair comparisons, we use the I3D backbone in both cases, at a video resolution of $540 \times 960 \times 8$. We demonstrate better accuracies with our Fourier Attention formulation at lower FLOPs and GPU memory.

Comparisons with transformer/self-attention methods. We compare against self-attention based transformer methods. Specifically, we compare against (i) TimesFormer [174] (ICML 2021) - a self-attention video recognition method based on joint space-time self attention, and (ii) MViT (ICCV 2021) - a transformer based method that combines multi-scale feature hierarchies. We demonstrate much better performance at lower number of FLOPs, GPU memory and inference time. Another benefit is that FAR does not add any new parameters to the neural network and uses the same number of parameters as its backbone network. In contrast, MVIT and TimeSformer use much higher number of parameters.

Table 4.3: **Results on more UAV datasets.** We demonstrate that FAR improves the state-of-the-art accuracy by **8.02%-17.61%** across popular UAV datasets.

Method	Frames	Input Size	Init.	Top-1
(i) Dataset: UAV Human Night [22]				
I3D [24]	8	480×640	Kinetics	28.72
FAR	8	480×640	Kinetics	33.78
(ii) Dataset: Drone Action [23]				
HLPF [243]	All	1920×1080	None	64.36
PCNN [244]	-	1920×1080	None	75.92
X3D-M [7]	16	224×224	Kinetics	83.4
FAR	16	224×224	Kinetics	92.7
(iii) NEC Drone [197]				
X3D-M [7]	8	960×540	Kinetics	66.15
FAR	8	960×540	Kinetics	71.46

4.4.4 Results: More UAV Datasets

We demonstrate the effectiveness of FAR on multiple UAV benchmarks. We demonstrate that FAR outperforms prior work by **17.61%**, **11.15%** and **8.02%** on UAV Human Night, Drone Action, and NEC Drone respectively.

4.5 Conclusions, Limitations and Future Work

We present a new method for UAV Video Action Recognition. Our method exploits the mathematical properties of Fourier transform to automatically disentangle object from the background, and to encode long-range space-time relationships in a computationally efficient manner. We demonstrate benefits in terms of accuracy, computational complexity and training time on multiple UAV datasets. Our method has a few limitations. The sampling strategy based on randomly initialization is a naive method to span all video frames. It might be interesting to explore the usage of better video sampling strategies [213, 214]. Next, we assume that the input videos contain only one human agent performing action. Multi-action videos could be a potential extension of our method. Moreover, we believe that FAR can be extended to other tasks such as video object segmentation and video generation, front-camera action recognition, graphics

and rendering [245, 246].

Chapter 5: DifFAR: Differentiable Fourier Aerial Video Recognition

UAVs are being increasingly used for surveillance, security, inspection, search and rescue, agriculture and videography. Action recognition [247] for videos captured by these UAVs is a crucial video analysis task. While there has been immense work on videos obtained using front-view cameras, research on UAV videos is limited. Deep neural networks for human activity recognition [7, 205] have been widely used for ground-camera scenes [24, 198], where the human actors take a high fraction of the pixels in the video scenes. While these methods work well on ground-camera videos, their accuracy on UAV videos can be low. This is mainly because UAV videos consist of human actors that occupy less than 10% of the spatial resolution of the video. Moreover, the action captured by these cameras may be low-resolution, and there is background and camera movement.

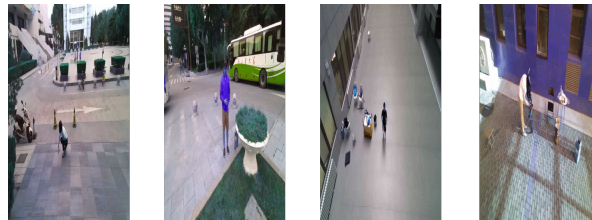


Figure 5.1: Human activity recognition in videos recorded using dynamic cameras at oblique and aerial angles typically consist of human actors that occupy less than 10% of the spatial resolution of the video. Background motion, noise, and low-resolution activities add to the complexities faced by video recognition systems. Our method achieves upto **41.86%** and **80.75%** top-1 accuracies on UAV Human and NEC Drone, respectively.

Moving objects in videos can be detected using object detection and tracking [177], optical

flow [186], motion segmentation [190], etc. However, many issues arise in terms of using them on UAV videos: unavailability of ground-truth object detection labels, domain gap between scenes with access to object localization and scenes on which action recognition needs to be performed [179, 180, 181], the need for a separate learning technique for modeling motion [202], etc. Thus, the network needs to automatically disentangle regions of the feature map, which are vital for video recognition, from entangled representations encompassing all aspects of the video.

The design of a differentiable method to compute disentangled representations enables the neural network to intrinsically learn disentangled regions, from motion cues and learned saliency information. A differentiable formulation between the neural network and motion representations establishes a bridge between classical and learning-based techniques, harnessing the power of both. Differentiable programming has been used in diverse tasks such as simulation [248] and control [249]. However, it has not been explored in action recognition literature.

5.0.1 Main Contributions

We present a novel method, DiffFAR, for activity recognition in UAV videos with dynamic backgrounds and moving cameras. Our approach is designed to handle *oblique* and *aerial* camera views with *low actor-background pixel ratios*, which are common characteristics in videos captured using aerial cameras. We take advantage of frequency domain representations [250] of the feature maps that provide an alternate complementary formulation to holistically model the time-evolution of pixels corresponding to various entities in the scene. We propose a differentiable static-dynamic frequency mask that (i) empowers the network to learn disentangled feature maps via a differentiable loss function and (ii) samples crucial frames from the video at

test time. The novel components of our work include:

1. **Differentiable static-dynamic mask:** “Dynamic” saliency mask denotes important regions of the scene corresponding to moving pixels. Similarly, “static” saliency mask represents important regions of the scene corresponding to stationary pixels. We propose a *differentiable* frequency-based “static” saliency mask that complements the frequency-based “dynamic” saliency mask. While the “dynamic” mask highlights dynamic salient regions of the scene and suppresses static non-salient regions, our “static” mask amplifies static salient regions and suppresses dynamic non-salient regions. Our approach uses a combination of *both masks to handle dynamic and static salient regions* of the scene (including *moving human actors*) and suppresses dynamic backgrounds, noise, and static non-salient regions.
2. **Differentiable disentanglement:** We bestow the network with the ability to inherently “learn” disentangled feature representations that are crucial for the underlying task of action recognition, via an identity loss function using the “*differentiable*” *static-dynamic frequency mask*. This empowers layers of the network to serve as the disentangler and directly compute salient features in its forward pass.
3. **Mask sampling:** We propose a frame sampling method using our *static-dynamic frequency mask*, to select the most informative frame within uniformly spaced non-overlapping video segments. These frames serve as a complementary descriptor to frames chosen using uniform sampling.

DifFAR can be embedded within any 3D action recognition backbone. It is parameter-free i.e. it does not add any learnable layers or parameters to the backbone neural network, which

simplifies initialization and training. We conduct extensive experiments on two UAV datasets: UAV Human [22] and NEC Drone [197]. These datasets are captured at varying altitudes, angles and lighting conditions. We demonstrate a considerable improvement of **2.84%** – **13.05%** in Top-1 accuracy over the corresponding baselines. We demonstrate state-of-the-art performance on UAV Human and NEC Drone by at least **3.26%** and **7.74%**, respectively, over the state-of-the-art methods. Moreover, DiffFAR is on par with the state-of-the-art transformer architectures while using far fewer computations (up to **61.8×**) and network parameters (up to **31.94×**).

5.1 Related Work

Aerial Video Recognition. The availability of UAV datasets [22, 251, 252] has fostered aerial video research [253] pertaining to person reidentification [254], human detection [255], tracking [256, 257], pose estimation [258], few-shot learning [217], drone detection [259] and path planning [260]. Many architectures have been proposed to specifically tackle aerial video action recognition [218, 219, 220, 261], in addition to generic action recognition [7, 205]. Recently, FAR [250] proposed a frequency-based method to disentangle moving objects by modulating feature maps. However, FAR does not place any explicit emphasis on static regions of the scene, a complementary descriptor. Moreover, FAR applies the mask as a transformation over feature maps, denying the network the opportunity to directly ‘learn’ disentangled feature maps in a differentiable manner.

Differentiability and Deep Learning. Differentiability has emerged as a powerful tool towards making neural networks learn tailored feature representations for diverse tasks such as simulation [248], control [249], video structure from motion [262], and physics [263]. In this chapter, we present a differentiable frequency-domain method for aerial video action recognition.

Motion Representation Methods. A popular motion representation method is optical flow [182], which is very expensive. Deep learning based optical flow computation [184, 185, 186] has a lower computational complexity, but requires a two-stream NN [200] for action recognition that increases the network complexity. Object tracking methods [264] require ground-truth bounding boxes that are difficult to obtain. Other alternatives such as motion feature networks [187] and ActionFlowNet [188] are less accurate optical flow. Techniques such as background subtraction [189] and motion segmentation [190] are not ideal for action recognition methods that rely heavily on object localization [191, 192]. In contrast, we present a computationally efficient frequency-based disentanglement technique to automatically highlight regions of the scene salient for action recognition.

Frame sampling. Frame sampling [265] has mainly been explored in the context of untrimmed videos [266] using Reinforcement learning [267], saliency information [214], audio information [268], etc. Most of these methods require the neural network to be retrained using the selected frames which is expensive. Moreover, CNN methods such as MG Sampler [213] use only prior frames, while it is beneficial to use future frames as well in a non-online setting.

5.2 DifFAR

We present a novel method, DifFAR, for video recognition in UAV videos. These videos, recorded using dynamic cameras placed at oblique angles obstructing the field of view of the action being performed, contain human actors occupying fewer than 10% of the pixels, along with background motion and light noise. Our method has multiple benefits: (i) It can be embedded within any 3D action recognition backbone without modifying the structure of the underlying

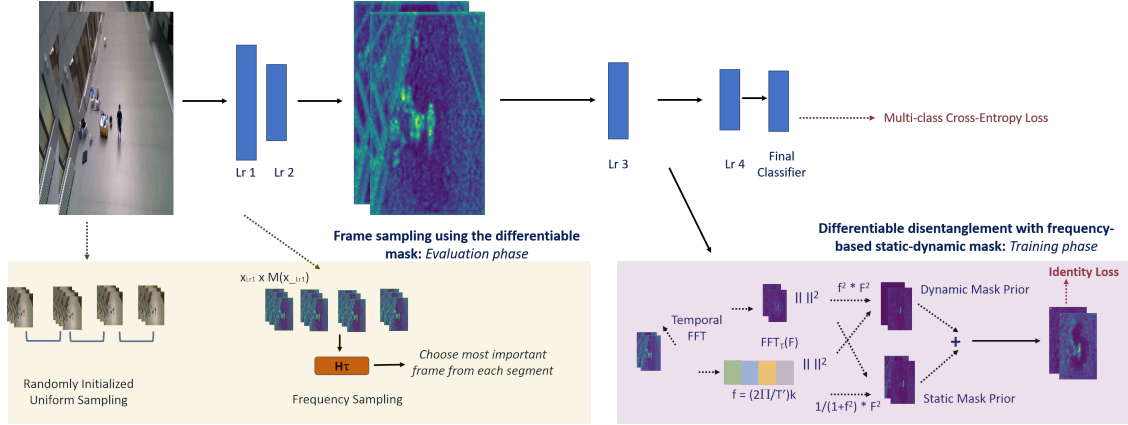


Figure 5.2: We present a differentiable frequency-based method for aerial video recognition. Our differentiable static-dynamic frequency mask provides a prior for disentangled regions relevant to action recognition. This mask is used to guide the learning of disentangled features within the layers of the neural network using an identity function. Furthermore, we propose a frame sampling strategy that chooses the best frame within each uniform video segment at test time, using the static-dynamic frequency mask and temporal difference.

neural network. (ii) It is parameter-free, i.e., it does not add any learnable layers or parameters to the backbone neural network. These properties allow us to initialize the neural network with weights pre-trained on large-scale ground-camera datasets such as Kinetics [24] and simplifies training.

We sample T frames from the video using uniform sampling [250], where T is the frame rate. The architecture involves a 3D CNN backbone such as I3D [24] or X3D [7]. 3D CNNs

Notation	Term
$M_{dynamic}$	Dynamic mask
M_{static}	Static mask
M	$M_{dynamic} + M_{static}$
f	Frequencies
F	Fourier transform
x	Feature representations
C_p	Prior frames cost
C_f	Future frames cost
T	Frame sampling rate

Table 5.1: Symbols used in the chapter

typically consist of an initial set of convolution layers, followed by 4 blocks of ResNet-type convolution layers. This is followed by a final prediction layer. Our architecture retains the structure of this backbone. We propose a differentiable static-dynamic frequency mask to highlight regions of the scene relevant for action prediction. This is computed using the outputs at the second block of the neural network. The neural network is trained using traditional multi-class cross entropy loss for action recognition and an identity loss. The identity loss function is applied at the outputs of the third block of the neural network, and bestows the network with the ability to learn disentangled feature maps in a differentiable manner via an identity loss function. We also propose a method for frame sampling using the differentiable static-dynamic frequency mask, which samples frames from the video at the evaluation phase. We now describe our method in detail.

5.2.1 Background: Differentiability

Differentiable programming allows for computing derivatives of functions or gradient operators to learn representations for the underlying task. For feature disentanglement, differentiable programming can be a more powerful tool than applying transformations over features [250]. This is because blocks of the network are optimized to directly compute disentangled features, where the entire set of layers serves as the disentanglement module.

Frequency domain representations contain an abundance of information for disentanglement that can improve the performance of action recognition neural networks. One way to do so is to apply the corresponding frequency-domain masks as a transformation over neural network features. However, enabling the neural network to learn disentangled representations within its

layers, using frequency-domain information as well as training data, creates an amalgamation of techniques from traditional signal processing and deep learning. This has the potential to empower the neural network with the benefits of both worlds.

However, the training of neural networks requires loss functions to be differentiable in order to perform backpropagation. This raises questions pertaining to frequency domain representations that can be used for training, appropriate loss function and the layers of the neural network that should be used for learning robust disentangled features.

5.2.2 Differentiable Static-Dynamic Frequency Mask

Consider a video depicting the action of human diving. The focus of the camera shifts from the diving springboard to the water beneath as the action is executed. Moreover, the springboard might move as the person jumps into the water, and there may be other humans moving around the swimming pool. In the absence of any external localization ground-truth, the neural network needs to innately disentangle the human actor performing the action of diving from the rest of the scene. In general, a video scene consists of four types of regions - dynamic salient regions, static salient regions, dynamic non-salient regions and static non-salient regions. The first two are very important for the underlying task of action recognition. A robust feature disentanglement technique should highlight pixels corresponding to static salient regions and dynamic salient regions and suppress the non-salient patches.

The frequency domain of a signal is a powerful representation that can quantify the temporal change of spatial pixels, and hence can be used for predicting static and dynamic regions of the scene. To identify dynamic pixels, we can formulate the “dynamic” mask $M_{dynamic}$ [250] as the

dot product of the temporal FFT F_T of the feature maps and the corresponding frequencies f , over the frequency spectrum $k = 1 \dots T$. For features x , mathematically,

$$M_{FAR}(x) = \underbrace{\|F_T(x)\|_2^2 \times \|f\|_2^2}_{M_{dynamic}} \odot x, \quad (5.1)$$

The formulation amplifies high amplitudes at high frequencies, and suppresses low amplitudes at low frequencies. The final dot product with the feature representation ensures that only salient moving regions are amplified, and moving background regions are suppressed. Since this equation only models dynamic regions of the scene, it causes inconsistencies in static parts of the scene corresponding to salient regions; it may also incorrectly disentangle dynamic backgrounds and noise. To alleviate this issue, we propose modeling the salient static parts of the video through a static mask M_{static} .

High amplitudes at low frequencies correspond to static regions of the scene. Given a feature representation x , we mathematically model this relation as follows:

$$M_{static} = \|F_T(x)\|_2^2 \odot (1/(1 + \|f\|_2^2)), \quad (5.2)$$

The inverse shifted quadratic formulation of the frequency component ensures that the computation of the mask is stable. A dot product of M_{static} and network features amplifies static, salient regions and heavily suppresses dynamic non-salient regions. M_{static} , in conjunction with the dynamic mask $M_{dynamic}$, forms a static-dynamic frequency mask prior M . Thus, the net mask encapsulating static, and dynamic parts of the scene is $M = M_{static} + M_{dynamic}$. M guides the network to *learn* object disentangled low-level feature representations, as described in the

following section.

5.2.3 Differentiable Disentanglement

Our next goal is to use the differentiable static-dynamic frequency masks to help the neural network learn disentangled feature representations. We guide the learning of disentangled feature representations at the penultimate layer of the network using the static-dynamic mask prior M from mid-level features. Since mid-level features provide a fine balance between encoding video features and maintaining the space-time resolution of the video, we use mid-level features x_{mid} to derive the disentanglement mask $x_{mid,M}$, computed by applying M on x_{mid} . We constrain the features at the penultimate layer, x_p to resemble disentangled feature representations using an identity loss function as follows:

$$L_{mask} = \lambda_{mask} \cdot MSE\{x_p, M(x_{mid}) \odot x_p\}, \quad (5.3)$$

where MSE is mean squared error. In the above equation, the dot product of mask $M(x_{mid})$ (M applied on x_{mid}) and x_p transforms x_p to the corresponding disentangled feature representation. Further, the application of the identity loss enables the network to intrinsically learn high-quality disentangled feature representations at the penultimate layer. During backpropagation, gradients flow from x_p as well as x_{mid} (through M) ensuring that neural network layers between the mid-level and penultimate layer learn disentangled features. The hyperparameter λ_{mask} , typically set to 0.1 in our benchmarks, scales the identity loss function. We use a combination of the classical multi-class cross entropy loss function [24] applied on the final softmax layer of the neural network, along with L_{mask} , to train the action recognition neural network.

5.2.4 Frame Sampling

Uniform sampling divides the video into uniform segments, equal to the number of frames N to be sampled, and selects the k_{th} from each segment where k is an integer lesser than or equal to the number of frames in each segment. Such a sampling strategy is commonly adopted during the training of action recognition architectures on trimmed videos and works well due to its ability to span the duration of the video. However, samples selected from each video segment using uniform sampling are not always the most salient frames in the corresponding segment. We present a method to choose the best frame within uniform video segments. Our solution can be applied at the evaluation phase and does not require the neural network to be retrained.

A good frame selection strategy should consider frames prior to the video segment under consideration as well as frames succeeding the video segment. Frames that provide maximal information w.r.t. prior as well as future frames can be regarded as salient frames. For the task of action recognition, movement or difference of feature representations between frames provides this information. Corresponding to segment i ($i = 1 \dots T$), where T is the number of frames, for frame j 's ($j = 1 \dots s$), where s is the number of frames within each segment (or total number of frames in the video divided by sampling rate), we denote the feature representations $x_{i,j}$. Since movement is best encapsulated by the salient regions, which are useful for predictions, we compute the saliency mask representations for features $x_{i,j}$ as $M \odot x_{i,j}$. We denote uniform frames for segment i as x_{U-i} , and the corresponding saliency mask representations as $M \odot x_{U-i}$. Weighted sum of mean squared difference between $M \odot x_{i,j}$ and prior and future frames gives the cost w.r.t. prior C_p and future frames C_f . The weights are designed to be inversely proportional to the distance between the video segment under consideration for frame sampling and the video

segment in the equation for computation for prior and future costs. Mathematically,

$$C_p = \sum_{k=1}^{i-1} \frac{T-i+k}{T} \times (M \odot x_{i,j} - M \odot x_{U-k})^2 \quad (5.4)$$

$$C_f = \sum_{k=i+1}^T \frac{T+i-k}{T} \times (M \odot x_{i,j} - M \odot x_{U-k})^2. \quad (5.5)$$

The net cost is a weighted sum of C_p and C_f where the weights, w_p and w_f , are proportional to the costs and time between the video segment under consideration for frame sampling, the beginning of the video, and the end of the video. Mathematically, $w_p = \{C_p/(C_p + C_f)\} \times i$ and $w_f = \{C_f/(C_p + C_f)\} \times (T - i)$.

Ensembling the predictions. We use the set of uniform frames F_U to sample the set of salient frames F_S . We use the trained action recognition neural network to compute the predictions for F_U as well as F_S . These predictions are summed to obtain the final prediction, which is used to determine the predicted action class.

5.3 Experiments

5.3.1 Network and Training Details

Backbone network architecture: We benchmark our models using two state-of-the-art 3D CNN based action recognition architectures :- I3D (CVPR 2017) and X3D (CVPR 2020). We use level 2 features (mid-level) of the underlying 3D ResNet-101 to derive object disentanglement masks. This mask is used to guide the learning of disentangled feature representations obtained after level 3. The cost functions for frequency-based frame sampling are computed using features

Model	Data Aug.	Input	Frames	Top-1 (%)
I3D Experiments				
I3D	No	(540, 960)	8	23.86
Ours (I3D Backbone)	No	(540, 960)	8	32.94
X3D Experiments				
X3D	No	(540, 960)	16	35.71
Ours (X3D backbone)	No	(540, 960)	16	40.81
X3D + Data Augmentation Experiments				
X3D	Yes	(540, 540)	8	36.6
X3D + FD	Yes	(540, 560)	8	41.86

Table 5.2: We evaluate our method on UAV Human. Across two 3D CNN backbones (I3D, X3D) and datasets, we demonstrate relative improvements of **14.28% - 38.05%** in the Top-1 accuracy over the corresponding baselines.

from level 1.

Training details: All our models were trained using NVIDIA GeForce 1080 Ti GPUs and NVIDIA RTX A5000 GPUs. Initial learning rates were set at 0.01 or 0.1. We use the Stochastic Gradient Descent (SGD) for optimization, with 0.0005 weight decay, and 0.9 momentum. We use cosine/poly annealing for learning rate decay and multi-class cross entropy loss to constrain the final softmax predictions of the neural network. We report top-1 accuracies for all our models.

Datasets: We benchmark our models on two UAV datasets:- UAV Human [22] and NEC Drone [197].

UAV Human is an outdoor activities dataset, containing low-resolution videos taken under adverse lighting and weather conditions. It contains 155 actions, many of which are similar and hard to distinguish. Moreover, the videos contain dynamic backgrounds, camera motions and noise. NEC Drone is an indoor dataset with single-human and two-person actions. It contains 16 actions performed in an unconstrained manner.

Experiment	Top-1 (%)
$M_{dynamic}$	28.1
$M_{dynamic} + L_{mask}$	29.68
$M_{static} + M_{dynamic} + L_{mask}$	31.95

Table 5.3: We demonstrate the effectiveness of each component of FD with ablation experiments on UAV Human. All our experiments use RIUS for training as well as testing, and I3D backbone, and a spatial and temporal resolution of 540×960 and 8 frames respectively.

Model	Backbone	Data Aug.	Input	Frames	Params (M)	FLOPs (G)	Top-1 (%)
UAV Human							
I3D (CVPR 2017) [24]	-	No	(540, 960)	8	28.0	108	21.06
MVIT (ICCV 2021) [242]	-	Yes	(224, 224)	16	36.6	70.8	24.3
X3D-M (CVPR 2020) [7]	-	Yes	(224, 224)	8	3.8	6.2	30.6
X3D-M (CVPR 2020) [7]	-	Yes	(540, 540)	8	3.8	36.03	36.6
X3D-M (CVPR 2020) [7]	-	No	(540, 960)	16	3.8	36.03	x
FNet (arXiv 2021) [241]	I3D	No	(540, 960)	8	28.0	108	24.39
FAR (ECCV 2022) [250]	I3D	No	(540, 960)	8	28.0	108	29.21
FAR (ECCV 2022) [250]	X3D-M	Yes	(540, 540)	8	3.8	36.6	38.6
Ours	X3D-M	Yes	(540, 540)	8	3.8	36.6	41.86
Ours	X3D-M	No	(540, 960)	16	3.8	130.13	40.81

Table 5.4: We compare our method against prior methods on UAV Human. We demonstrate state-of-the-art performance by at least 5.72%. Our method outperforms transformer architectures while using far fewer computations (upto **61.8** $\times$) and network parameters (upto **31.94** $\times$). Our method imposes minimal memory overhead over the backbone neural network.

Evaluation: We evaluate all action recognition models using the top-1 accuracy score, which is the ratio of the number of correctly classified samples to the total number of samples in the evaluation set. We also report top-5 accuracy which is the ratio of samples for which the top-5 classes (in terms of probability of predictions) contains the ground-truth class. We report all improvements in terms of relative performance. The relative improvement of method 1 over method 2, where the absolute accuracies of method 1 and method 2 are $x_1\%$ and $x_2\%$ respectively, is $(x_1 - x_2)/x_2 \times 100$

5.3.2 Results: UAV Human

Our method can be embedded within any state-of-the-art 3D CNN to improve the performance. We demonstrate the effectiveness of our method over various backbone architectures. We report relative improvements of 38.05% and 14.28% using the I3D and X3D backbones respectively.

The large improvements in top-1 accuracy are indicative of the ability of FD to inherently localize the human actor from the background and identify actions from obscured top-down and oblique views in the presence of dynamic UAV cameras and noise.

All models use uniform sampling (randomly initialized) [250] for training. In the first two sets of experiments, we do not perform any data augmentation. The differentiable static-dynamic frequency mask and identity loss are used for training and the ensemble frequency-based frame sampling strategy is used at evaluation for our method. In the third set of experiments, the X3D model is trained with spatial data augmentations, consistent with the publicly available official code for X3D [7] in Meta AI’s SlowFast GitHub repository. While we use the differentiable static-dynamic frequency mask and identity loss for training, the frame sampling and evaluation scheme are consistent with the SlowFast repository codebase i.e. our ensemble frame selection strategy is not used for testing.

Ablation experiments on the differentiable static-dynamic frequency mask and identity loss. We present ablation experiments on UAV Human. We use the I3D backbone at spatial and temporal resolutions of 540×960 and 8 frames respectively. All experiments use uniform sampling (randomly initialized) for training as well as testing.

- In the first experiment, we apply only the dynamic mask $M_{dynamic}$, as a transformation over mid-level features, as in FAR. The accuracy is 28.1%. [250].
- In the second experiment, we quantitatively demonstrate the benefits of differentiability. We use $M_{dynamic}$ to train the network to learn disentangled features in a differentiable manner using L_{mask} (Equation 3). The absolute improvement over the first experiment is 1.58%.

Model	Input	Frames	Top-1 (%)	Top-5 (%)
I3D Experiments				
w/o sampling	(540, 960)	8	31.75	52.92
with sampling	(540, 960)	8	32.94	54.40
X3D Experiments				
w/o sampling	(540, 960)	16	40.12	59.08
with sampling	(540, 960)	16	40.81	59.94

Table 5.5: We present ablation experiments on the frame sampling method. None of our models use data augmentation.

- In the third experiment, we show the effects of the static-dynamic mask. We use M_{static} along with the setup in the second experiment. The absolute improvement over the second experiment is 2.27%.

Effectiveness of the ensemble frame sampling method. We present experiments on the effectiveness of the frame sampling method. On I3D, the frame sampling method improves top-1 and top-5 accuracies by (relative) 3.74% and 2.79% respectively. On X3D, the frame sampling method improves top-1 and top-5 accuracies by (relative) 1.71% and 1.45% respectively.

State-of-the-art comparisons. We present comparisons against the state-of-the-art. We compare against 3D CNNs (I3D [24], X3D [7]), transformer models (MVIT [242]) and frequency-based methods (FNet [241], FAR [250]). Our method has benefits in terms of top-1 accuracy as well as computation (FLOPs, number of parameters). It has minimal memory overheads over the corresponding backbone neural network. We demonstrate (relative) improvements of at least 5.72% over the state-of-the-art method.

Ablation experiments on the frame sampling method. All experiments are conducted on the X3D backbone, without data augmentation and a spatial-temporal resolution of $16 \times 540 \times 960$.

Model	Backbone	Input	Frames	Params (M)	FLOPs (G)	Top-1 (%)
X3D-M (CVPR 2020) [7]	-	(960, 540)	8	3.8	64.05	66.15
FAR (ECCV 2022, Ours) [250]	X3D-M	(960, 540)	8	3.8	65.08	71.46
DiffFAR (ICRA 2023, Ours)	X3D-M	(960, 540)	8	3.8	65.08	80.75

Table 5.6: We present experiments on NEC Drone. We demonstrate improvements of **13% - 22%** over prior work.

The purpose of the disentanglement mask in frame sampling is to ensure samples are chosen based on the relevance of only salient features. When applied on a model trained for learning disentangled features, a good way to verify if the network has actually learnt disentangled features is to remove the mask in the calculation of cost functions for frame sampling. This results in an accuracy of 40.9%, proving that our differentiable method is indeed successful. Setting $w_p = w_f = 1$ results in an accuracy of 40.70%.

5.3.3 Results: NEC Drone

We present results on NEC-Drone. We use uniform sampling [250] to train all models. In the experiment corresponding to our method, we initialize the neural network with weights corresponding UAV Human. Following prior work [250], while testing, we select the most frequent class prediction among the networks’ result using all sets of Randomly Initialized Uniform Sampling frames.

We obtain a top-1 accuracy of 80.75%. We compare against 3D CNNs (X3D [7]) and frequency-based methods (current state-of-the-art for aerial video recognition - FAR [250]) and show relative improvements of **13.00% - 22.07%**.

5.4 Conclusion, Limitations and Future Work

In this chapter, we presented a differentiable disentanglement method that simultaneously exploits the potential of learning-based neural networks and traditional signal processing. The method includes a static-dynamic frequency mask and identity loss for learning the disentanglement, along with a frame selection method at the evaluation stage. Our method has a few limitations. First, it is completely supervised. This means it requires labeled data and annotation is a laborious process. Unsupervised, self-supervised and transfer learning methods can alleviate this issue. Second, while our method is more effective than transformers in terms of accuracy, speed and memory, it is still expensive due to 3D convolutions and requires multiple GPUs and iterations to train. Future work on methods that can be run on mobile GPUs will be of practical importance. In our experiments and datasets, we assumed that a single human-agent is performing the action. Our method can be incorporated within any 3D neural network to improve performance and it might be interesting to explore the usage of our method within multi-agent neural networks.

Part IV

Generating Aerial-view Images using a Pretrained Model and an Exemplar Front-view Image

While aerial views suffer from limited supervised data, there is an abundance of front view data, and generative approaches are gaining popularity. The question arises: Can we leverage front view data to generate aerial view data? Unfortunately, obtaining paired training data or comprehensive 3D information is challenging. However, we propose utilizing a pretrained image generation model to synthesize aerial view images from single front view images — using approaches presented in this section.

This section has two chapters. In the Chapter 6, we present a novel method, *Aerial Diffusion*, for generating aerial views from a single ground-view image using text guidance. Aerial Diffusion leverages a pretrained text-image diffusion model for prior knowledge. We address two main challenges corresponding to domain gap between the ground-view and the aerial view and the two views being far apart in the text-image embedding manifold. Our approach uses a homography inspired by inverse perspective mapping prior to finetuning the pretrained diffusion model. Additionally, using the text corresponding to the ground-view to finetune the model helps us capture the details in the ground-view image at a relatively low bias towards the ground-view image. Aerial Diffusion uses an alternating sampling strategy to compute the optimal solution on complex high-dimensional manifold and generate a high-fidelity (w.r.t. ground view) aerial image. We demonstrate the quality and versatility of Aerial Diffusion on a plethora of images from various domains including nature, human actions, indoor scenes, etc. We qualitatively prove the effectiveness of our method with extensive ablations and comparisons. To the best of our knowledge, Aerial Diffusion is the first approach that performs single image ground-to-aerial translation in an unsupervised manner.

In the Chapter 7, we present HawkI, again, for synthesizing aerial-view images from text and an exemplar image, without any additional multi-view or 3D information for finetuning or

at inference. HawkI uses techniques from classical computer vision and information theory. It seamlessly blends the visual features from the input image within a pretrained text-to-2D-image stable diffusion model with a test-time optimization process for a careful bias-variance trade-off, which uses an *Inverse Perspective Mapping (IPM) homography transformation* to provide subtle cues for aerial-view synthesis. At inference, HawkI employs a unique *mutual information guidance* formulation to steer the generated image towards faithfully replicating the semantic details of the input-image, while maintaining a realistic aerial perspective. Mutual information guidance maximizes the semantic consistency between the generated image and the input image, without enforcing pixel-level correspondence between vastly different viewpoints. Through extensive qualitative and quantitative comparisons against text + exemplar-image based methods and 3D/ multi-view based novel-view synthesis methods on proposed synthetic and real datasets, we demonstrate that our method achieves a significantly better bias-variance trade-off towards generating high fidelity aerial-view images.

Chapter 6: Aerial Diffusion: Text Guided Ground-to-Aerial View Translation from a Single Image using Diffusion Models

The analysis of aerial image and video [269] plays a pivotal role in different applications, such as search and rescue, aerial photography, surveillance, movie production, etc. However, the paucity of aerial data [22] and the complexities associated with data capture from aerial cameras/UAVs makes it difficult and costly to train large neural networks [270] for these applications. Hence, the synthesis of aerial-view images [271] offers a promising alternative to address these challenges. Conditional image synthesis [272, 273] allows for control over the generation process.

Ground-view annotated datasets [21, 24, 198, 274] are readily available for many tasks. Hence, a method that transforms ground-view images to aerial views (or cross-view synthesis [26, 27]) could allow the reuse of annotated metadata for a variety of aerial-view applications, e.g., classification [275], segmentation [43], action recognition [276], representation learning [277], domain adaptation [197], augmentation [278], etc. However, cross-view synthesis requires the network to learn a very large non-trivial translation. The network needs to hallucinate a new and enormously different view of all entities in the scene and the background, while being consistent with the details including the semantics, colors, relations between various parts of the scene, pose, etc.

Prior work [26, 27, 279, 280] on ground-to-aerial generation use NeRFs and GANs. However,

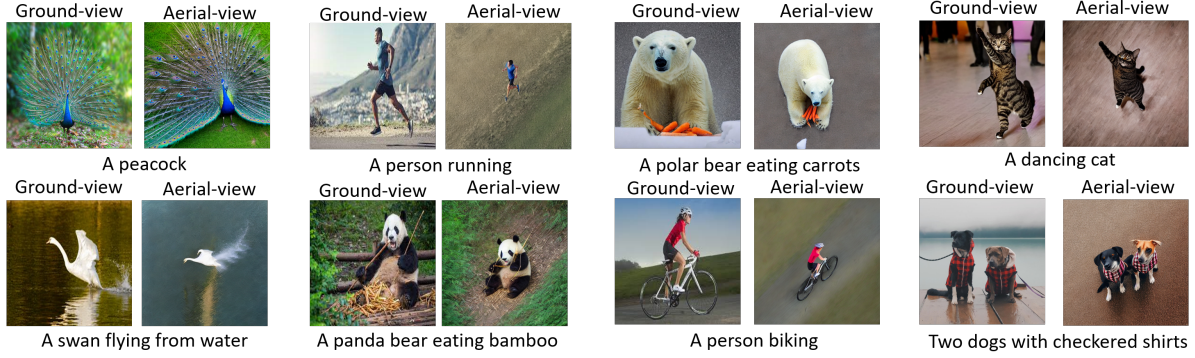


Figure 6.1: Given a single ground-view image and the corresponding text description as input, Aerial Diffusion generates corresponding aerial-view image. Our method does not require any supervision from aerial-view data, pairs of ground-aerial view, depth maps, semantic maps, multi-views, etc. It is one of the first approaches to achieve ground-to-aerial view translation in an unsupervised manner.

all of these methods use paired data for ground-view and the corresponding aerial views, which is seldom available. Moreover, training on a specific dataset limits the application to specific scenes similar to the training data; necessitating the collection of paired data for widely different distributions. Instead, our goal is to develop a generic method for generating aerial views from ground-views without any paired data or other auxiliary information such as multi-views, depth, 3D mapping, etc.

While there are many diverse datasets of ground images, there are not many such good quality aerial datasets [22] - hence, unpaired image-to-image translation [83] is not a viable solution. On the contrary, text is an auxiliary modality that can be easily obtained using off-the-shelf image/video captioning tools [281] Moreover, text provides a natural representation space describing images. Consequently, our goal is to use the text description of a ground-view image to generate its corresponding aerial view.

Recently, diffusion models have emerged as state-of-the-art architectures for text-to-image [8, 273, 282, 283] high-quality realistic image synthesis. The availability of immense prior knowledge

via large-scale robust pretrained text-to-image models [284], motivates us to pose ground-to-aerial view translation as text-guided single-image translation [8, 283]. Text-guided single-image translation methods finetune the diffusion model to the input image and then perform linear interpolation in the text embedding space to generate the desired output. However, direct application of these methods [8, 273, 283] to ground-to-aerial translation either generates high-fidelity non-aerial images or low-fidelity aerial images.

Main contributions. We present two postulates for text-guided image translation, for ground-to-aerial translation: (i) domain gap between the finetuning task (w.r.t. ground view) and target task (aerial view generation) hinders the diffusion model from generating accurate target views and introduces bias towards the source view, (ii) a finetuned diffusion model cannot generalize well to the target prompt, if the text embedding and image spaces corresponding to the source and the target are far apart on the nonlinear text-image manifold.

Based on these findings, we propose “*Aerial Diffusion*”, a simple, yet effective, method for generating aerial views, given a single ground-view image and the corresponding text description as input. The novel elements of our algorithm include:

- Instead of directly finetuning the diffusion model with the ground-view image, we apply a *homography based on inverse perspective mapping* on the ground-view image to obtain a homography projected image prior to the finetuning. This reduces the bias of the diffusion model towards the input image while finetuning.
- To finetune the diffusion model, we use the *source text* corresponding to the ‘ground-view’ as the *guiding factor*, instead of the target text (‘aerial view’). This helps the diffusion model search for an optimized text embedding in the vicinity of a text space close to the image space, enabling the learning of a ‘good’ optimized text embedding. This also

prevents the diffusion model from developing a bias towards an incorrect aerial view.

- To obtain a high-fidelity aerial image (w.r.t. ground-view), at inference time, we manipulate the text embedding layer, such that it prioritizes fidelity and the aerial viewpoint in an alternating manner. *Alternating between text embeddings corresponding to the viewpoint and fidelity switches the denoising direction*, such that the backward diffusion takes one step towards preserving fidelity followed by another step towards generating an aerial view. As noises are gradually removed, the process ends up with a high-fidelity aerial-view image on a manifold with a better fidelity-viewpoint trade-off than linear interpolation.

We apply our method on numerous in-the-wild images from various domains such as nature, animals and birds, human actions, indoor objects, etc. Our method is able to generate high-quality aerial view images that preserve the details contained in the source-view image. We conduct extensive ablation studies highlighting the benefits of each element of our method; and demonstrate the trade-off between fidelity (w.r.t. source image) and faithfulness to target view via hyperparameter tuning. We compare with the state-of-the-art diffusion model-based text-guided editing approach and show far superior qualitative performance for ground-to-aerial translation. Comparison to other text-embedding manipulation approaches also shows that our alternate prompting strategy works better.

6.1 Related work

There has been immense work on *image-to-image translation* [83, 285, 286, 287, 288, 289, 290] using GANs [291], transformers [292], diffusion models [284, 293, 294], etc. for problems, such as style transfer [295], image restoration [296], and multimodal style translation [297].

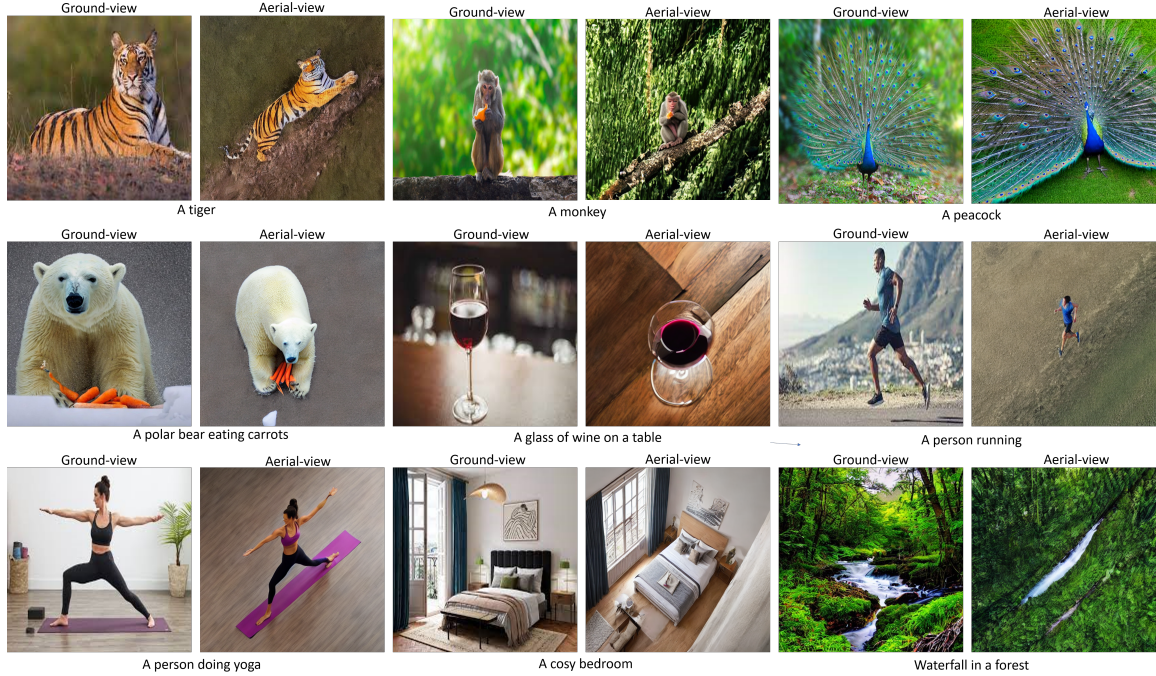


Figure 6.2: **Results.** We apply Aerial Diffusion on diverse images such as animals/ birds, natural scenes, human actions, indoor settings, etc and show that our method is able to generate high-quality high-fidelity aerial images.

Many of these methods are capable of performing these tasks using paired/ unpaired data [298]. Recently, diffusion models [273, 299, 300, 301, 302, 303, 304, 305] have been successful in performing non-trivial operations, such as posture changes and multiple objects editing. Prior work on *cross-view synthesis* [26, 27, 279, 280, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316] generally use paired data and other complex auxiliary modality such as semantic maps, depth, multi-views, etc within various generative approaches.

A closely related problem is *novel view synthesis* [317, 318] where the goal is to generate new views of the scene. However, most novel-view synthesis methods including GANs [319], NeRFs [320], diffusion models [321] use multiple views of the scene for training, even while they may be capable of performing single-view evaluation [322, 323]. Again, this is prohibitive since it requires multiple views of the scene/ depth information [324] for training. On the other hand,

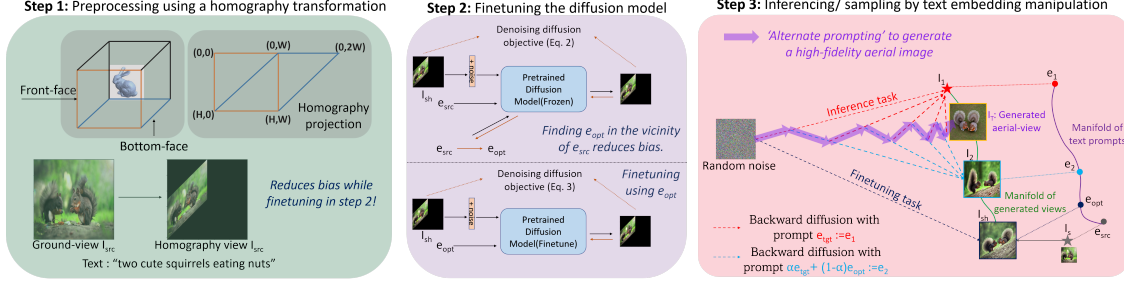


Figure 6.3: **Aerial Diffusion.** In step 1, we apply a homography transformation to the ground-view image I_S . This creates a gap between I_{Sh} and txt, which in turn reduces the bias of the model towards I_{Sh} in step 2. In step 2, we use the e_{src} to optimize to e_{opt} and finetune the diffusion model to reconstruct I_{Sh} for the e_{opt} . Using the e_{src} to find e_{opt} reduces the bias of the model towards I_{Sh} due to the disparity between I_{Sh} and the e_{src} . In stage 3, we manipulate the text embedding by using an alternating strategy to find the optimal solution in a higher-dimensional non-linear space to generate a high-fidelity aerial image I_T .

3D reconstruction methods [325, 326, 327] rely on depth information or auxiliary data such as shape priors. Moreover, 3D reconstruction is a complex and expensive task, which is redundant when the goal is to just obtain a 2D aerial view of the scene.

Text, which is easily available, has been widely used as a guiding factor for image translation [328, 329] and image editing [8, 282, 283, 330, 331, 332, 333, 334, 335], particularly in the context of diffusion models recently. The availability of large databases of image-text pairs [13], open-source pretrained models [336], and the fact that text is a natural representation of the world makes it conducive to use text as an auxiliary modality to provide guidance. Many *text-based image editing* approaches operate on a single real image [8, 337] and perform inference time optimization, making them easy to generalize across diverse images. *Single-image approaches* [338, 339, 340] have been proposed for image manipulation tasks without text-guidance as well. Generally, they aim to learn useful representations by finetuning a pretrained model on a single image for reconstruction. The inference then controls the feature space [341, 342] to achieve the desired changes.

6.2 Aerial Diffusion

We present Aerial Diffusion for view translation from a single real ground-view source image I_S to its aerial-view (or target image I_T), given a text description txt of the image. The text description can be obtained using an off-the-shelf image captioning tool [281]. We assume no access to any paired data or other modalities such as depth, semantic maps, other views of the scene, etc. Corresponding to the ground-view, we use the source text description $\text{txt}_G = \text{'front view of'} + \text{txt}$ with text embedding e_{src} . Similarly, for the aerial-view, we use the target text description $\text{txt}_A = \text{'aerial view of'} + \text{txt}$ with text embedding e_{tgt} .

6.2.1 Postulates

In this section, we analyze text-guided single image translation in the context of ground-to-aerial view synthesis and present two postulates. A common strategy adopted for text-based single image translation is to use a robust text-to-image pretrained model in a two-stage process. The first step finds the ‘optimized text embedding’ e_{opt} (in the vicinity of e_{tgt}) that best generates the ‘source’ image I_S and subsequently finetune the diffusion model to generate the ‘source’ image I_S using e_{opt} . In the second step, a linear interpolation of e_{tgt} and e_{opt} are used to generate the edited image I_T from the finetuned neural network, i.e., the backward diffusion process is

$$x_{t-1} = x_t - f(x_t, t, \alpha e_{tgt} + (1 - \alpha)e_{opt}), \quad t = T, \dots, 0. \quad (6.1)$$

A text-based single image translation approach [8, 282, 330, 331] for ground-to-aerial generation overcomes multiple limitations in terms of data availability and generalization. However,

the challenges involved in ground-to-aerial translation inhibit the direct application of existing text-based single image translation methods for ground-to-aerial generation. We present two postulates for text-based single-image translation in the context of ground-to-aerial generation.

Postulate 1. Domain gap between the finetuning task (e.g., ground view generation) and target task (aerial view generation) hinders the diffusion model from generating accurate target views and introduces bias towards the source view.

Diffusion models are probabilistic models. They are trained [343] by optimizing the negative log-likelihood of the model distribution under the expectation of the data distribution. Further simplification of the equation for formulating the training loss function involves variance reduction. In the first step of finetuning, the diffusion model is being trained to reproduce the source image given the optimized text embedding, irrespective of the input random noise. Hence, it has a natural bias towards the source image.

When the image space corresponding to the target text embedding is in the vicinity of the image space corresponding to the optimized text embedding, consistent with the variance within which the neural network was trained to generate, the generated target image is a high fidelity image consistent with the target text. When the desired transformation is large (ground-to-aerial), outside the limits of the variance, the diffusion model is unable to generate an aerial image.

Postulate 2. A finetuned diffusion model cannot generalize well to the target prompt if the text embedding and image spaces corresponding to the source and the target are very different and far away from each other on the nonlinear text-image embedding manifold.

The embedding space and the corresponding image representation space are locally linear.

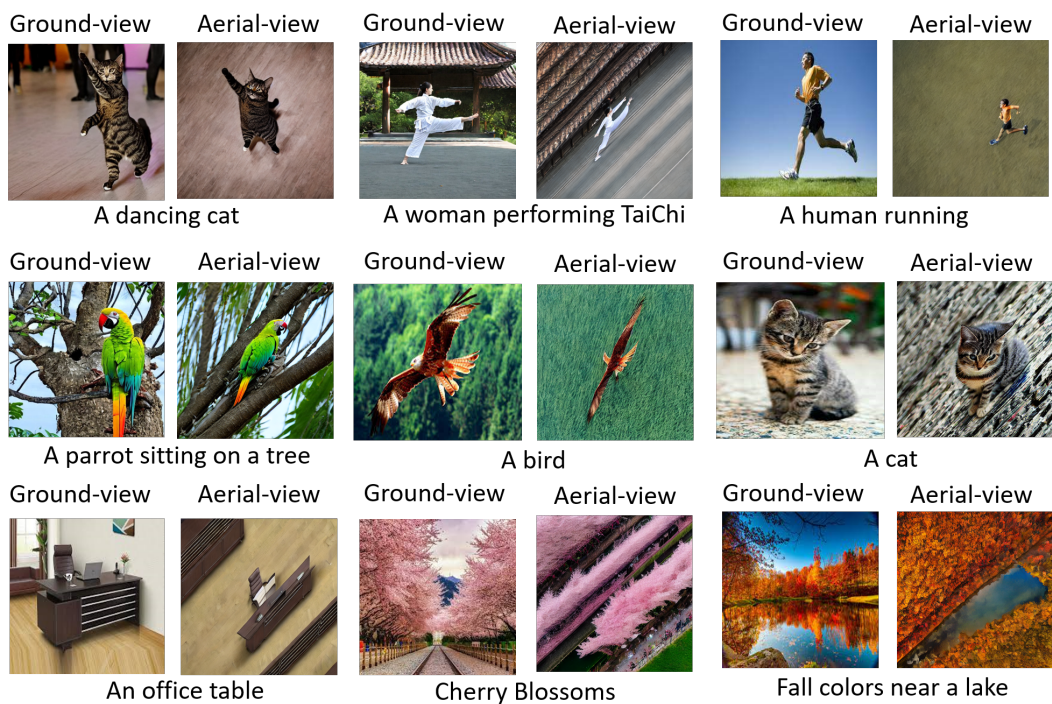


Figure 6.4: **Results.** We apply Aerial Diffusion on diverse images such as animals/ birds, natural scenes, human actions, indoor settings, etc and show that our method is able to generate high-quality high-fidelity aerial images.

Hence, when the target text embedding dictates a relatively small change to the source image, a linear combination between the optimized text embedding and the target text embedding generates a high-fidelity target image, faithful to the target text. In contrast, when a linear interpolation of the text prompts is applied to ground-to-aerial translation, depending on α , the images generated are either high fidelity (but low target text faithfulness) or high target text faithfulness (but low fidelity). Moreover, the ground-view image doesn't gradually change to an oblique-view image followed by aerial-view image, the manifold is not smooth. Rather, the change is quite drastic and it is difficult to find an optimal solution in the linear interpolation space. Essentially, when there is a large perspective changes from the source to target images (i.e. involving large rotation of camera poses), the image representation space is no longer "locally linear", thereby linear interpolation is no longer adequate to generate high-fidelity images.

6.2.2 Method

Motivated by the challenges described above, we propose Aerial Diffusion for text guided single-image ground-to-aerial translation. An overview of our solution is as follows. We start with a pretrained robust stable diffusion [284] model as the backbone. Our method has three stages. In the first step, we preprocess the ground-view image I_S with a carefully crafted homography transformation to generate I_{Sh} . This reduces the bias in the finetuning step. In the second step, we finetune the diffusion model by first optimizing the text-embedding within the vicinity of e_{src} to find e_{opt} that best generates I_{Sh} . Subsequently, we finetune the diffusion model to reconstruct I_{Sh} , given e_{opt} . In the third step on inferencing/sampling, we use an alternating strategy to manipulate the text embedding layer to generate a high-fidelity aerial image I_T . Next we describe each step

in detail.

Step 1: Preprocessing using a homography transformation. The bias acquired by the diffusion model during the second step of finetuning inhibits large transformations. One way to decrease the bias is to reduce the number of iterations while finetuning. However, this leads to unsurprisingly low quality generated images. To decrease the bias while finetuning, we preprocess the ground-view image by transforming it with a 2D *homography transformation* [344] (inverse perspective mapping). This homography projects the ground-view image to its rough 2D projected aerial view. Note that we are unable to use a 3D homography mapping to obtain the 3D aerial view projection, a better pseudo estimate of the aerial view, due to the unavailability of camera matrix, multi-views, depth information, etc. On the other hand, depth estimation methods [345] increase the complexity of the problem.

Without loss of generality, the 2D image captured by a ground-camera can be regarded as the projection of the scene in the front-face of the cube. A camera facing the top face of the cube will be able to capture the accurate 2D aerial view of the scene. Since we have no knowledge of the camera parameters corresponding to the ground-view image, we are unable to shift the camera to obtain a different view of the scene. With respect to the ground-camera, the 2D projection of the front-face of the cube on the bottom face of the cube is the best ‘aerial projection’ that we can get (inverse perspective mapping [344]). This aerial projection is nowhere close to the true aerial view and does not resemble the ground-view either. Hence, when the diffusion model is finetuned, the bias is much lower than what it would have been if the optimization/finetuning were done directly with the ground-view image. This is because of the disparities between the image space of I_{Sh} and e_{src}/e_{tgt} , ingrained in the pretrained network. Moreover, it provides a

pseudo estimate of the direction in which the image needs to be transformed in order to generate its aerial view at the inference stage.

We maintain the horizontal and vertical distance between the edges of the faces in the ground view and its projected aerial view, to better preserve the resolution and the aspect ratio. Formally, the coordinates (in order) of the corners of the ground-view image and the homography projected image are $\{(0, 0), (H, 0), (H, W), (0, W)\}$ and $\{(0, W), (H, 0), (H, W), (0, 2W)\}$ respectively. The homography can then be computed and applied.

Step 2: Finetuning the diffusion model. We first optimize the text-embedding [8, 283] to generate I_{Sh} and subsequently finetune the diffusion model using e_{opt} to generate I_{Sh} . In contrast to popular text-based image editing approaches that find the optimized text embedding e_{opt} in the vicinity of the target text embedding e_{tgt} , we find e_{opt} in the vicinity of the source text embedding e_{src} . This is due to two reasons: (i) the disparity between the homography transformed view and the target text is still large (though much smaller than the disparity between the ground-view and target text). Hence, it is unlikely that a good e_{opt} will be obtained when the optimization is run (around e_{tgt}) for a limited number of iterations. (ii) we do not want the network to develop a bias towards the homography image as the ‘aerial view’.

To find e_{opt} , we freeze the parameters of the generative diffusion model f_θ and optimize e_{src} using the denoising diffusion objective [343]. This optimization is run for a small number of iterations, in order to remain close to e_{src} for meaningful embedding space manipulation at inferencing.

$$\min_{e_{opt}} \sum_{t=T}^0 L(f(x_t, t, e_{opt}; \theta), I_{Sh}), \quad (6.2)$$

where $f(x, t, y)$ is the t -th backward diffusion step, θ denotes the U-net parameters, and

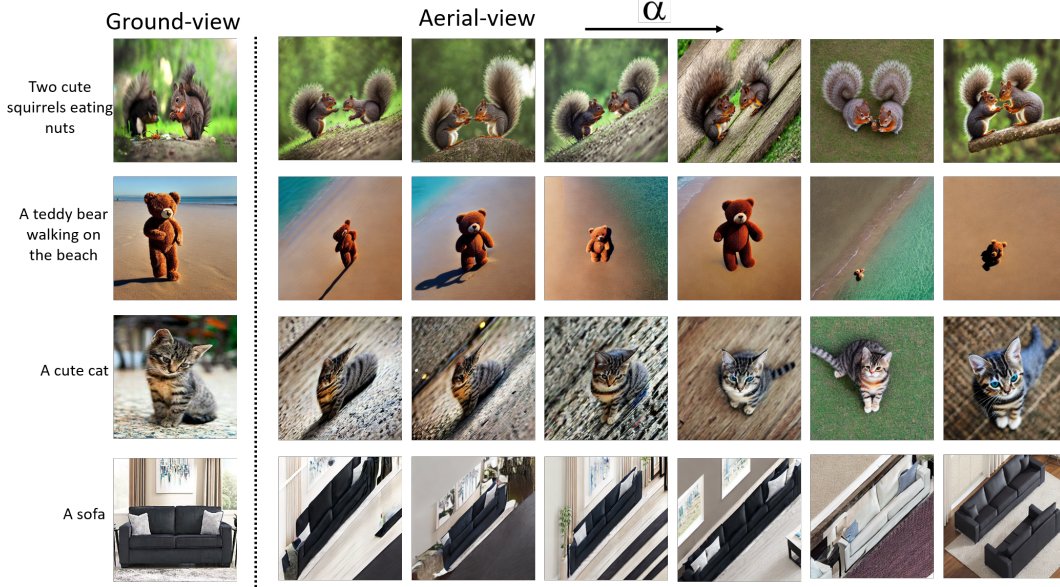


Figure 6.5: **Effect of α .** Low values of α generate images that are less aerial, high values of α generate low-fidelity images. A trade-off between the viewpoint and fidelity generates high-fidelity aerial images. The transformation, with α , is not smooth, reinforcing Postulate 2.

L is the denoising diffusion objective. To enable e_{opt} reconstruct the I_{Sh} with high fidelity, we finetune the diffusion model, again using the denoising diffusion objective [8, 303, 343]:

$$\min_{e_{theta}} \sum_{t=T}^0 L(f(x_t, t, e_{opt}; \theta), I_{Sh}). \quad (6.3)$$

Step 3: Inferencing/ sampling by text embedding manipulation. Our next step is to use the finetuned diffusion model to generate a high-fidelity aerial image. Prior work [8, 283] use linear interpolation between the optimized text embedding e_{opt} and the target text embedding e_{tgt} . Linear interpolation is not the best solution for large transformations such as ground-to-aerial generation and is unable to generate high-fidelity aerial images.

Sampling from stable diffusion [284] involves iteratively denoising the image for T steps conditioned by text, starting with random noise. To deal with the aforementioned issues, we

propose to alternate between two text embeddings e_1 and e_2 , starting with e_1 . We designate e_1 as the target text embedding e_{tgt} . This imposes a strong constraint on the diffusion model to generate an aerial view image corresponding to the text description. The bias of the diffusion neural network motivates the network to generate an image whose details are close to the ground-view image. However, merely relying on the bias of the neural network to capture all details of the scene is severely insufficient. Hence, we designate e_2 to be the linear interpolation of e_{opt} and e_{tgt} , controlled by the hyperparameter α . The linear interpolation can be mathematically represented as $e_2 = \alpha * e_{tgt} + (1 - \alpha) * e_{opt}$. e_2 enables the network to generate a high fidelity image while retaining the aerial viewpoint. For very low values of α , the generated image is less aerial, despite reinforcing the viewpoint to be aerial by applying e_1 alternatingly. This is because of the bias of the neural network. Very high values of α result in low fidelity images, some details of the generated aerial image are not consistent with the ground-view image. An optimal solution is by tuning α .

Linear interpolation enforces the generation of an image consistent with a text embedding in the linear space between e_{opt} and e_{tgt} . This is a reasonable when the desired change is small: when the image spaces corresponding to e_{opt} and e_{tgt} are closeby, linear interpolation works due to local linearity. When the desired change is large (such as ground-to-aerial translation), the image spaces corresponding to e_{opt} and e_{tgt} are not nearby. Since the representation spaces are not globally linear, it becomes essential to search for the solution in a much higher dimensional non-linear space. This is achieved by our alternating strategy. The pseudo code for the alternating strategy is given below.

While the sampling repetitively alternates between e_1 and e_2 , it is more beneficial to use e_1 (over e_2) at the first iteration. When the diffusion process starts with e_1 , the network generates

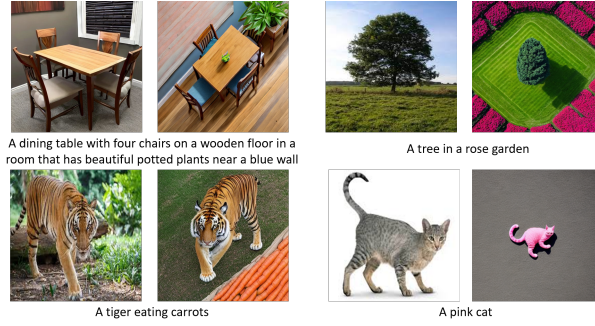


Figure 6.6: Based on the text description, Aerial Diffusion can generate aerial views with scene entities slightly different from the ground-view. The hallucination of background and unseen parts of the scene can also be controlled by text.

starts by generating an aerial image with details weakly dictated by its bias. Subsequent iterations that alternate between e_2 and e_1 fortify the generation of a high fidelity aerial image. On the contrary, when the diffusion process starts with e_2 , the generated image in the first iteration is less aerial though with very high fidelity. The bias, along with e_{opt} serve as a strong prior towards a non-aerial viewpoint. Subsequent iterations that use e_1 are unable to overcome this strong prior to alter the viewpoint to aerial view. Hence, we start inferencing with e_1 .

6.3 Experiments and Results

6.3.1 Implementation details

We use the stable diffusion [284] text-to-image model as the backbone architecture. It has been pretrained on the massive text-image LAION-5B dataset (laion2B-en, laion-high-resolution, laion-improved-aesthetics). Prior to the homography transformation, we resize all images to a resolution of 256×256 . We use OpenCV to apply the homography transformation on the image to generate a homography transformed image of resolution 512×512 , which is used to optimize the text embedding and the diffusion model in the next step. We finetune the text embedding for

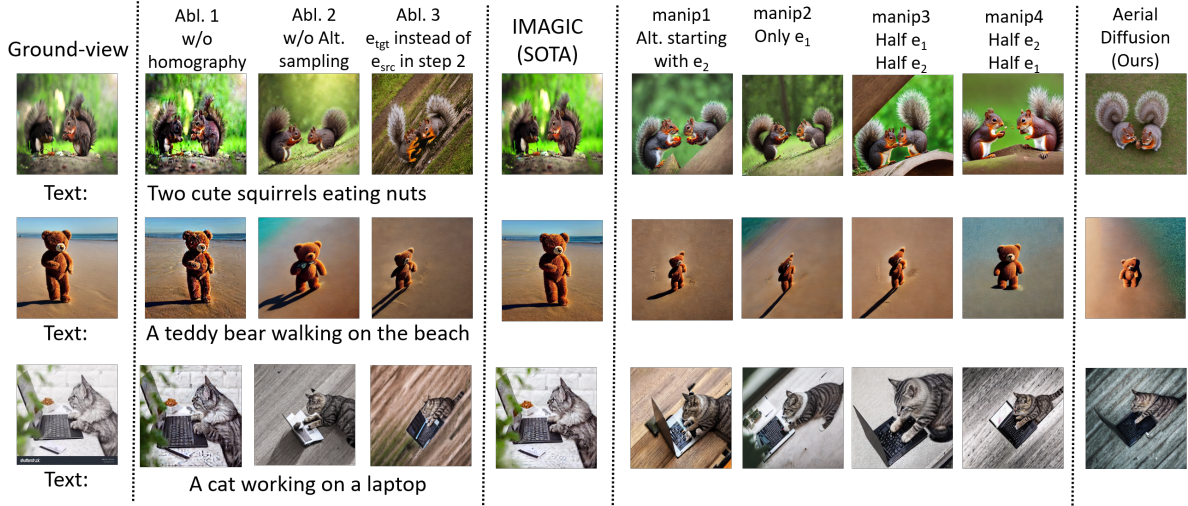


Figure 6.7: **Ablations and Comparisons.** Ablations: we prove the effectiveness of the homography, our alternating strategy over linear interpolation and finetuning with e_{src} instead of e_{tgt} . SOTA Comparisons: IMAGIC [8] (CVPR 2023) is unable to generate aerial views due to high bias towards the input image, domain gap and restricting the solution search to the linear interpolation manifold. Comparisons with other embedding space manipulation strategies that utilize both e_1 and e_2 reveal that our Alternating strategy is better.

500 iterations with a learning rate of $1e - 3$ using the Adam optimizer, and the diffusion model for 1000 iterations at a learning rate of $2e - 6$. For each image, this entire optimization takes between 8 and 9 minutes on one NVIDIA RTX A5000 GPU with 24GB memory.

While sampling/inferencing using the finetuned diffusion model, we iteratively refine the image, starting with random noise, for $T = 50$ iterations. At every iteration, the diffusion model is applied on the refined image from the previous iteration, as per the standard procedure [343] in sampling from diffusion models. The text embedding condition, while sampling at each iteration, alternates between e_1 and e_2 , starting from e_1 . The guidance scale is set to 7.5.

6.3.2 Qualitative evaluation

We apply our method on a number of real images from various domains including nature, human actions, buildings, etc. We collect most of these images from Google Images/ Flickr.

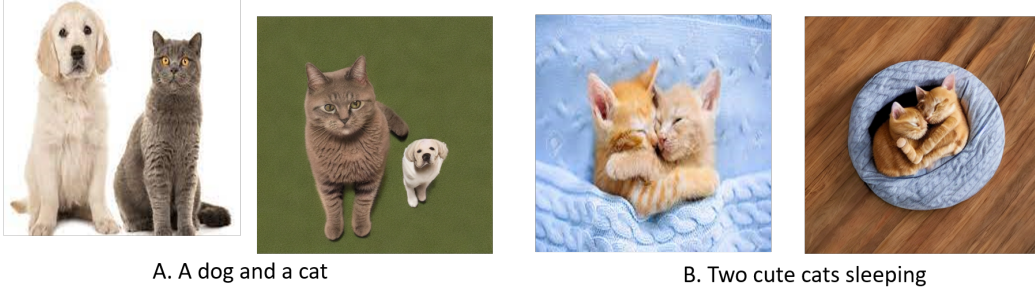


Figure 6.8: **Failure cases.** A. The identity of the dog and cat is interchanged. B. The fidelity of the background (bed) is lost.

Aerial images can be captured from varying heights and angles (including side/oblique views). We do not constrain the network on the height/ angle, hence, the diffusion model generates an aerial image of arbitrary height/ angle dictated by the random noise. For each image, we choose the best result corresponding to $0 < \alpha < 1$. Aerial Diffusion is successful in generating the aerial view, given its ground-view. It is able to hallucinate the aerial view of the entities in the scene (encompassing unseen aspects) as well as the background; while being faithful to the details in the ground-view. Since the underlying diffusion model is probabilistic, we get different results for different random seeds, all the generated images are faithful to the details in the ground-view as well as conform to the viewpoint being aerial. This diversity allows users to choose among a variety of options and is also a useful property in curating synthetic datasets.

More applications. We show that our method is capable of generating aerial views consistent with the text description, even when the text dictates an aerial view with scene entities slightly different from the ground-view. Moreover, background hallucination can be controlled by the text description as well.

Effect of α . The finetuned diffusion model has high bias towards the ground-view image. Hence, the value of α needs to be carefully tuned in order to generate a high fidelity aerial image. A low value of α implies higher weight to the optimized text embedding and low weight to ‘aerial view’.

This leads to the generated image being very similar to the homography image, the viewpoint of the generated image is less aerial. A high value of α implies higher weight to ‘aerial view’. The generated image is an aerial image, though with low fidelity. While the details in the generated aerial image are not completely different from the details in the ground-view image, due to the bias of the finetuned diffusion model, the fidelity or conformance to details contained in the ground-view image is still low. A good trade-off between fidelity w.r.t. ground-view image and ‘aerial view’ is achieved at mid-values of α . The change from ground-view to aerial-view as α varies from 0 to 1 is not gradual, reinforcing Postulate 2.

6.3.3 Ablations and comparisons

Ablating the model. Ablation 1 uses the ground-view image instead of the homography transformed image to finetune the diffusion model. It proves that the homography transformation is successful in significantly lowering the bias of the model and helps generate aerial views, reinforcing our solution to Postulate 1. Ablation 2 uses linear interpolation for sampling instead of our alternating strategy. Results with the alternating strategy are better (high fidelity images with faithfulness to ‘aerial view’) than the results with linear interpolation, justifying our solution to Postulate 2. Ablation 3 finds e_{opt} in the vicinity of e_{tgt} instead of e_{src} . Generated images are more aerial when e_{opt} is optimized in the vicinity of e_{src} , proving that it helps in reducing the bias.

SOTA Comparisons. We compare with IMAGIC [8] (CVPR 2023), DreamBooth [340] (CVPR 2023), SINE [283], SOTA text-based single image translation methods. Our method is far superior than prior art for ground-to-aerial translation, which are is unable to effectively perform ground-to-aerial translation due the high bias incurred while finetuning and searching for the

solution in a limited linear interpolation space.

Comparisons with other text embedding manipulation methods. We compare with other strategies to manipulate the text embedding space using e_1 and e_2 . In manip1, we condition on e_2 and e_1 alternately, starting from e_1 . Clearly, it is more beneficial to start sampling from e_1 as explained in Section 6.2.2. In manip2, we use just e_1 (text embedding corresponding to aerial view) to sample and rely on the bias of the network to generate the aerial image. Our alternating sampling method is able to generate higher fidelity aerial images. In manip3, for the first $T/2$ iterations, we sample using e_1 and for the second $T/2$ iterations, we sample using e_2 . In manip4, for the first $T/2$ iterations, we sample using e_2 and for the second $T/2$ iterations, we sample using e_1 . These experiments prove that our alternating sampling strategy works best.

Quantitative evaluation - user study. Text guided single image ground-to-aerial translation is a recent development, and Aerial Diffusion is the first solution towards this goal. As such, no standard benchmark (and ground-truth) or quantitative metrics exist for evaluation. We evaluate Aerial Diffusion via human perceptual evaluation and observe that Aerial Diffusion is able to generate high-fidelity aerial images. We conduct the following types of evaluation:

1. **Image Quality:** Given a ground-view image and an aerial-view image generated using Aerial Diffusion, we ask participants to determine if the generated aerial-view image is a high-fidelity (w.r.t. ground-view image) aerial-view image; and rate the image on the 5-point Likert scale. The average rating over 10 images (rated by 49 participants) is 3.289.
2. **Alternating Sampling:** Given a ground-view image and two aerial-view images generated using Aerial Diffusion and Ablation 2 (i.e. Aerial Diffusion without the Alternating Sampling method) respectively, we ask participants to choose the better high-fidelity aerial-view

image. **83.05%** of the participants rate the image generated using Aerial Diffusion as the one with higher quality.

3. **Reference for Aerial Diffusion:** Given a ground-view image and two aerial-view images generated using Aerial Diffusion and Ablation 3 (Aerial Diffusion with e_{tgt} instead of e_{src} while training) respectively, we ask participants to choose the higher fidelity aerial-view image. **78.125%** of the participants rate the image generated using Aerial Diffusion as the one with higher quality.

Quantitative evaluation - metrics. Our method uses prior knowledge (from robust pretraining), along with the knowledge gained while finetuning, to generate aerial images. It *hallucinates* large parts and views of the scene that it has not encountered before. Hence, comparisons against other ground-to-aerial methods [26, 279, 307, 314], that learn a specific data distribution by training on (an entire dataset with) paired data (and auxiliary information such as depth, semantic maps) is not relevant. In line with prior work [8, 273, 282, 283, 340] on text-based image translation, we report two metrics: (i) LPIPS to evaluate fidelity of aerial image w.r.t. ground-view image - the average score for $1 - \text{LPIPS}$ is 0.352 (higher, the better). However, similarity scores (such as FID, LPIPS) compute patch-wise image similarity and the aerial view is much different from the ground-view. No better evaluation strategy is available and is a direction for future work. (ii) CLIP score to evaluate the alignment of the generated aerial image with the text (dictating the viewpoint to be aerial) - the average CLIP score is 0.3233 (higher, the better). For reference, text-based image-editing methods such as Imagic [8] (CVPR 2023) and DreamBooth [340] (CVPR 2023) achieve CLIP scores of 0.25 to 0.3 on image editing dictating far lesser change than front-to-aerial translation.

6.4 Conclusion, Limitations and Future Work

In this chapter, we introduce a novel method, Aerial Diffusion, for generating aerial views from a single ground-view image using text as a guiding factor. We use homography as guidance and a diffusion model on the generated image with an alternating denoising direction based on switching between viewpoint text embedding and fidelity of the generated images. Our method has some limitations. Our method has some limitations: (i) the homography transformation results in a directional (diagonal) bias in the generated aerial image in many cases; (ii) it is limited to the knowledge contained in the pretrained stable diffusion model and is unable to hallucinate scenes [346] outside of this domain; (iii) the value of α in sampling needs to be manually tuned; (iv) the problem domain of unpaired ground-to-aerial does not have concrete quantitative analysis metrics. Future work can focus on the development of methods that overcome these limitations. Other directions include extending Aerial Diffusion to complex scenes with multiple objects (and an intricate background), generating higher-fidelity images, extending the method to videos, using the synthetic aerial data for aerial video analysis, detection, and recognition tasks.

Chapter 7: HawkI: Homography and Mutual Information Guidance for 3D-free Single Image to Aerial View

Widely available text-to-image models such as Stable Diffusion [284], trained on large-scale text and 2D image data, contain rich knowledge of the 3D world. They are capable of generating scenes from various viewpoints, including aerial views. However, due to limitations of the expressiveness of the text, we may not be able to completely describe the precise scene that we wish to generate. Moreover, the generative capabilities of the pretrained model is constrained by the aerial-view images in the dataset that it was trained on, which is typically limited. Consequentially, along with text, it is beneficial to use an easily available representative front-view image describing the aerial view of the scene we wish to generate. The task of generating aerial-view images from a given input image and its text description finds applications in the generation of realistic diverse aerial view synthetic data for improved aerial view perception tasks [22, 197, 216, 276, 347], and weak supervision for cross-view synthesis applications [308] such as localization and mapping [348], autonomous driving [349], augmented and virtual reality [350], 3D reconstruction [351], medical imaging [352], drone-enabled surveillance [353].

Aerial-view images corresponding to text and an input image can be sampled using text-to-3D and novel view synthesis (NVS) [9, 354]. These methods sample different camera viewpoints by explicitly specifying the camera angle. However, they often need to be trained on enormous,

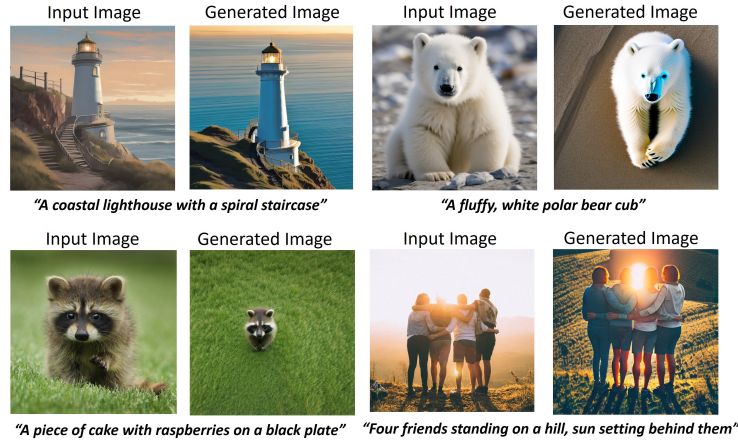


Figure 7.1: HawkI generates aerial-view images from a text description and an exemplar input image. It builds on a text to 2D image stable diffusion model and does not require any additional 3D or multi-view information at fine-tuning or inference.

large-scale datasets with 3D details and scenes from multiple views. Is it possible for text-to-image(2D) diffusion models to generate aerial-view images without any multi-view or 3D information?

Another closely related task is image editing [355] and personalization [28], where the goal is to use an input image and a target text to generate an image consistent with both inputs. These methods are generally successful in performing a wide range of non-rigid transformations including text-controlled view synthesis. However, the large translation required for aerial view synthesis makes them sub-optimal [354]. This is due to bias-variance trade-off issues, even more amplified when only a single input image is provided. Aerial Diffusion [356] attempted to alleviate this bias-variance trade-off, but at the cost of per-sample hyperparameter tuning, residual diagonal artifacts in many of the generated images arising from direct finetuning on sub-optimal homography projections and severe performance drops on complex scenes with multiple objects.

Main contributions. We propose HawkI for aerial view synthesis, guided by text and a single input image. Our method leverages text-to-image diffusion models for prior knowledge and does not require any 3D or multi-view data. Since explicitly specifying camera details in text descriptions isn't always possible, similar to prior work on text-based viewpoint generation [28, 356], we consider any generated image with a significantly higher viewpoint and altitude compared to the original image to be an aerial view. HawkI fuses techniques from classical computer vision and information theory within a stable diffusion backbone model to guide the synthesis of the aerial-view image. The key novel components of our algorithm include:

1. **Test-time optimization:** This step enables the model to acquire the characteristics of the input image, while maintaining sufficient variability in the embedding space for aerial-view synthesis. We condition the embedding space by sequentially optimizing the CLIP text-image embedding and the LoRA layers corresponding to the diffusion UNet on the input image and its Inverse Perspective Mapping (IPM) homography transformation in close vicinity. In addition to creating variance, IPM provides implicit guidance towards the direction of transformation for aerial-view synthesis.
2. **Mutual Information Guided Inference:** This step generates a semantically consistent aerial-view image while accounting for viewpoint differences. Unlike conventional approaches [357, 358] that rely on restrictive pixel-level constraints (often ineffective for different viewpoints), we propose a mutual information vastly guidance formulation. Mutual information guidance, rooted in information theory, ensures consistency between the contents of the generated image and the input image by maximizing the information contained between the probability distributions of the input image and the generated aerial image.

Our method performs *inference-time* optimization on the given text-image inputs and does not require a dataset to train on, hence, it is easily applicable to any in-the-wild image. To test our method, we collect a diverse set of synthetic images (from Stable Diffusion XL) and real images (from Unsplash), spanning across natural scenes, indoor scenes, human actions and animations. Qualitative and quantitative comparisons with prior work, on metrics such as CLIP [336] (measuring viewpoint and text consistency) and SSCD [359], DINOv2 [360] (measuring consistency w.r.t. input image), demonstrate that HawkI generates aerial-view images with a significantly better viewpoint-fidelity (or bias-variance) trade-off. We also present extensive ablation experiments and comparisons with 3D-based novel-view synthesis methods highlighting the benefits of our 3D-free classical guidance approaches. Our method can also be extended to generate more views that can be text-controlled (such as ‘side view’, ‘bottom view’, ‘back view’), as evidenced by our results.

7.1 Related work

3D and novel view synthesis. Novel view synthesis [322, 323, 361] from a single image is an active area of research in generative AI. Many methods [362, 363, 364, 365] use NeRF based techniques. Nerdi [366] use language guidance with NeRFs for view synthesis. Many recent methods use diffusion [9, 10, 367, 368, 369, 370, 371, 372] to sample different views. 3D generation methods [354, 373, 374, 375, 376] use text to guide the reconstruction. All of these methods use large amounts of multi-view and 3D data for supervised training. Methods like Zero-1-to-3 [9] and Zero-123++ [10] use a pretrained stable diffusion [284] model, along with large data for supervised training, to learn different camera viewpoints. 3D-free methods such as

Free3D [377] still require multi-view and 3D information while training.

Warping, scene extrapolation and homography. Scenescape [378], DiffDreamer [379] and similar methods [322, 380, 381] estimate a depth map, reproject the pixels into the desired camera perspective and outpaint the scene. Again, these methods require 3D and multi-view information at training stage. Using a homography to estimate the scene from an aerial perspective is highly inaccurate, hence, attempting to create realistic aerial view images by simply filling in missing information based on the homography (outpainting) leads to poor outcomes. Homography maps have also been used in various deep learning based computer vision solutions [382, 383, 384, 385, 386].

Image editing/ personalization. Diffusion models have emerged as successful tools for single image editing and personalization. Methods such as DreamBooth [28], DreamBooth LoRA [333], HyperDreamBooth [387], Textual Inversion [334], Custom Diffusion [388] are able to generate personalized images of subjects. Image editing methods such as Imagic [355], Paint-by-Example [389], ControlNet [390], DiffEdit [391] are able to edit images to perform non-rigid transformations and also use exemplar signals for guidance. However, these methods can either generate aerial images with low fidelity w.r.t. the input image or generate high-fidelity images with viewpoints very close to the input image.

Cross-view synthesis. Prior work on cross-view synthesis [26, 27, 279, 280, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316] are data intensive - they use paired data and modalities such as semantic maps, depth, multi-views, etc within their architectures. Aerial Diffusion [356] uses text and an exemplar image for the task by alternating sampling between viewpoint and homography projects. However, the generated images have diagonal artifacts with poor quality results for complex scenes that typically contain more than one object and requires manual per-

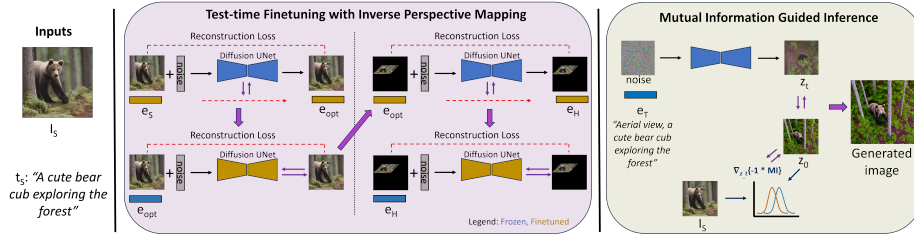


Figure 7.2: **Overview.** HawkI generates aerial-view images, using a text description and a single image I_S as supervisory signals. It builds on a pretrained text-to-image diffusion model, and does not use any 3D or multi-view information. It performs test-time finetuning to optimize the text embedding and the diffusion model to reconstruct the input image and its inverse perspective mapping in close vicinity. Such a mechanism enables the incorporation of image specific knowledge within the model, while retaining its imaginative capabilities (or variance). At inference, HawkI uses mutual information guidance to maximize the information between the probability distributions of the generated image and I_S , to generate a high-fidelity aerial-view image.

sample hyperparameter tuning.

Guidance techniques in diffusion. Guidance methods [293, 357, 392, 393] have been used to control and guide diffusion denoising towards semantic maps, image classes, etc. These guidance techniques cannot enforce view-invariant image-image similarity, critical for aligning the contents in two images with vastly different viewpoints.

7.2 Method

We present HawkI to generate aerial view images using a single input image I_S and its text description t_S (e.g. ‘a cosy living room’, can be obtained using the BLIP-2 model [394]). We do not use any training data or 3D/multi-view data. We leverage the pretrained text-to-2D image stable diffusion [284] model to serve as a strong prior, and utilize classical computer vision and information theory principles to achieve the desired goal in a holistic manner.

- *Test-time optimization:* We perform multi-step test-time optimization to incorporate the input image I_S within the pretrained model, at an appropriate bias-variance trade-off.

Specifically, we optimize the CLIP text-image embedding and the LoRA layers in the diffusion UNet sequentially on the input image and its inverse perspective mapping, in close vicinity. This additionally conditions the embedding space viewpoint transformations, along with acquiring the characteristics of the input image.

- *Inference*: To generate the aerial-view image, we use the target text description t_T , that takes the form ‘aerial view, ’ + t_S (e.g. ‘aerial view, a cosy living room’). To ensure that the generated aerial image is semantically close to the input image, we use mutual information guidance.

Next, we describe our method in detail.

7.2.1 Test-time optimization

The text-to-2D image stable diffusion model has knowledge of the 3D world as a consequence of the large amount of diverse data it has been trained on. It understands [13] different viewpoints, different styles, backgrounds, etc. Image editing and personalization methods such as DreamBooth [28], DreamBooth LoRA [333], Imagic [355], SVDiff [395] exploit this property to perform transformations such as making a standing dog sit and generating its image in front of the Eiffel tower. At a high level, the standard procedure adopted by these methods to generate edited or personalized images is to finetune the model on the input image, followed by inferencing. These methods are however not very successful in text-guided aerial view synthesis, which demands a large transformation. Specifically, directly finetuning the diffusion model on e_S to reconstruct I_S results in severe overfitting, where e_S is the CLIP text embedding for t_S . This makes it difficult for the model to generate large variations to the scene required for aerial view synthesis.

We propose a four-step finetuning approach to enable the model to learn the characteristics of I_S , while ensuring sufficient variance for aerial view generation.

7.2.1.1 Optimization using I_S :

In the first step, we start from e_S and compute the optimized CLIP text-image embedding e_{opt} to reconstruct I_S using a frozen diffusion model UNet using the denoising diffusion loss function L [343].

$$\min_{e_{opt}} \sum_{t=T}^0 L(f(x_t, t, e_{opt}; \theta), I_S), \quad (7.1)$$

where t is the diffusion timestep and x_t is the latents at time t . This formulation allows us to find the text embedding that characterizes I_S better than the generic text embedding e_S .

Next, to enable e_{opt} accurately reconstruct I_S , we optimize the diffusion UNet using the denoising diffusion objective function. Note that we insert LoRA layers within the attention modules in the diffusion UNet and finetune only the LoRA layers with parameters θ_{LoRA} , the rest of the UNet parameters are frozen,

$$\min_{\theta_{LoRA}} \sum_{t=T}^0 L(f(x_t, t, e_{opt}; \theta), I_S). \quad (7.2)$$

While optimizing e_{opt} instead of e_S to reconstruct I_S ensures lesser bias (or more variance), the embedding space is still not sufficiently conditioned to generate an aerial view of the image.

7.2.1.2 Optimization using inverse perspective mapping.

Inverse perspective mapping (IPM) [344] is a homography transformation from classical computer vision to generate the aerial-view of an image from its ground-view. Despite not being accurate, it can provide pseudo weak supervision for the generation of the aerial image and also add more variance to the embedding space. We denote the inverse perspective mapping of the input image by I_H , computed following [356]. We perform the following optimization steps to condition the embedding space towards the desired viewpoint transformation. To find the text embedding e_H that best characterises I_H , we start from e_{opt} and optimize the text embedding with a frozen diffusion model. Finding e_H in the vicinity of e_{opt} instead of e_S ensures that the text-image space corresponding to e_S doesn't get distorted to generate the poor quality IPM image. Next, we finetune the diffusion model using the denoising diffusion objective function to reconstruct I_H at e_H . Again, only the LoRA layers are finetuned, the rest of the UNet is frozen.

Note that we find e_{opt} and e_H by optimizing e_S and e_{opt} , respectively for a small number of iterations. We need to ensure that e_S , e_{opt} and e_H are all in close vicinity. Our finetuning approach conditions the embedding space to encapsulate the details of I_S and viewpoint, while having sufficient variance to generate large transformations required for the generation of the aerial image.

7.2.2 Mutual Information Guided Inference

Our next step is to use the finetuned diffusion model to generate the aerial view image for the text prompt t_T . The text embedding for t_T is e_T . Diffusion denoising, conditioned on e_T is capable of generating aerial images corresponding to I_S . However, oftentimes, the contents of the

generated aerial view image does not align well with the contents of I_S . Consequently, to ensure high fidelity generations, our goal is to guide the contents of the aerial view image towards the contents of I_S .

Similarity measures such as L1 distance, cosine similarity are capable of providing this guidance. However, they are not invariant to viewpoint/ structure. Since we want the two images to be similar (while observed from different viewpoints), using metrics that impose matching at the pixel (or feature) level is not the best approach. Rather, it is judicious to use the probability distribution of the features.

In information theory, mutual information quantifies the ‘amount of information’ obtained about one random variable by observing the other random variable. Mutual information has been used [396, 397, 398, 399] to measure the similarity between images in various computer vision tasks such as medical image registration, frame sampling, etc. It yields smooth cost functions for optimization [400]. The mutual information between two probability distribution functions (pdf) $p(x), p(y)$ for two random variables $\mathcal{X}, \mathcal{Y}$ is defined as $I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X}, \mathcal{Y})$ where $H(\mathcal{X}), H(\mathcal{Y})$ are the entropies of $p(x), p(y)$ and $H(\mathcal{X}, \mathcal{Y})$ is the joint entropy. Entropy of a random variable X is a measure of its uncertainty, $H(\mathcal{X}) = -\sum_{x \in \mathcal{X}} p_X(x) \log(p_X(x))$;

$$H(\mathcal{X}, \mathcal{Y}) = - \sum_{(x,y) \in \mathcal{X}, \mathcal{Y}} p_{XY}(x, y) \log(p_{XY}(x, y)).$$

Thus,

$$I(\mathcal{X}, \mathcal{Y}) = - \sum_{(x,y) \in \mathcal{X}, \mathcal{Y}} p_{XY}(x, y) \frac{\log(p_{XY}(x, y))}{p_X(x)p_Y(y)}.$$

Hence, mutual information, in some sense, measures the distance between the actual joint distribution

between two probability distributions and the distribution under an assumption that the two variables are completely independent. Thus, it is a measure of dependence [401] and can be used to measure the information between two images. In order to maximize the similarity in content between I_S and the generated aerial image, we maximize the mutual information between them. We define our mutual information guidance function as follows.

Let z_t denote the predicted latents at timestep t . We denote $z_{0,t}$ as the latents of the final predicted image extrapolated from z_t i.e. if the denoising were to proceed in a vanilla fashion in the same direction that computed z_t , the latents of the final predicted image would be $z_{0,t}$. At every step of sampling (except the final step), we wish to maximize the mutual information between $z_{0,t}$ and z_S where z_S are the latents corresponding to I_S . Hence, the guidance function we wish to maximize is, $G_{MI} = I(z_{0,t}, z_S)$.

The computation of mutual information requires us to compute the marginal and joint probability density functions (pdf) of $z_{0,t}$ and z_S . We construct 2D histograms of the latents (by reshaping the latents of size $C \times H \times W$ into $C \times HW$) and compute their marginal pdfs. The joint pdfs can then be computed from the marginal pdfs, which can be plugged into the formula for mutual information. Our next task is to use G_{MI} to guide the generation of the aerial image.

Guidance techniques such as classifier-free guidance [392], universal guidance [357] and steered diffusion [393] modify the sampling method to guide the image generation with feedback from the guidance function. The gradient of the guidance function w.r.t. the predicted noise at timestep t is an indicator of the additional noise that needs to be removed from the latents to steer the generated image towards the guidance signal. Synonymously, the gradient of the guidance function w.r.t. the predicted latents is an indicator of the direction in which the latents

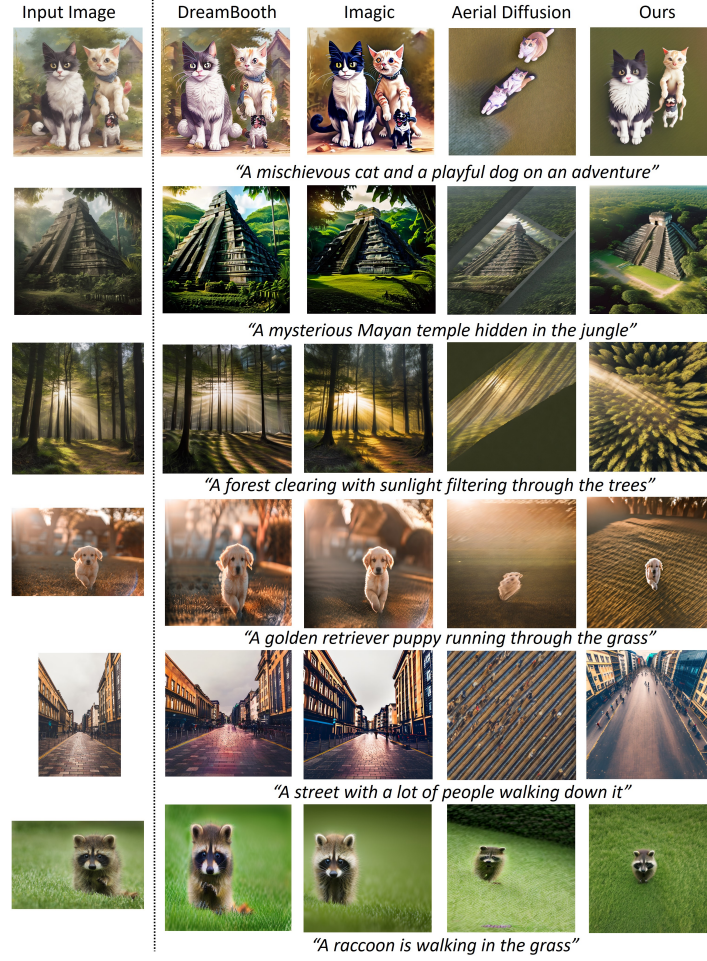


Figure 7.3: Compared to state-of-the-art text + exemplar image based methods, HawkI is able to generate images that are “more aerial”, while being consistent with the input image. The top three images are from the HawkI-Syn dataset, the bottom three images are from the HawkI-Real dataset.

need to move in order to maximize their alignment with the guidance function. Specifically, at every step of sampling (except the final step), we modify the predicted latents z_t as $\hat{z}_t = z_t - \lambda_{MI} \nabla_{z_t} (-1 * G_{MI})$. Note that we use the negative of the mutual information to compute the gradients since we want to maximize the mutual information between the generated latents and the input image.

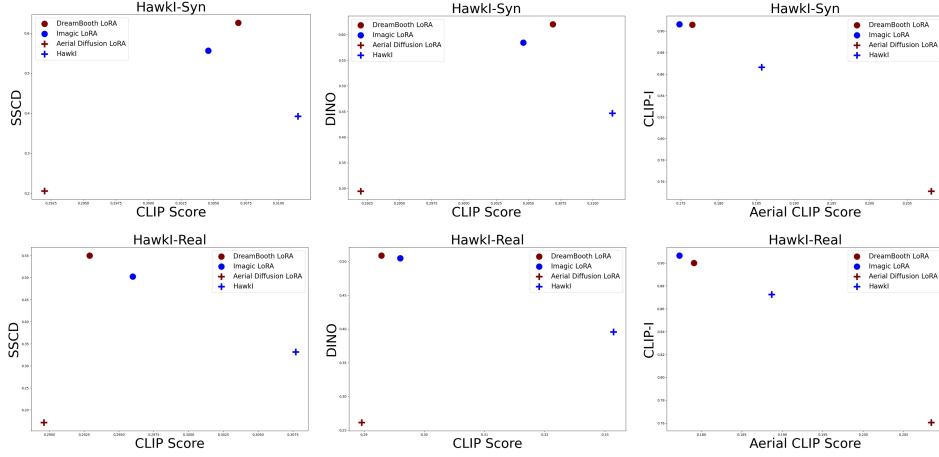


Figure 7.4: HawkI achieves the best viewpoint-fidelity trade-off amongst prior work on text + exemplar image based aerial-view synthesis, on various quantitative metrics indicate of text-alignment (for viewpoint and a broad description of the scene) and image alignment (for fidelity w.r.t. input image).

7.3 Experiments and Results

Data. We collect a synthetic dataset, HawkI-Syn and a real dataset, HawkI-Real. Both datasets contain images across a wide variety of categories including indoor scenes, natural scenes, human actions, birds/ animals, animations, traffic scenes and architectures. HawkI-Syn contains 500 images that were generated using Stable Diffusion XL [402]. To generate the text prompts for the generated images in HawkI-Syn, we used Large Language Models (LLMs) such as ChatGPT and Bard. HawkI-Real contains 139 images downloaded from the Unsplash website, the text descriptions for these images were obtained using the BLIP-2 [394] model.

Training details. We use the stable diffusion 2.1 model in all our experiments, ablations and comparisons. All our images (except for images in HawkI-Real) are at a resolution of 512×512 . With respect to I_S , we train the text embedding and the diffusion model for 1,000 and 500 iterations respectively, at a learning rate of $1e - 3$ and $2e - 4$ respectively. Using 1000 iterations to optimize the text embedding ensures that the text embedding e_{opt} at which I_S is reconstructed

is not too close to e_S , which would make it biased towards I_S otherwise. Similarly, it is not too far from e_S either, hence the text embedding space learns the characteristics of I_S . With respect to I_H , we train the text embedding and the diffusion UNet for 500 and 250 iterations respectively. We want e_H to be in close vicinity of e_{opt} ; we train the diffusion model for just 250 iterations so that the model does not completely overfit to I_H . The role of I_H is to create variance and provide pseudo supervision, it is not an accurate approximation of the aerial view. We set the hyperparameter for mutual information guidance at $1e - 5$ or $1e - 6$, the inference is run for 50 steps.

Computational cost. For each input image, HawkI takes 3.5 minutes to perform test-time optimization on one NVIDIA A5000 GPU with 24 GB memory. The inference time is consistent with that of Stable Diffusion, about 7 seconds to generate each sample with 50 denoising steps. The computational cost is on par with text-based image personalization [28, 355] and text-based aerial-view synthesis [356] methods. The number of network parameters is also consistent across all these models, we use the Stable Diffusion v2.1 + LoRA backbone across methods.

Quantitative evaluation metrics. We follow prior work on text-based image editing/ personalization and text-based view synthesis to evaluate our method:

- **Viewpoint and text alignment:** We use the text description ‘aerial view,’ + t_S and ‘aerial view’, along with the generated image, to compute the CLIP-Score [336] and the A-CLIP Score respectively. The former indicates alignment of the generated image with the detailed textual description of the image describing the contents along with the viewpoint; the latter focuses more on the viewpoint.

- Image fidelity and 3D coherence: To evaluate the overall alignment of the contents of the generated aerial-view image with the input image, we compute the CLIP-I score [28] which measures the cosine similarity between the embeddings of the aerial-view image and the input image in the CLIP space. For a better indicator of the fidelity and 3D coherence between the two images, we also use the self-supervised similarity detection metrics, DINOv2 [360, 403] and SSCD [359].
- Top-1 accuracy on a downstream UAV task.
- Ground-truth comparison with images sampled from 3D models using CLIP, LPIPS, and DINO metrics.

Higher values are desired for each of these metrics. Viewpoint faithfulness and fidelity w.r.t input image are a direct result of the bias-variance trade-off of the model, and high values for both are desired. However, as noted by Blau et. al. [404], maximizing both is not straightforward; inevitably one of the factors will degrade in response to the improvement in the other. For each input image, we generate 5 aerial images, with random noise initializations, and choose the image with the highest CLIP + SSCD score (since CLIP is an indicator of viewpoint + content alignment and SSCD score measures the fidelity w.r.t. the input image).

7.3.1 Comparisons against text + exemplar image based methods

We compare our method with DreamBooth LoRA [28], a text-based image personalization method; Imagic LoRA [355], a text-based image editing method; and Aerial Diffusion LoRA [356], a method for text-based aerial image generation from a single image. We keep the backbone stable diffusion model, image prompts, training details, and evaluation method consistent across



Figure 7.5: **(Left figure.)** Ablation experiments show that Inverse Perspective Mapping helps in the generation of images that are aerial, mutual information guidance helps in preserving the contents w.r.t. input image. **(Right figure.)** We compare with latest related work on novel view synthesis: Zero-1-to-3[9] and Zero123++ [10]. Both methods use the pretrained text-to-2D-image stable diffusion model along with the 800k 3D objects dataset, Objaverse [11], for training. Our method uses just the pretrained text-to-2D-image stable diffusion model to generate better results for the task of aerial view synthesis, guided by text and a single input image.

all comparisons.

We show qualitative results in Figure 9.3. Our method is able to generate aerial views as per input image guidance across a diverse set of scenes. Our method generates results that are more aerial in viewpoint than DreamBooth and Imagic, while being largely consistent with the contents of the input image. Aerial Diffusion is unable to generate good quality images of scenes that have many objects. Our method is able to deal with complex scenes as well as modify the viewpoint.

We show the quantitative results in Figure 7.4. Our method achieves a higher CLIP score than all prior work, indicating that it is able to generate an aerial view of the scene with contents dictated by the text better than prior work. The A-CLIP score achieved by our method is higher than that of DreamBooth and Imagic, indicating better conformance to the aerial viewpoint. Even though the A-CLIP score of HawkI is lower than than of Aerial Diffusion, Aerial Diffusion generates poor quality images for scenes with more than one object and also has diagonal artifacts in its generated images, as we observe from the other metrics and qualitative results, thus,

offsetting its high A-CLIP Score.

CLIP-I and self-supervised metrics such as SSCD and DINO are not viewpoint invariant. In many cases, since Imagic and DreamBooth generate views close to the input view, rather than aerial views, it is natural for them to have higher CLIP-I, SSCD and DINO Scores. Our method has a much higher CLIP-I, SSCD and DINO score than Aerial Diffusion, showing considerable improvement over prior work in retaining the fidelity and 3D consistency w.r.t. the input image, while modifying the viewpoint. In summary, our specialized aerial-view synthesis method achieves the best viewpoint-fidelity trade-off amongst all related prior work.

7.3.2 Comparisons against 3D based novel view synthesis (NVS) methods

We compare with state-of-the-art benchmark methods on stable diffusion based novel view synthesis from a single image. Zero-1-to-3[9] and Zero123++ [10], both, train on large amounts of multi-view and 3D data from Objaverse [11] contain 800K+ 3D models; in addition to leveraging a pretrained text-to-2D image stable diffusion model. In contrast, our method does not use any multi-view or 3D information and is capable of generating better results in multiple cases. Another task-level difference between our method and prior work on NVS is that the latter aim to explicitly control the camera angle and generate 3D objects in Zero-1-to-3[9], the camera-angle generated by our method is arbitrary within the realms of the text control. The CLIP scores on HawkI-Syn for Zero123++ and HawkI are 0.3071 and 0.3115 respectively, the DINO scores are 0.4341 and 0.4466 respectively. On HawkI-Real, the CLIP score for Zero123++ and HawkI are 0.2908 and 0.3077 respectively, the DINO scores are 0.3916 and 0.3956 respectively. Our aerial-view synthesis method, even without any 3D/ multi-view information and large dataset

training, is better than or comparable to 3D-based NVS methods.

7.3.3 Downstream application.

We performed a proof-of-concept experiment using HawkI to generate synthetic images for key-frames of scene-based human actions in the UAV Human [22] dataset. By computing the L2 distance between self-supervised features of UAV Human video frames and synthetic images, we achieved a 32.14% accuracy in zero-shot action recognition, representing a 7.14% (absolute) improvement.

7.3.4 Ground-truth comparison

We obtained 3D models and text descriptions from Dreamfusion [354] to extract the front-view and top-view (ground-truth or GT). We evaluated Zero123++ and HawkI using CLIP Scores (higher the better), and the numbers were 0.2991, 0.3316, respectively. The DINO score (higher the better), which measures self-supervised similarity between the generated images and the GT, are 0.3912, 0.3710 for Zero123++ and HawkI respectively. The LPIPS scores (lower the better) are 0.5801, 0.6373 for Zero123++ and HawkI respectively - our method generates images that are higher in elevation due to ‘aerial view’ being the text-control. Overall, our 3D-free method, built on stable diffusion, is comparable to 3D methods such as Zero123++ which uses stable diffusion + 800k+ 3D objects.

7.3.5 Ablation studies

Please refer to the figure. In the second column, we show results of our model where it is neither finetuned on the homography image nor uses mutual information guidance for sampling. Thus, the text embedding for the input image, followed by the diffusion UNet are finetuned and the diffusion model generates the aerial image by diffusion denoising, without any mutual information guidance. Many of the generated images either have low fidelity or have low correspondence to the aerial viewpoint. In the experiment in the third column, we add mutual information guidance to the model in column 2. We see higher fidelity (than column 2) of the generated images w.r.t. the input image. In the fourth column, we add the homography image finetuning step to the model in column 2, but do not use mutual information guidance at inference. The generated images, in many cases, are aerial, but have lower fidelity w.r.t. the input image. In the final column, we show results with our full model. The generated images achieve the best trade-off between the viewpoint being aerial and fidelity w.r.t. the input image, in comparison to all ablation experiments. Quantitative analysis:

- **Effect of I_H :** To study the effect of I_H , we compare the CLIP and A-CLIP scores for the full model vs full model w/o I_H . The scores in all cases where I_H is not present are lower, indicating lower consistency with the viewpoint being ‘aerial’. For instance, on HawkI-Real, the CLIP scores for the full model and full model w/o G_{MI} are 0.3077 and 0.3040 respectively. The A-CLIP scores for the full model and full model w/o I_H are 0.1887 and 0.1842 respectively.
- **Effect of G_{MI} :** To study the effect of I_H , we compare the SSCD and DINO scores for the full model vs model w/o G_{MI} . The scores in all cases where G_{MI} is not present are lower.

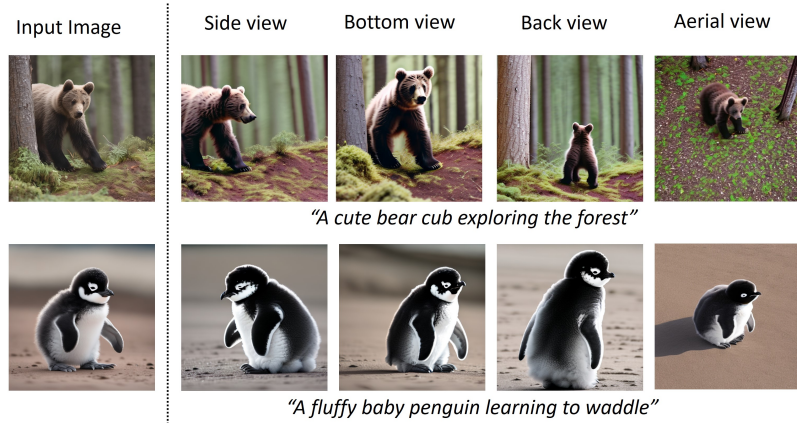


Figure 7.6: HawkI can be extended to generate other text-controlled views as well. For instance, on HawkI-Real, the SSCD scores for the full model and model w/o G_{MI} are

0.3314 and 0.3204 respectively. The DINO scores for the full model and model w/o G_{MI} are 0.3956 and 0.39 respectively.

7.3.6 Comparison with other metrics for guidance.

We compare with two other metrics for diffusion guidance at inference: (i) $L2$ distance between the features of the generated image and the input image, (ii) a metric inspired by Wasserstein distance or Earth Mover’s distance, for which we compute the distance between the histograms of the probability distributions of the two images. Our mutual information guidance method is *better at preserving the fidelity w.r.t the input image*, as evidenced by *higher SSCD scores*. The SSCD score on HawkI-Real for Wasserstein guidance, $L2$ guidance, and mutual information guidance are 0.3137, 0.3204 and 0.3314 respectively. The DINO score on HawkI-Real for Wasserstein guidance, $L2$ guidance, and mutual information guidance are 0.3847, 0.3858 and 0.3956 respectively.

7.3.7 Text controlled view synthesis: Other views

HawkI can be extended to generate other text-controlled views such as side view, bottom view and back view. We modify the target text t_T to indicate different viewpoints, and retain the other finetuning/ inference details.

7.3.8 3D-free HawkI + 3D priors?

Our 3D-free approach **complements** 3D-based methods [10, 373]. While **3D data collection and large-scale training are expensive and unsustainable**, front-view 2D images are more readily available[35]. Hence, it is beneficial to solve the fundamental problems associated with the task in a data-efficient (or 3D-free) manner. Our goal is to push the frontiers of 3D-free aerial-view generation from a single image, achieving results comparable to methods [10] using stable diffusion + 800k 3D objects. That being said, HawkI be combined with 3D approaches, to multiply the benefits of 3D and 3D-free approaches. To validate this, we replaced the homography prior with the image from the 3D-based Zero123++ [10] approach and evaluated our HawkI model on images from HawkI-Syn. The CLIP scores for Zero123++, HawkI, and “HawkI with Zero123++ prior” are 33.17, 33.17, and 33.81, respectively. The DINO scores for Zero123++, HawkI, and “HawkI with Zero123++ prior” are 0.3613, 0.3977, and 0.4612, respectively, thus demonstrating our claim!. Also, using 3D priors with HawkI allows finer camera control.

7.4 Conclusions, Limitations and Future Work

We present a novel method for aerial view synthesis. Our goal is to leverage pretrained text-to-image models to advance the frontiers of text + exemplar image based view synthesis *without*

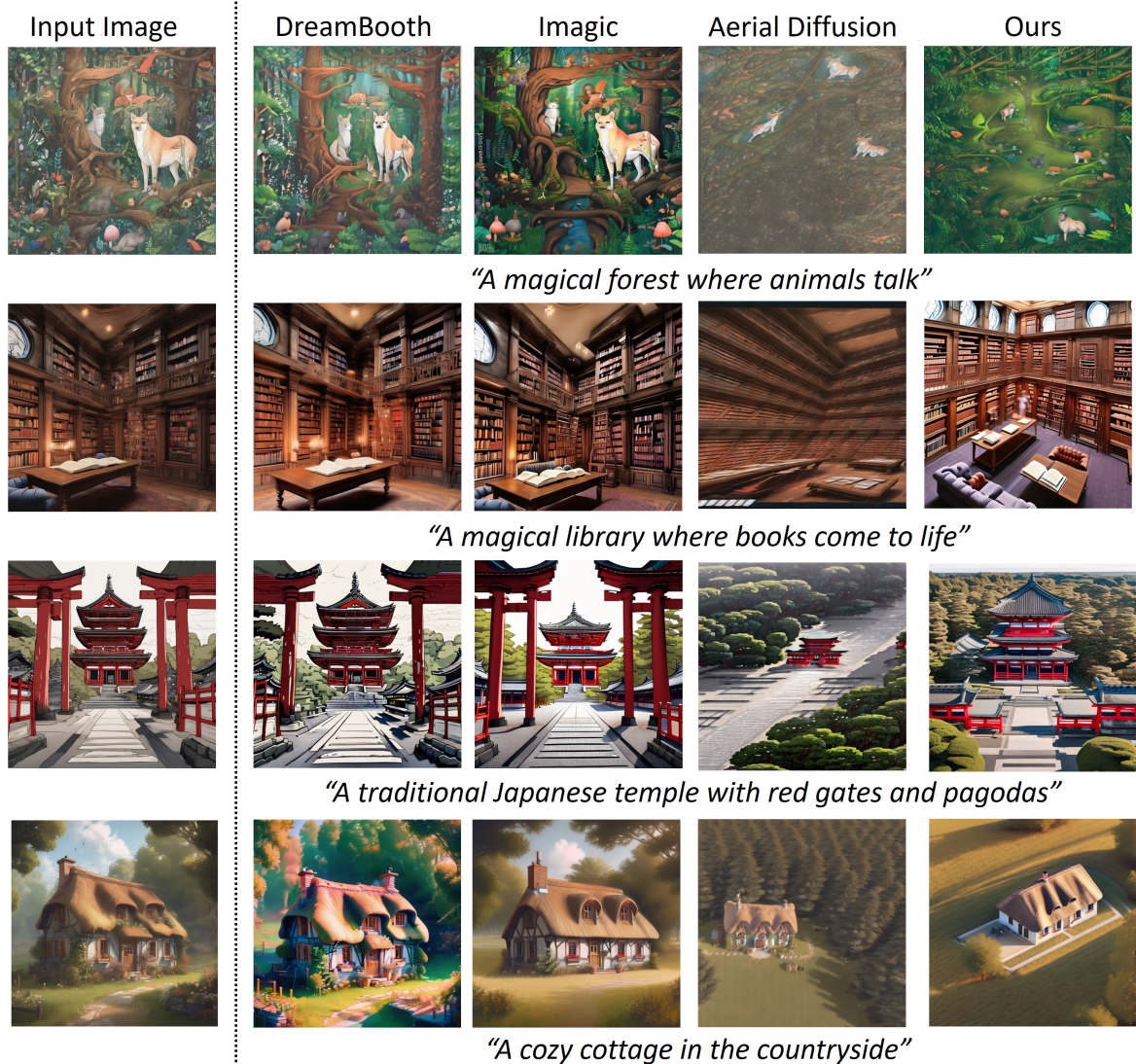


Figure 7.7: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.



Figure 7.8: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

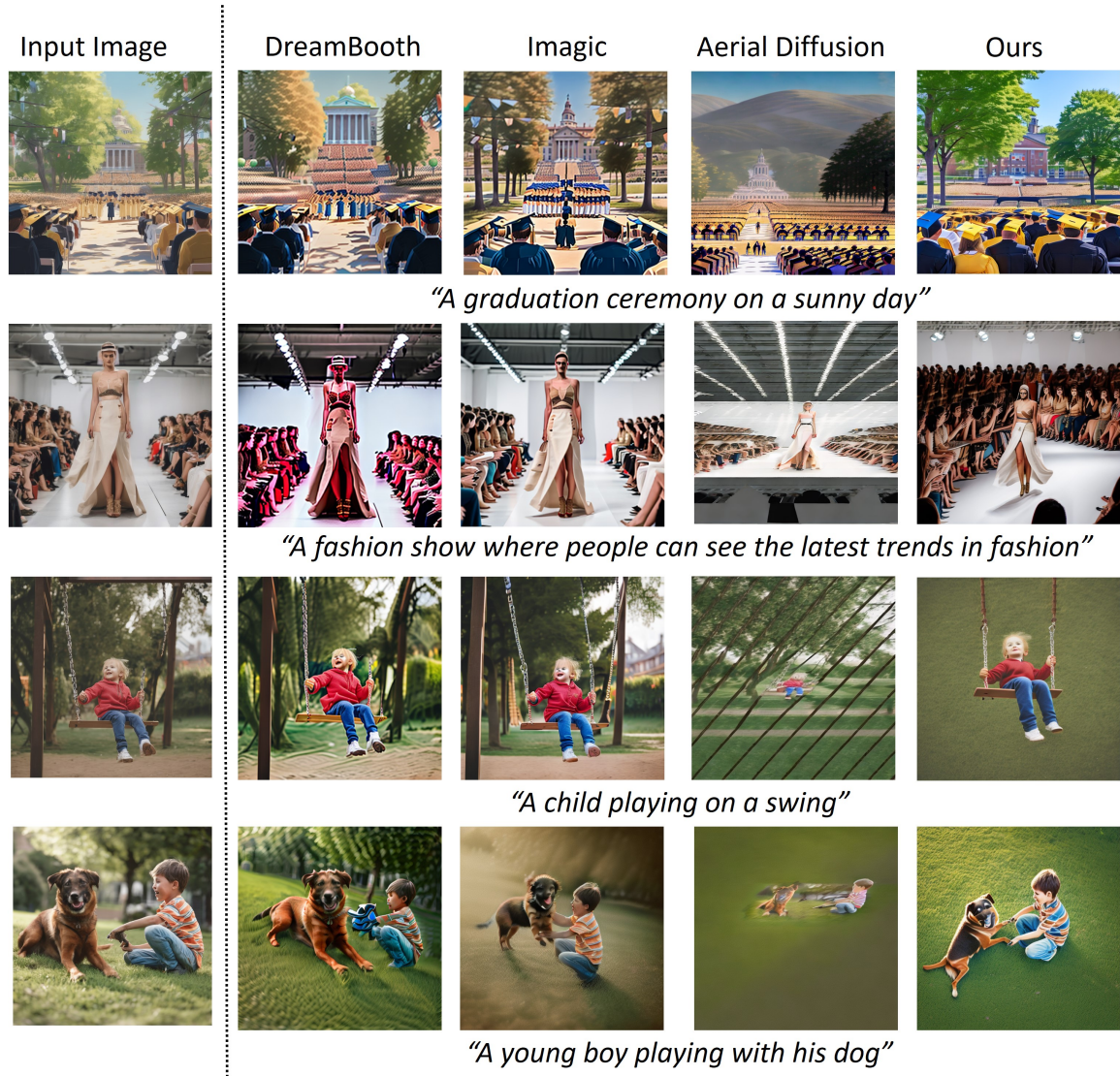


Figure 7.9: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

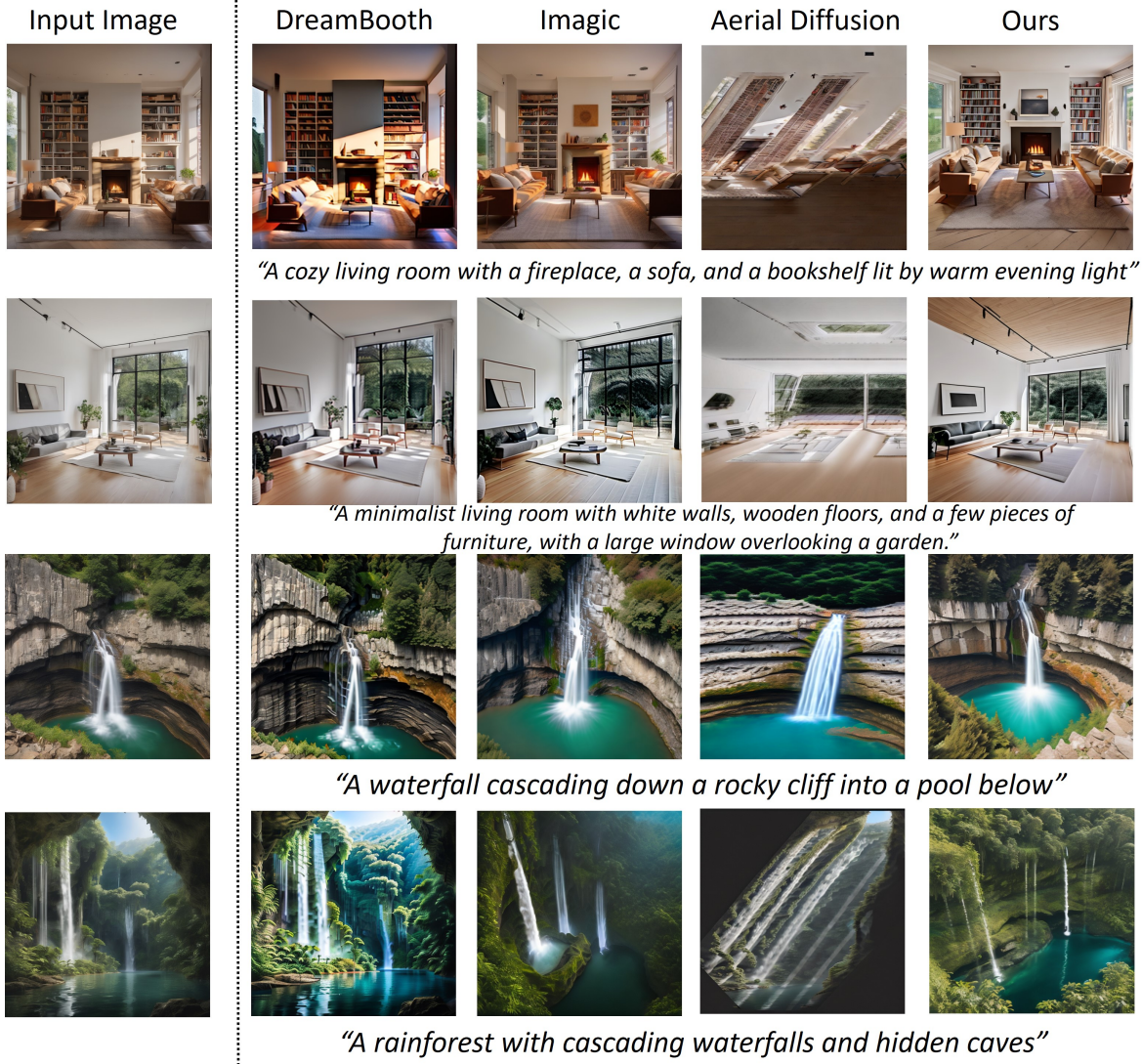


Figure 7.10: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Syn dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

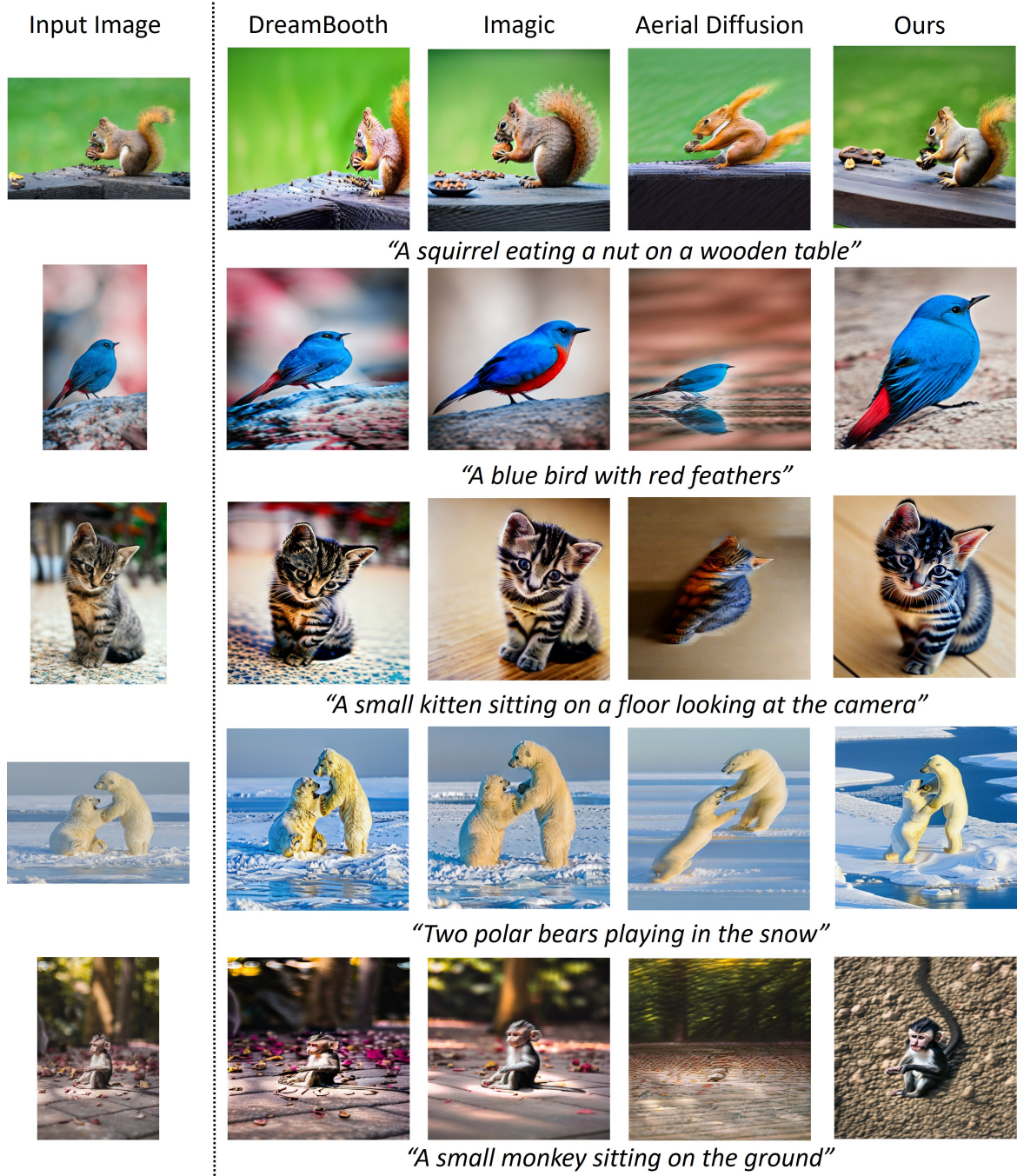


Figure 7.11: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

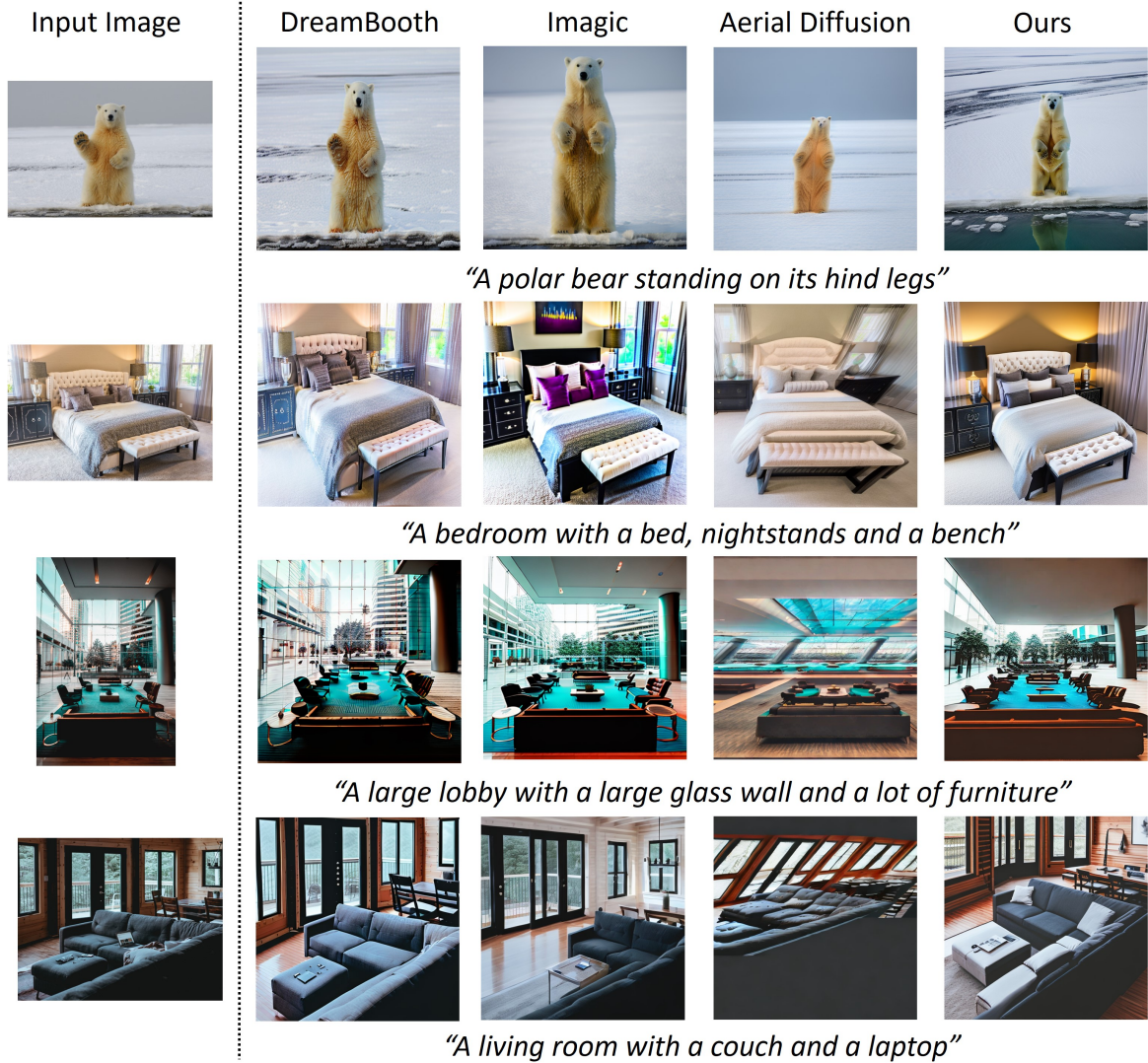


Figure 7.12: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

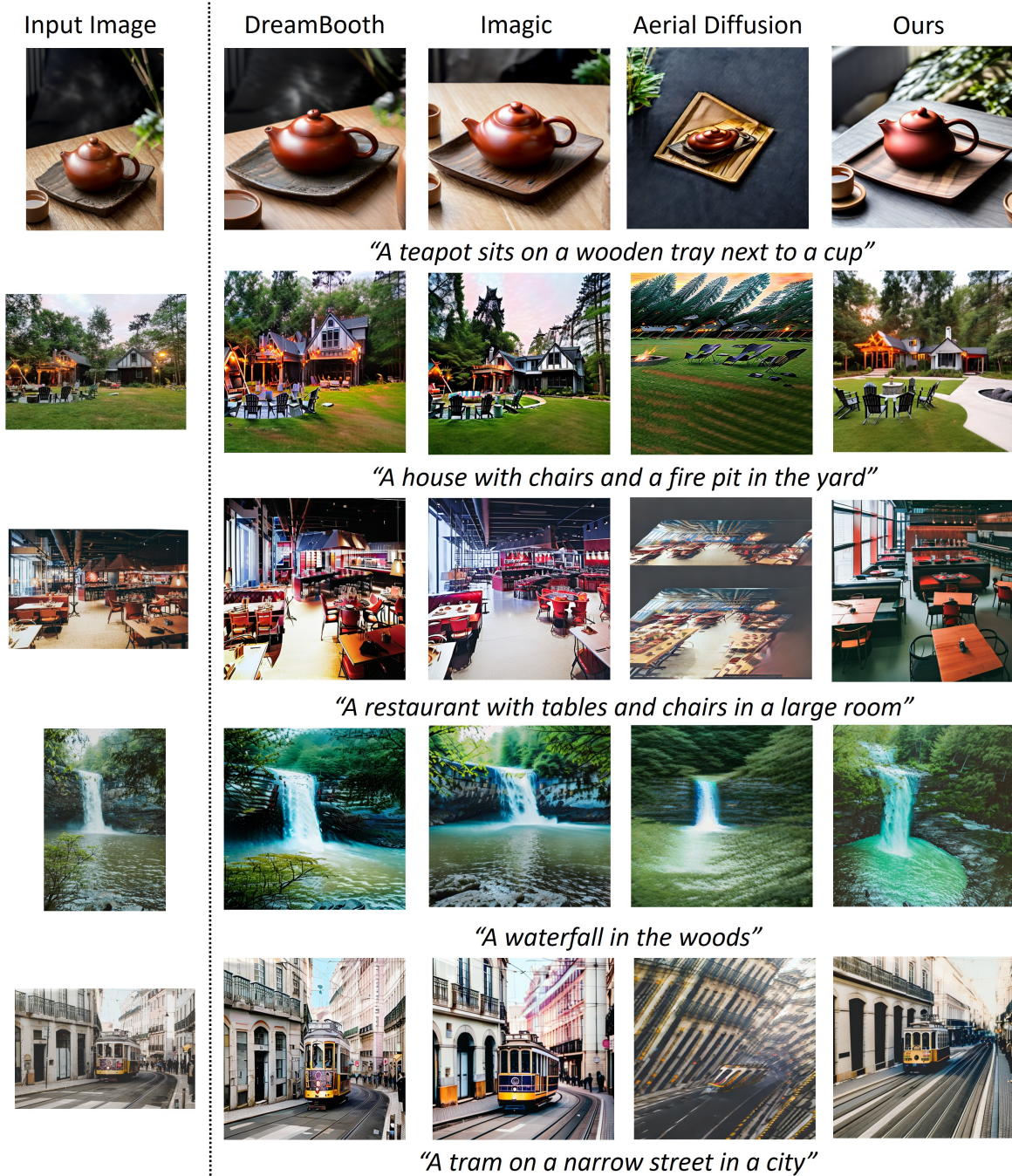


Figure 7.13: We show more examples for comparisons with the state-of-the-art image editing, personalization and text-controlled aerial view synthesis methods on the HawkI-Real dataset. Our methods is able to generate images that are “more aerial”, while preserving the fidelity w.r.t. the input image.

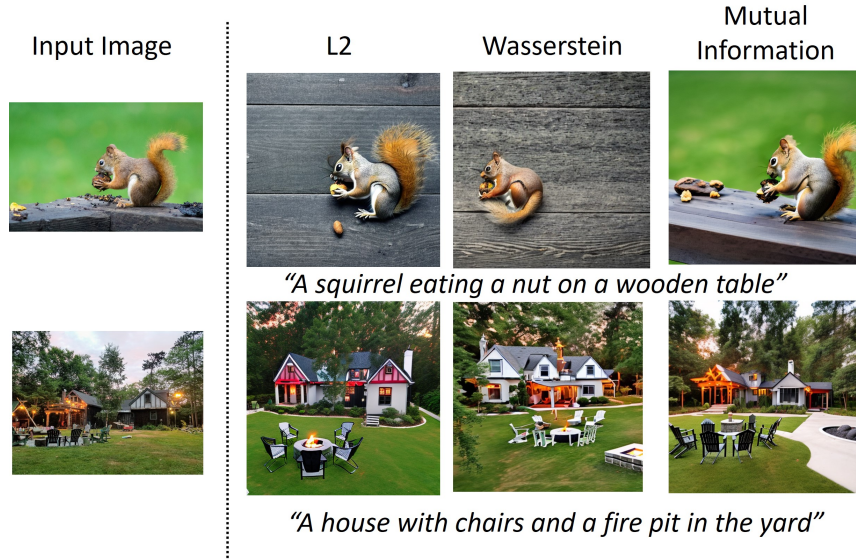


Figure 7.14: We show a few examples for comparisons with other metrics for diffusion guidance such as L2 distance and Wasserstein distance. Our mutual information guidance method is better at preserving the fidelity w.r.t the input image, as also evidenced by higher SSCD scores. The SSCD score for Wasserstein guidance, L2 guidance, and mutual information guidance are 0.3181, 0.3224 and 0.3345 respectively, averaged over all images in the HawkI-Real dataset.



Figure 7.15: We compare with latest related work on novel view synthesis: Zero-1-to-3 and Zero123++ on images from HawkI-Syn. Both of these methods use the pretrained stable diffusion model and the 3D objects dataset, Objaverse with 800k+ 3D objects, for training. Our method uses just the pretrained stable diffusion model for the task of aerial view synthesis from a single image.

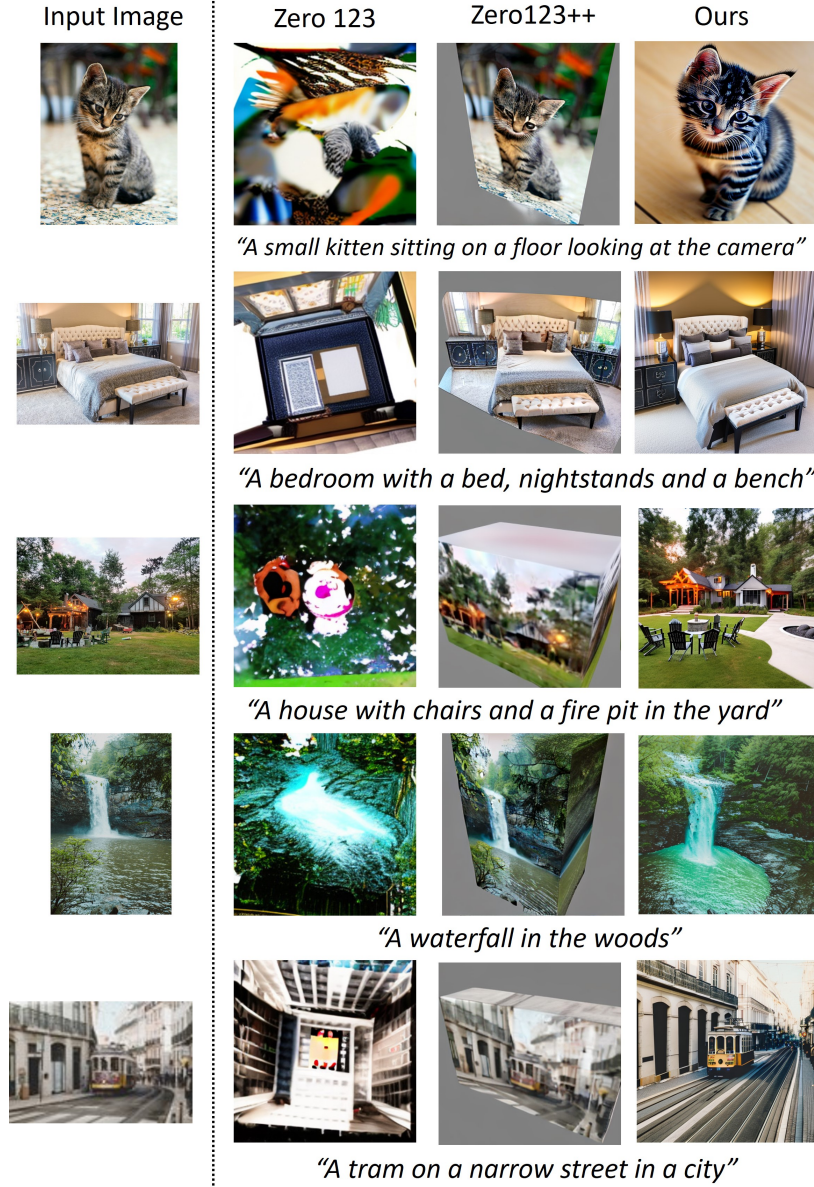


Figure 7.16: We compare with latest related work on novel view synthesis: Zero-1-to-3 and Zero123++ on images from HawkI-Real. Both of these methods use the pretrained stable diffusion model and the 3D objects dataset, Objaverse with 800k+ 3D objects, for training. Our method uses just the pretrained stable diffusion model for the task of aerial view synthesis from a single image.

3D generation methods like Zero123++ are capable of generating different views with high fidelity by using pretrained stable diffusion models to finetune on large-scale 3D objects datasets. However, their generalization capabilities are limited. Our method is able to generate high quality aerial images for the given input images without any 3D data and using just the pretrained text-to-2D image stable diffusion model, however, there is scope for improving the fidelity of the generated aerial image w.r.t the input image. Moreover, our method controls the viewpoint via text and does not provide the provision to quantitatively control the camera angle. Both of these limitations of our method can be alleviated by exploring the combination of pretrained Zero123++ models (or other 3D models) and our method, as a part of future work.

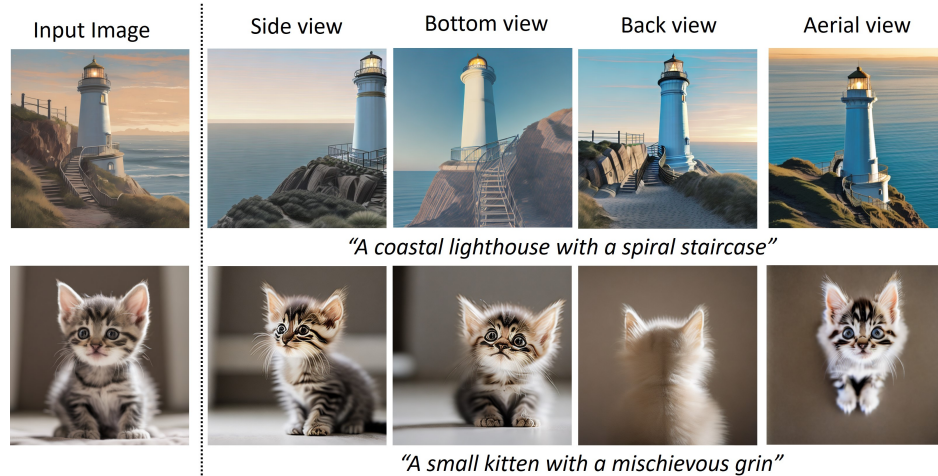


Figure 7.17: Additional results on extending HawkI to generate other text-controlled views.

any additional 3D or multi-view data at train/ test time. Our method has a few limitations, which form an avenue for future work - (i) Further exploration of combining our 3D-free approach with 3D priors can enable the generation of camera-controlled views. (ii) Alleviating hallucination, a widespread issue in all diffusion models, will enable the generation of images with higher fidelity w.r.t. input images. Other directions for future work are - (i) IPM is as an important tool for aerial view generation – similar homography projections to synthesize other views can be explored, (ii) our mutual information guidance formulation can be used in other image editing and personalization problems, (iii) using generated data for various downstream UAV and aerial-view applications such as cross-view mapping, 3D reconstruction and synthetic data augmentation.



Figure 7.18: **Can any data augmentation be used in place of Inverse Perspective Mapping (IPM)?** One question that arises from the usage of Inverse Perspective Mapping is related to whether it actually provides pseudo weak guidance, in addition to increasing variance (or reducing bias) in the representation space that is being conditioned for aerial view generation. The latter can be achieved with any random data augmentation. To understand this, we use a 45 degrees rotated image in place of the image corresponding to the Inverse Perspective Mapping in the second stage of finetuning the text embedding and the diffusion UNet. We do not use mutual information guidance in any of our experiments. Our finding is that results with models that use Inverse Perspective Mapping are generally better in terms of the viewpoint being aerial, while preserving the fidelity with respect to the input image, than models that use the 45 degree rotated image. Thus, we conclude that rather than using any random data augmentation technique, it is beneficial to use IPM as it is capable of providing pseudo weak guidance to the model for aerial view synthesis. This finding also paves direction for future work on using carefully crafted homography priors for view synthesis corresponding to different camera angles and viewpoints.

Part V

Generating Personalized Images using Pretrained Text to Image Diffusion Models and Exemplar Control Images

While the generation of aerial view images from a pretrained model using single front view images is intriguing, let’s delve further into controllable image and video generation. Can we manipulate the background, action, and subject using just a single exemplar image to create diverse images and videos? This is precisely the exploration undertaken in this section. This section has three chapters.

In the Chapter 8, we present ImPoster, a novel algorithm for generating a target image of a ‘source’ subject performing a ‘driving’ action. The inputs to our algorithm are a single pair of a source image with the subject that we wish to edit and a driving image with a subject of an arbitrary class performing the driving action, along with the text descriptions of the two images. Our approach is completely unsupervised and does not require any access to additional annotations like keypoints or pose. Our approach builds on a pretrained text-to-image latent diffusion model and learns the characteristics of the source and the driving image by finetuning the diffusion model for a small number of iterations. At inference time, ImPoster performs *step-wise text prompting* i.e. it denoises by first moving in the direction of the image manifold corresponding to the driving image followed by the direction of the image manifold corresponding to the text description of the desired target image. We propose a novel diffusion guidance formulation, *image frequency guidance*, to steer the generation towards the manifold of the source subject and the driving action at every step of the inference denoising. Our frequency guidance formulations are derived from the frequency domain properties of images. We extensively evaluate ImPoster on a diverse set of source-driving image pairs to demonstrate improvements over baselines. To the best of our knowledge, ImPoster is the first approach towards achieving both subject-driven as well as action-driven image personalization.

In the Chapter 9, we present a method for multi-concept customization of pretrained text-

to-video (T2V) models. Intuitively, the multi-concept customized video can be derived from the (non-linear) intersection of the video manifolds of the individual concepts, which is not straightforward to find. We hypothesize that sequential and controlled walking towards the intersection of the video manifolds, directed by text prompting, leads to the solution. To do so, we generate the various concepts and their corresponding interactions, sequentially, in an autoregressive manner. Our method can generate videos of multiple custom concepts (subjects, action and background) such as a teddy bear running towards a brown teapot, a dog playing violin and a teddy bear swimming in the ocean. We quantitatively evaluate our method using videoCLIP and DINO scores, in addition to human evaluation.

While diffusion models often benefit from training on large datasets, text prompting strategies offer a way to strategically control the generation of new concepts without requiring additional training data. In Chapter 10, we delve into the specifics of how this can be achieved. We introduce a novel approach for prompt mixing, aiming to generate images at the intersection of multiple text prompts using pre-trained text-to-image diffusion models. At each time step during diffusion denoising, our algorithm forecasts predictions w.r.t. the generated image and makes informed text conditioning decisions. To do so, we leverage the connection between diffusion models (rooted in non-equilibrium thermodynamics) and the Black-Scholes model for pricing options in Finance, and draw analogies between the variables in both contexts to derive an appropriate algorithm for prompt mixing using the Black Scholes model. Specifically, the parallels between diffusion models and the Black-Scholes model enable us to leverage properties related to the dynamics of the Markovian model derived in the Black-Scholes algorithm. Our prompt-mixing algorithm is data-efficient, meaning it does not need additional training. Furthermore, it operates without human intervention or hyperparameter tuning. We highlight the benefits of our approach

by comparing it, qualitatively and quantitatively using CLIP scores, to other prompt mixing techniques, including linear interpolation, alternating prompts, step-wise prompt switching, and CLIP-guided prompt selection across various scenarios such as single object per text prompt, multiple objects per text prompt and objects against backgrounds.

Chapter 8: ImPoster: Text and Frequency Guidance for Subject Driven Action Personalization using Diffusion Models

Monkeys seldom meditate! Neither do they play the violin! Generative AI has enabled the creation of images that cannot be easily enacted or photographed in real life. More often than not, applications such as animation, movie creation, etc, require a particular subject performing a specific action. For example, given a ‘source image’ of a dog and a ‘driving image’ of a cat drinking water from a mug, we might want to generate an image of the dog drinking water from a mug in exactly the same manner as the cat. Thus, the generated dog needs to have the same identity as the ‘source subject’. Its pose or the style with which it drinks water from the mug needs to resemble the ‘driving pose’, i.e. the pose of the cat in the driving image. Consequently, the specific action depicted in the generated image and the identity of the subject performing the action are not arbitrary - rather, they are dictated by the source and the driving images.

Prior work on the closely related problem of motion transfer [405, 406, 407, 408, 409] condition on SMPL pose information [410], keypoints, etc in a supervised/ unsupervised manner and train on large datasets to transfer the action depicted by a subject in a driving image or video to the subject in the source image. Such methods however do not generalize well to arbitrary pairs of source and driving images and often require a lot of data for training. Thus, it is useful to develop an approach that can directly work on a single pair of source-driving images in a

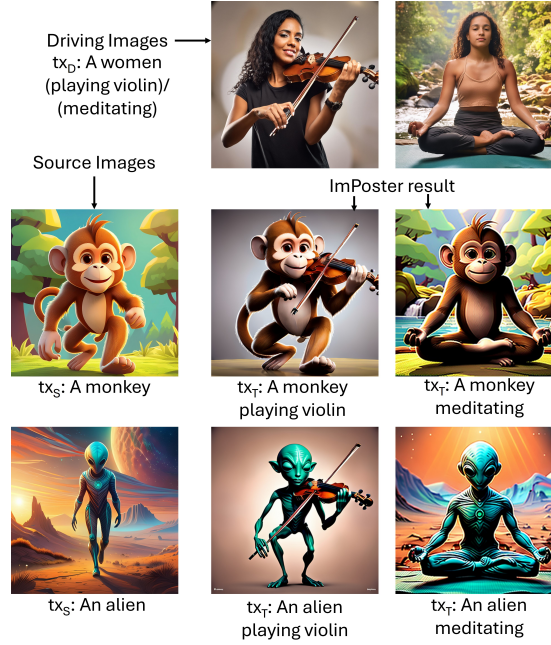


Figure 8.1: Given a single “source” image and a “single” driving image and the corresponding text descriptions, ImPoster generates an image of the source subject performing driving action. We show how ImPoster is able to make a monkey and an alien meditate and play violin.

completely unsupervised manner.

Text is an excellent auxiliary modality to guide the generation process, and the recent progress in text-to-image diffusion models [284, 302, 411] motivates their usage for controlled image generation. Prior work on personalized text-guided image editing such as DreamBooth [340], IMAGIC [8], custom diffusion [412], InstantBooth [413], [414], ELITE [415] and SUTI [416] are able to add a specific subject to the model while finetuning the diffusion model to generate various actions dictated by the text input. However, the user does not have any control over the exact manner in which they might like the subject to perform the action in the generated image. For instance, a dog can drink water from a mug in various ways - we want the model to be able to generate a specific pose corresponding to the action, as defined by the driving image. On the other hand, prior work such as blended diffusion [417] and DiffusionCLIP [331] are able to generate a specific action and change the style of the image. However, they are unable to make a source

subject perform a specific driving action.

Main contributions. We propose an algorithm, ImPoster, for generating an image of a specific subject from a source image performing a specific action described in the driving image. Given a single source-driving image pair and the corresponding text descriptions, our method can perform ‘body’ transformations to the source subject, as dictated by the action depicted in the driving image. Our method builds on a pre-trained text-to-image diffusion model to perform test-time optimization/ inference and does not require any additional information such as pose or keypoints. Our method, ImPoster , first learns the characteristics of the source and the driving image by finetuning the pretrained diffusion model on the source and driving corresponding text-image pair. Our inference guidance methods, which form the novel contributions of this chapter, enable the generation of the desired target image:

1. **Stepwise text prompting.** ImPoster generates the desired target image of the source subject performing the specific driving action using a *stepwise text prompting* strategy. Since geometric or structural information is much harder to generate, ImPoster lays the foundation for generating the driving action by first moving towards the direction of the manifold of the driving image. This step is followed by switching directions and denoising by moving in the direction of the text description corresponding to the desired target image.
2. **Image frequency guidance.** While the stepwise text prompting method provides a strong prior for the driving action, it is insufficient to generate an accurate image of the source subject. To alleviate this issue, we delve into the frequency domain representation of an image, a powerful space for understanding how the human brain interprets natural images [16, 17, 418, 419]. Motivated by guidance techniques [357, 392] in diffusion

models, we harness the amplitude map of the Fourier transform of image features and present frequency amplitude guidance to preserve the characteristics of the source subject. In order to prevent the loss of the specific driving action (or structure) while extracting source subject characteristics, frequency amplitude guidance is complemented by frequency phase guidance. This is inspired by the fact that the phase of the Fourier transform of an image is representative of its geometry and structure. Frequency amplitude guidance and frequency phase guidance, termed as *frequency guidance*, guide the diffusion denoising towards generating the source subject performing the specific driving action.

We apply ImPoster on a wide variety of source-driving image pairs on a curated dataset with 120 source-driving image pairs. ImPoster can make an elephant read a book (described by a human reading a book), a monkey meditate and perform push-ups (described by a human performing meditating and doing pushups), a teddy bear play guitar (described by a human playing guitar), etc. We also show the effectiveness of stepwise inference and our frequency guidance method, along with qualitative and quantitative comparisons against prior work metrics such as CLIP Score, SSCD, DINO and a new metric to quantify the alignment with driving action, phase score, defined using the phase of the Fourier transform.

In summary, the contributions of this chapter are as follows: (i) We formalize a novel task of generating images consistent with a source subject and driving action; (ii) We propose ImPoster, a novel diffusion models based method that can address this pragmatic task; (iii) We curate a dataset of 120 source-driving image pairs; (iv) In order to quantitatively establish the correspondence between the generated image and driving action, we propose a new metric ‘Phase score’ based on the phase of the Fourier transform; (v) Finally, our exhaustive experimental results reveal large

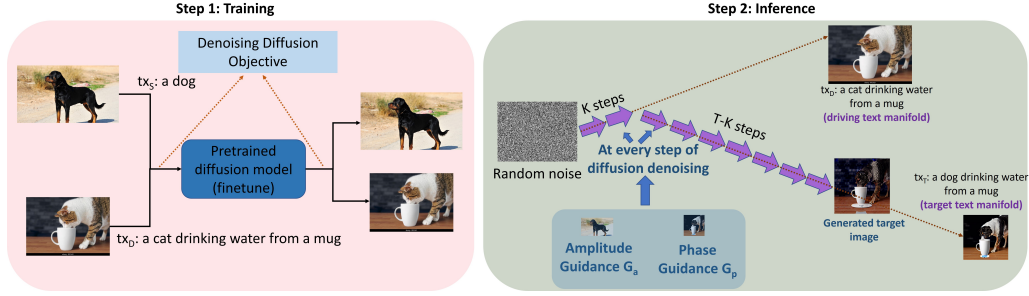


Figure 8.2: Given a single source-driving image pair, ImPoster generates an image of the source subject performing the action depicted in the driving image. ImPoster first finetunes the text-to-image diffusion model on the source-driving image pair. At inference, ImPoster begins by first denoising in the direction of the driving image manifold followed by moving towards the manifold corresponding to the desired target image. At every step of the inference, frequency guidance steers the generation of an image with source subject characteristics and driving action.

gains over baselines; thereby setting a new benchmark for the task.

8.1 Related work

Exemplar image animation. A great amount of literature has been dedicated to exemplar image animation [405, 406, 406, 407, 408, 409, 420, 421] transfer motion characteristics from a driving video to a source image. These methods at the core of them transfer motion at frame level from a driving frame onto the source image, similar to our method. However, these models are restricted in their domain to only human and almost always require additional annotations like keypoints, 2D poses, 3D poses or require computation of optical flow.

Diffusion models for Text-based image editing/ personalization. Recent progress in generative AI has shown promising results using diffusion models [273, 299, 300, 301, 302, 303, 304, 305, 390] for performing non-trivial operations, such as posture changes and multiple objects editing. Text-based image editing and personalization approaches [8, 283, 330, 334, 340, 356, 395, 412, 412, 413, 413, 414, 415, 416, 422, 423, 424, 425, 426, 427, 428, 429] add a subject to the diffusion

using a few images of the subject followed by using text to manipulate the image to obtain the desired output. The methods for personalized or subject-driven based image editing can be categorized into two broad categories. First category of works [8, 304, 340, 412, 414, 415] perform fine-tune a pre-trained model on a small number of images that involve the subject and then perform inference optimization with the text depicting the action as the input. While these methods are successful in generating images with the subject performing an action depicted by the text, the user does not have control over the specific pose or imitable characteristic. The second category of works [413, 416] address the shortcomings of fine-tuning the model and propose methods that are purely based on inference-time optimization and hence much faster. However, even they suffer from the same drawbacks in not being able to give users the control of the action depicted. Multi-concept customization methods [395, 412, 428, 429] are not very effective either. For instance, one of the closest works to ours is Custom Diffusion [412] that enable generating images by fusing multiple concepts from multiple images. However, this method is not well suited to generate images involving concepts of actions.

Recent methods on motion customization [430, 431, 432] of diffusion models are able to transfer the motion from a video to subjects. However, these methods rely on temporal properties of videos to transfer the motion. Our problem statement is tangential to video motion customization, wherein the goal is to customize the action from an image.

Guidance methods for diffusion. Guidance methods [14, 293, 392, 433, 434, 435, 436, 437, 438] have been used to control and guide diffusion denoising. One of the first guidance methods is classifier and classifier-free guidance [293, 392] that reinforce the class of the object in the generated image. Bansal et. al. [357] proposed universal guidance to guide the generation using

segmentation maps, sketches, etc. Kothandaraman et. al. [14] proposed a mutual information based guidance method to generate high fidelity images from various viewpoint.

8.2 Method

Given a source image I_S and a driving image I_D , ImPoster generates an image of the source subject performing the driving action. We assume access to the text descriptions tx_S and tx_D for the source and driving images respectively. The text description, tx_T , corresponding to the desired target image I_T is a modified combination of tx_S and tx_D . For example, if tx_S is “A dog” and tx_D is “A cat drinking water from a mug”, tx_T would be “A dog drinking water from a mug”. We assume no access to any training dataset, pose information (such as SMPL), keypoints, etc. Our method is completely unsupervised and works on a single source-driving image pair.

An overview of ImPoster is as follows. To add knowledge pertaining to the source and the driving image, we finetune the diffusion model on I_S and I_D using tx_S and tx_D , respectively. During inferencing, we begin by denoising by first moving in the direction of the manifold corresponding to I_D followed by moving in the direction of the manifold corresponding to tx_T . Such a mechanism, which we term **stepwise text prompting** creates a prior for pose information followed by denoising to generate the desired subject performing the action. Further, to reinforce the generation of an image of the source subject performing the driving action, at every inference step, we apply a novel **image frequency guidance** strategy to explicitly steer the denoising towards the desired driving action and source subject characteristics. We now turn to describe our method in detail.

8.2.1 Training

A pretrained text-to-image diffusion model such as stable diffusion is trained on massive text-image data. However, it does not contain information specific to the source and driving image. In order to enable the diffusion model reconstruct I_S and I_D from tx_S and tx_D , respectively, we finetune the diffusion model for n_{tr} iterations on the source and driving text-image pairs using the denoising diffusion objective function [343]. Specifically, let θ denote the U-net parameters, L denote the denoising diffusion objective, and e_S, e_D denote the text embeddings corresponding to the source and driving texts. At each iteration $i < n_{tr}$ of finetuning, we perform the following optimizations:

$$\begin{aligned} \min_{\theta} \sum_{t=T}^0 L(f(x_t, t, e_S; \theta), I_S), \\ \min_{\theta} \sum_{t=T}^0 L(f(x_t, t, e_D; \theta), I_D). \end{aligned} \tag{8.1}$$

8.2.2 Inference

8.2.2.1 Stepwise text prompting

A key question that we need to keep in mind while developing an effective inference strategy is: *How should the noisy image be perturbed so that it better corresponds with the driving action and the source subject ?* Moving in the direction of the driving image manifold will ensure the reconstruction of the driving action accurately, however, biased towards the driving subject. Similarly, moving in the direction of the source image manifold will ensure the reconstruction of the source subject accurately, however, biased towards the source action. The image manifold

corresponding to tx_T generally does not contain the desired target image – i.e. given tx_T , it is unable to generate the desired target image by naive diffusion denoising. On one hand, a pretrained diffusion model conditioned on tx_T is inclined to generate a wide variety of images, most not fully conformant with the specific characteristics of the source subject as well as the specific pose delineated by the driving action. On the other hand, there could be a bias towards reconstructing the source or driving image as well.

To solve the aforementioned issues, we propose a stepwise inference strategy. Let the number of inference steps be denoted by T_i . Starting from random noise, we denoise using tx_D for the first K iterations following by denoising using tx_T for the next $T - K$ iterations. This implies that the generation process begins by first moving towards the driving image manifold followed by the manifold corresponding to the target text. In the first few iterations, the sampler denoises to reconstruct the driving image, creating a strong prior for the driving action. In the subsequent iterations, denosing to move towards the target text manifold ensures that the generated image contains the source subject, as dictated by the target text.

On one hand, the target text points towards the driving action. On the other, denoising using tx_D in the first few iterations creates a strong prior for the structure of the generated image. This ensures that the generated image has the driving action. Note that we denoise with tx_D first followed by tx_T and vice versa. Diffusion works by sequentially denoising from random noise and is a slow process i.e. the updates in each step are small. Therefore, it is easier for the model to first generate the pose and then denoise to obtain subject characteristics. It is much more difficult to modify the pose than it is to modify the characteristics or identity of the subject in the image.

Despite the stepwise inference strategy, the generated image might still contain certain characteristics of the driving subject or source action. To alleviate this issue, we propose a novel

frequency guidance method, which reinforces the model to stick to the driving action and source subject at every step of the generation process.

8.2.2.2 Frequency guidance

The frequency domain representation of an image provides rich information [418] about the image. The amplitude of the 2D spatial Fourier transform of an image is representative of the intensities of different frequencies in the image, and represents changes in the spatial domain. It contains information about the geometrical structure of features in the image. The phase of the 2D spatial Fourier transform of an image represents the location of these features which help the human eye understand the image; it contains information regarding the edges, contours, etc. It is possible to reconstruct the grayscale counterpart of an image with just its phase representation. The amplitude information, along with the phase, helps in reconstructing the characteristics (colors, attributes, texture, identity) of the scene entities. However, the amplitude alone cannot generate a meaningful image; phase information is very important.

In light of this, it can be inferred that the phase of the Fourier transform of the driving image is indicative of the driving action. Similarly, the amplitude of the Fourier transform of the source image is indicative of the knowledge required to reconstruct the source image. We use the Fourier domain representation to **guide** the generation process. Frequency domain properties of an image propagate to its feature space (computed by a neural network). Hence, we can apply frequency guidance on the latent feature representation space of stable diffusion [284]. Classifier-free guidance [392] and universal guidance [357] mathematically derive guidance as,

$$\tilde{\epsilon}_{\theta}(z_t, t) = \epsilon_{\theta}(z_t, t) + s(t) \times \nabla_{z_t} G. \quad (8.2)$$

where z_t is the denoised latents at timestep t , $\nabla_{z_t} G$ is the gradient of the guidance function and $s(t)$ controls the strength of the guidance for each sampling step. An appropriately weighted version of this noise is subtracted from the latents computed in the previous timestep to obtain updated latents. Consequentially, rather than adjusting the noise predicted by the diffusion model using the gradients of the guidance function, we can directly modify the computed latents using a suitable guidance function using the gradient of the computed amplitude and phase functions w.r.t. the latents. At every step of the inference sampling process, we modify the computed latents z_t as,

$$\tilde{z}_t = z_t - s_a \times \nabla_{z_t} G_a - s_p * \nabla_{z_t} G_p, \quad (8.3)$$

where s_a and s_p are the scaling factors for the frequency amplitude guidance and frequency phase guidance functions, G_a and G_p , respectively. Note that z_t is computed using the noise predicted by the diffusion model at each timestep. G_a is the $L2$ distance between the amplitude of the generated latents and the amplitude of the latents of the source image. G_p is the $L2$ distance between the phase of the generated latents and the phase of the latents of the driving image.

G_a drives the generation towards the source subject at every step of the sampling. G_p reinforces the driving pose at every step of the sampling and prevents any distortion in pose that may be caused by G_a .

8.3 Experiments and Results

Dataset and implementation details. Since there is no dataset for this task, we collect a dataset with 15 driving actions and 8 source subject, resulting in a total of 120 source-driving image pairs. The source subjects and driving actions are generated using Adobe Firefly. This dataset

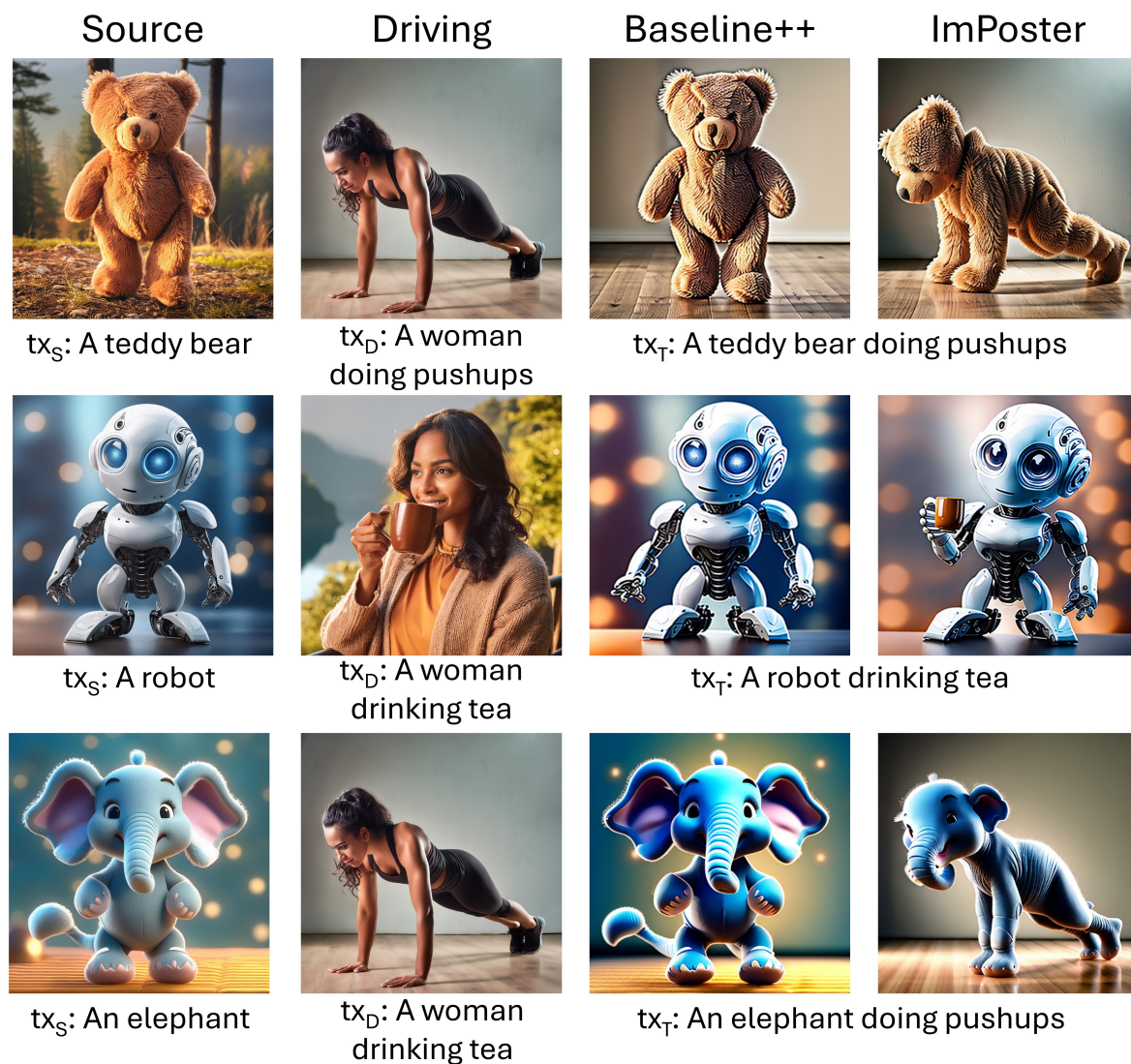


Figure 8.3: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues. Please see the appendix for more results generated using ImPoster and comparisons with Baseline++.

curation is inspired by prior methods that defined new tasks in diffusion personalization such as DreamBooth [340](30 subjects), Custom Diffusion [412](5 concepts, 8 prompts for the multi-concept setting), and Concept Decomposition [439](15 pairs for application 1, 13 concepts for application 2) and is also comparable in size to these prior datasets. Consistent with prior work, we generate results corresponding to 5 different random seeds for each source-driving pair, resulting in a total of 600 images being used for all quantitative analysis.

We finetune the diffusion model for $n_{tr} = 500$ iterations. We set s_a to $1e-6$, $s_p = 1e-3$, $k = 5$ and $T = 50$. We use an image size of 512×512 . Our model takes about 51/2 minutes per image pair on one NVIDIA RTX A5000 GPU. We use the Stable Diffusion 2.1 model as the backbone. We add LoRA [333] layers to the backbone model for the finetuning step, the rest of the model is frozen.

Qualitative results.

State-of-the-art comparisons. Closest related to our work are image editing or personalization approaches. The most recent and effective methods in this domain are DreamBooth [340](CVPR 2023) and Custom Diffusion [412](CVPR 2023), inspired by which, we compare with an enhanced baseline, Baseline++. Baseline++ takes in a single source image and driving image along with their corresponding text prompts and finetunes the diffusion model (similar to DreamBooth [340] + LoRA [333]). Next, it uses a pragmatic combination of the text prompts corresponding to the source and the driving image (tx_T) to generate the target image (similar to Custom Diffusion [412]).

We show comparisons with Baseline++. Baseline++, due to bias issues with respect to the pose in the source image, and inability to effectively generate the driving action, is unable to

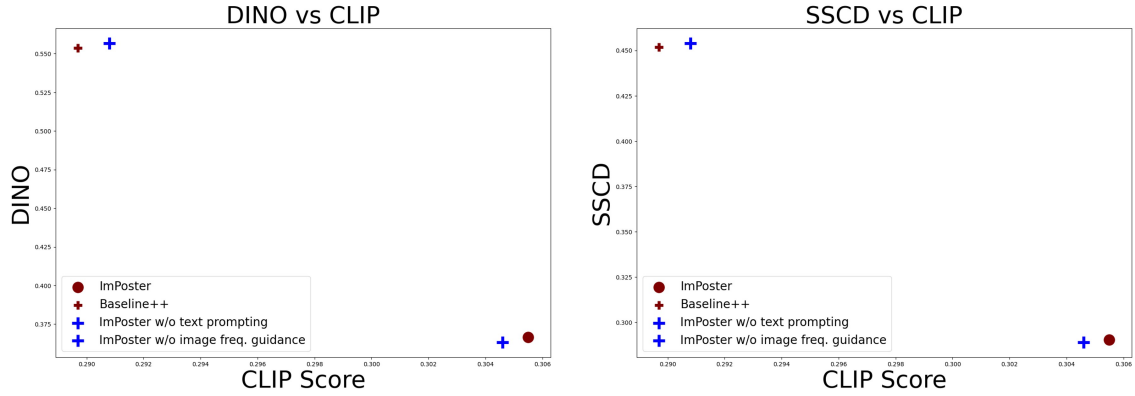


Figure 8.4: ImPoster is able to successfully transfer the driving motion while retaining the characteristics of the source subject, and achieves a better trade-off between driving action (CLIP/Phase score) and source subject (SSCD/DINO) than prior work, as also evidenced by our qualitative results. Stepwise text prompting creates a prior for the driving action to enable the model generate the driving action. The image frequency guidance formulations help the model in preserving the characteristics of the source subject, while reinforcing the driving action.

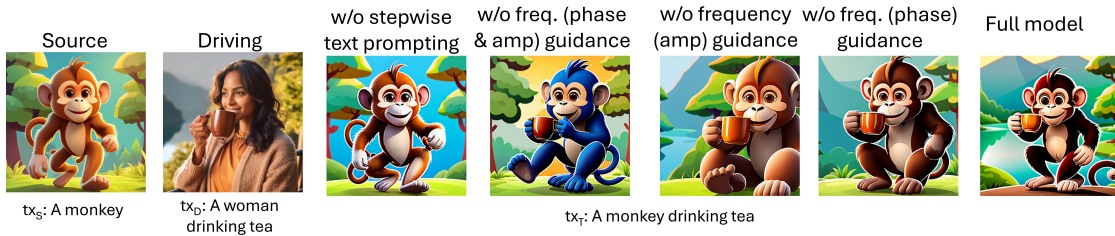


Figure 8.5: **Ablations.** Without stepwise text prompting, there is no prior for the driving action, which inhibits the model from generating the driving action accurately. The frequency (amplitude and phase) guidance methods help in generating the characteristics of the source subject (here, monkey) accurately - notice that there are changes to the color of the monkey (column 6), missing details in the fingers (column 4), changes in the size of the monkey (column 5).

generate the source subject performing driving action. In contrast, ImPoster holistically distills the driving action and the characteristics of the source subject through its stepwise text prompting and image frequency guidance strategies to achieve a good bias-variance trade-off and generate the source subject performing driving action.

Quantitative Results: Comparisons with Baseline++ and Ablations In concordance with prior work [8, 340, 412] on diffusion models for text-based image editing/ personalization, we evaluate

our method using the following quantitative metrics. We compute the averages using all 600 generated images.

1. **Text alignment:** We evaluate the alignment of the generated target image with the target text using the CLIP score [340]. Higher CLIP scores indicate higher alignment with text. Here, CLIP Score gives us a broad overview of the alignment of the generated image with the subject and action. ImPoster achieves a far higher CLIP score than Baseline++, indicating its ability to effectively transfer the driving action to the source subject.
2. **Fidelity:** We evaluate the fidelity of the subject in the generated target image the source subject using self-supervised similarity metrics - SSCD [359] and DINO scores [403]. Baseline++ is unable to generate the driving action, and simply replicates the source image. This results in it having a higher value of SSCD and DINO score than ImPoster, which is able to transfer the driving action, as well as preserve the characteristics of the source image.
3. **Phase score:** To evaluate the correctness of the action or pose generated in the target image (as dictated by the driving image), we define a new metric called the phase score. As per classical computer vision literature [418], the phase of the Fourier transform of the image is indicative of the action depicted in the image. We compute phase score as the cosine similarity between the phase of the Fourier transform of the driving image and the generated target image. Higher cosine score indicates higher similarity. ImPoster achieves a higher phase score of 0.7529 as compared to Baseline++'s phase score of 0.7515, indicating its ability to transfer the driving action effectively.

Ablation analysis. **a. Overall alignment with driving action and source subject:** ImPoster achieves a higher CLIP score than all ablations, indicating the usefulness of each component of our model towards transferring the driving action to the source subject, while preserving its characteristics. The stepwise text prompting and frequency guidance methods (including phase frequency guidance and amplitude frequency guidance) are complementary to each other and work in a holistic manner to achieve the desired goal. **b. Effectiveness of stepwise text prompting:** The stepwise text prompting strategy provides a crucial signal for generating the driving image, without which, the model achieves an overall low CLIP score as well as Phase score. Similar to Baseline++, the model without the stepwise text prompting strategy has a tendency to replicate the source image without the driving action, resulting in a higher SSCD or DINO score. **c. Effectiveness of image frequency guidance:** Without the frequency guidance functions, the SSCD and DINO scores drop, indicating its effectiveness in preserving fidelity w.r.t. source subject while transferring the driving action.

8.4 Conclusions

We propose the usage of infusing text and image frequency for the task of method for the image generation task of subject-driven action personalization. To the best of our knowledge, ours is the first approach towards this task. Our method is successful in achieving a wide range of image animations such as making a monkey meditate and play the violin, as also validated by our quantitative results on our curated dataset using various metrics, including the newly proposed ‘phase score’. We hope this chapter inspires further research in the area.

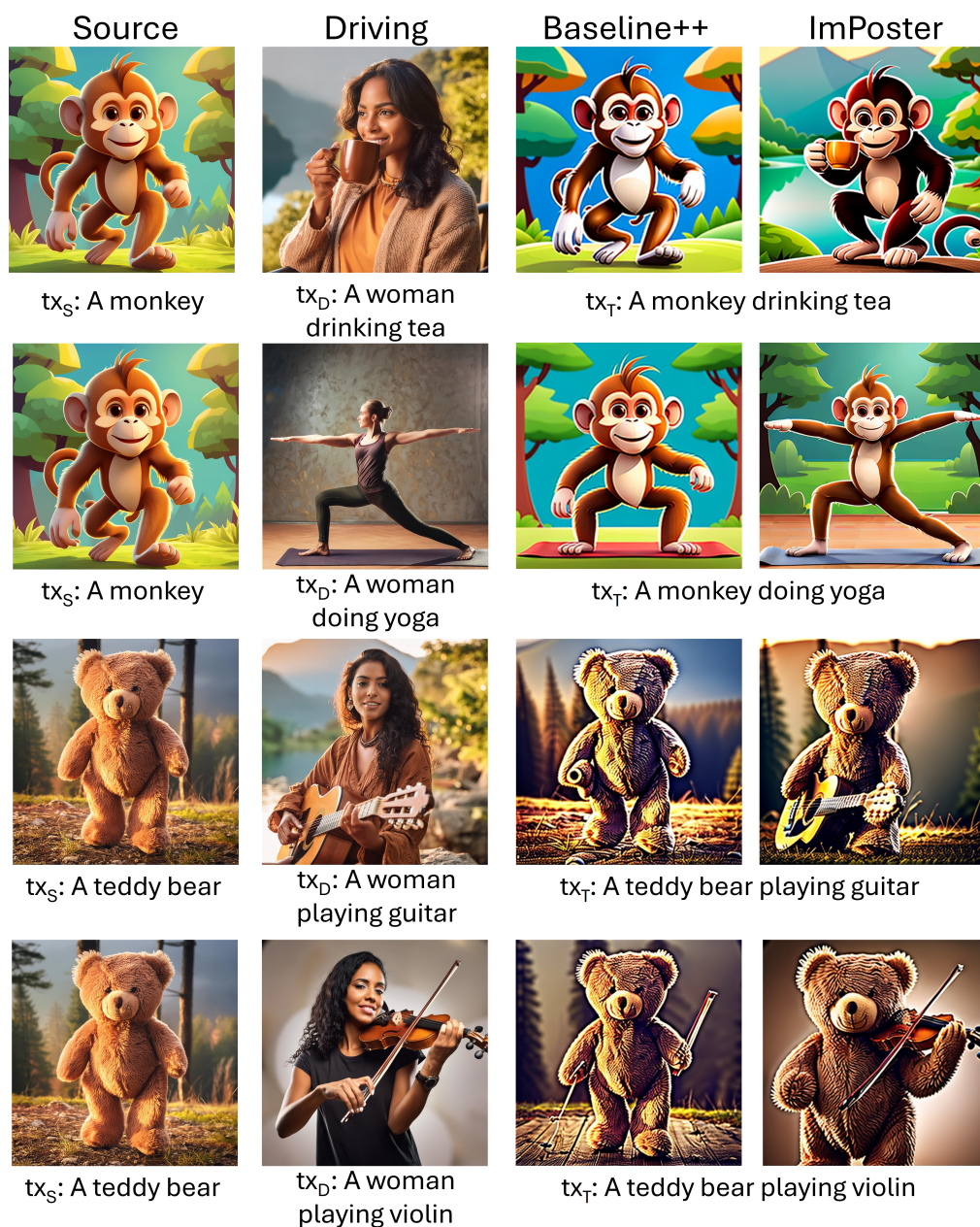


Figure 8.6: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

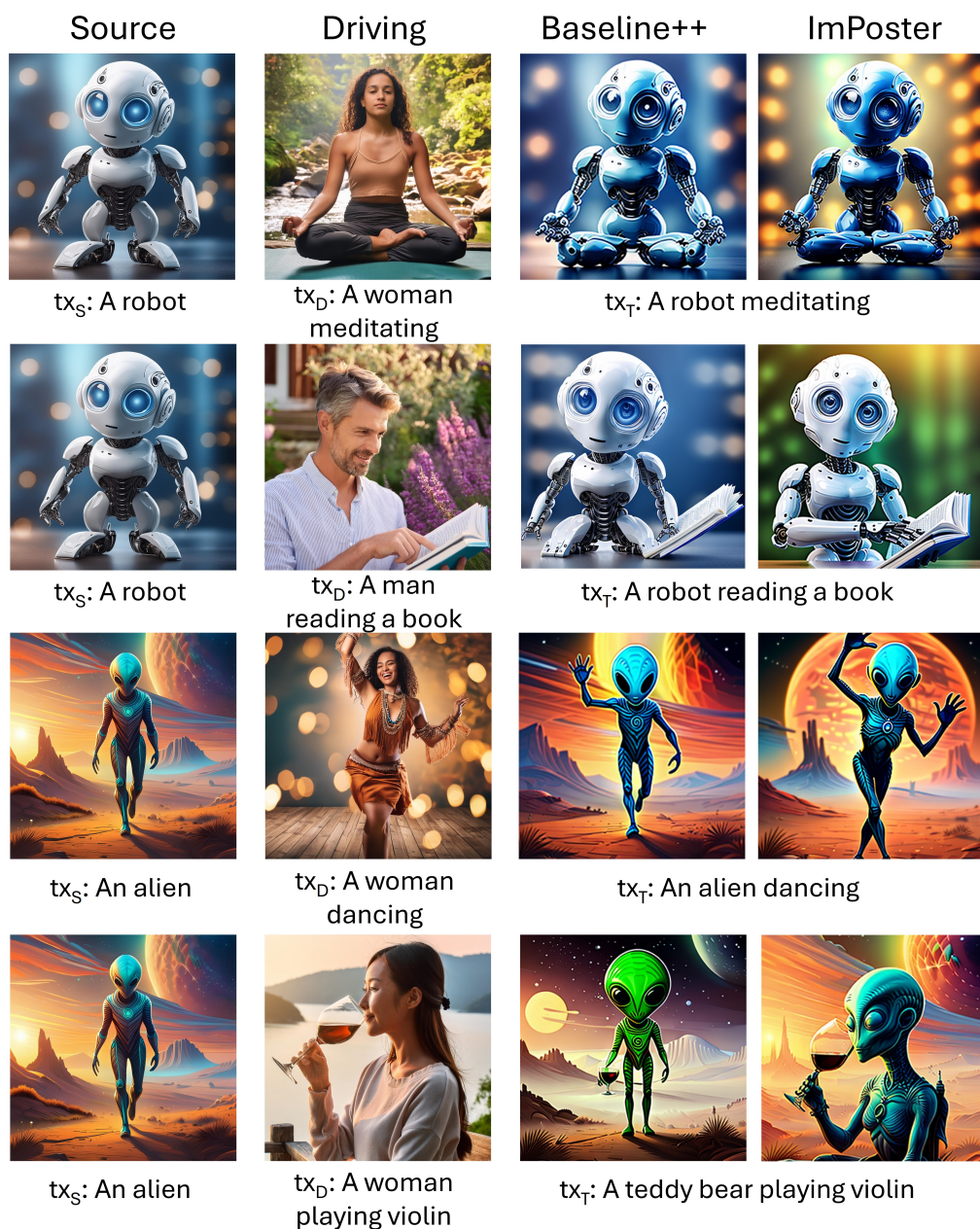


Figure 8.7: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

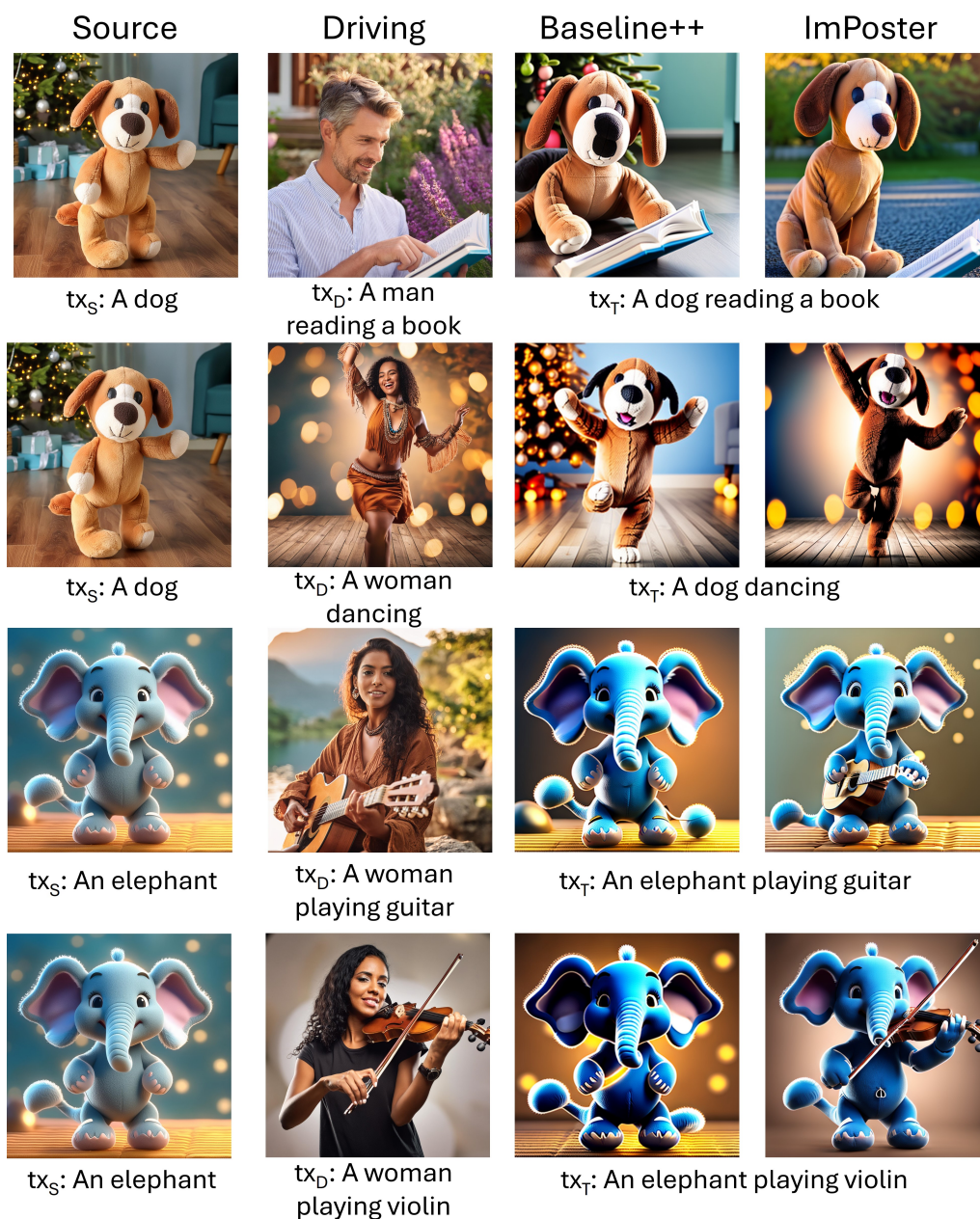


Figure 8.8: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.



Figure 8.9: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

8.5 Limitations and Future Work

While our method is able to execute large non-trivial non-rigid pose transfers to specific subjects, as defined by a driving image, it is still limited in the ability of change that it can bring forth. For instance, it is unable to make a rabbit or a cat do yoga, indicating that there is scope for further research. Another direction for future research, also a widespread issue in the text-based image personalization literature, is the development of better quantitative metrics for comprehensively measuring subject fidelity and the desired edit. Since self-supervised methods such as SSCD and DINO measure image-level similarity, they result in a high value even if there is absolutely no action transfer – this leads to inappropriate evaluation of the models’ capability in performing the desired edit (driving action transfer in our case), while maintaining source



Figure 8.10: **Failure cases.** Our method is ineffective in a few cases - it is unable to make a rabbit perform yoga or dance, and a cat perform yoga or meditate. We believe that this is due to bias-variance trade-off issues, that remain unresolved even after the application of our stepwise text prompting and image frequency guidance strategies. Usage of stronger vision and language backbones in the text to image generation pipeline, as they are made open-source to the community, can help the model disentangle features of the image better to alleviate these bias-variance trade-off issues. Besides, further research on this problem can lead to the development of newer methods that can help improve these results.

subject fidelity. More directions for future work including investigating our frequency guidance strategy for other image editing and personalization applications, extension to video applications and scenarios that involve more than one source subject or driving action in the scene.

Chapter 9: Text Prompting for Multi-Concept Video Customization by Autoregressive Generation

Text-to-video generation models such as Video Diffusion Models [440], Make-A-Video [441], Imagen Video [411], Phenaki [442] CogVideo [443], VideoFusion [444], Gen-1 [445] are capable of generating visually appealing videos. However, their application is constrained due to the inability to generate long videos (beyond a few seconds) [446, 447] and limited control [432]. These limitations are problematic mainly because when generating a long video that includes multiple subjects, we expect the same subjects to appear later in the video. For instance, if we generate a video of a specific *Teapot* boiling under a specific *Tree*, later in the video we expect to see the same *Teapot* and the same *Tree* and not another randomly generated one. One way to achieve such results with the current text to video models is via providing extreme level of details in the captions, so that the model generates the same *Tree* and the same *Teapot* everytime. However, this is usually impractical [28, 440] because the models may not be sensitive to all the details and the concepts may be personal and unseen by the model during large-scale pre-training [440, 441, 442].

Another way of getting consistent subjects in consecutive video frames is that video generation should have the capability to generate clips conditioned on one or more given subjects that may or may not be a part of the training data. Model customization [28, 334] methods such as

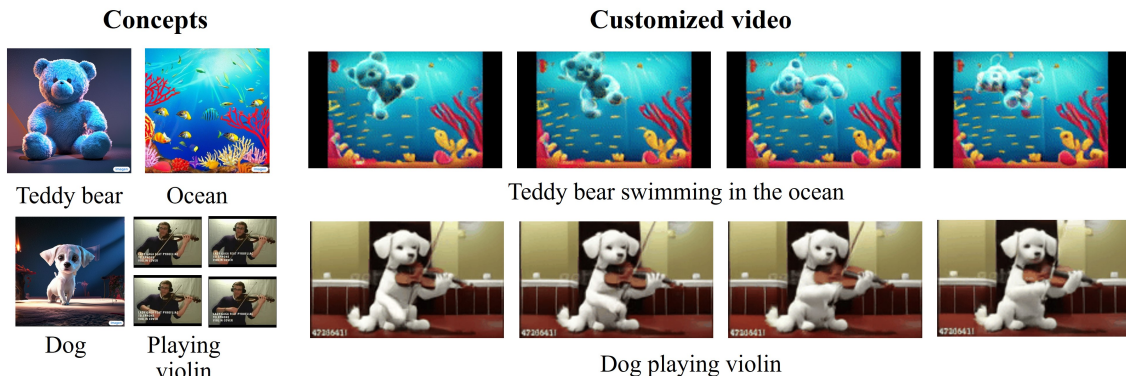


Figure 9.1: We present a method for multi-concept customization of text-to-video models. Our method only relies on a pre-trained text to video model and concept specific data in the form of image(s) or video(s). In the illustration above, our model uses an image of a teddy bear and ocean background (top row) and an image of a dog and video of playing violin (bottom row) to generate the customized videos.

DreamBooth [28] and DreaMix [448] work well for single concept video customization such as generating a random video of a customized subject or making a random subject perform a custom action. However, they are not as effective in the multi-concept customization regime [432]. We hypothesise that this is due to the inability of the model in understanding the interactions between different custom concepts, that it may have not encountered during training [388]. Also, learning characteristics of concepts from limited data per concept results in severe overfitting [395]. Hence, the variance of the model w.r.t. the custom concept is low and it is predominant to use the variance of the model to overcome bias issues w.r.t. the custom concept. Additionally, when multiple concepts are customized in a video, there is a mixup of their corresponding attributes [449].

Prior work on multi-concept customization of text to image models [388, 395] and concurrent work on multi-concept video customization [432] are able to put together multiple subjects in the scene in a realistic manner. However, they often train individual adapter models [388] for each concept before projecting to a common space, making it memory inefficient for video models,

and also rely on expensive data augmentation from the original training dataset. Moreover, the text-to-image customization methods suffer from attribute binding problems and artifacts [432] when extended to the video space.

Main contributions. In this chapter, we present a simple, yet effective solution for multi-concept video customization based on “teaching” the model interactions between various concepts one-by-one in order to generate the custom video.

We show that sequential and controlled traversal through the manifolds of each subject (each understood by the model) towards the intersection space of the manifolds, while generating video frames sequentially, leads to the generation of the custom video. The main intuition behind our proposed method is that while sequentially generating the frames of a video, the model needs to remember what it previously generated and synthesize the interactions accordingly while staying causal in time. Our solution for multi-concept video customization exploits the inherent strong causality in autoregressive T2V models to generate the interactions between the different concepts in the video. Moreover, at each stage of generation after the first step, the availability of prior knowledge through previously generated frames helps in increasing the variance of the model, which is useful for generating the interactions between various concepts.

We qualitatively evaluate our method on diverse scenarios of generating multi-concept customized videos such as *a dog playing violin*, *a dog and a teddy bear eating together* and *a teddy bear walking in a living room*; and report quantitative metrics such as videoCLIP and DINO, along with human evaluation. We show improved results over baselines such as DreamBooth style finetuning. We also show the usefulness of our method for single concept customization in cases where compositionality is desired.

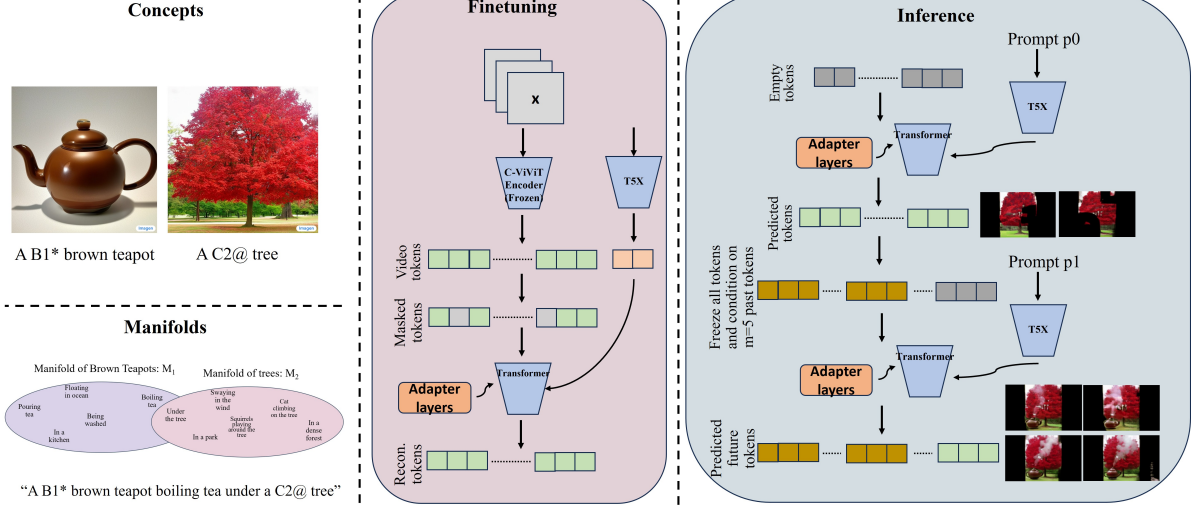


Figure 9.2: Proposed method for multi-concept video customization. First, we add adapter layers to the transformer architecture in an autoregressive T2V model and finetune these additional layers on the given images or videos associated with the N custom concepts and their corresponding text prompts. The goal is to find the solution at the intersection of the video manifolds corresponding to various custom concepts. Then, we condition on m ($= 5$) prior frames and sequentially generate the custom concepts and their interactions in a controlled manner using a set of prompts $p_0 \dots p_N$. The prompts $p_0 \dots p_N$ are designed to represent the scene in a top-down manner, each prompt adding a custom concept and the associated interaction.

9.1 Related work

9.1.1 Text to Video Models

Recent advances in generative AI has led to the development of text-to-video models such as Video Diffusion Models [440], Make-A-Video [441], Phenaki [442], CogVideo [443], ImagenVideo [411], PyoCo [450], PVDM [451], etc. These pretrained text-to-video models, along with pretrained text-to-image models, have been used for video generation, video editing and personalization (including motion/ action personalization) by a number of approaches such as Dreamix [448], Tune-A-Video [430], Pix2Video [452], Align Your Latents [453], CCEdit [454], Fatezero [455], video-p2p [456], Wang et. al. [457], ConditionVideo [458], infusion [459], Zhang

et. al. [460]. However, these methods are usually limited to performing style transfer, concept editing and single subject personalization in videos where a single subject is involved. Methods such as Esser et. al. [445] are able to control the structure and content individually, but require a large amount of data to train on. The goal of this chapter is to generate a multi-concept customized video, consistent with the input text description, using a limited number of data samples per image/ video concept (1 in this chapter) and a pretrained text-to-video model.

9.1.2 Personalization of generative models

Image and video personalization using generative AI models has seen tremendous progress in the recent past for single concept personalization with the development of methods such as DreamBooth [28], ControlNet [390], Instructionpix2pix [461], Imagic [355], Null text inversion [462], Video ControlNet [463], Attend-and-excite [3], paint by example [389], Free-bloom [464]. Recent times have also seen multi-concept customization methods for text-to-image models such as Custom Diffusion [388], FastComposer [429], SVDiff [395], Subject Diffusion [428]. However, multi-concept customization methods developed for text-to-image models are not very effective [432] for video customization and cause artifacts, attribute binding problems, memory and data issues. In contrast, our method trains a single lightweight adapter jointly on all concepts and does not require any data regularization for inducing variance. VideoDreamer [432] is a multi-subject video customization method concurrent to our work. VideoDreamer leverages a pretrained text-to-image stable diffusion model to extend to custom video generation. Its use of a text-to-image model limits its application to cases where there is minimal subject-subject interaction or limited motion. Moreover, it cannot transfer or customize motion information. In contrast, our goal is

to leverage pretrained text-to-video models to perform multi-concept customization, where the custom concept can be a subject or motion, or even background.

9.2 Method

9.2.1 Problem Formulation

We are given a pretrained text-to-video model $\mathcal{M}$ and image(s) or video(s) $c_i, i = 1 \dots N$ corresponding to N concepts with text descriptions $t_{c_i}, i = 1 \dots N$, where N is the number of concepts. Our goal is to generate a video V_{gen} that contains the custom concepts, as described by a query text t_{gen} . We assume that the model is capable of auto-regressively generating future frames.

9.2.2 Proposed Method

As our first step, we incorporate the knowledge of the custom concepts $c_i, i = 1 \dots N$ within the text-to-video model by finetuning the diffusion model. Our next step is to use the finetuned model to generate the custom video, corresponding to t_{gen} , by adding subjects one-by-one, in an autoregressive manner. Each step of the generation progressively adds various custom concepts and their corresponding interactions, as dictated by the query text t_{gen} . We now turn to describe our method in detail.

9.2.2.1 Multi-concept finetuning

Given $c_i, i = 1 \dots N$, we fine-tune $\mathcal{M}$. Following [333], we do parameter efficient finetuning and specifically, adapter tuning, as is shown effective for generative image transformers [465].

Note that we jointly train a single adapter for all concepts instead of training an individual adapter for each concept as in [388].

Our next step is to use the finetuned text-to-video model to generate personalized videos with custom concepts, dictated by a query text t_{gen} . Consider a *Teapot* and a *Tree*. The video manifold M_1 of *Teapots* contains videos of the *Teapot* pouring tea, floating in the ocean, being washed, boiling tea and in a kitchen, dictated by text. Similarly, the video manifold M_2 of *Trees* contains videos of the tree swaying in the wind, the relational notion of under the tree, in a park, squirrels playing around the tree, cat climbing the tree and the tree being located in a dense forest, dictated by text. Thus, each of the manifolds contains what the model is familiar with in terms of generation. The video corresponding to "a *Teapot* boiling tea under a *Tree*" is contained in the (non-linear) intersection of the two manifolds $M_1 \cap M_2$, generation of which is not straightforward. In the next section, we present a method to generate the multi-concept customized video.

9.2.2.2 Causal generation, one subject at a time

We hypothesize that sequential and controlled traversal through the various manifolds (each understood by the model) towards the intersection space of the manifolds, while generating video frames sequentially, leads to the solution. Our solution is based on multi-step generation, each step incrementally generating what the model understands, while remembering what it has already generated. Thus, we begin at the manifold corresponding to one of the custom concepts to generate a few frames corresponding to the first concept by using an appropriate text description. Next, to generate video frames with multiple concepts, we condition on the generated video

frames of the first concept and use text descriptions describing the other concepts to incrementally add them and their interactions.

For instance, in the case of "a *Teapot* boiling tea under a *Tree*", we begin at the manifold of *Trees* to generate a few frames of the *Tree* using the text prompt "a *Tree*". Next, move towards the manifold of *Teapots* to generate a video of the *Teapot* boiling tea under a *Tree*. Consequently, to generate the next video frames of "a *Teapot* boiling tea under a *Tree*", we use the text prompt "a *Teapot* boiling tea under a *Tree*", conditioned on the previously generated frames of the *Tree*. Thus, the first few frames of the video depict the *Tree*, and the subsequent frames show the *Teapot* boiling tea under the *Tree*.

At each step of the sequential generation, the existence of a strong prior (in terms of video frames generated upto that stage), along with the variance arising from the pretrained model, alleviates bias related issues. Consequentially, at each step of the sequential generation, the model needs to be conditioned on an appropriate number of prior frames it previously generated and generate the subsequent frames in an autoregressive manner. If too few frames are used for the conditioning, the model cannot remember prior knowledge. On the other hand, if too many frames are used, the model may not be able to bring in new concepts due to bias issues. Moreover, it is important to keep the sequential generation controlled, in order to make sure that the generated video stays within the intersection manifold i.e. the number of frames generated for each concept is a key hyperparameter. In summary, causality is key to maintain prior context.

Algorithmically, let F be the total number of frames that are being generated and $p_i, i = 1 \dots N$ be the prompts describing the various concepts and their interactions. Tokens corresponding

to all F frames are initialized to be empty.

$$\text{for } k \text{ in range } (0, F) : z_k \leftarrow \mathcal{M}(z_k) | (z_{k-1:k-m}, p). \quad (9.1)$$

As per this equation, the generation of each frame (except the first m frames) is conditioned on the past m frames. The prompt p is sequentially set to $p_1 \dots p_N$. Since the model can generate 11 frames in one go, we set $p = p_1$ for $k \leq 11$. In the 2-concept customization case, we set $p = p_2$ for the rest of the frames.

Structuring the concepts. Sequential generation requires the various concepts to be structured in an appropriate manner – it is important to carefully design $p_i, i = 1 \dots N$. We propose to structure the concepts in a top-down manner - consistent with the natural representation of a scene. For example,

- Custom subject and custom background - first generate the background. Next, condition on the frames containing the custom background to add the custom subject performing action. For eg. given concepts ‘a B1* futuristic restaurant’ and ‘a C2@ cat’, and $t_{gen} = \text{‘a C2@ cat eating noodles in a B1* futuristic restaurant’}$, we set
 - $p_1 = \text{‘a B1* futuristic restaurant’}$
 - $p_2 = \text{‘a C2@ cat eating noodles in the futuristic restaurant’}$.
- Custom subject and custom action (from a video) - first generate a random (related subject) performing the custom action. Next, steer the model towards the manifold of the custom subject performing action. For eg. given concepts ‘a person playing tennis in a tennis court’ and ‘a C2@ cute cat’, and $t_{gen} = \text{‘a C2@ cute cat playing tennis in a tennis court’}$,

we set

- $p_1 = \text{'a person playing tennis in a tennis court'}$,
 - $p_2 = \text{'a C2@ cat playing tennis in a tennis court'}$.
- Custom subject 1 and custom subject 2 - generate one of the custom subjects. Condition on this to generate the other custom subject interacting with it. For eg. given concepts ‘a B1* brown teapot’ and ‘a C2@ tree’, and $t_{gen} = \text{'a B1* brown teapot boiling tea under a C2@ tree'}$, we set

- $p_1 = \text{'a C2@ tree'}$,
- $p_2 = \text{'a B1* brown teapot boiling tea under a C2@ tree'}$.

9.3 Experiments

9.3.1 Setup

Our proposed method for multi-concept video customization is built on autoregressive generation. Consequentially, we use Phenaki [442], a text-to-video model with autoregressive capabilities, as our backbone architecture. Phenaki learns video representations in a causal manner, compressing the video into a small representation of discrete tokens using a spatial transformer and a causal transformer. A MaskGiT transformer model [466] is then trained to reconstruct the video tokens, conditioned on a text prompt. A bidirectional text transformer is used to compute video tokens from text. At inference, starting from empty tokens and conditioned on the input text, video tokens are generated, which are then detokenized to generate a video. At each step of the inference, Phenaki can freeze video tokens corresponding to past frames and generate video tokens corresponding to future frames in an autoregressive manner. Classifier-

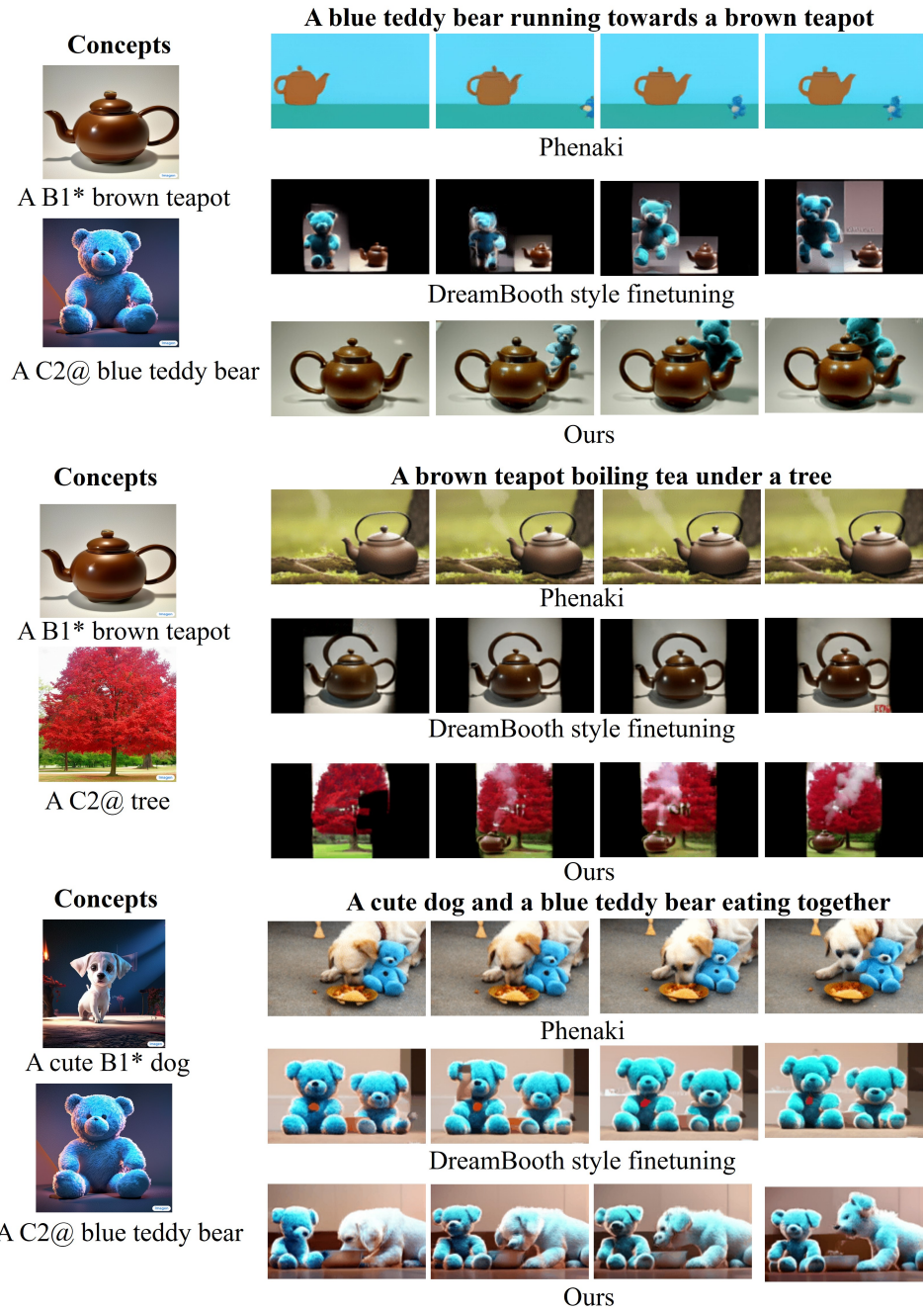


Figure 9.3: **Two subject customization.** While Phenaki contains some prior knowledge of the interactions between different kinds of subjects, the finetuned model is unable to generate the interactions between the custom subjects. Through controlled and sequential autoregressive generation of concepts and their interactions, our method is able to generate customized videos with two subjects interacting with each other.

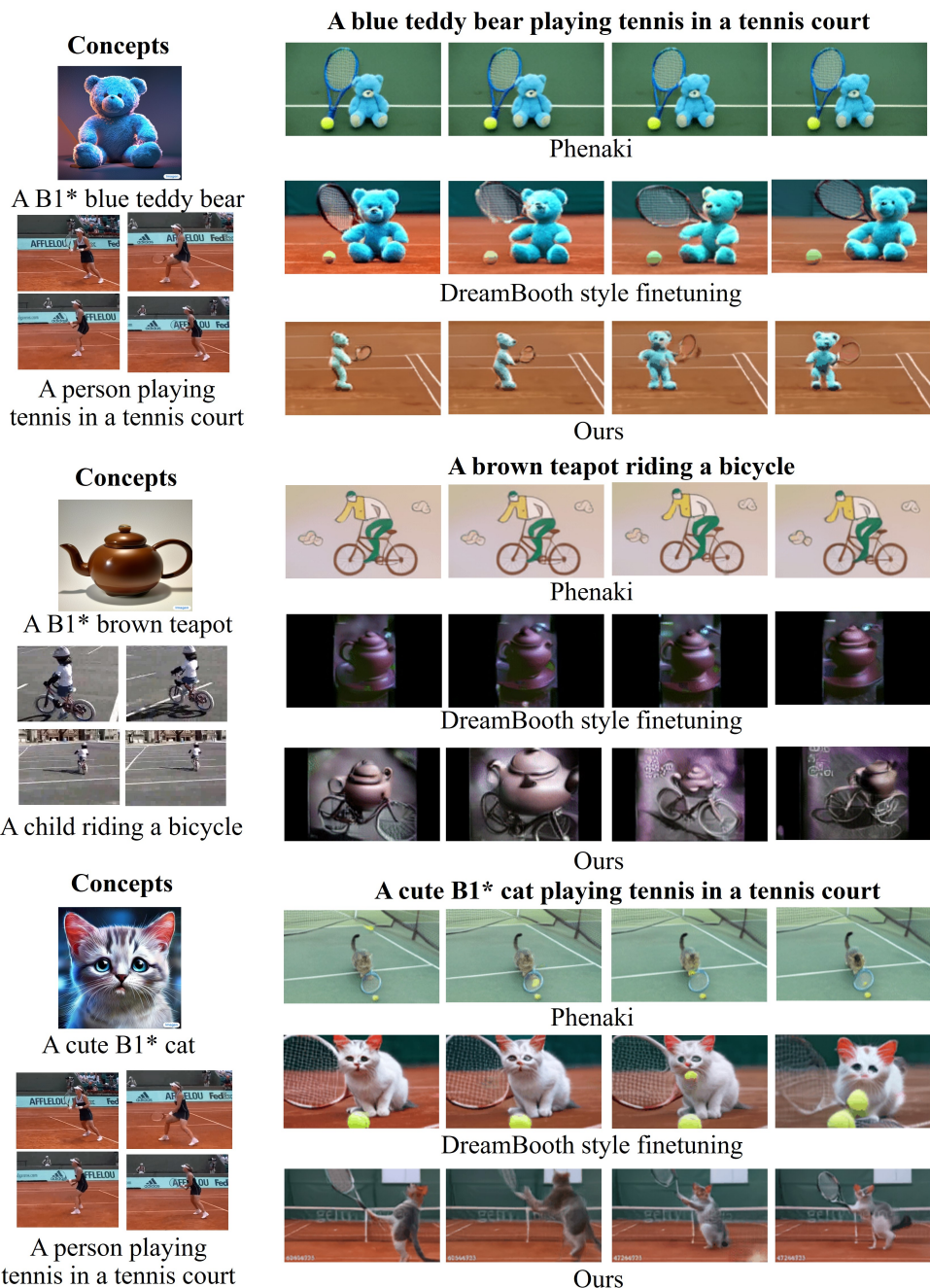


Figure 9.4: **Subject Action customization.** Phenaki, due to lack of prior knowledge, is unable to generate certain actions. The finetuned model is able to generate the custom subject, but not the precise action, even after finetuning. Generating the motion corresponding to the action first, followed by generating the custom subject performing action (conditioned on the motion) enables the generation of the custom subject performing the custom action. Thus, we are not only able to ‘teach’ the action to the model, but are also able to customize the subject performing the action.

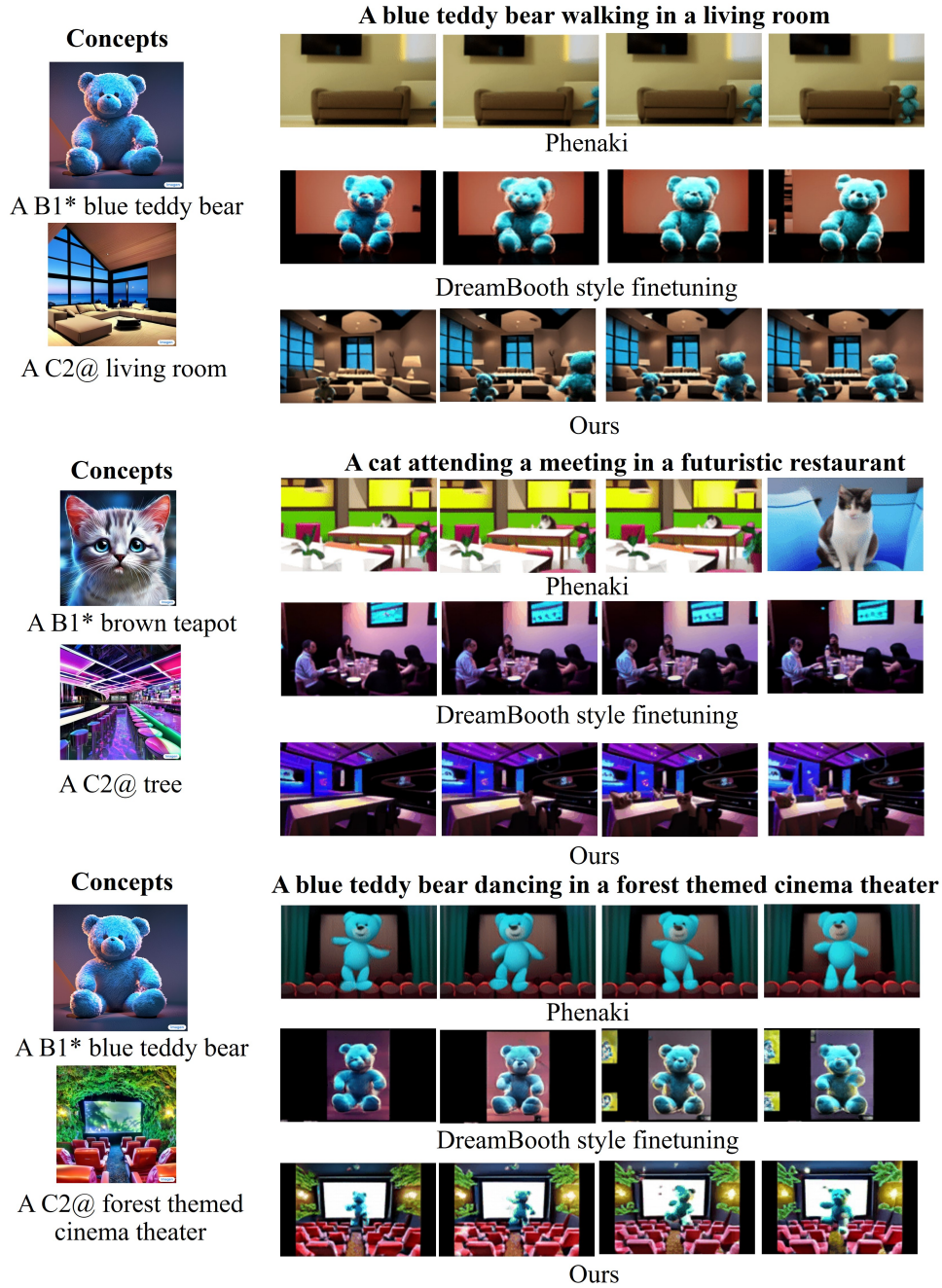


Figure 9.5: **Subject BG customization.** While subject-background customization is relatively easy for the Phenaki model, the finetuned model is unable to emphasize on the background due to bias issues. This is resolved by our model, which first generates the background and then generates video frames of the custom subject performing action, conditioned on the background.

free guidance [392] controls the alignment between the text and the generated video. Phenaki is pretrained on a large-scale text-image and text-video dataset, which we use for multi-concept video customization. We use the low-resolution Phenaki model in all experiments.

We finetune the adapter weights jointly on all image concepts, followed by video concepts, for 1000 steps with a learning rate of $1e - 4$ using the Masked Visual Token Modeling (MVTM) loss[466], along with classifier-free guidance. The adapter has a hidden size of 2. Our generated videos have a resolution of 160×96 . At inference time, to condition on prior frames, we set $m = 5$. Conditioning on 5 frames, determined empirically, is consistent with the fact that Phenaki is trained to generate 11 frames at a time and conditioning on half the frames achieves the right trade-off between the amount of prior knowledge required for the generation of future frames. For each example, we present comparisons of our method with two baselines:

- DreamBooth style finetuning: In this case, the model is finetuned and inferred using t_{gen} . This is inspired by the DreamBooth [28] model with adapter [333] layers, extended to Phenaki video customization. This is our closest baseline for multi-concept video customization.
- Phenaki: This baseline corresponding to Phenaki as it was trained with no additional finetuning.

We use Imagen [411] to generate the image for each custom concept, concept-specific videos are from UCF-101 [467]. Custom concepts can take multiple forms, e.g. background, visuals of a character or subject and a custom performed action. We experiment with various types of customization such as subject-subject, subject-action, subject-background. The design of the prompts was a purely creative task, we designed close to 100 prompts and attempted to keep it as diverse as possible. Our data set-up is as follows:

- Subject-subject customization: We use 14 examples corresponding to the 3 examples shown, 9 in the supplementary and ‘a cute dog and a blue teddy bear in a running race’, ‘a cute dog washing a blue plate’.
- Subject-action customization: We use 4 subjects: a small cute fox, a small cute dog, a small cute cat and a blue teddy bear and 8 actions: playing violin, playing hula hoop, playing drums, playing tennis, crawling, doing pushups, riding bicycle, doing TaiChi. This results in a total of 32 videos.
- Subject-background customization: We use 4 subjects: a small cute fox, a small cute dog, a small cute cat and a blue teddy bear, 4 backgrounds: living room, futuristic restaurant, forest themed cinema theater, ocean with seaplants and fishes. For each background, we generate 2 actions: (dancing, walking), (eating noodles, attending a meeting), (dancing, clapping hands), (playing with fishes, swimming) in order. This results in a total of 32 videos.

9.3.2 Qualitative Results

9.3.2.1 Two-concept customization

Two-subject customization: In the first case, while Phenaki is able to generate the video of a generic teddy bear running towards a generic teapot, the teddy bear is much smaller than the teapot. The finetuned model generates both subject in the video but is unable to generate the correct interaction (of the teddy bear running *towards* the teapot). Our method structures the generation as $p_0 = \text{‘a B1* brown teapot’}$ and $p_1 = \text{‘a C2@ blue teddy bear running towards’}$

the B1* brown teapot’ to generate a video of the C2@ blue teddy bear running towards the B1* brown teapot; the static object (teapot) is generated first followed by the moving teddy bear. In the second example, Phenaki is able to generate the video with generic subjects. The finetuned model has a bias towards the teapot and fails to generate the tree. Our method structures the generation as $p_0 = \text{‘a C2@ tree’}$ and $p_1 = \text{‘a B1* brown teapot boiling tea under a C2@ tree’}$ to generate a customized video of the teapot boiling tea under the tree; the background, static subject (tree) is generated first followed by the teapot boiling tea. In the third example, Phenaki generates a toy teddy bear, that is not eating. The finetuned model mixes the attributes of the concepts. Our method structures the generation as $p_0 = \text{‘a cute B1* dog eating together with the C2@ blue teddy bear’}$ and $p_1 = \text{‘a C2@ blue teddy bear eating together with a cute B1* dog’}$ to generate the dog and teddy bear eating together; the emphasis is on the dog first, followed by the teddy bear. In summary, controlled and sequential generation of the custom concepts and their interactions is effective in generating a multi-subject customized video.

Subject-action customization: We first generate the motion depicting the action, followed by the subject performing the action conditioned on the motion frames. Hence, p_0, p_1 take the general structure of $p_0 = \text{‘a random subject performing the action’}$, $p_1 = \text{‘custom subject performing the action’}$. In the first case, the pretrained model Phenaki does not understand what a teddy bear playing tennis looks like and is unable to generate the video. In the result of the fine-tuned model, there is some movement in terms of the teddy bear holding the tennis bat but the video does not depict a teddy bear playing tennis. Our method is able to generate a video of the custom teddy bear playing tennis. Thus, our method is not only able to teach the model the action ‘playing tennis’, but is also able to effectively condition on the motion to customize the subject

performing action. Similarly, in the third example, neither Phenaki nor the finetuned model are able to generate a cat playing tennis. Our method is able to generate a video of a custom cat playing tennis, thus teaching the model the action, as well as customizing the subject. In the second example, again, the Phenaki model does not know what a teapot riding a bicycle is. The finetuned model is able to generate some movement of the teapot but the generated video isn't really of a teapot riding a bicycle.

Subject-background customization: Our next task is to customize the subject and the background (and generate a random action). We first generate the custom background, followed by the subject performing a generic action in the background. Hence, p_0, p_1 take the general structure of $p_0 = \text{'custom background'}$, $p_1 = \text{'custom subject performing action in the background'}$. In the first example, while Phenaki understands what a teddy bear walking in a living room is, when finetuned, due to bias issues, the finetuned model is unable to generate a custom teddy bear walking in the custom living room. Our method first generates the custom living room, conditioned on which, the custom teddy bear walking in the living room is generated. Similarly, in the second and third examples, while Phenaki understands the interactions between the subject and the background, after finetuning, due to bias issues, the model is unable to generate the background well. Our method is able to generate videos of custom cats attending meetings in futuristic restaurants, and a custom teddy bear dancing in a forest themed cinema theater.

9.3.2.2 Three-concept customization

We show generated videos with 3 custom concepts. In the first example, we have three concepts: a teddy bear, a fox and a teapot and we want to generate a video of the two animals

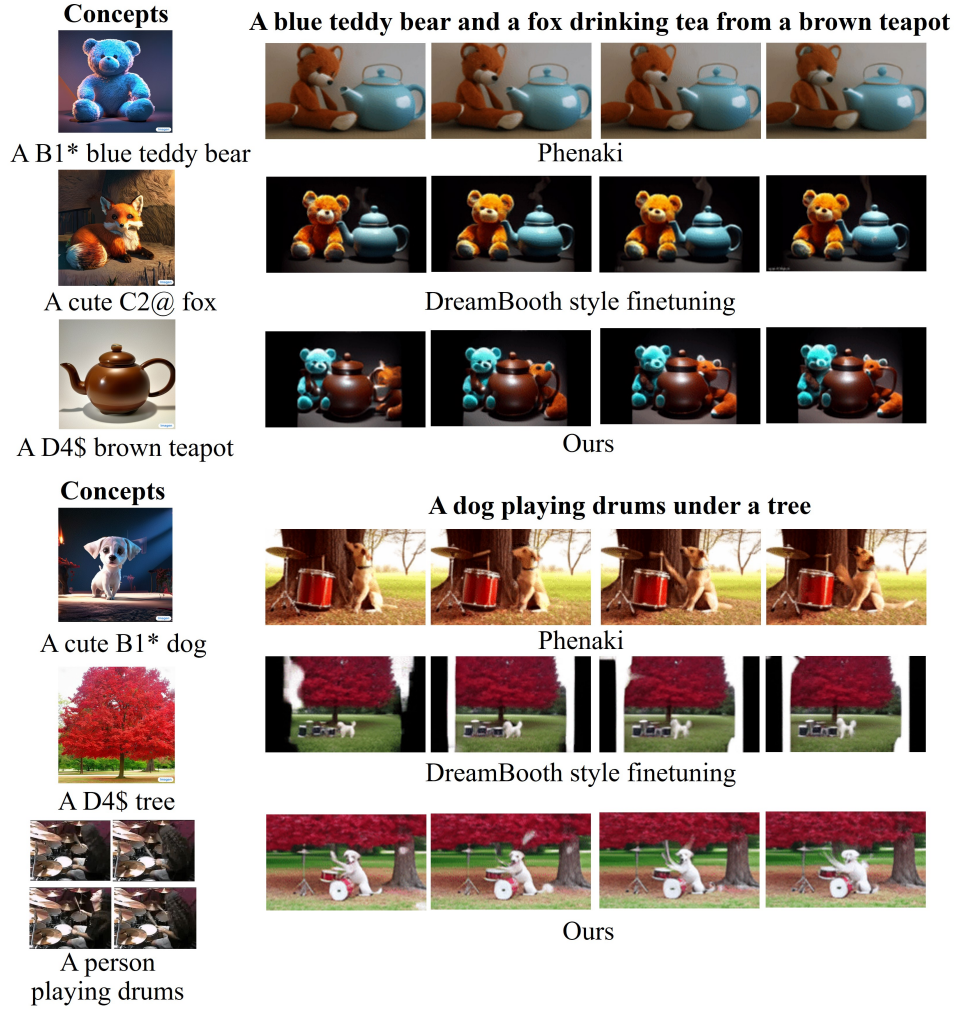


Figure 9.6: **Three-concept customization.** In the first case, the baseline models mix the attributes of the various concepts. Our method, by sequential generation of concepts and interactions, is able to generate the complex video with all desired concepts and interactions in the scene. Similarly, in the second example, while the finetuned model is able to bring in all concepts into the video, it is unable to generate the interaction of playing drums. Conditioning on motion information helps our method do so.

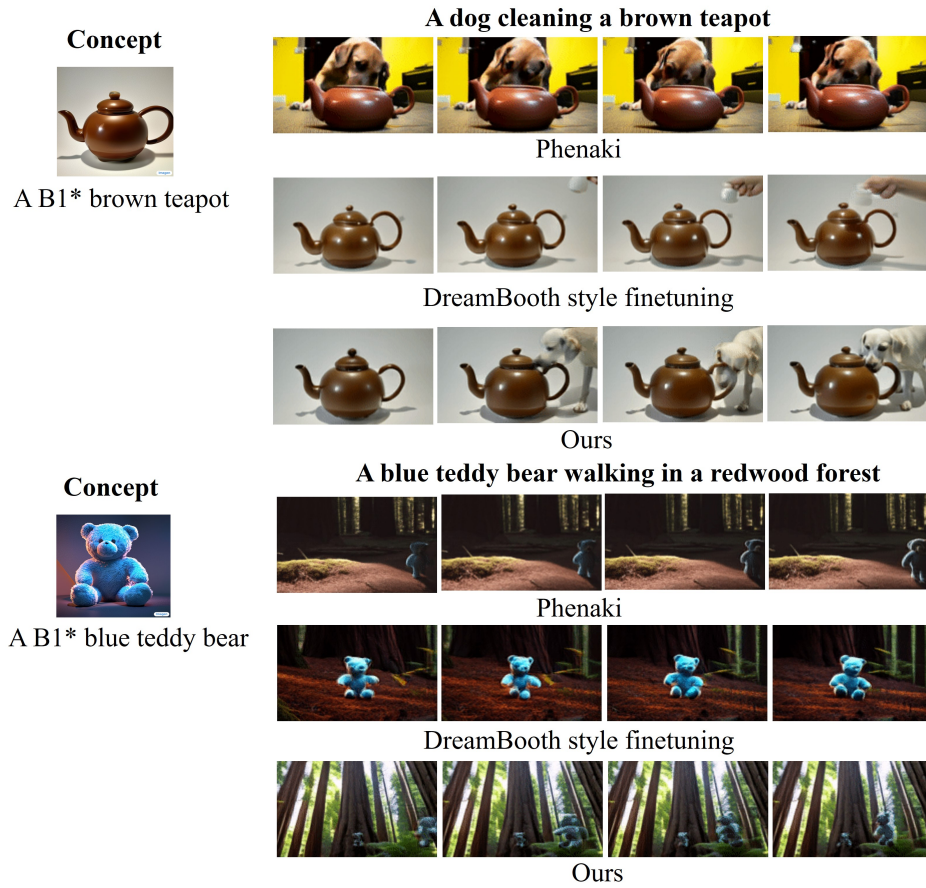


Figure 9.7: **Single concept customization and compositionality.** Causality can be a useful tool for generating videos where compositionality is desired, an interesting direction for future work.

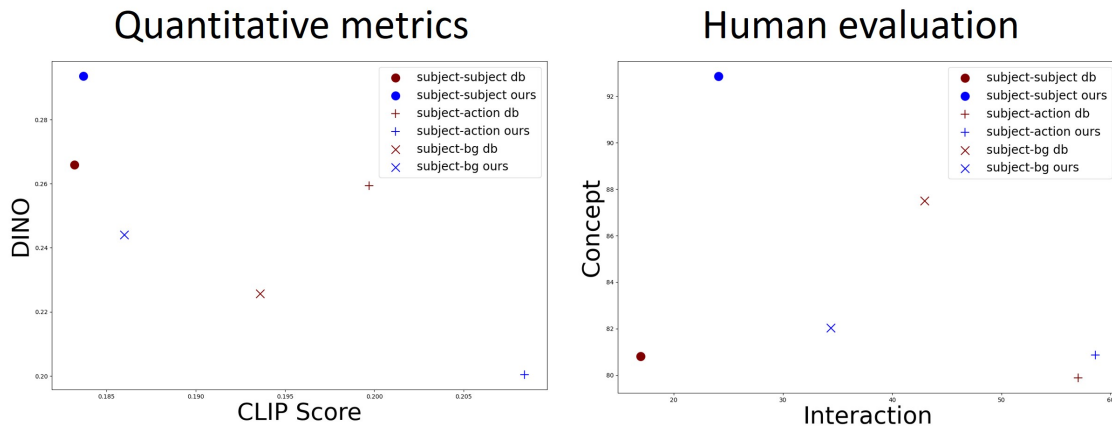


Figure 9.8: In the first figure, we show the videoCLIP and DINO scores for our method and finetuning baseline. In the second figure, we show results of the human evaluation.

drinking tea together from the teapot. Phenaki is unable to generate the video (even without customization) - we see that the attributes of the three concepts are mixed up. The model generates a blue teapot instead of a brown teapot and a hybrid of the teddy bear and the fox. In the result of the finetuned model, a teddy bear in the color of the fox is generated, and the teapot is in the color of the teddy bear. There is no fox in the scene. By bringing in the concepts and the interactions systematically, our method is able to generate the custom blue teddy bear drinking tea from the custom brown teapot with the custom fox. In the second example, we want to generate a dog playing drums under a tree. The dog, the tree and the action of playing drums are customized. Phenaki understands the scene and is able to generate a generic video. The finetuned model brings in all concepts into the scene but is unable to generate the action of ‘playing drums’. Our method, by conditioning on the motion tokens (of playing drums), is able to generate a video of the custom dog playing drums under the custom tree.

9.3.2.3 Compositional single-concept customization

While the focus of our method is on multi-subject customization, it is also useful for single-concept customization where the model does not understand compositionality. In such cases naive finetuning may not work due to bias issues and leveraging causality to generate compositional scenes can lead to better results. These results provide a direction for future work on using causality to generate compositional videos.

9.3.3 Quantitative results

We quantitatively evaluate our method using the Video-CLIP score [468] and the self-supervised similarity score, DINO [403]. The former computes the alignment with text, signifying if the concepts and interaction described by the text are present in the generated video. To compute the DINO score, we compute the average over the individual DINO score w.r.t. each concept in the video, also averaged over all frames of the video. Blau et. al. [404] state that generative models face difficulty in trading off perception-distortion. Similarly, we argue that videoCLIP-DINO (or text describing concept-concept interaction vs fidelity w.r.t. each concept) is a difficult trade-off to solve as it relates to the bias-variance trade-off.

We also analyze our results with a human evaluation study. For each video, we ask participants three questions: (i) Is concept 1 present in the video? (ii) Is concept 2 present in the video? (iii) Is concept 1 interacting with concept 2 as described by the text? We report the percentage of affirmative responses for each question and for each method.

We use 14 examples for subject-subject customization, 32 examples for subject-motion customization and 32 examples for subject-background customization. To compute the DINO and videoCLIP scores, for each example, we generate 8 results and choose the generated video with the best videoCLIP score as the final result. For the human evaluation, we show all 8 results (for each example) to human raters and compute the average.

Subject-subject customization. We show improved DINO score, there is a small improvement in the video-CLIP score as well. Human raters also rate our method much higher than the baseline in terms of both concept interaction, as well as alignment of the characteristics of the concept with

the input images. This indicates our method’s ability to better generate multiple custom subjects interacting with each other.

Subject-action customization. We show improved videoCLIP score indicating our methods’ ability to generate the action better. The DINO score of our method is lower than that of the baseline. The baseline method is unable to generate good actions and in many cases reproduces the input concept image with mild movement, hence, its resemblance to the input concept image is better leading to a higher DINO score. Human raters rate our method better in terms of both interaction as well as concept fidelity.

Subject-background customization. We show improved DINO score, indicating our methods’ ability to generate the subject as well as the background with high fidelity. The video-CLIP score of our method is lower than that of the baseline. This is because the baseline method generates a video that focuses on the subject performing the action. This also implies higher resemblance to the input subject image or a higher individual DINO score w.r.t. the subject. We believe this also leads to human raters preferring the baseline method more than our method. However, the background is not very well depicted. The individual DINO score for background for our method is 0.3496, while that of the baseline is 0.1896. Our method focuses on generating the entire background depicted in the input concept image, along with the subject performing action; thus indicating higher consistency with both input concepts while generating the interaction.

Other comparisons. Custom diffusion [388] requires us to mine images from the original training dataset (on which Phenaki was trained) for regularization, which was unavailable. Other multi-concept customization approaches, with the exception of DreamBooth, were simply not feasible

to expand to video customization using Phenaki.

9.4 Limitations and Future Work

Our method has a few limitations: (i) extension beyond three concepts is non-trivial because the model may not have the ability of remember what it generated in the first step, and structuring the concepts/ interactions is complex. (ii) Controlling the interactions through text is not easy, it may be useful to have more control signals that enable the model to generate interactions respecting the 3D world (such as contact points between subject and ground, relative size between objects). (iii) We use the low resolution model of Phenaki to generate the results. This was the only autoregressive model we had access to at the time of working on this method. The development of better video foundation models and superresolution can enable the generation of better results. (iv) The quality of the generated videos is highly dependent on the base model, with the development of richer models that understand temporal coherence and 3D relations in the world, the artifacts in the generated videos can be removed. Furthermore, the recent development of autoregressive diffusion transformer models for video generation can enable the application of our multi-concept video customization method. (v) Our method takes in a single image per subject to generate personalized videos. This leads to low variance, and the model hallucinates a lot of details corresponding to the subject. This causes low subject fidelity in some generated videos, a direction for future work. Future work can focus on overcoming these limitations, in addition to further improving the text alignment and fidelity w.r.t. input concepts. Another direction for future work is to automate the structuring of the prompts using LLMs and simplify hyperparameter tuning related to finding the number of future frames (and conditioning

on prior frames) to be generated using each prompt. Also, incorporating causality within non-causal text-to-video models such as video diffusion models to use them for multi-concept video customization is an avenue to explore. This chapter provides a direction for the importance of causality in multi-concept customization and can also be extended to long video generation.

Chapter 10: Prompt Mixing in Diffusion Models using the Black Scholes Algorithm

Text-to-image diffusion models [343, 469] such as Imagen [470] by Google, Firefly [471] by Adobe, DALL-E [472] by Microsoft, and Stable Diffusion [284, 473] by Stability AI can produce photorealistic images from textual descriptions and are typically trained on vast amounts of data using substantial computational resources. The latest versions of these models (as of May 2024) have seen significant improvements in comparison to their initial versions, attributed to various factors, such as increased training data, enhanced text understanding networks, improved image generation networks, and optimized strategies. Data efficient solutions [474], leveraging pre-trained diffusion models, can enhance results without requiring additional data collection. Prompt engineering [475] is one effective approach, involving techniques such as rephrasing prompts to improve model generalization. Another method is using large language models (LLMs) [476, 477] to parse complex prompts and determine useful priors for image generation. Attention maps [3] and prompt mixing techniques [478, 479] also contribute to better results.

Prompt mixing becomes particularly valuable when aiming to generate images that blend different text-image concepts i.e. when aiming to generate images at the intersection [478, 480, 481] of different text-image concepts. For instance, consider generating an image resembling a hybrid of a pink cat and a dog. Current diffusion models [470, 471, 472, 473] excel at straightforwardly combining individual text prompts to create visually appealing combinations.

However, these capabilities are often restricted to advanced model versions (version 3) and rely on large-data techniques.

Data efficient prompt mixing techniques can enhance image generation without additional data. While linear interpolation [355] of text embeddings from individual prompts is a simple approach, it may not be optimal due to the highly non-linear nature of the text-image manifold and potential bias issues. Alternatives include switching between prompts during diffusion denoising, either by alternating [356] or using a step-wise technique [478]. However, these methods often require human involvement and careful prompt engineering for optimal results.

The question arises: what’s the best way to switch between text prompts for prompt-mixing? An automated approach considers the model’s varying capabilities with different prompts. While attention maps [3, 482] and layout guidance![483, 484] methods excel at guiding the model toward distinct scene entities, they may not be as effective for blending concepts within the same entity. To address this, during each step of diffusion denoising, the network should automatically focus on the text prompt that corresponds to the image’s deficiencies.

Main contributions. In this chapter, we introduce a novel approach to prompt mixing by integrating concepts from economics and finance. Our objective is to leverage pretrained diffusion models for prompt-mixing to generate images at the intersection of various text-image manifolds. During each step of the diffusion denoising process, our method dynamically conditions on the text prompt that requires the highest level of attention. To achieve this, our algorithm assesses the ‘cost’ associated with each relevant text prompt and selects the optimal conditions for the subsequent step of image generation based on the ‘cost’ that requires maximum optimization.

Key Insight: The inspiration for our approach lies in non-equilibrium thermodynamics [485], which is at the core of the development of diffusion models. These models share a conceptual foundation with the Black Scholes model[486], a Nobel Prize-winning mathematical framework extensively employed in financial markets for pricing European call options[29]. The denoising steps within a diffusion model, aimed at image generation, constitute a Markovian time-series. In our analogy, the image itself represents a valuable ‘stock’ or ‘asset’ that we aim to ‘purchase’ (or generate) at the most favorable ‘cost’ (or alignment with the text prompts). During each step of the diffusion denoising process, we *extrapolate the generate latents to compute the ‘stock’ prices at the corresponding time-step, and use the properties of diffusion models to compute the various variables involved in the Black Scholes algorithm.* Consequently, we *leverage the Black Scholes model to predict a score for each text prompt.* This score serves as *an indicator of how the image should be conditioned in the subsequent timestep.* By adopting this approach, our model dynamically focuses on aspects that require *attention*, ultimately generating an image that optimally satisfies all relevant text prompts.

We perform experiments on prompts with varying complexities to assess our method’s ability to seamlessly blend different objects and backgrounds across scenarios. Through CLIP-based quantitative and qualitative comparisons against several baselines such as vanilla stable diffusion, linear interpolation, alternating sampling, prompt switching, and CLIP guided prompt switching, we highlight the superiority of our approach.

10.1 Related work

Prompt mixing. Prompt mixing is a technique where different text prompts are used at various steps during the diffusion denoising process. Patashnik et al. [478] follow a step-wise approach, while Aerial Diffusion [356] alternates between prompts to create semantically consistent aerial-view images. Tools such as Image Mixer Diffusion [481] and CLIP Guided Image Mixing [480] follow similar techniques of prompt switching to blend multiple text prompts. However, these methods often require complex hyperparameter tuning.

Mixing step. The mixing time or mixing step [487] of a Markov chain refers to the duration it takes for the chain to reach its steady-state distribution. Mixing time has found applications in image editing [488] and synthesizing out-of-distribution (OOD) images [489]. When switching between prompts using mixing time, a step-wise approach is followed, akin to the work by Patashnik et al. [478]. The advantage of mixing-time approaches over the method proposed by Patashnik et al. [478] is that it mathematically determines the optimal switching time without requiring complex hyperparameter tuning. To approximate the mixing time, it estimates the radius of the latent space. However, there are a few limitations to these methods. Firstly, most of these approaches are applied to diffusion models trained on relatively small, problem-specific datasets. Secondly, identifying clear boundaries on large-scale foundation models remains an unsolved challenge. Additionally, these methods lack the flexibility to choose the most optimal prompt at each timestep.

Understanding the latent space. Tangential to our work, there has been substantial work [490, 491, 492] on exploring the latent space. These investigations have yielded valuable insights and

practical solutions for downstream tasks such as image editing and manipulation applications [305, 341, 493, 494, 495]. Furthermore, a deeper understanding of the latent space within diffusion models has led to advancements in various methods [284, 494, 496].

10.2 Background

10.2.1 The Black Scholes Model

In this section, we provide a brief overview of the Black Scholes pricing model [486] used to determine the price of European call options of assets. which is used to determine the price of European call options on assets. In simple terms, a European call option allows an investor to lock in the price of an asset at any time but permits stock purchase (if desired) only upon expiration. Regardless of whether the stock price moves favorably or unfavorably over time, this option structure remains consistent. Investors rely on the Black-Scholes model to predict stock prices over time and make informed decisions about the optimal timing for stock purchases. The Black Scholes formula involves 5 key variables:

1. Underlying stock price or spot price S : This represents the current price of the asset.
2. Strike price K : The strike price is the cost of the asset at the time of expiry.
3. Time to expiration t : It measures the time difference between the current moment and the expiry time.
4. Volatility σ : Volatility reflects the variation in prices of the asset.
5. Risk free rate r : The risk-free rate is the minimum return on an investment when the investor faces zero risks.

To obtain the **Black Scholes score** of purchasing an asset, the spot price S is first multiplied by the standard normal probability distribution function. From this result, to obtain the final cost C , the strike price K multiplied by the cumulative standard distribution function is subtracted. Mathematically,

$$SN(d_1) - Ke^{-rt}N(d_2), \quad (10.1)$$

where

$$d_1 = \frac{\log \frac{S}{K} + (r + \frac{\sigma^2}{2})(t)}{\sigma\sqrt{t}}, d_2 = d_1 - \sigma\sqrt{t}. \quad (10.2)$$

Assumptions made by the Black Scholes Model. The Black Scholes model makes five assumptions:

(i) No dividends are paid out during the life of an option, **(ii) Market movements are somewhat random**, (iii) There are no transaction costs in buying the asset, **(iv) The volatility and risk free rate of the underlying asset are known and constant**, **(v) The returns of the underlying asset are normally distributed.**

10.2.2 Relation to Diffusion Models

The main concept behind diffusion models [343] involves iteratively adding small amounts of random Gaussian noise to transform an initial photorealistic image x_0 to noise $x_T \sim \mathcal{N}(0, I)$ in T steps. This process is known as the *forward process*. The converse of the theorem states that by starting from random noise $x_T \sim \mathcal{N}(0, I)$ and refining it iteratively for T steps, we can generate a photorealistic image x_0 . Since diffusion is a gradual process, the value of T is typically large. At each intermediate timestep, $t \in \{0 \dots T\}$, x_t satisfies $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon_t$. $0 = \alpha_T < \alpha_{T-1} \dots < \alpha_1 < \alpha_0 = 1$ are the hyperparameters of the diffusion schedule [343] and $\epsilon_T \sim$

$\mathcal{N}(0, I)$. To obtain x_{t-1} at each refinement step, the neural network $f_\theta(x_t, t)$ is applied along with the corresponding random Gaussian noise perturbation. In other words, during each step of diffusion denoising, the known variance in added noise follows a Gaussian distribution [343]. In finance, the volatility refers to the standard deviation in the way stock prices change from one time instant to the next. **The mathematical description of diffusion models, derived from non-equilibrium thermodynamics [485], shares similarities with the derivation of the Black-Scholes model [486] used in pricing European call options within financial markets. Both models emerge from similar assumptions and conditions, underpinned by a shared mathematical structure.** Assumptions (i) and (iii) made by the Black Scholes model, which talk about dividends and transaction costs, are irrelevant to diffusion models.

10.3 Prompt Mixing using Black Scholes Model

10.3.1 Problem Formulation

Consider a set of N text prompts denoted as $P_1, \dots, P_N$ and T diffusion denoising steps. Note that text-to-image diffusion architectures often include a text encoder that maps the text prompt to a joint text-image space. Starting from random noise and conditioned on the embedding in the text-image space, the neural network generates the final image.

Our objective is to generate an image that aligns with multiple text prompts simultaneously i.e. at the intersection of the various text-image manifolds of the diffusion models. We pose this task as a prompt mixing [478] problem. During each step of diffusion denoising, our aim is to select the most relevant text prompt (or the text prompt with respect to which the image requires further refining) for conditioning the model, ensuring the generation of an optimal image

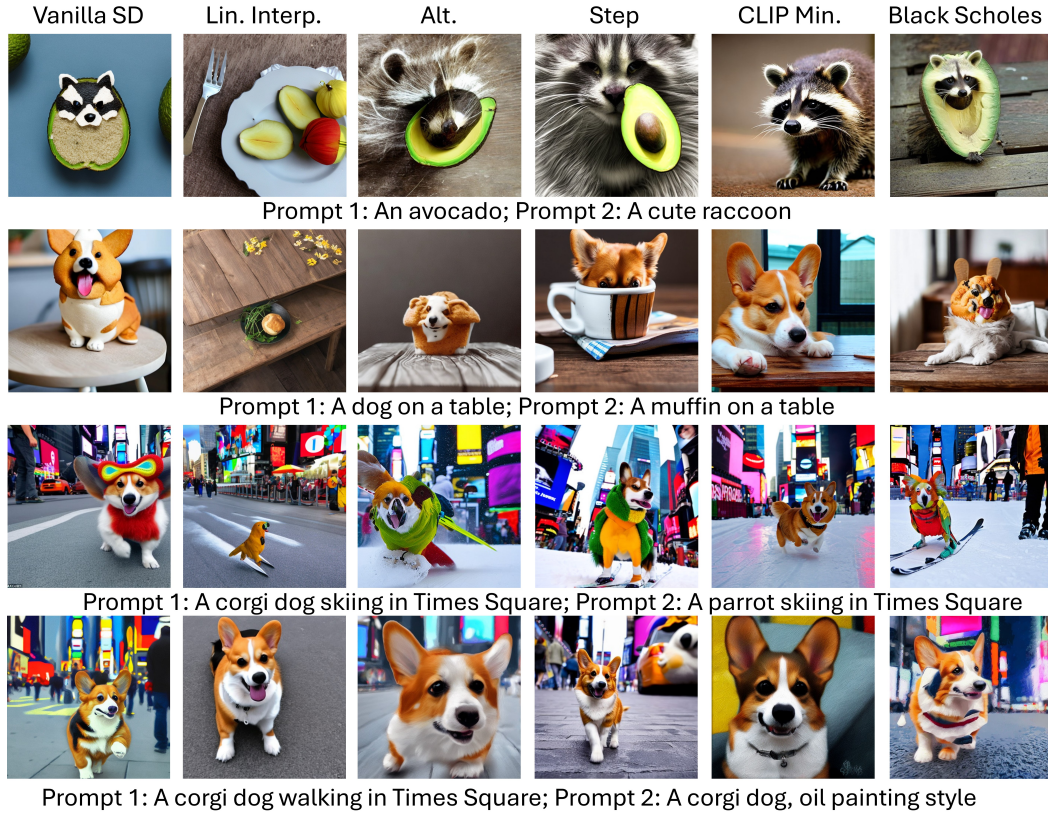


Figure 10.1: We present more results of our method along with comparisons. Vanilla SD fails to capture clear characteristics of individual text prompts, omitting distinct features such as those related to avocado/raccoon, muffin, parrot, and oil painting style. Linear interpolation generates images not consistent with the prompts, due to issues with non-linear manifolds. CLIP Min. generates images biased towards one of the prompts. Alt. and Step prompt selection methods suffer from artifacts and are not very successful in blending the characteristics of objects corresponding to the individual text prompts - the avocado/raccoon, muffin/dog are not blended well. In the parrot/dog image, the characteristics of the parrot are missing. Alt. generates artifacts in the Times Square/ oil painting image, while Times Square is not characterized well in the image generated by Step. In contrast, the Black-Scholes model adeptly overcomes these limitations, generating realistic images that meticulously balance and preserve the unique characteristics of each individual text prompt. The images are from set 1, set 2, set 3 and set 4 respectively.

satisfying all individual text prompts.

10.3.2 Method

Selecting the most effective text prompt to condition a text-to-image diffusion model can be approached in various ways. Naively switching [356, 478] between prompts is sub-optimal due to the significant human effort involved. Moreover, it fails to address the critical issue of prioritizing the most relevant prompt in an automatic manner. Alternatively, one can compute the CLIP score at each step and move toward the text prompt with the lowest CLIP score. While this approach is reasonable, it overlooks essential factors related to the dynamics [343] of the diffusion denoising process, which significantly impact the final image generation. Leveraging the Black Scholes model allows us to model these dynamics effectively.

We utilize the CLIP score to quantify the “price” of the generated image, treating it as analogous to a stock. The CLIP score has been widely used to measure text-image alignment, making it a suitable metric for evaluating how well our generated image aligns with each text prompt. Consider diffusion denoising at step i out of total steps T . We derive the relevant variables for Black Scholes formula as follows:

1. Underlying stock price S : The underlying stock price corresponds to the current value of the asset. Let z_t represent the predicted latents at timestep t . We denote $z_{0,t}$ as the latents of the final predicted image extrapolated from z_t . In other words, if the denoising process were to proceed in a straightforward manner following the same direction as computed for z_t , the latents of the final predicted image would be $z_{0,t}$. The underlying stock price, denoted as S , is determined by the CLIP score with respect to text prompt P_p . This CLIP

score can be computed using the image decoded from $z_{0,t}$. To maintain values within the range of 0 to 100, we multiply the CLIP score by 100.

2. Strike price K : This represents the estimated asset price at expiry, following the last step of diffusion denoising, which predicts the final image. Based on our experience with diffusion models, where we found that a CLIP score of approximately 0.25 indicates reasonable text-image alignment, we opted for a constant value of 0.25.
3. Time to expiration t : $t = T - i$.
4. Volatility σ : We compute the volatility as the square root of the variance used by the diffusion denoising scheduler at timestep i .
5. Risk-free rate r : $r = 1/T$.

During each iteration i of diffusion denoising, we calculate the Black-Scholes score $b_{i,p}$ for each text prompt P_p , $p = 1 \dots N$ using the given variables and the Black Scholes equations. Subsequently, during the next step of diffusion denoising, we condition on the text prompt with the lowest Black-Scholes score. This approach ensures that the denoising process prioritizes the text prompt associated with the minimum Black-Scholes score for generating an optimal image consistent with all prompts.

10.4 Experiments and Results

Metrics. In line with previous research [3], we assess performance using the CLIP Score. We use two variants of the CLIP Score: (i) CLIP-combined, which computes the score by comparing the generated image against a combination of individual text prompts; and (ii) CLIP-add, which

Algorithm 1 Black Scholes prompt mixing in backward diffusion enables the diffusion model generate an optimal image at the intersection of multiple text-image manifolds.

```

1: Initialize latents to random Gaussian noise.  $z_T \sim \mathcal{N}(0, I)$ ; T is the total number of diffusion
   denoising steps.
2: Use the text encoder  $\epsilon$  to compute the CLIP embeddings of a (linguistic) combination of the
   text prompts  $\{P\}$ .  $e \leftarrow \epsilon_p(\cup\{P\})$ .
3: Initialize Black Scholes variables, strike price  $K = 0.25 \times 100$ ,  $rater = 1/T$ .
4: // Diffusion denoising - image prediction/generation loop
5: for  $t \leftarrow T$  to 0 do
6:    $z_{t-1} = z_t - f(z_t, t, e)$ ; f is the diffusion UNet.
7:   Use  $z_{t-1}$  to compute  $z_{0,t-1}$ .
8:   Variance  $\sigma$  at step  $t - 1 \leftarrow$  computed using the scheduler of the diffusion model.
9:   for  $i \leftarrow 1$  to  $N$  do
10:    Spot price  $S \leftarrow$  CLIP score With respect to text prompt  $P_i$ 
11:    Time to expiration  $\leftarrow t$ 
12:    Black Scholes score  $b_{t,i}$  with respect to prompt  $P_i$ , at timestep  $t$ ,  $\leftarrow$  use Equation 1,2.
13:   end for
14:    $P_{min} \leftarrow$  Text prompt corresponding to  $\min\{B_{t,i}\}, i = 1 \dots N$ 
15:    $e \leftarrow \epsilon_p(P_{min})$ 
16: end for

```

averages the CLIP scores for the generated image across each individual text prompt. To compute our metrics, we generate five images for each text prompt and use them all in the computation of our metrics.

Baselines. Our study examines several baselines: (i) **Vanilla SD:** Prompt engineering, we use the vanilla stable diffusion model and condition it on a single text prompt effectively describing all individual text prompts, (ii) **Lin. Interp.:** Direct combination of text embeddings, achieved through linear interpolation between text embeddings [355]. This method equally weights the text embeddings associated with each text prompt. (iii) Switching between text prompts. We consider two variations here: (iii-a) **Step:** Following Patashnik et. al. [478], we use one text prompt for the initial 7th to 17th denoising steps and then switch to the other text prompt for the remaining steps, (iii-b) **Alt.:** Following Kothandaraman et. al. [356], we alternate between

Method	CLIP-combined	CLIP-add
Vanilla	0.3463	0.3012
Linear Int. [355]	0.2885	0.2778
Alt. Samp. [356]	0.3445	0.3098
CLIP Min.	0.3195	0.2955
Step [478]	0.3220	0.2997
Black Scholes	0.3469	0.3112

Table 10.1: We follow prior work [3, 4] and analyze using CLIP Scores. The Black Scholes algorithm for prompt mixing in diffusion models achieves superior results, compared to other prompt-mixing techniques, as also evidenced by the qualitative results.

the two text prompts. (iv) **CLIP Min.:** Score-based combination, denoted as CLIP-min, where we select the text prompt corresponding to the lowest CLIP score from the previous denoising iteration.

Backbone architecture. We use the Stable Diffusion 2.1 backbone model [284] in all our experiments. There is no training involved, we use the pretrained model to directly perform inference. Our experiments are run on one NVIDIA A5000 GPU with 24 GB memory, and takes 14 seconds to generate each image using the Black Scholes model, the diffusion denoising is performed for 100 steps. For comparison, the vanilla stable diffusion model takes 13 seconds for 100 steps on the same GPU.

10.4.1 Experimental Settings

We consider the following experimental settings:

Set 1: Simple text prompts with single objects. In this experiment, we validate our approach using straightforward scenarios, we use two text prompts, each describing one class. Specifically, we consider 17 classes: [‘a rock’, ‘a coffee mug’, ‘a cute dog’, ‘a pink cat’, ‘a teddy bear’, ‘a robot’, ‘an alien’, ‘an avocado’, ‘a cute raccoon’, ‘a corgi dog’, ‘a parrot’, ‘a car’, ‘a squirrel’, ‘a

cute rabbit’, ‘pizza’, ‘muffin’, ‘icecream’], and construct $17c_2 = 136$ text prompts.

Set 2: Multiple objects per text prompt. To add complexity, blend multiple objects, in the presence of additional objects in the scene. The goal of this experiment is to assess the models’ capabilities in blending the right objects in the scene, and preserving the characteristics of the other objects. We work with three classes of objects: [‘a basket’, ‘a teapot’], [‘apples’, ‘bananas’, ‘a cute cat’, ‘a cute dog’, ‘muffins’], [‘a table’, ‘a carpet’, ‘a bed’] and construct combinations using the first and second set and the second and third set. Additionally, we include the data point “a cat in a bathtub” and “a corgi dog in a bathtub,” resulting in a total of 71 prompts.

Set 3: Object actions against backgrounds (prompt mixing w.r.t. object). We investigate the model’s ability to morph objects while considering object-level action information and scene or background context. We consider the following backgrounds, [‘walking in Times Square’, ‘skiing in Times Square’, ‘walking in a beautiful garden’, ‘surfing on the beach’, ‘eating watermelon on the beach’, ‘sitting on a sofa on the beach’, ‘watching northern lights’, ‘watching sunset at a beach’, ‘admiring the opera house in Sydney’, ‘sleeping in a cozy bedroom’, ‘admiring a beautiful waterfall in a forest’, ‘walking in a cherry blossom garden’, ‘walking in a colorful autumn forest’, ‘flying in the sky at sunset’]. The objects are [‘a kangaroo’, ‘a cute cat’, ‘a corgi dog’, ‘a parrot’, ‘a teddy bear’, ‘a penguin’]. In total, we create 210 prompts.

Set 4: Object performing action against a background (prompt mixing w.r.t. background). In this experiment, we explore how the model blends global information and backgrounds while considering objects in various scenarios. We consider the objects [‘a kangaroo’, ‘a cute cat’, ‘a corgi dog’, ‘a parrot’, ‘a teddy bear’, ‘a penguin’]. The backgrounds and style prompts are

[‘walking in Times Square’, ‘sunset’, ‘van gogh style’, ‘oil painting’, ‘a garden with tulips’, ‘northern lights’]. Overall, we have 90 prompts in this experiment.

10.4.2 Analysis and Comparisons

More qualitative results can be found in the appendix. The quantitative comparisons are in Figure 10.1. Our quantitative results are averaged across the metrics for each of the four sets, with 680, 355, 1050, 450 images respectively. A detailed analysis of our method and benefits over prior art is as follows:

- **Vanilla stable diffusion (SD):** Vanilla stable diffusion v2.1 can create plausible combinations of the provided text prompts. However, due to hallucination issues, it is constrained by the distributions it learned during training. As a result, it struggles to capture clear characteristics of individual text prompts. For example, in the figure, the distinct features of the parrot, cat, dog/penguin, and sunset/penguin are missing. Similarly, the avocado/raccoon, muffin, parrot, and oil painting style lack distinct characteristics. Despite this, the generated images align with the broad text prompt descriptions, resulting in a considerable CLIP score.
- **Linear interpolation [355]:** Linear interpolation is ill-suited for generating an image that combines two text prompts. This limitation arises from the non-linear nature of the text-image space. While linear interpolation assumes a simple linear relationship, the actual mapping between textual descriptions and image features is intricate. Furthermore, linear interpolation can only traverse a straight path between two points in the latent space, failing to capture nuanced variations or produce novel features beyond the endpoints.

- Alternating Sampling [356],: Alternating sampling generates features related to both text prompts in the final output image. However, the objects produced suffer from poor definition, unrealistic appearance, and low quality in many instances. Artifacts are prevalent, and several generated images appear implausible. The underlying issue lies in the algorithm's routine alternation between the two text prompts during diffusion denoising, without adequately emphasizing the most relevant prompt at each step. Despite these limitations, the model achieves a reasonable CLIP Score by broadly aligning with the overall image description, akin to vanilla denoising.
- Step-wise switching [356]: Step-wise switching is similarly ineffective for the same reasons as Alternating sampling.
- CLIP-min: The CLIP-min results exhibit bias toward one of the text prompts in most cases, preventing the model from generating an image that aligns with the intersection of both prompts. This bias arises because the model overlooks the dynamics of diffusion denoising and instead selects the text prompt with the lowest CLIP score at each step, resulting in bias-related issues.
- Black Scholes algorithm: Our approach based on the Black-Scholes algorithm effectively generates realistic images that align with the intersection of the two text prompts. These images exhibit minimal unrealistic artifacts. Additionally, our algorithm successfully preserves individual characteristics corresponding to each text prompt. By modeling the dynamics of diffusion denoising, the algorithm strategically selects the optimal text prompt at each step, achieving a balanced synthesis of both prompts. Notably, our model attains superior quantitative results.

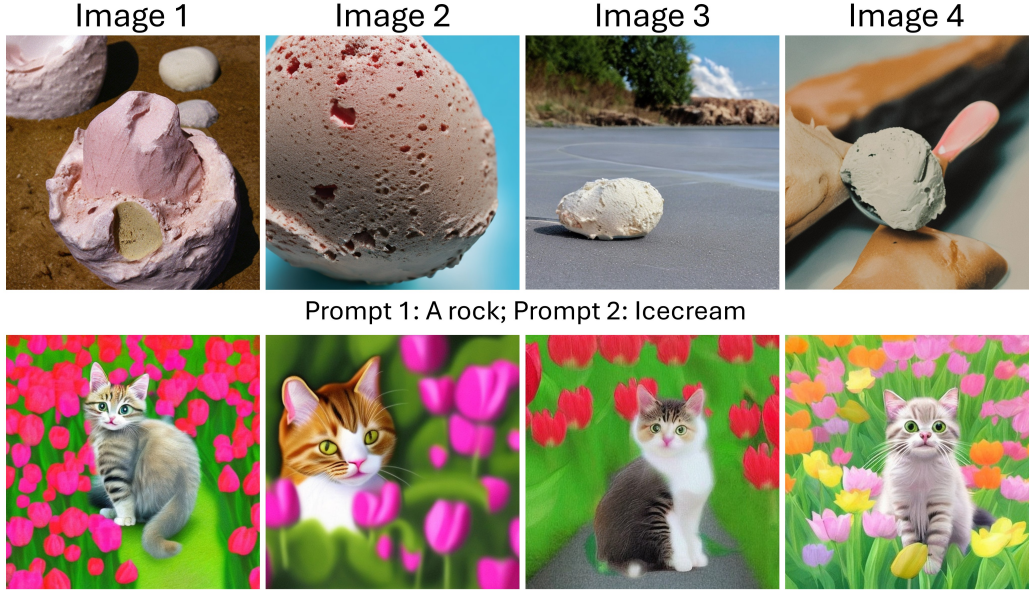


Figure 10.2: Image variations from our method using the Black Scholes model for a single prompt, starting from different random Gaussian noise initializations.

In summary, the Black-Scholes model outperforms all previous methods for prompt mixing by generating realistic images that meticulously preserve and balance the characteristics of each individual text prompt.

10.5 Conclusions, Limitations, and Future Work

In this chapter, we introduce a novel perspective on prompt mixing, inspired by a probabilistic prediction model and its theoretical framework from the domain of Finance. We draw parallels between the Black-Scholes algorithm used in financial markets and diffusion models. From this comparison, we derive an algorithm for prompt mixing, aiming to generate images that align with multiple text prompts. Our extensive analysis showcase both qualitative and quantitative advantages of our approach. Our study has a few limitations, which opens avenues for future work: (i) We follow prior work [3, 4] by using CLIP scores for quantitative evaluation. However, CLIP scores may not be the optimal evaluation strategy, as they can yield similar scores for

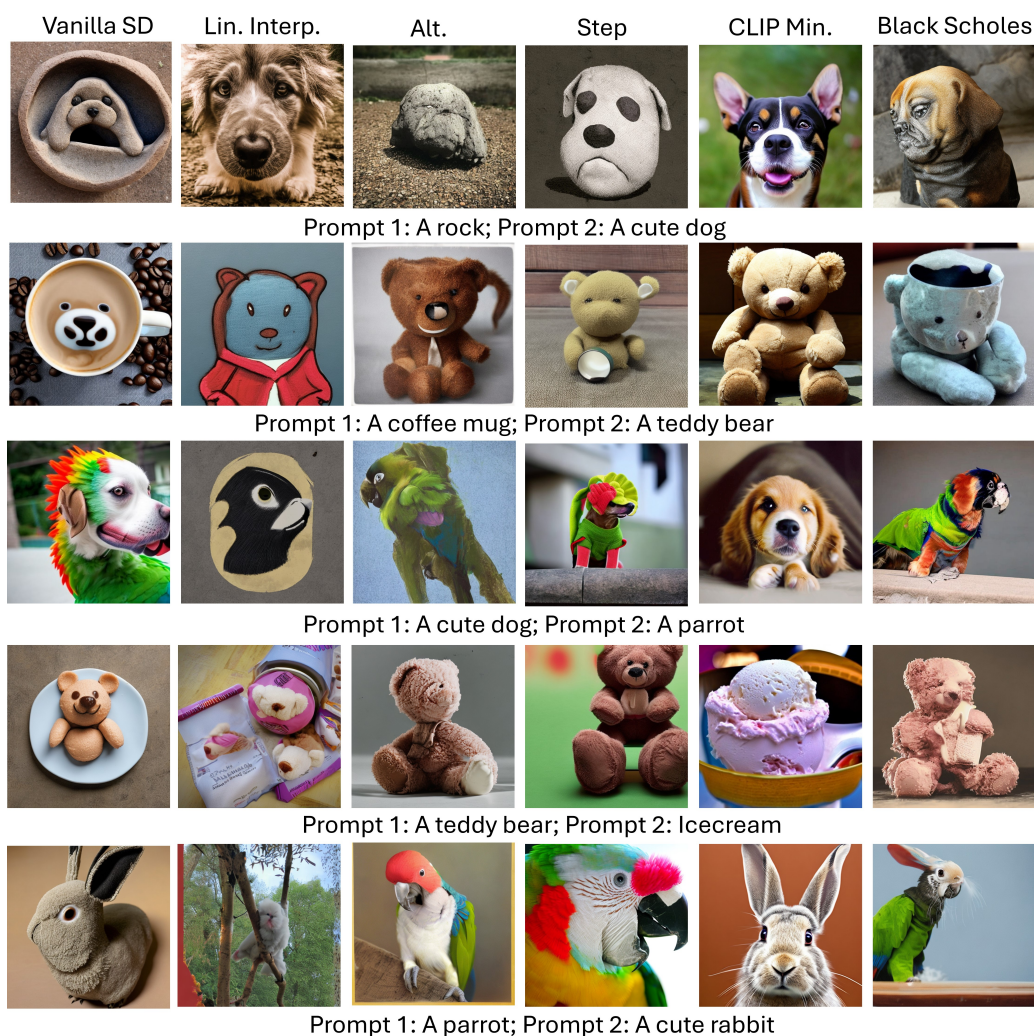


Figure 10.3: Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, it misses out on the characteristics of the dog, parrot and ice-cream. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min. generates images biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. For instance, in order, the issues are: missing characteristics of dog, missing characteristics of coffee mug and artifacts in teddy bear, artifacts and missing characteristics of parrot, missing characteristics of teddy bear and artifacts in teddy bear, missing characteristics of rabbit. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 1.

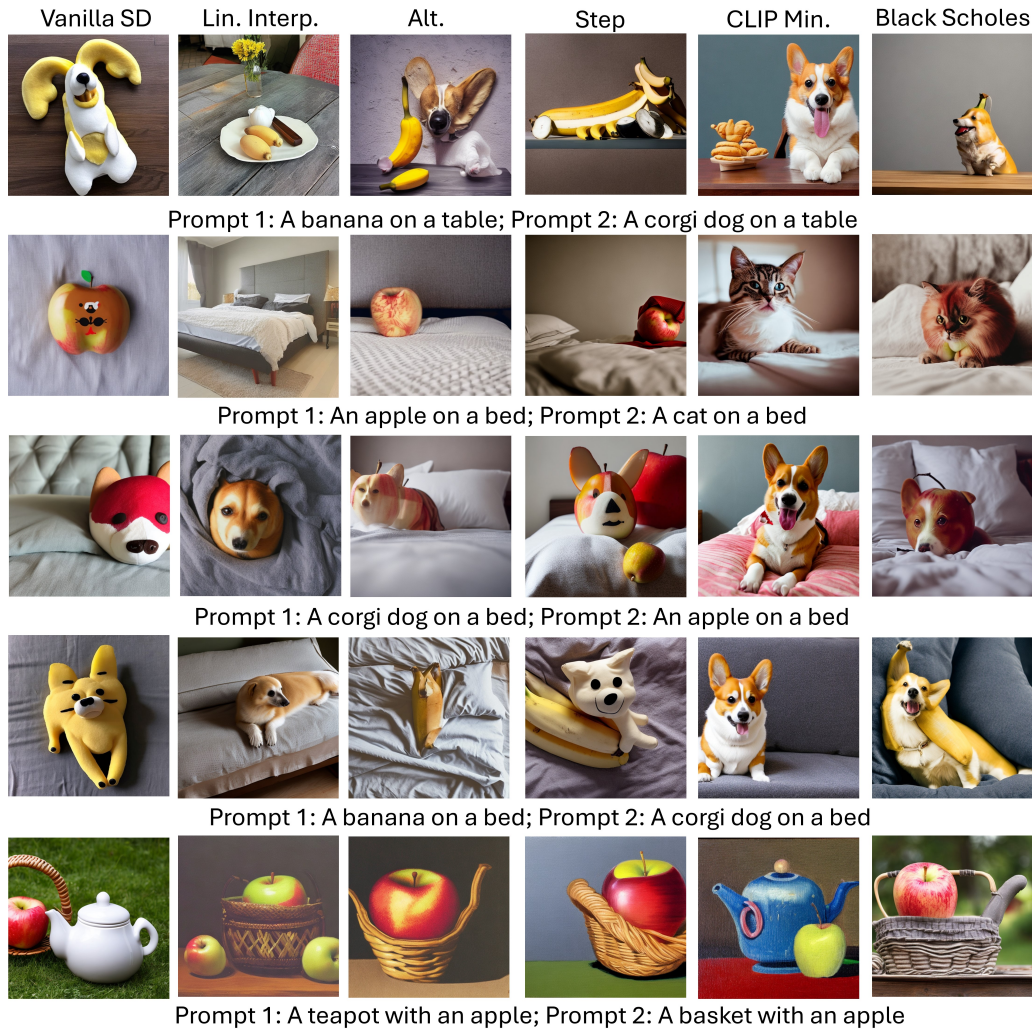


Figure 10.4: Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, it misses out on the characteristics of the dog, cat, corgi dog and basket/ teapot mixing. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min. is biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. For instance, in order, the issues are: artifacts in dog/banana and missing characteristics of dog, missing characteristics of cat, the corgi dog in the dog/apple image does not describe the dog as well as our Black Scholes method does, the dog is not well described in the fourth example, the teapot is not well described in the final image. Our method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 2.

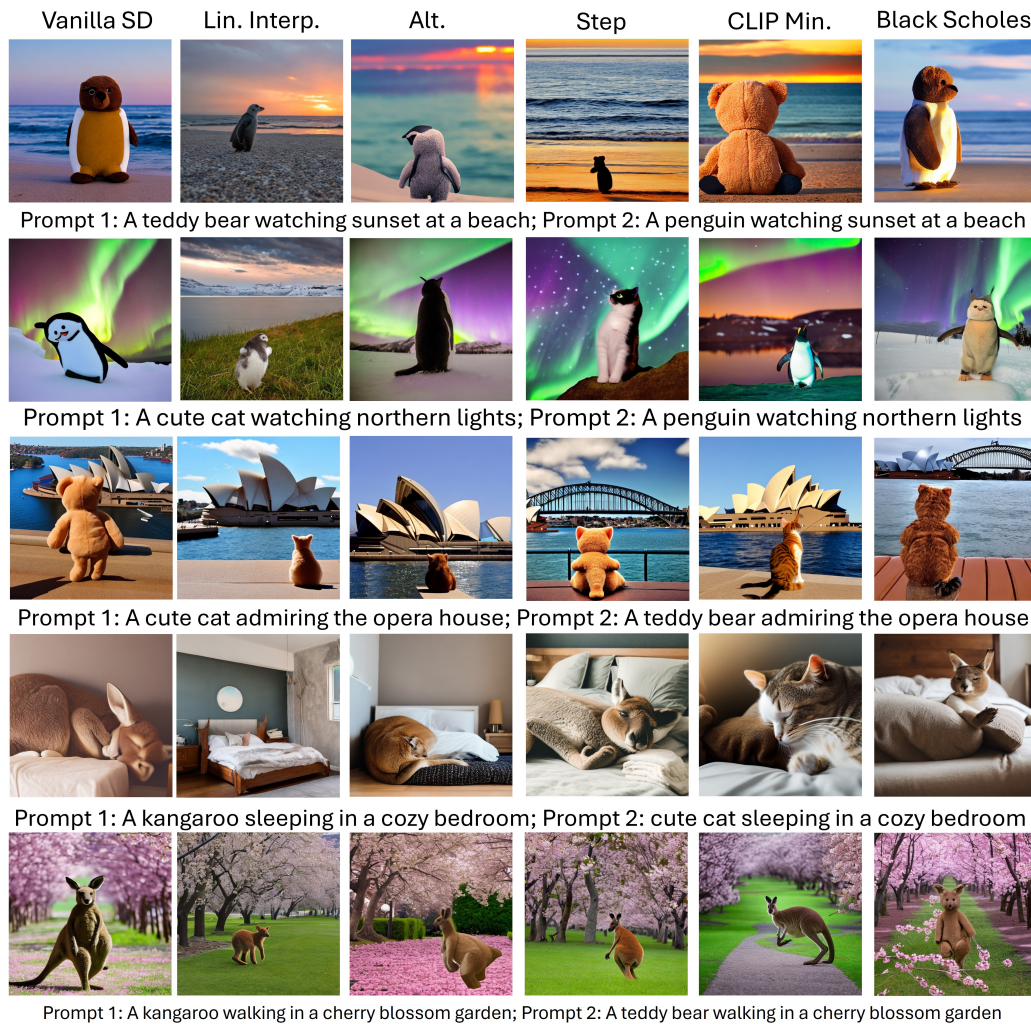


Figure 10.5: Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. For instance, in the first and fifth image, the characteristics of the teddy bear are missing, the second, third and fourth image do not describe the cat well, Linear interpolation generates images that are not consistent with the text prompts. CLIP Min is again biased towards one of the text prompts. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. Specifically, the first image does not describe the subjects well, the second and third images are reasonably generated but our Black Scholes method generates a clearer image of the penguin blended cat, and teddy bear blended cat, the results of Alt. for the fourth case has artifacts and Step does not describe the cat well, the teddy bear is not well described in the fifth image. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 3.

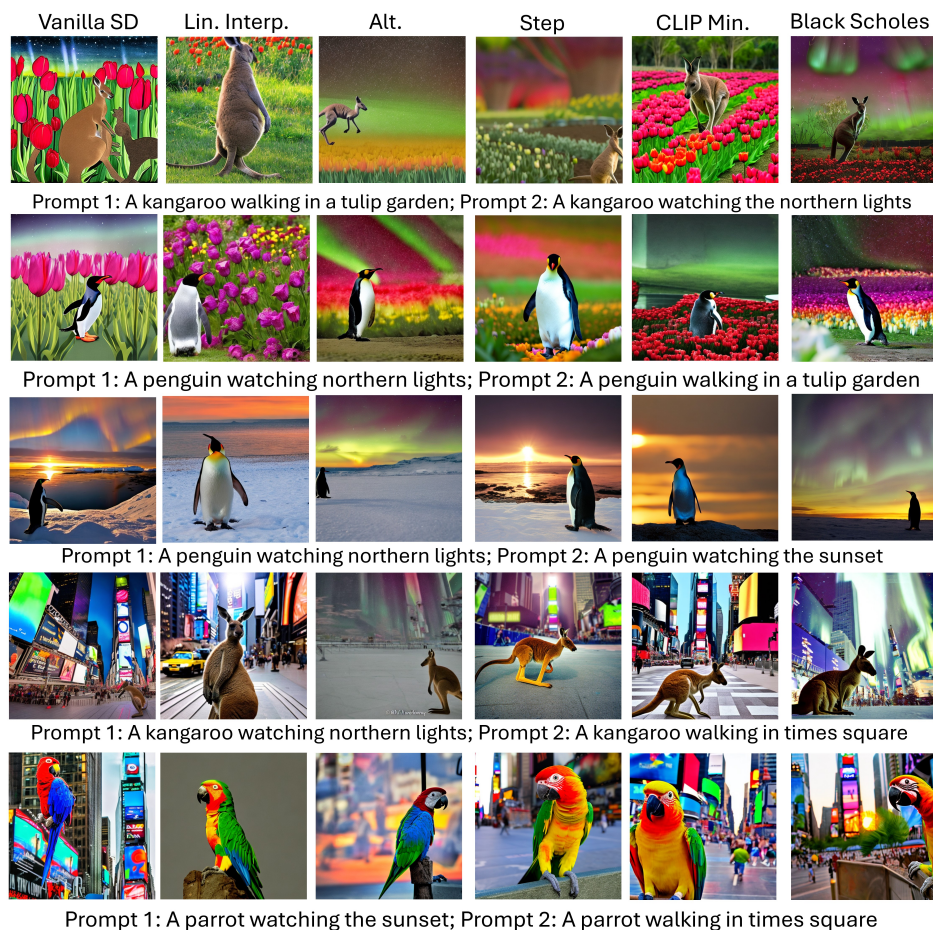


Figure 10.6: Vanilla stable diffusion (SD) is able to generate images satisfying both text prompts in many cases, however it misses out on the fine-grained characteristics of the individual text prompts. The first two images are cartoon-ish, and Northern lights is not clearly depicted. The third image also misses a clear description of Northern lights. In the fourth image, some of the billboards in Times Square are green, but there is no sign of Northern lights. In the fifth image, the sunset is missing. Linear interpolation generates images that are not consistent with the text prompts. CLIP Min results are biased towards one of the prompts for all images, except the second one. Alternating sampling and Step wise inference strategies generate images with a lot of artifacts, and the generated images are perceptually implausible in many cases. In the first image, the kangaroo is not generated well (almost flying in Alt's result, and the characteristics of the kangaroo are not well generated by Step). In the second image, the characteristics of the tulip garden are not well represented. In the third image, notice how our model generates the characteristics of the sunset and Northern lights better, while ensuring that they are blended well to form a realistic image. In the fourth image, Alt and Step miss out on the characteristics of Times Square and Northern lights respectively. In the fifth image, Alt and Step miss out on the characteristics of Times Square and the sunset respectively. Our Black Scholes method is able to generate images that capture the fine-grained characteristics of objects corresponding to both text prompts, and the generated images look realistic with minimal artifacts. The text prompts are from set 4.

images with significant quality differences. Due to its learned nature, CLIP is vulnerable to hallucination and bias issues related to its training data. Consequently, it may assign similar scores to images that exhibit general characteristics of the text prompt, without focusing on specific details and features of the scene entities described in the text prompt. To address this, future research could explore more advanced image-language models [497] for evaluation; the development of good metrics for the evaluation of generative models is a widespread issue, and could help better quantify the performance of our method. (ii) Our study focuses on prompt mixing with two prompts. Investigating the impact of using more than two prompts could be a promising direction for future work. (iii) While our method leverages diffusion models, it may not be applicable to non-Gaussian [498] diffusion models or one-step diffusion models [499], as demonstrated in recent research.

Moreover, future work on integrating our approach with recent advancements in attention guidance, layout modeling, and related areas could extend its applicability to critical downstream tasks such as image editing [355, 389, 417], compositionality [4, 449], handling complex prompts [476], text-based view synthesis [14, 356] and personalized image generation [28, 388]. While recognizing the success of recent text-to-image models like Dalle-3, Stable Diffusion 3, Firefly 3, and Imagen 3 in generating images from diverse and complex text prompts, we introduce a data-efficient approach to prompt mixing. Our method offers advantages over previous techniques for prompt-mixing, and future extension could explore integrating our model’s benefits into state-of-the-art text-to-image architectures for data efficient prompt mixing and downstream applications.

Part VI

Conclusions and Future Directions

In the earlier chapters, we explored learning from limited data. The central concept in our dissertation involves harnessing existing data from related domains, pretrained models, and classical computer vision techniques to improve data efficiency across perception and synthesis tasks. These ideas are summarized below:

- Domain adaptation for autonomous driving: Leveraging related domain data for semantic segmentation in unstructured traffic environments (Chapter 1), Utilizing pretrained models and minimal data for road segmentation in hazardous weather (Chapter 2), Proposing a generic approach that combines pretrained models with intelligent sample selection in unlabeled domains (Chapter 3).
- Video action recognition for autonomous aerial vehicles: Employing Fast Fourier Transform (Chapter 4) and Differentiability (Chapter 5) techniques to learn effective video representations with limited data compared to traditional front-view images.
- Aerial view synthesis: Using homography and diffusion sampling principles (Chapter 6) and mutual information-based techniques (Chapter 7) to synthesize aerial views from a single exemplar front-view image using pretrained models.
- Controllable image and video synthesis: Employing frequency guidance (Chapter 8) and text prompting strategies (Chapter 9) to generate customized multi-concept images and videos from pretrained models and single exemplar images.
- Text prompt mixing for image synthesis: Leveraging the connection between image diffusion models and financial models (Chapter 10) to create images at the intersection of different text prompts using pretrained diffusion models without additional training data.

Our research focuses on developing data-efficient state-of-the-art methods. Throughout this dissertation, we’ve advanced the field in computer vision-based perception and synthesis tasks. Specifically, the applications of my work are as follows:

- Domain adaptation and aerial video recognition benefit autonomous ground and aerial vehicles.
- Synthesizing aerial views from front images aids autonomous aerial vehicles, cross-view mapping/navigation, and AR/VR.
- Controllable image and video synthesis, along with prompt mixing, has applications in movie creation, animation, and graphics.

Importantly, the fundamental techniques that we have proposed can also extend beyond the specific contexts presented in this dissertation. While our research makes encouraging progress toward data efficiency, there remain unexplored directions. While not exhaustive, there are several intriguing directions that can be explored:

- Data efficient and domain-Generalizable Multimodal Foundational Models: Current foundational models across text, image, audio, and multimodal domains are heavily reliant on extensive datasets. For instance, consider generating a video of two pink cats dancing in a garden under the northern lights, all from a simple text prompt. Today’s image generation models can accomplish this due to the vast amounts of data they’ve been trained on. But what if we want to extend this capability to video or audio? Collecting data at such a scale is impractical, which has led to the exploration of data-efficient multimodal foundational model techniques. Additionally, these models often struggle with generalizability beyond

their training domains. Therefore, researching domain-generalizable, data-efficient multimodal models presents an exciting opportunity for future exploration. Some directions include using classical techniques, physics priors, pretrained models, data from related domains, etc.

- **Real-World Applications:** Embodied AI, robot and autonomous driving applications operate in complex real-world environments. Gathering data for every conceivable scenario is not practical, and scaling up to accommodate multiple modalities—like various sensors and LiDAR—is simply unfeasible. Addressing this challenge represents another critical area of exploration.
- **Sustainable AI and social good:** Can we train models that conserve GPU hours, mitigate global warming, and promote sustainable deep learning? Data-efficient techniques play a crucial role in social good applications, such as satellite imagery analysis, wildlife conservation, and medical imaging, where data collection is challenging.

My long-standing research goal is to develop intelligent, data-efficient solutions that address impactful real-world problems. My dissertation represents a positive step forward, and I remain committed to advancing this goal in the future.

Bibliography

- [1] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *ICCV*, 2019.
- [2] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 936–937, 2020.
- [3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.
- [4] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023.
- [5] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7374–7383, 2019.
- [6] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018.
- [7] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.
- [8] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

- [9] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.
- [10] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.
- [11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.
- [12] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.
- [13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [14] Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. Aerialbooth: Mutual information guidance for text controlled aerial view synthesis from a single image. *arXiv preprint arXiv:2311.15478*, 2023.
- [15] Divya Kothandaraman, Rohan Chandra, and Dinesh Manocha. Ss-sfda: Self-supervised source-free domain adaptation for road segmentation in hazardous environments. *arXiv preprint arXiv:2012.08939*, 2020.
- [16] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.
- [17] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.
- [18] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017.
- [19] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 2018.
- [20] Alexandra S Mueller and Lana M Trick. Driving in fog: The effects of driving experience and visibility on speed compensation and hazard avoidance. *Accident Analysis & Prevention*, 48:472–479, 2012.

- [21] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [22] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16266–16275, 2021.
- [23] Asanka G Perera, Yee Wei Law, and Javaan Chahl. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones*, 3(4):82, 2019.
- [24] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [25] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.
- [26] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788, 2019.
- [27] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019.
- [28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
- [29] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. In *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*, pages 3–21. World Scientific, 2019.
- [30] Yilei Sun Brenda Goh. Tesla 'very close' to level 5 autonomous driving technology. link, 2020.
- [31] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4), 2017.
- [32] Markus Hofmarcher, Thomas Unterthiner, José Arjona-Medina, Günter Klambauer, Sepp Hochreiter, and Bernhard Nessler. Visual scene understanding for autonomous driving using semantic segmentation. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 285–296. Springer, 2019.

- [33] Jianxiong Xiao, Bryan C Russell, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Basic level scene understanding: From labels to structure and beyond. In *SIGGRAPH Asia 2012 Technical Briefs*, pages 1–4. 2012.
- [34] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020.
- [35] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [36] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*. IEEE, 2019.
- [37] Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928), 2010.
- [38] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.
- [39] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*. Springer, 2016.
- [40] Magnus Wrenninge, , , and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv*, 2018.
- [41] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Ttraffic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8483–8492, 2019.
- [42] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *NIPS*, 2019.
- [43] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406:302–321, 2020.
- [44] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238, 2019.
- [45] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

- [46] Zhiyang Guo, Yingping Huang, Xing Hu, Hongjian Wei, and Baigan Zhao. A survey on deep learning based approaches for scene understanding in autonomous driving. *Electronics*, 10(4):471, 2021.
- [47] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *CVPR*, 2020.
- [48] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020.
- [49] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.
- [50] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuster, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.
- [51] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O’Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *CVPR*, 2020.
- [52] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *arXiv*, 2019.
- [53] Yixin Zhang and Zilei Wang. Joint adversarial learning for domain adaptation in semantic segmentation. In *AAAI*, 2020.
- [54] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020.
- [55] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, 2020.
- [56] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *AAAI*, 2020.
- [57] Taekyung Kim and Changick Kim. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. *arXiv*, 2020.
- [58] Da Li and Timothy Hospedales. Online meta-learning for multi-source and semi-supervised domain adaptation. *arXiv*, 2020.
- [59] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 2019.

- [60] Can Qin, Lichen Wang, Qianqian Ma, Yu Yin, Huan Wang, and Yun Fu. Opposite structure learning for semi-supervised domain adaptation. *arXiv*, 2020.
- [61] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Buda: Boundless unsupervised domain adaptation in semantic segmentation. *arXiv*, 2020.
- [62] Divya Kothandaraman, Athira Nambiar, and Anurag Mittal. Domain adaptive knowledge distillation for driving scene semantic segmentation. In *Proceedings of the IEEE/CVF WACV*, pages 134–143.
- [63] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *arXiv preprint arXiv:2005.10876*, 2020.
- [64] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, 2017.
- [65] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018.
- [66] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *CVPR*, 2019.
- [67] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*, 2020.
- [68] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. In *AAAI*, 2020.
- [69] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. *arXiv preprint arXiv:1911.11554*, 2019.
- [70] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. *arXiv preprint arXiv:2007.08801*, 2020.
- [71] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. *arXiv preprint arXiv:2007.01261*, 2020.
- [72] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *NIPS*, 2018.
- [73] Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.
- [74] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6), 1996.

- [75] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [76] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2019.
- [77] Maxime Bucher, VU Tuan-Hung, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NIPS*, 2019.
- [78] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [79] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *ICCV*, 2019.
- [80] Yuhua Chen, Wen Li, , and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018.
- [81] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. 2020.
- [82] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *ICCV*, 2019.
- [83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [84] Dirty Tesla. Tesla autopilot works on dirt roads! <https://www.youtube.com/watch?v=S7eVqQsnj4U>, 2019.
- [85] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.
- [86] Rohan Chandra, Tianrui Guan, Srujan Panuganti, Trisha Mittal, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. *IEEE Robotics and Automation Letters*, 5(3): 4882–4890, 2020.
- [87] Rohan Chandra, Uttaran Bhattacharya, Christian Roncal, Aniket Bera, and Dinesh Manocha. Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In *ACM Computer Science in Cars Symposium*, pages 1–9, 2019.
- [88] Rohan Chandra, Uttaran Bhattacharya, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Roadtrack: Realtime tracking of road agents in dense and heterogeneous environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1270–1277. IEEE, 2020.

- [89] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [90] Michael A Cusumano. Self-driving vehicle technology: progress and promises. *Communications of the ACM*, 63(10):20–22, 2020.
- [91] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *European Conference on Computer Vision*, pages 340–356. Springer, 2020.
- [92] Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko. Reverse and boundary attention network for road segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [93] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.
- [94] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*. PMLR, 2018.
- [95] Youngeun Kim, Sungeun Hong, Donghyeon Cho, Hyoungeob Park, and Priyadarshini Panda. Domain adaptation without source data. *arXiv preprint arXiv:2007.01524*, 2020.
- [96] Arun Reddy Nelakurthi, Ross Maciejewski, and Jingrui He. Source free domain adaptation using an off-the-shelf classifier. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 140–145. IEEE, 2018.
- [97] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.
- [98] Yunzhong Hou and Liang Zheng. Source free domain adaptation with image translation. *arXiv preprint arXiv:2008.07514*, 2020.
- [99] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 615–625, 2021.
- [100] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021.
- [101] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

- [102] Qi Wang, Junyu Gao, and Yuan Yuan. Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):230–241, 2017.
- [103] Farnoush Zohourian, Borislav Antic, Jan Siegemund, Mirko Meuter, and Josef Pauli. Superpixel-based road segmentation for real-time systems using cnn. In *VISIGRAPP (5: VISAPP)*, pages 257–265, 2018.
- [104] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.
- [105] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv preprint arXiv:1711.06969*, 2:2, 2017.
- [106] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7364–7373, 2019.
- [107] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017.
- [108] Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7087–7093. IEEE, 2019.
- [109] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128(5):1182–1204, 2020.
- [110] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.
- [111] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2990–2998, 2020.
- [112] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.
- [113] Shiqi Yang, Yaxing Wang, Joost van de Weijer, and Luis Herranz. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 2020.
- [114] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

- [115] Masato Ishii and Masashi Sugiyama. Source-free domain adaptation via distributional alignment by matching batch normalization statistics. *arXiv preprint arXiv:2101.10842*, 2021.
- [116] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018.
- [117] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Amrith Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.
- [118] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Context-reinforced semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4046–4055, 2019.
- [119] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–499. Springer, 2020.
- [120] Xinzhe Li, Qianru Sun, Yaoyao Liu, Shibao Zheng, Qin Zhou, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *arXiv preprint arXiv:1906.00562*, 2019.
- [121] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001*, 2020.
- [122] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *European Conference on Computer Vision*, pages 532–548. Springer, 2020.
- [123] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.
- [124] Jaehoon Choi, Minki Jeong, Taekyung Kim, and Changick Kim. Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv preprint arXiv:1908.00262*, 2019.
- [125] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3130–3139, 2020.

- [126] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.
- [127] Divya Kothandaraman, Rohan Chandra, and Dinesh Manocha. Bomuda: Boundless multi-source domain adaptive segmentation in unconstrained environments. *arXiv preprint arXiv:2010.03523*, 2020.
- [128] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005.
- [129] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017.
- [130] Frederick Tung, Jianhui Chen, Lili Meng, and James J Little. The raincouver scene parsing benchmark for self-driving in adverse weather and at night. *IEEE Robotics and Automation Letters*, 2(4):2188–2193, 2017.
- [131] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- [132] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2019.
- [133] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *arXiv preprint arXiv:2101.11605*, 2021.
- [134] Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I Morariu, Varun Manjunatha, Tong Sun, and Yun Fu. Cross-domain document object detection: Benchmark suite and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12915–12924, 2020.
- [135] Diede Rusticus, Lutz Goldmann, Matthias Reisser, and Mauricio Villegas. Document domain adaptation with generative adversarial networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1432–1437. IEEE, 2019.
- [136] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.
- [137] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. *arXiv preprint arXiv:2010.08666*, 2020.
- [138] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

- [139] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.
- [140] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021.
- [141] Victor Bouvier, Philippe Very, Clément Chastagnol, Myriam Tami, and Céline Hudelot. Stochastic adversarial gradient embedding for active domain adaptation. *arXiv preprint arXiv:2012.01843*, 2020.
- [142] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.
- [143] Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, and Luc Van Gool. Unsupervised robust domain adaptation without source data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2009–2018, 2022.
- [144] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [145] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.
- [146] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006.
- [147] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.
- [148] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.
- [149] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.
- [150] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.
- [151] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

- [152] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [153] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.
- [154] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29:343–351, 2016.
- [155] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pages 597–613. Springer, 2016.
- [156] Xudong Mao and Qing Li. Unpaired multi-domain image generation via regularized conditional gans. *arXiv preprint arXiv:1805.02456*, 2018.
- [157] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.
- [158] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Buló. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5067–5075, 2017.
- [159] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.
- [160] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.
- [161] Fan Zhou, Changjian Shui, Shichun Yang, Bincheng Huang, Boyu Wang, and Brahim Chaib-draa. Discriminative active learning for domain adaptation. *Knowledge-Based Systems*, 222:106986, 2021.
- [162] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021.
- [163] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.
- [164] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

- [165] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.
- [166] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [167] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.
- [168] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [169] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
- [170] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. S4t: Source-free domain adaptation for semantic segmentation via self-supervised selective self-training. *arXiv preprint arXiv:2107.10140*, 2021.
- [171] Marvin Klingner, Jan-Aike Termöhlen, Jacob Ritterbach, and Tim Fingscheidt. Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 210–220, 2022.
- [172] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021.
- [173] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [174] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.
- [175] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.
- [176] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can’t i dance in the mall? learning to mitigate scene bias in action recognition. *arXiv preprint arXiv:1912.05534*, 2019.

- [177] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.
- [178] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.
- [179] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.
- [180] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.
- [181] Bilel Benjdira, Yakoub Bazi, Anis Koubaa, and Kais Ouni. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369, 2019.
- [182] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995.
- [183] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9945–9953, 2019.
- [184] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [185] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.
- [186] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.
- [187] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.
- [188] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. ActionflowNet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018.

- [189] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004.
- [190] Luca Zappella, Xavier Lladó, and Joaquim Salvi. Motion segmentation: A review. *Artificial Intelligence Research and Development*, pages 398–407, 2008.
- [191] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020.
- [192] Marc Ellenfeld, Sebastian Moosbauer, Ruben Cardenes, Ulrich Klauck, and Michael Teutsch. Deep fusion of appearance and frame differencing for motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4339–4349, 2021.
- [193] Yuehua Zhu, Cheng Deng, Huanhuan Cao, and Hao Wang. Object and background disentanglement for unsupervised cross-domain person re-identification. *Neurocomputing*, 403:88–97, 2020.
- [194] Aysegul Dundar, Kevin J Shih, Animesh Garg, Robert Pottorf, Andrew Tao, and Bryan Catanzaro. Unsupervised disentanglement of pose, appearance and background from images and videos. *arXiv preprint arXiv:2001.09518*, 2020.
- [195] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.
- [196] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.
- [197] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020.
- [198] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.
- [199] Jinwoo Choi. Action recognition list of papers. In <https://github.com/jinwchoi/awesome-action-recognition>.
- [200] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

- [201] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.
- [202] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Two stream lstm: A deep fusion framework for human action recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 177–186. IEEE, 2017.
- [203] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.
- [204] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.
- [205] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.
- [206] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.
- [207] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.
- [208] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. *arXiv preprint arXiv:2012.06399*, 2020.
- [209] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.
- [210] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time human action recognition. *arXiv preprint arXiv:2107.00606*, 2021.
- [211] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
- [212] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. *arXiv preprint arXiv:2012.10671*, 2020.
- [213] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. *arXiv preprint arXiv:2104.09952*, 2021.

- [214] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019.
- [215] Brent A Griffin and Jason J Corso. Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8914–8923, 2019.
- [216] Mohammadamin Barekatin, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 28–35, 2017.
- [217] Waqas Sultani and Mubarak Shah. Human action recognition in drone videos using a few aerial training examples. *Computer Vision and Image Understanding*, 206:103186, 2021.
- [218] Han Peng and Abolfazl Razi. Fully autonomous uav-based action recognition system using aerial imagery. In *International Symposium on Visual Computing*, pages 276–290. Springer, 2020.
- [219] Meng Ding, Ning Li, Ziang Song, Ruixing Zhang, Xiaxia Zhang, and Huiyu Zhou. A lightweight action recognition method for unmanned-aerial-vehicle video. In *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, pages 181–185. IEEE, 2020.
- [220] Anwaar Ulhaq, Xiaoxia Yin, Yunchan Zhang, and Iqbal Gondal. Action-02mcf: A robust space-time correlation filter for action recognition in clutter and adverse lighting conditions. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 465–476. Springer, 2016.
- [221] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.
- [222] Payal Mittal, Raman Singh, and Akashdeep Sharma. Deep learning-based object detection in low-altitude uav datasets: A survey. *Image and Vision computing*, 104:104046, 2020.
- [223] H Buijs, A Pomerleau, M Fournier, and W Tam. Implementation of a fast fourier transform (fft) for image processing applications. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(6):420–424, 1974.
- [224] B Srinivasa Reddy and Biswanath N Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8):1266–1271, 1996.
- [225] Zhong Zhang, Jianhui Zhao, Dengyi Zhang, Chengzhang Qu, Youwang Ke, and Bo Cai. Contour based forest fire detection using fft and wavelet. In *2008 International conference on computer science and software engineering*, volume 1, pages 760–763. IEEE, 2008.

- [226] Byung Tae Chun, Younglae Bae, and Tai-Yun Kim. Automatic text extraction in digital videos using fft and neural network. In *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315)*, volume 2, pages 1112–1115. IEEE, 1999.
- [227] A. Lavin and S. Gray. Fast algorithms for convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4013–4021, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.435. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.435>.
- [228] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4479–4488. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/2fd5d41ec6cfab47e32164d5624269b1-Paper.pdf>.
- [229] K. Xu, M. Qin, F. Sun, Y. Wang, Y. Chen, and F. Ren. Learning in the frequency domain. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1737–1746, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00181. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00181>.
- [230] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf>.
- [231] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4084–4094, 2020.
- [232] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539, 2021.
- [233] Sinong et al Wang. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020.
- [234] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.
- [235] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nystöm-based algorithm for approximating

- self-attention. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 35, page 14138. NIH Public Access, 2021.
- [236] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pages 9355–9366. PMLR, 2021.
 - [237] Young Jin Kim and Hany Hassan Awadalla. Fastformers: Highly efficient transformer models for natural language understanding. *arXiv preprint arXiv:2010.13382*, 2020.
 - [238] Rui Li, Jianlin Su, Chenxi Duan, and Shunyi Zheng. Linear attention mechanism: An efficient attention for semantic segmentation. *arXiv preprint arXiv:2007.14902*, 2020.
 - [239] Feng Shi, Chonghan Lee, Liang Qiu, Yizhou Zhao, Tianyi Shen, Shivran Muralidhar, Tian Han, Song-Chun Zhu, and Vijaykrishnan Narayanan. Star: Sparse transformer-based action recognition. *arXiv preprint arXiv:2107.07089*, 2021.
 - [240] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1381–1384. IEEE, 1998.
 - [241] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms, 2021.
 - [242] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.
 - [243] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.
 - [244] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.
 - [245] D Brandon Lloyd, Naga K Govindaraju, Cory Quammen, Steven E Molnar, and Dinesh Manocha. Logarithmic perspective shadow maps. *ACM Transactions on Graphics (TOG)*, 27(4):1–32, 2008.
 - [246] Don P Mitchell and Arun N Netravali. Reconstruction filters in computer-graphics. *ACM Siggraph Computer Graphics*, 22(4):221–228, 1988.
 - [247] Itsaso Rodríguez-Moreno, José María Martínez-Otzeta, Basilio Sierra, Igor Rodriguez, and Ekaitz Jauregi. Video activity recognition: State-of-the-art. *Sensors*, 19(14):3160, 2019.
 - [248] Yiling Qiao, Junbang Liang, Vladlen Koltun, and Ming Lin. Differentiable simulation of soft multi-body systems. *Advances in Neural Information Processing Systems*, 34:17123–17135, 2021.

- [249] Tetsuya Takahashi, Junbang Liang, Yi-Ling Qiao, and Ming C Lin. Differentiable fluids with solid coupling for learning and control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6138–6146, 2021.
- [250] Divya Kothandaraman, Tianrui Guan, Xijun Wang, Sean Hu, Ming Lin, and Dinesh Manocha. Fourier disentangled space-time attention for aerial video recognition. *arXiv preprint arXiv:2203.10694*, 2022.
- [251] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Qinqin Nie, Hao Cheng, Chenfeng Liu, Xiaoyu Liu, et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.
- [252] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020.
- [253] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Yingli Tian, Xiaoming Liu, Feng Liu, and Arun Ross. The state of aerial surveillance: A survey. *arXiv preprint arXiv:2201.03080*, 2022.
- [254] Matthew C Fysh and Markus Bindemann. Person identification from drones by humans: insights from cognitive psychology. *Drones*, 2(4):32, 2018.
- [255] Jan Portmann, Simon Lynen, Margarita Chli, and Roland Siegwart. People detection and tracking from aerial thermal views. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 1794–1800. IEEE, 2014.
- [256] Yujie He, Changhong Fu, Fuling Lin, Yiming Li, and Peng Lu. Towards robust visual tracking for unmanned aerial vehicle with tri-attentional correlation filters. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1575–1582. IEEE, 2020.
- [257] Matthias Mueller, Gopal Sharma, Neil Smith, and Bernard Ghanem. Persistent aerial tracking system for uavs. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1562–1569. IEEE, 2016.
- [258] Alessandro Benini, Matthew J Rutherford, and Kimon P Valavanis. Real-time, gpu-based pose estimation of a uav for autonomous takeoff and landing. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3463–3470. IEEE, 2016.
- [259] Muhammad Waseem Ashraf, Waqas Sultani, and Mubarak Shah. Dogfight: Detecting drones from drones videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7067–7076, 2021.
- [260] Samira Hayat, Evşen Yanmaz, Timothy X Brown, and Christian Bettstetter. Multi-objective uav path planning for search and rescue. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 5569–5574. IEEE, 2017.

- [261] Ketan Kotecha, Deepak Garg, Balmukund Mishra, Pratik Narang, and Vipul Kumar Mishra. Background invariant faster motion modeling for drone action recognition. *Drones*, 5(3):87, 2021.
- [262] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018.
- [263] Jonas Degraeve, Michiel Hermans, Joni Dambre, et al. A differentiable physics engine for deep learning in robotics. *Frontiers in neurorobotics*, page 6, 2019.
- [264] Yucheng Zhang, Tian Wang, Kexin Liu, Baochang Zhang, and Lei Chen. Recent advances of single-object tracking methods: A brief survey. *Neurocomputing*, 455:1–11, 2021.
- [265] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019.
- [266] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6222–6231, 2019.
- [267] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1451–1459, 2021.
- [268] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.
- [269] Rakesh Kumar, Harpreet Sawhney, Supun Samarasekera, Steve Hsu, Hai Tao, Yanlin Guo, Keith Hanna, Art Pope, Richard Wildes, David Hirvonen, et al. Aerial video surveillance and exploitation. *Proceedings of the IEEE*, 89(10):1518–1539, 2001.
- [270] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022.
- [271] Antonella Barisic, Frano Petric, and Stjepan Bogdan. Sim2air-synthetic aerial dataset for uav monitoring. *IEEE Robotics and Automation Letters*, 7(2):3757–3764, 2022.
- [272] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [273] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

- [274] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [275] Zhengyu He. Deep learning in image classification: A survey report. In *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pages 174–177. IEEE, 2020.
- [276] Divya Kothandaraman, Tianrui Guan, Xijun Wang, Shuowen Hu, Ming Lin, and Dinesh Manocha. Far: Fourier aerial video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 657–676. Springer, 2022.
- [277] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.
- [278] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020.
- [279] Hao Ding, Songsong Wu, Hao Tang, Fei Wu, Guangwei Gao, and Xiao-Yuan Jing. Cross-view image synthesis with deformable convolution and attention mechanism. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part I 3*, pages 386–397. Springer, 2020.
- [280] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021.
- [281] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- [282] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [283] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*, 2022.
- [284] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

- [285] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [286] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.
- [287] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5524–5532, 2018.
- [288] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2021.
- [289] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.
- [290] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–25, 2022.
- [291] Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, Huiyu Zhou, Ruili Wang, M Emre Celebi, and Jie Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 72:126–146, 2021.
- [292] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [293] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [294] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022.
- [295] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.
- [296] Jingwen Su, Boyan Xu, and Hujun Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022.
- [297] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.
- [298] Aziz Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705, 2020.

- [299] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.
- [300] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2022.
- [301] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.
- [302] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.
- [303] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [304] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022.
- [305] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizatwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.
- [306] Bin Ren, Hao Tang, Yiming Wang, Xia Li, Wei Wang, and Nicu Sebe. Pi-trans: Parallel-convmlp and implicit-transformation based gan for cross-view image translation. *arXiv preprint arXiv:2207.04242*, 2022.
- [307] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021.
- [308] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022.
- [309] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10009–10022, 2022.
- [310] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1843–1847. IEEE, 2020.

- [311] Bin Ren, Hao Tang, and Nicu Sebe. Cascaded cross mlp-mixer gans for cross-view image translation. *arXiv preprint arXiv:2110.10183*, 2021.
- [312] Gaowen Liu, Hugo Latapie, Ozkan Kilic, and Adam Lawrence. Parallel generative adversarial network for third-person to first-person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1917–1923, 2022.
- [313] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 2022.
- [314] Yan Shen, Meng Luo, Yun Chen, Xiaotao Shao, Zhongli Wang, Xiaoli Hao, and Ya-Li Hou. Cross-view image translation based on local and global information guidance. *IEEE Access*, 9:12955–12967, 2021.
- [315] Syed Ammar Abbas and Andrew Zisserman. A geometric approach to obtain a bird’s eye view from an image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [316] Yun Zhao, Yu Zhang, Zhan Gong, and Hong Zhu. Scene representation in bird’s-eye view from surrounding cameras with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4511–4519, 2022.
- [317] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.
- [318] Amit More and Subhasis Chaudhuri. Deep learning based novel view synthesis. *arXiv preprint arXiv:2107.06812*, 2021.
- [319] Xiaogang Xu, Ying-Cong Chen, and Jiaya Jia. View independent generative adversarial network for novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7791–7800, 2019.
- [320] Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022.
- [321] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.
- [322] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

- [323] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.
- [324] Yuxin Hou, Arno Solin, and Juho Kannala. Novel view synthesis via depth-guided skip connections. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3119–3128, 2021.
- [325] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1578–1604, 2019.
- [326] Anny Yuniarti and Nanik Suciati. A review of deep learning techniques for 3d reconstruction of 2d images. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pages 327–331. IEEE, 2019.
- [327] George Fahim, Khalid Amin, and Sameh Zarif. Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021.
- [328] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1357–1365, 2020.
- [329] Bowen Li, Xiaojuan Qi, Philip HS Torr, and Thomas Lukasiewicz. Image-to-image translation with text guidance. *arXiv preprint arXiv:2002.05235*, 2020.
- [330] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.
- [331] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.
- [332] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *arXiv preprint arXiv:2102.01187*, 2021.
- [333] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [334] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [335] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European*

- Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer, 2022.
- [336] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
 - [337] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022.
 - [338] Jihyeong Yoo and Qifeng Chen. Sinir: Efficient general image manipulation with single image reconstruction. In *International Conference on Machine Learning*, pages 12040–12050. PMLR, 2021.
 - [339] Yael Vinker, Eliahu Horwitz, Nir Zabari, and Yedid Hoshen. Deep single image manipulation. *arXiv preprint arXiv:2007.01289*, 2020.
 - [340] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.
 - [341] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.
 - [342] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.
 - [343] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
 - [344] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.
 - [345] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.
 - [346] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.
 - [347] Divya Kothandaraman, Ming Lin, and Dinesh Manocha. Diffar: Differentiable frequency-based disentanglement for aerial video action recognition. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8254–8261. IEEE, 2023.

- [348] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.
- [349] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.
- [350] René Emmaneel, Martin R Oswald, Sjoerd de Haan, and Dragos Datcu. Cross-view outdoor localization in augmented reality by fusing map and satellite data. *Applied Sciences*, 13(20):11215, 2023.
- [351] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021.
- [352] Gijs van Tulder, Yao Tong, and Elena Marchiori. Multi-view analysis of unregistered medical images using cross-view transformers. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 104–113. Springer, 2021.
- [353] Shervin Ardeshtir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 285–300, 2018.
- [354] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [355] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.
- [356] Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. Aerial diffusion: Text guided ground-to-aerial view translation from a single image using diffusion models. *arXiv preprint arXiv:2303.11444*, 2023.
- [357] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.
- [358] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

- [359] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022.
- [360] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [361] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3500–3509, 2017.
- [362] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [363] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021.
- [364] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023.
- [365] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023.
- [366] Congyue Deng, Chiyu Jiang, Charles R Qi, Xincheng Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023.
- [367] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.
- [368] Yukai Shi, Jianan Wang, He Cao, Boshi Tang, Xianbiao Qi, Tianyu Yang, Yukun Huang, Shilong Liu, Lei Zhang, and Heung-Yeung Shum. Toss: High-quality text-guided novel view synthesis from a single image. *arXiv preprint arXiv:2310.10644*, 2023.
- [369] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

- [370] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.
- [371] James Burgess, Kuan-Chieh Wang, and Serena Yeung. Viewpoint textual inversion: Unleashing novel view synthesis with pretrained 2d diffusion models. *arXiv preprint arXiv:2309.07986*, 2023.
- [372] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023.
- [373] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.
- [374] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023.
- [375] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023.
- [376] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023.
- [377] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. *arXiv preprint arXiv:2312.04551*, 2023.
- [378] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [379] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2139–2150, 2023.
- [380] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021.
- [381] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

- [382] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 653–669. Springer, 2020.
- [383] Changxing Ding and Dacheng Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017.
- [384] Jiazhen Liu and Xirong Li. Geometrized transformer for self-supervised homography estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9556–9565, 2023.
- [385] Jiaqi Gu, Bojian Wu, Lubin Fan, Jianqiang Huang, Shen Cao, Zhiyu Xiang, and Xian-Sheng Hua. Homography loss for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1080–1089, 2022.
- [386] Giacomo D’Amicantonio, Egor Bondarev, et al. Automated camera calibration via homography estimation with gnns. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5876–5883, 2024.
- [387] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023.
- [388] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.
- [389] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.
- [390] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [391] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.
- [392] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [393] Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M Patel, and Tim K Marks. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20850–20860, 2023.

- [394] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [395] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdif: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.
- [396] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.
- [397] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.
- [398] Stefan Klein, Marius Staring, and Josien PW Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE transactions on image processing*, 16(12):2879–2890, 2007.
- [399] Ruiqi Xian, Xijun Wang, Divya Kothandaraman, and Dinesh Manocha. Pmi sampler: Patch similarity guided frame selection for aerial action recognition. *arXiv preprint arXiv:2304.06866*, 2023.
- [400] Joy A Thomas. Elements of information theory, 1991.
- [401] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
- [402] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [403] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [404] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.
- [405] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.
- [406] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3637–3646, 2022.

- [407] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021.
- [408] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [409] Yoav Shalev and Lior Wolf. Image animation with perturbed masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3647–3656, 2022.
- [410] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021.
- [411] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [412] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.
- [413] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.
- [414] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023.
- [415] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.
- [416] Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023.
- [417] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [418] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

- [419] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020.
- [420] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.
- [421] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.
- [422] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [423] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.
- [424] Divya Kothandaraman, Ming Lin, and Dinesh Manocha. Prompt mixing in diffusion models using the black scholes algorithm. *arXiv preprint arXiv:2405.13685*, 2024.
- [425] Wanfeng Zheng, Qiang Li, Xiaoyan Guo, Pengfei Wan, and Zhongyuan Wang. Bridging clip and stylegan through latent alignment for image editing. *arXiv preprint arXiv:2210.04506*, 2022.
- [426] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023.
- [427] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023.
- [428] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023.
- [429] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.
- [430] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

- [431] Divya Kothandaraman, Kihyuk Sohn, Ruben Villegas, Paul Voigtlaender, Dinesh Manocha, and Mohammad Babaeizadeh. Text prompting for multi-concept video customization by autoregressive generation. *arXiv preprint arXiv:2405.13951*, 2024.
- [432] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023.
- [433] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.
- [434] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.
- [435] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.
- [436] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.
- [437] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022.
- [438] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022.
- [439] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13, 2023.
- [440] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.
- [441] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
- [442] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.
- [443] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

- [444] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023.
- [445] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.
- [446] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.
- [447] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.
- [448] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.
- [449] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.
- [450] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.
- [451] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023.
- [452] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.
- [453] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.
- [454] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccredit: Creative and controllable video editing via diffusion models. *arXiv preprint arXiv:2309.16496*, 2023.

- [455] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.
- [456] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.
- [457] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.
- [458] Bo Peng, Xinyuan Chen, Yaohui Wang, Chaochao Lu, and Yu Qiao. Conditionvideo: Training-free condition-guided text-to-video generation. *arXiv preprint arXiv:2310.07697*, 2023.
- [459] Anant Khandelwal. Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2023.
- [460] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *arXiv preprint arXiv:2305.17431*, 2023.
- [461] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.
- [462] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.
- [463] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023.
- [464] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibe Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *arXiv preprint arXiv:2309.14494*, 2023.
- [465] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19840–19851, 2023.
- [466] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

- [467] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [468] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
- [469] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- [470] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [471] Adobe. Adobe firefly. <https://firefly.adobe.com/>, 2024.
- [472] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [473] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.
- [474] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [475] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022.
- [476] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.
- [477] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. *arXiv preprint arXiv:2311.16090*, 2023.
- [478] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023.

- [479] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [480] GitHub Repo. Clip guided images mixing with stable diffusion. https://github.com/TheDenk/images_mixing, 2023.
- [481] Justin Pinkney. Image mixer diffusion. <https://www.justinpinkney.com/blog/2024/image-mixer-diffusion/>, 2024.
- [482] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023.
- [483] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023.
- [484] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.
- [485] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [486] Robert C Merton. Option pricing when underlying stock returns are discontinuous. *Journal of financial economics*, 3(1-2):125–144, 1976.
- [487] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- [488] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Boundary guided learning-free semantic control with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [489] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Unseen image synthesis with diffusion models. *arXiv preprint arXiv:2310.09213*, 2023.
- [490] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [491] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019.

- [492] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.
- [493] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016.
- [494] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.
- [495] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.
- [496] Yongqi Yang, Ruoyu Wang, Zhihao Qian, Ye Zhu, and Yu Wu. Diffusion in diffusion: Cyclic one-way diffusion for text-vision-conditioned generation. *arXiv preprint arXiv:2306.08247*, 2023.
- [497] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [498] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36, 2024.
- [499] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.