

ABSTRACT

Title of Thesis: : SYNPLAY: IMPORTING
REAL-WORLD DIVERSITY FOR A
SYNTHETIC HUMAN DATASET

Jinsub Yim
Master of Science, 2024

Thesis Directed by: Professor Shuvra S. Bhattacharyya
Department of Electrical and Computer Engineering

In response to the growing demand for large-scale training data, synthetic datasets have emerged as practical solutions. However, existing synthetic datasets often fall short of replicating the richness and diversity of real-world data. Synthetic Playground (SynPlay) is introduced as a new synthetic human dataset that aims to bring out the diversity of human appearance in the real world.

In this thesis, We focus on two factors to achieve a level of diversity that has not yet been seen in previous works: i) realistic human motions and poses and ii) multiple camera viewpoints towards human instances. We first use a game engine and its library-provided elementary motions to create games where virtual players can take less-constrained and natural movements while following the game rules (*i.e.*, rule-guided motion design as opposed to detail-guided design). We then augment the elementary motions with real human motions captured with a motion capture device.

To render various human appearances in the games from multiple viewpoints, we use seven virtual cameras encompassing the ground and aerial views, capturing abundant aerial-*vs*-ground and dynamic-*vs*-static attributes of the scene. Through extensive and carefully-designed experiments, we show that using SynPlay in model training leads to enhanced accuracy over existing synthetic datasets for human detection and segmentation. Moreover, the benefit of SynPlay becomes even greater for tasks in the data-scarce regime, such as few-shot and cross-domain learning tasks. These results clearly demonstrate that SynPlay can be used as an essential dataset with rich attributes of complex human appearances and poses suitable for model pretraining.

SYNPLAY: IMPORTING REAL-WORLD DIVERSITY FOR A SYNTHETIC HUMAN DATASET

by

Jinsub Yim

Thesis 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
Masters of Science
2024

Advisory Committee:

Professor Shuvra S. Bhattacharyya, Chair/Advisor

Dr. Heesung Kwon

Professor Jonathan Z. Simon

© Copyright by
Jinsub Yim
2024

Acknowledgments

First and foremost I would like to thank my advisor, Prof. Shuvra S. Bhattacharyya for working with him and for his unwavering support, and encouragement throughout this journey. Working alongside him, I discovered that this period has been the most valuable time of my life.

Secondly, I extend gratitude to Dr. Heesung Kwon for his invaluable contribution to this project. His generous support and assistance have been instrumental throughout its development. Without his extraordinary insights and expertise, this thesis would have remained a distant dream. I'd also like to thank Prof Simon Z. Jonathan for providing insightful and valuable feedback.

I would like to thank our research group members, especially Dr. Hyungtae Lee and Dr. Sungmin Eum. From the inception of brilliant idea to the culmination of this project, their direct support has played a pivotal role in achieving the results we see today. Additionally, I would like to express thanks to Yi-Ting and Yan Zhang for their contribution to this project.

Lastly, I express my gratitude to my wife Songhee Jung. Her love and support have meant everything to me. I'm so grateful to have her by my side. And to our beloved daughter, Thierry Yim, who joined our lives this year, thank you for being our special gift.

This research was sponsored in part by the Army Research Office and Army Research Laboratory (ARL) and was accomplished under Grant Number W911NF-21-1-0258. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, Army Research Laboratory (ARL)

or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Table of Contents

1	Introduction	1
2	Related Works	7
3	SynPlay Dataset: Methods and Diversity	10
3.1	Diverse yet realistic human motion	11
3.2	Multiple viewpoints	17
3.3	Scenario Design	18
3.4	Other Design Factors	21
3.4.1	3D Animated Characters	21
3.4.2	Background	22
3.5	SynPlay Statistics	24
3.5.1	Character Distribution	24
3.5.2	Altitude Distribution	24
3.5.3	Perspective Distribution	25
3.5.4	Bbox Size Distribution	26
3.5.5	Bbox Heatmap Distribution	26
3.5.6	Dataset Comparison	28
3.6	SynPlay Sample Images	30
4	Task Evaluation	33
4.1	General tasks: detection and segmentation	33
4.1.1	Experiment Setting	34
4.1.2	Aerial-view tasks	35
4.1.3	Ground-view tasks	37
4.1.4	Combination with MS-COCO for pre-training dataset	37
4.2	Data-scarce tasks: Few-shot and cross-domain learning	38
4.2.1	Experiment Setting	39
4.2.2	Comparison with other synthetic data	40
4.2.3	Scaling behaviors	41
4.3	Image Quality Evaluation	42
5	Discussion and Conclusion	43
5.1	Discussion	43
5.2	Conclusion	45
	Bibliography	46

Chapter 1

Introduction

Large-scale synthetic datasets, known for their scalability, provide a practical solution to the increasing demand for training large-capacity models. These datasets are crucial as they enable the development and fine-tuning of advanced machine-learning algorithms without the constraints of limited real-world data. By offering a vast amount of diverse and high-quality data, synthetic datasets ensure that models can generalize better and perform more accurately in various scenarios. Moreover, the ability to generate tailored datasets for specific applications accelerates the innovation process, allowing researchers to test and implement novel ideas quickly. This scalability and flexibility make synthetic datasets an indispensable tool in the advancement of machine learning technologies.

Recently developed rendering engines (*e.g.*, Unity [49], Unreal [16], *etc.*) have significantly enhanced the realism of synthetic data, broadening its applicability across various computer vision tasks. Despite efforts to scale synthetic data to match the extensive curation of real-world data, the desired level of diversity has not yet been achieved. This insufficiency in diversity is largely due to the inadequate consideration and integration of key factors that are essential to real-world diversity in the process of creating synthetic data.

In the past few years, several attempts have been made to increase human

appearance diversity by controlling innate characteristics (*e.g.*, race, gender) [5], body shape (*e.g.*, height) [5, 36], or clothing [5]. These datasets have demonstrated their effectiveness in tasks aimed at identifying human characteristics from close-up images, *e.g.*, human body/pose estimation [5, 36] and shape reconstruction [33]. However, these datasets have seldom yielded a discernible positive impact on computer vision tasks aimed at identifying humans *from a distance*, *e.g.*, human detection and segmentation.

When it comes to recognizing the overall human appearance from a distance, the motions and poses exhibited by the individuals play more vital roles than other characteristics. Despite prior attempts to synthesize various human poses, the quality of the rendering remained suboptimal, lacking in realism [39] and diversity [45]. AMASS [33] was the output of an early endeavor aimed at achieving both realism and diversity, where a motion scanner was utilized to collect real human motions. Before capturing these motions, detailed descriptions were provided to articulate specific movements — *e.g.*, *5 seconds waving above the head with both arms*¹, while adhering to physical constraints that limit large movements in motion-capture environments. This *detail-guided motion design* often results in capturing a restricted range of motions tied to specific descriptions while missing out on all the motions that defy easy description.

We claim that providing relatively high-level, less-detailed guidance greatly helps in breaking out from the aforementioned limitations and provides more freedom

¹This motion description was used to construct the Mocap Database HDM05 in AMASS (<https://resources.mpi-inf.mpg.de/HDM05/05-01/index.html>)

towards the expansion of the diversity in human motions. In constructing our dataset, we follow a new **rule-guided motion design** approach providing game “rules” or winning strategies for the virtual players to follow, which serve as a set of significantly coarser guidelines when compared to detail-guided approaches. In this way, the motions that they manifest are not confined to predetermined/easily-describable motions. As for the “rules,” we opted to borrow them from the six traditional Korean games that were also played in the Netflix TV series “Squid Game” [23]. These games involve substantial amounts of physical movements, which naturally provide room for a diverse set of human poses and motions. The diversity is further influenced by in-game factors such as the uniquely defined rules of each game, the number of players, and the interactions between them.

Under our rule-guided motion design approach, each *scenario* run (i.e., one round of a game played with specific settings) in the virtual environment is initialized by carrying out a scenario design step which is followed by the incorporation of real-world motions. The scenario design involves the setup of all the parameters that control the appearance, players (winners and losers), game dynamics (e.g., how/when each game ends), and the human motion evolutions for each specific scenario. This is where the high-level rules of a given game are defined, and the coarse boundary of how human motions can evolve within the game is set. The incorporation of real-world motion is the phase where a rich variety of motions truly comes to life. Details on the entire pipeline will be elaborated in Chapter 3.

In addition, we also took into account that human appearance can vary greatly depending on the perspective from which it is viewed. Accordingly, we capture



Figure 1.1: **SynPlay** dataset is constructed while players play six traditional games in a virtual playground also introduced in the Netflix TV show “Squid game” [23]. We have diversified the human appearances in the scenes by focusing on two factors: i) leveraging real-world human motions and ii) adopting multiple viewpoints.

every scene from **multiple viewpoints** by implementing several image-capturing devices to take advantage of different perspective-related characteristics: three Unmanned Aerial Vehicles (UAVs), three Closed-Circuit Televisions (CCTVs), and one Unmanned Ground Vehicle (UGV). The three UAVs fly with random trajectories at different altitudes, the three CCTVs are located at the front, side, and back of the game playground, and the UGV moves randomly within the playground where the game is being played. These devices offer a variety of image-capturing properties, including aerial-*vs*-ground and dynamic-*vs*-static. Our strategy, designed to provide very diverse viewpoints in the scene capture process, not only serves to ensure that the dataset includes more diverse human appearances but also broadens the potential tasks (*e.g.*, re-identification, multi-view applications, aerial-to-ground scene matching, *etc.*) for which the dataset can be used.

By leveraging the aforementioned human appearance-diversifying strategies, we construct a large-scale synthetic human dataset called *SynPlay* that contains

more than *73k images with 6.5M human instances*, see sample images in Fig 1.1. To demonstrate SynPlay’s ability to represent a variety of human appearances to the extent seen in the real world, we conduct a series of experiments where we evaluate the impact of leveraging SynPlay alongside real (non-synthetic) datasets curated for a variety of human-related tasks, *i.e.*, aerial-view/ground-view human detection and segmentation. For all the tasks, training with SynPlay outperforms its counterparts (*i.e.*, training-from-scratch, using other synthetic data) across a variety of datasets. Experiments also demonstrate that the SynPlay dataset significantly improves model performance on data-scarce tasks, highlighting its value in scenarios that require substantial supplementary training data.

Research Questions To effectively evaluate SynPlay’s role as supplemental training data, we designed our experiments around three key questions:

- **Can SynPlay impact General Computer Vision Tasks?** This question delves into SynPlay’s ability to improve the generalizability of computer vision models across a range of common tasks. It assesses whether SynPlay effectively handles tasks that require identifying and understanding diverse human appearances, which can be challenging due to variations in pose, clothing, and lighting conditions.
- **Can SynPlay perform effectively for data-scarce tasks (few-shot, cross-domain learning)?** This question focuses on SynPlay’s ability to supplement real-world data in situations where data is scarce or limited.

- **How does SynPlay's data quality compare to other synthetic datasets?**

This question breaks down the comparison into two key aspects: Data quality metrics and correlation with performance.

Chapter 2

Related Works

Synthetic human data. The creation of various synthetic human datasets has been facilitated by the advancement of modern synthetic data rendering engines such as Blender [6], Unity [49], and Unreal [16], alongside human modeling tools like MakeHuman [34] and Character Creator [47]. These rendering engines enable a realistic representation of humans in 3D virtual environments, while the modeling tools give creators precise control over the design of virtual characters. The creators of these datasets leveraged these tools to meticulously control key design factors, ensuring suitability for specific tasks, *e.g.*, SOMAset [3], PersonX [48], UnrealPerson [53], CARGO [52] for Re-Identification, SURREAL [51] for pose estimation, GTA5 [39] for semantic segmentation, and Archangel-Synthetic [45] for detection.

Recently, several attempts have been made to enhance the realism of virtual human models, with the aim of bringing them closer to the quality of their real-world counterparts. SMPL-X [37] and AMASS [33] used motion-capturing devices to capture natural human motions. BEDLAM [5] tried to improve the diversity of various factors such as skin tones or clothing that affect human outer appearance, still relying on SMPL-X. However, motion scanners impose constraints on the environment, particularly in capturing large motions or events involving multiple humans. While AGORA [36] and ScoreHMR [46] sought to digress from the usage

of the motion scanners by fitting human body models to the human motions in real-world images/videos, the quality of the fitted human models declined drastically on the images taken from a distance. One of our goals for our dataset was to incorporate multiple viewpoints, including distant views of the target scene. To avoid compromising the quality of human motions/poses, we chose to use motion capture devices, while implementing our approach to ensure that final results in the dataset are not limited by the environments in which the devices were used.

Natural human motion acquisition. Whether we are creating a real or synthetic dataset, images of motions captured by directing humans to perform specific actions based on a description often appear awkward rather than natural. Because of that, most datasets aim to include humans engaged in daily activities (*e.g.*, MS COCO [29], MPII Human Pose [2]), performing tasks such as sports (*e.g.*, UCF-Sports [41], SoccerNet [10], SportsMOT [13]) or art (*e.g.*, Human-Art [26]) to capture their motions and poses in the most natural states possible. However, it is self-contradictory to artificially create a virtual event to capture natural motions associated with the event. In this thesis, we aim to avoid this self-contradict by initially designing the virtual events (*i.e.*, aforementioned games) using existing but non-natural virtual motions, which are then replaced by real-world motions captured using a motion capture device.

Supplemental datasets for training. Enhancing model performance by supplementing the training with additional data has been a common strategy [27, 38]. Initially, this involved combining datasets constructed with the same purpose, like

MS COCO [29] and PASCAL VOC [17], for tasks such as object detection. Some approaches utilized large-scale datasets (*e.g.*, ImageNet [14] or Instagram [32]), which were not necessarily designed for the target task, to build foundational features, followed by a transfer learning such as pretrain-finetune [19] or PTL [43] to adjust the model on the target dataset. As models have grown in size and complexity (*e.g.*, ViT [15]), the demand for large-scale, high-quality datasets has increased, but the high costs of annotation present a significant barrier. To address this, label-agnostic training methods like self-supervised learning [7, 8, 20, 21] and synthetic dataset generation with cost-free annotations have emerged as viable solutions. In response, we have specifically designed SynPlay to supplement various computer vision tasks that require a large-scale, highly-diversified human appearance set.

Chapter 3

SynPlay Dataset: Methods and Diversity

In creating our synthetic dataset, we focus on two key factors: **rule-guided motion design** and **multiple viewpoints**. These are crucial for capturing the wide range of real-world variations, enabling us to mirror this diversity in our dataset. By employing rule-guided motion design, we establish high-level game rules that allow virtual players to generate diverse and realistic movements, avoiding the limitations of predetermined motions. Capturing scenes from multiple viewpoints with various image-capturing devices ensures diverse perspectives. This approach enhances the diversity of human poses and appearances and broadens the dataset’s applicability for tasks like re-identification, multi-view applications, and aerial-to-ground scene matching. We drew inspiration from traditional Korean games featured in the Netflix series “Squid Game” [23], leveraging their diversity in physical activities and interactions.

Using these methodologies, We created ten different scenarios for each game, resulting in 60 game scenarios in total. Frames were rendered from seven different camera viewpoints at 1 fps with a resolution of 1920×1080 , resulting in a total of 73,892 images with more than 6.5M human instances. The frame generation rate was selected as 1 fps to avoid including highly redundant human poses. Taking full advantage of the game engine’s ability to generate annotations while rendering the

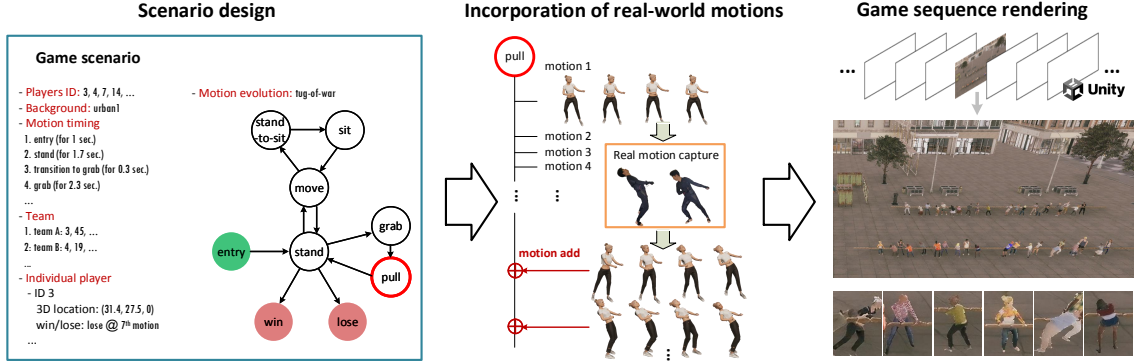


Figure 3.1: **Game sequence generation pipeline.** This illustrates how we create a sequence for a **tug-of-war** game which includes an example of how we incorporate real-world motions towards the elementary motion state of **pull**. In the motion evolution graph, the start and end nodes are indicated by green and red circles, respectively. A diverse set of **pull** motion instances is shown below the image of the rendered scene.

scenes, we provided various types of ground truth annotations useful for various computer vision tasks: 2D/3D bounding boxes, instance-level segmentation masks, depth maps, and human keypoint locations.

3.1 Diverse yet realistic human motion

We use *rule-guided motion design* in our SynPlay dataset, borrowing “rules” from six traditional Korean games, also featured in the Netflix series “Squid Game” [23]¹. This approach offers coarser motion guidance for the virtual players, facilitating the generation of a wide spectrum of natural motions, even including the ones that defy detailed description.

Our rule-guided motion design is effectively baked into the overall sequence-generating design pipeline, as shown in Figure 3.1, that consists of the *scenario design*

¹Our game scenarios are designed based on the traditional game rules, without taking any specific situations from the show.

followed by the *incorporation of real-world motions*. The scenario design involves the setup of all the parameters that control the appearance, players (winners and losers), game dynamics (*e.g.*, how/when each game ends), and the human motion evolutions for each specific scenario. All the items within each scenario that do not have to be hard-coded (*e.g.*, game rules) are selected randomly when designing each scenario. The motion evolution of each virtual player in a specific game is governed by a graph structure where all possible elementary motions and their potential transitions are represented as nodes and directed edges, respectively. Each node is tied with a pool of motions that fall under the same elementary motion state (*e.g.*, move, sit). As the game progresses, a virtual player evolves its motion following the directed edges and stays there according to the “motion timing”, also defined in the scenario design. At each state, the virtual human randomly chooses to exhibit one of the motions in the corresponding pool. Note that, while a uniquely designed scenario is used for a unique sequence, the same motion evolution graph [50] is used for all the sequences captured under the same game rule.(need to explain more detail)

Fig 3.2 shows motion evolution graphs used in designing the game scenarios for the SynPlay dataset. Even within the same game, the scenario may change, but the motion evolution graph will remain consistent. It is noteworthy to mention that, despite the wide range of situations and a variety of motions involved in the games, the motion evolution graph for each game consists of only a few motion nodes and their transitions. Given that each node encompasses a range of motions, this illustrates the essence of a rule-based design approach where only basic game rules are provided to freely allow the diverse array of human motions to be manifested.

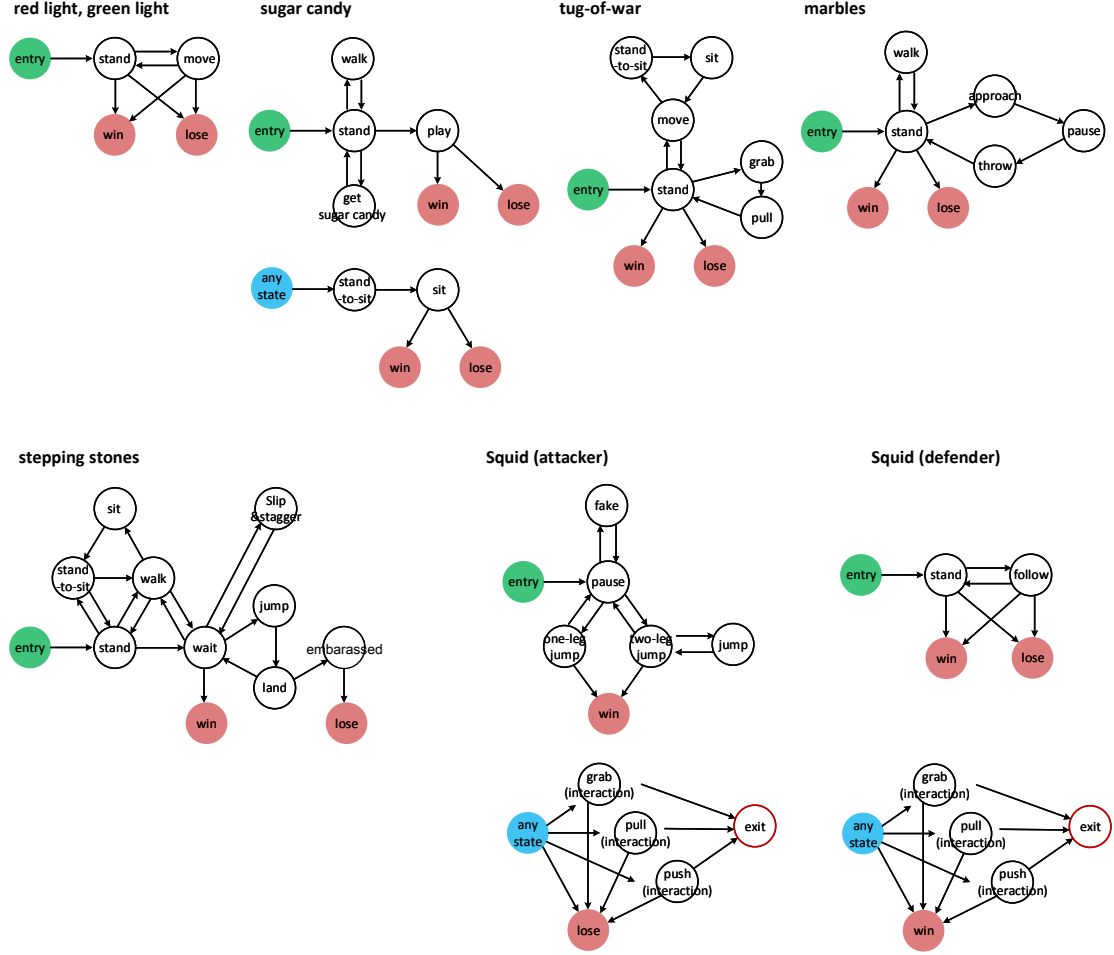


Figure 3.2: **Motion evolution graphs.** The start node (‘entry’) and the end nodes (‘win’ or ‘lose’) are indicated by green and red circles, respectively. For the games where secondary graphs are available (i.e., sugar candy or squid), at any given time (except at the start or end node), the current state in the main graph can move to the ‘any state’ node (blue-filled circle) in the secondary graph. When the ‘end’ node (red-bordered circle) is reached within the secondary graph, the current state moves its way back to the latest node that was touched in the main graph before entering the secondary graph.

Before incorporating the real-world motions, we leverage two techniques to pre-diversify the elementary human motions readily available in human motion libraries such as Mixamo [1]: i) dynamically blending two existing motions of similar types to generate a new motion type (*e.g.*, blending slow-walking and running to generate hasty-walking), ii) using elementary motions as animation layers to

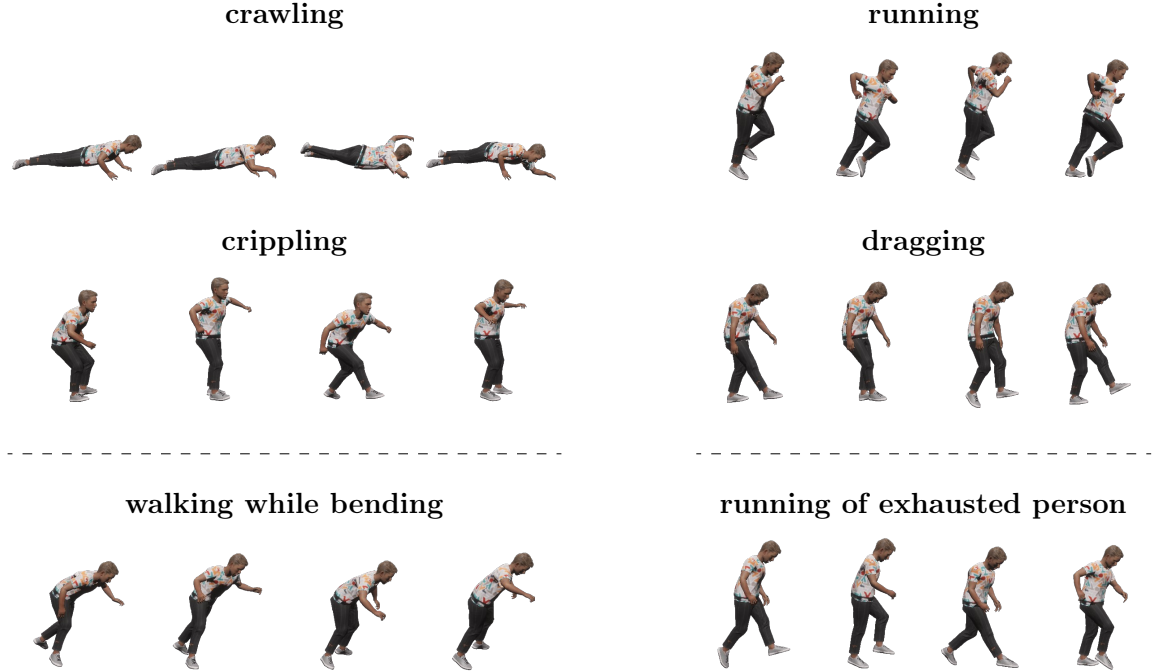


Figure 3.3: **Two motion blending examples.** For each example (left or right column), the two motions (top and middle row) are blended together to generate a new motion (bottom row). The blending ratio between the two input motions can be controlled. The blending process does not depend on the specific names given to the motions.

make a new motion (*e.g.*, raising hands while walking). Fig 3.3 and Fig 3.4 show several examples of the blending process and how the animation layers are leveraged: the two techniques for expanding human motions within the virtual environments, respectively. Interestingly, the motions created by blending are largely different from their corresponding input motions, while the motions created via leveraging the animation layers still exhibit the appearances and dynamics resembling both the input motions. These two techniques are readily available for use within the Unity environment.

On top of the pre-diversified set of motions for each game, a real human player



Figure 3.4: **Three examples of leveraging animation layers.** For each example (left, middle, or right), the resulting motion of leveraging the animation layers over two input motions (top and middle rows) is shown in the bottom row. Note that the semantic labels (e.g., walking, cheering) were not provided at the time of capture; they are included in the figure only for the convenience of the reader.

wearing a motion capture device ² is asked to either similarly mimic or newly create motions that align with the given game rule (thus, *rule-guided*). For example, for the game of tug-of-war, human players were provided with the game *rule*, and then asked to reenact any possible motion with the freedom of choosing the winning or losing side. For some games, more than one player were asked to play the game together to capture the motions that can naturally arise at the time of physical interactions. As the result of *incorporating the real-world motions*, the total number of unique motions in SynPlay increased from 104 to 257.

²Each real player used a SmartSuit Pro II and a pair of Smartgloves from Rokoko (<http://rokoko.com>)

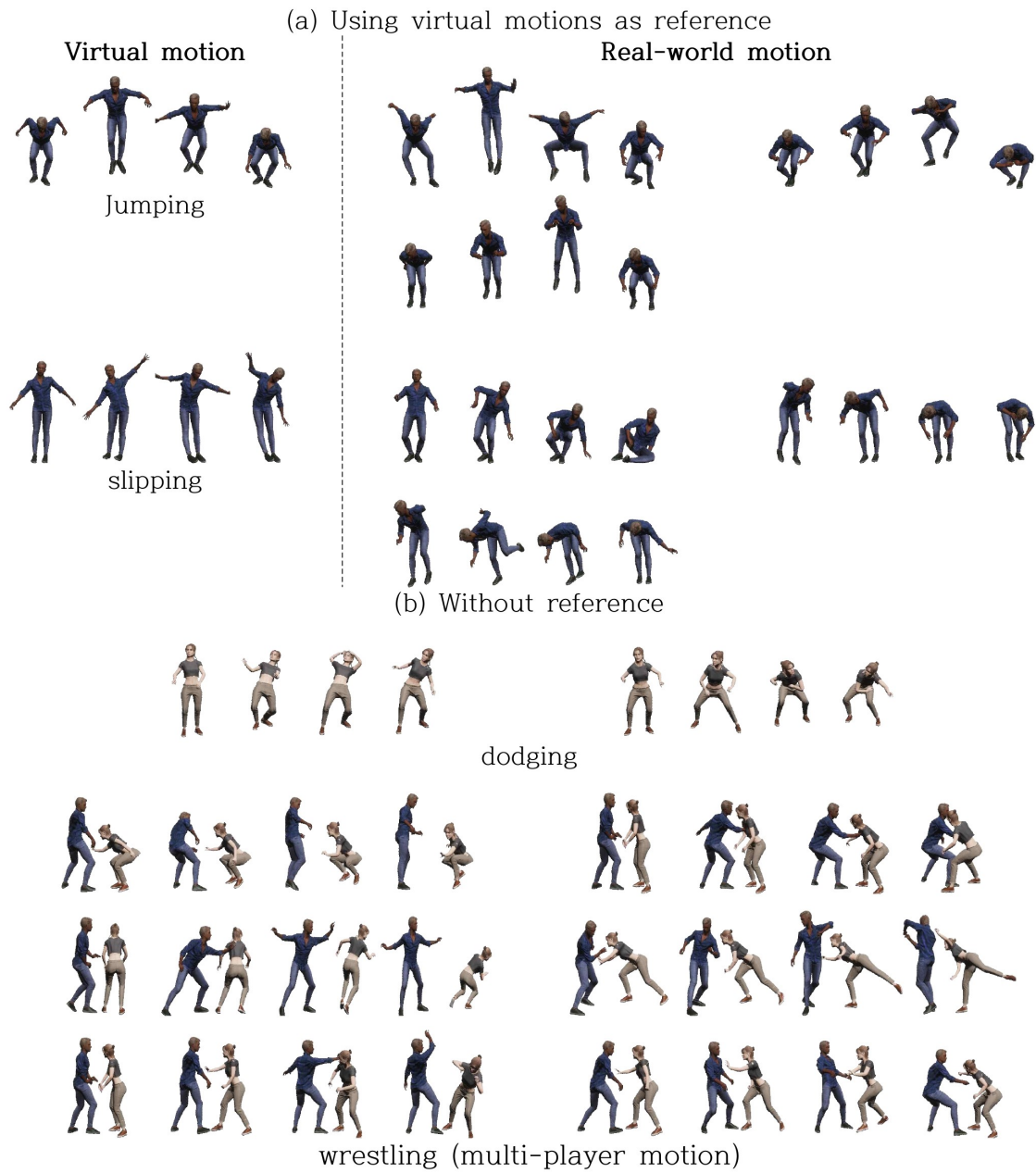


Figure 3.5: **Real-world motion examples.** Real-world motions are acquired either (a) by mimicking reference motions or (b) by exhibiting potential in-game motions without any references that align with the given game rules. Wearable motion scanners are used for all the cases.

Fig 3.5 shows several examples of real-world motions. Real-world motions are created either by having the real human wearing the motion capture device mimic the pre-provided reference motions or by demonstrating potential in-game motions under the given game rules. It is observed that real-world motions can express a wider range of specific actions while maintaining a sense of realism. Moreover, motions that are difficult to pinpoint or describe can also be created, *e.g.*, multi-person wrestling motions.

3.2 Multiple viewpoints

The camera viewpoints within SynPlay are diversified by implementing three widely used types of image-capturing platforms in the real world: UAV, UGV, and CCTV. They cover a variety of image-capturing properties like static/dynamic and ground/aerial. Viewpoint diversity is acquired by controlling the locations and the focal points of the cameras. Three UAVs, three CCTVs, and one UGV have been deployed (Figure 3.6), resulting in seven unique viewpoints for every game sequence. The UAVs are deployed to fly at various random locations while maintaining altitudes of low ($\sim 30m$), medium ($\sim 50m$), and high ($\sim 100m$). CCTVs are located at a height of $15m$ at the front, back, and either side of the game playground. UGV images are captured assuming that a vehicle is randomly roaming on the ground. The focal points are set at several locations close to the area where the game usually takes place. For the UAVs, the focal point is changed to a random location every $10\ sec$, where each change takes $5\ sec$ to be fixed at a location for another $5\ sec$. Focal

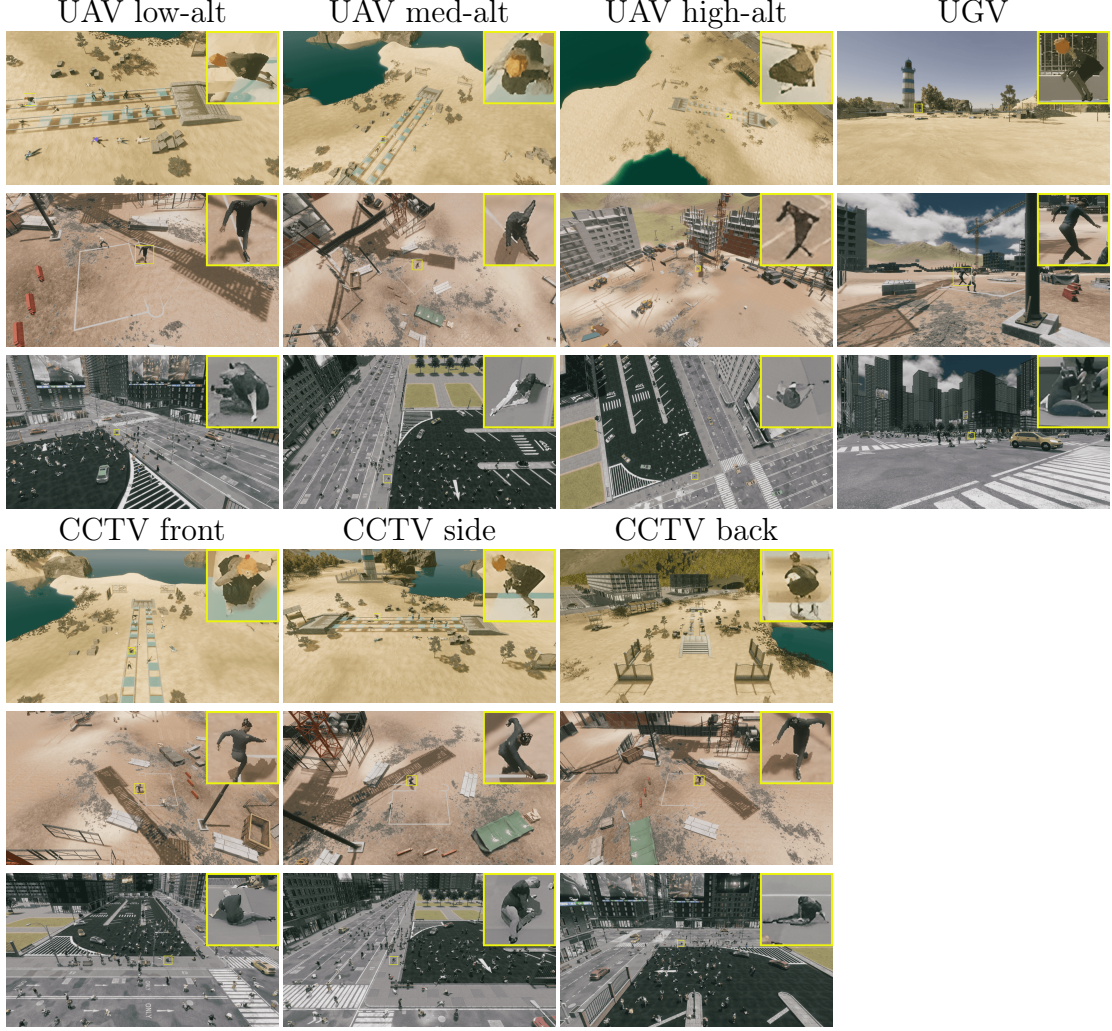


Figure 3.6: **Multiple viewpoints used in SynPlay.** On the top-right corner of each image, we place the enlarged crop of one human instance who is visible from all seven viewpoints in each scenario. Multiple camera viewpoints allow substantial variations in appearance for the same human subject with identical pose.

points of the CCTVs and the UGV do not change once determined. Figure 3.7 shows illustration of camera movement and perspective.

3.3 Scenario Design

For naturalness and connection, we used six traditional Korean games (Red Light Green Light, Honeycomb, Tug of War, Marbles, Glass Stepping, and Squid Game)

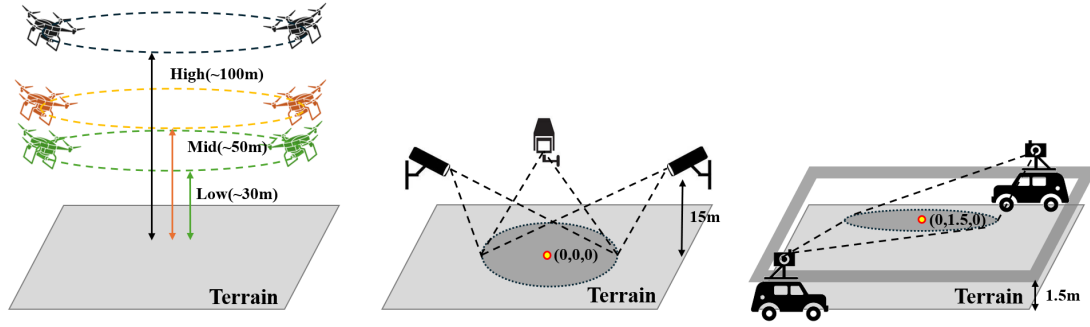


Figure 3.7: **Illustration of Camera Perspectives.** From left to right: UAV capturing frames from aerial perspectives at varying altitudes, CCTV positioned at different angles around the game ground, and UGV moving randomly within the game ground. These perspectives provide diverse viewing angles for comprehensive data collection.

in the Netflix series Squid Game as the motif for each scenario. The objective across all games was to authentically portray the required movements of the characters without direct user intervention. Each game sequence captured unique challenges and interactions, contributing to a diverse and extensive exploration of human behavior under pressure.

- **Red Light, Green Light:** In this game, contestants must advance towards a finish line during “Green Light” but freeze in place as soon as “Red Light” is announced. The challenge lies in stopping abruptly, Participants exhibit varying speeds and movements as they race forward, facing obstacles and navigating through the crowd. However, failure to stop results in immediate elimination, with contestants collapsing to the ground.
- **Sugar Candy:** Participants in this game are tasked with delicately carving out a specific shape from a honeycomb candy without causing it to break. Each participant chooses a symbol and lines up in front of it, then receives the honey-

comb from the NPC and places it in a random location. Once all participants have their honeycomb, they can each play the game. Throughout the game, participants exhibit a range of natural postures, seamlessly integrated into the gameplay. They can be seen standing patiently in line, sitting comfortably on chairs as they engage with the task, or even reclining on the floor.

- **Tug of War:** In the “Tug of War” game, two teams exert determined strength, each applying random force to the rope in a bid to surpass a critical victory threshold. Participants, organized into teams and lined up, showcase a variety of poses as they engage in the intense struggle, depicting dynamic motions during the pulling action.
- **Marbles:** This game involves players taking turns to flick marbles into a central hole. Each player throws their marbles into the central hole with random angles and force, aiming to score points. As with other games in the scenario, the diverse poses as they flick their marbles contribute to the atmosphere.
- **Stepping Stones:** This game involves crossing glass-covered bridges where pre-set fragile glass panels are placed. Players must find a safe path to reach the opposite end of the bridge. The game incorporates dynamic movements including preparatory actions before jumping, various jumping and landing movements, slipping movements after landing, expressions of surprise upon confirming glass breakage, and utilizing the rag-doll effect when players hit the ground for naturalistic body movements.

- **Squid Game:** This game takes place in a squid-shaped arena where two players face off against each other. One player acts as the defender, remaining within the squid-shaped area, while the other player acts as the attacker, moving to catch the defender. The attacker can stand with both feet inside the circle and must jump with one foot until crossing the defender’s torso. The game emphasizes interactions between players, including movements to catch or evade, pulling or pushing, and deceptive motions to confuse the opponent.

In conclusion in the Scenario Design section, we examined six traditional Korean games used as motifs for each scenario in the “Squid Game” series. The goal was to have the characters participating in each game naturally exhibit the movements required for the game while following the rules set for each game. The six games formed one sequence, and a total of 10 sequences were captured in different environments with different seeds for random number generation

3.4 Other Design Factors

3.4.1 3D Animated Characters

We have designed 456 virtual characters using the Character Creator [47](See Fig. 3.8), where each character is involved in multiple game scenarios. To vary the appearance of the characters and avoid generating biases, each character was uniquely designed with gender, skin color, age, height, obesity (body type), hair (styles and colors), and outfits. We kept the gender ratio between male and female at 1:1 and the ratio of skin color among white, black, yellow, and brown at 1:1:1:1. For age, each character

was designed to fall into one of three categories: child, middle-aged, and elderly, and the ratio was set at 1:2:1. For each gender and age group, heights were modeled to follow a bell-shaped distribution, resulting in an overall dataset range of 140 to 190 cm. We manually designated every character with a unique outfit, while setting the hair and obesity aspects as diverse as possible.



Figure 3.8: **456 virtual players in SynPlay** created using Character Creator.

3.4.2 Background

For each scenario, we set different environmental factors: sites, lighting conditions, and weather. Utilizing assets available on the Unity Asset Store and leveraging the Unity Terrain Tool, we developed five urban environments, encompassing three typical city locations, a construction area, and a factory site. Additionally, we simulated five natural environments, including a green area, a snowy field, a desert, a meadow, and a beach(See Fig. 3.9). Multiple locations within each site map can be used as local playgrounds.

The weather conditions were described as clear skies, foggy conditions, and

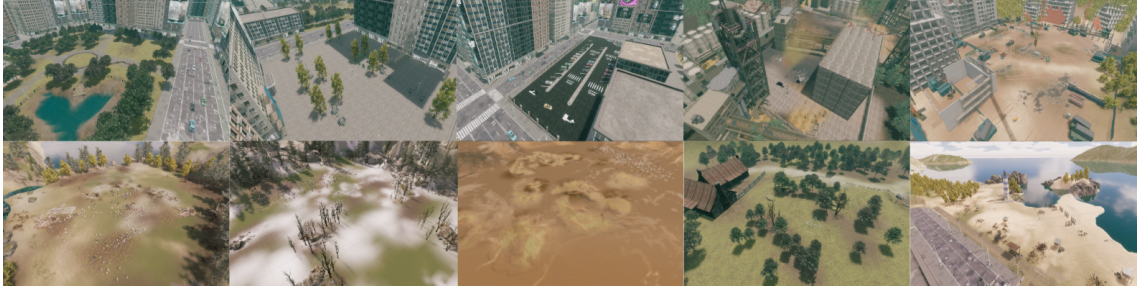


Figure 3.9: **Various Background Environments.** From top left to bottom right: Three typical city locations, Construction site, Factory cite, Green, Snowy, Desert, Meadow, and Beach.



Figure 3.10: **Various Lighting Conditions.** From left to right: Dawn, Morning, Noon, Afternoon, Sunset.

foggy with rain. This was done to enhance the model’s ability to perform accurately in a wider range of conditions. To achieve this, the Unity fog volume was utilized to depict foggy situations, while particle systems were employed to simulate rainfall.

The lightning environment were characterized by five distinct time periods: dawn, morning, noon, afternoon, and sunset(See Fig. 3.10). Initially, to simulate the sun’s position in the Unity environment, the rotation of directional light was precisely calculated and adjusted to correspond with each time period.

Each scenario involves randomly determined sites, lighting conditions, and weather.

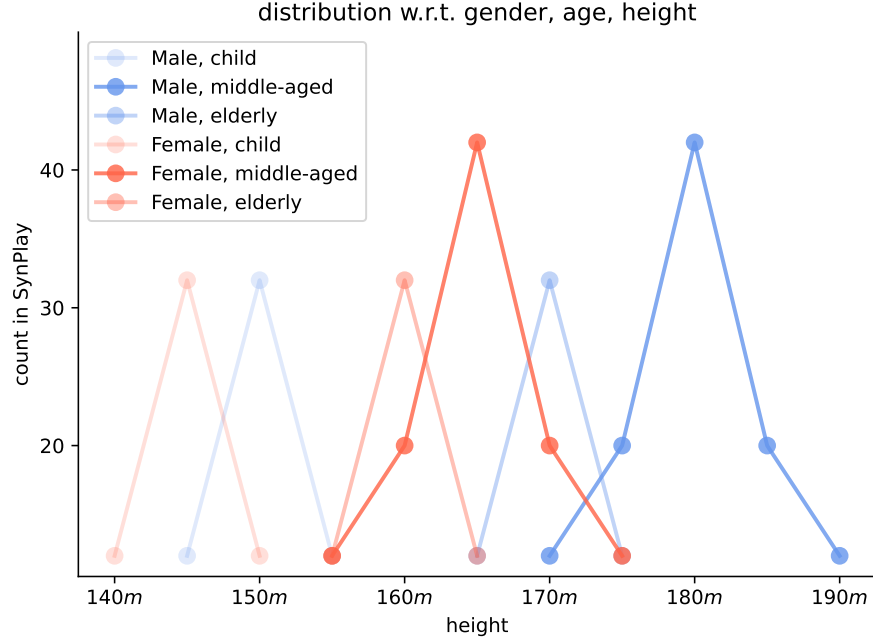


Figure 3.11: **Character height distribution** according to gender and age.

3.5 SynPlay Statistics

3.5.1 Character Distribution

Fig 3.11 displays bell-shaped distributions of human height categorized by gender and age. Leveraging these distributions, we generated 456 virtual characters that accurately reflect demographic characteristics. Beyond height, gender, and age, we diversified attributes like skin color, body type, hair style, and attire.

3.5.2 Altitude Distribution

Fig 3.12 depicts the altitude distribution within the dataset, showing a standard distribution pattern across UAV sensors, excluding fixed-range CCTVs and UGV. The data reveals a randomized yet consistent spread from minimum to maximum

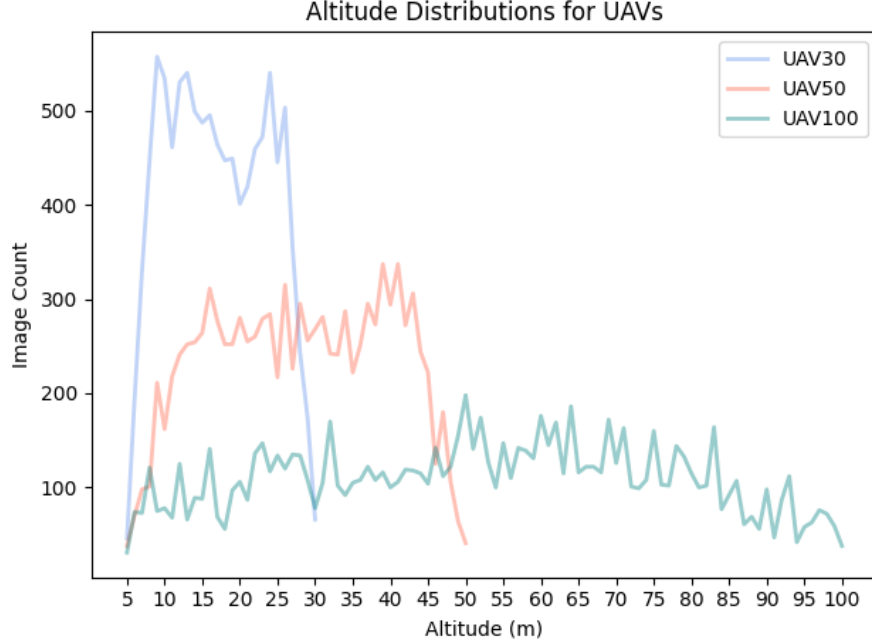


Figure 3.12: **Altitude distribution** for each image-capturing device.

altitudes, reflecting the dataset’s extensive coverage of human instances captured under various aerial conditions. This distribution underscores the dataset’s ability to represent diverse real-world surveillance scenarios, essential for training robust detection models capable of handling altitude variations effectively.

3.5.3 Perspective Distribution

Fig 3.13 shows the angle distribution ranging from 0 degrees(perpendicular to the ground) to 90 degrees(parallel to the ground). This indicates that the dataset reflects not only nadir-view perspectives but also various UAV perspectives. This diversity in perspectives enhances the robustness of object detection models by exposing them to a wide range of viewpoints, preparing them to recognize and localize objects under different spatial orientations.

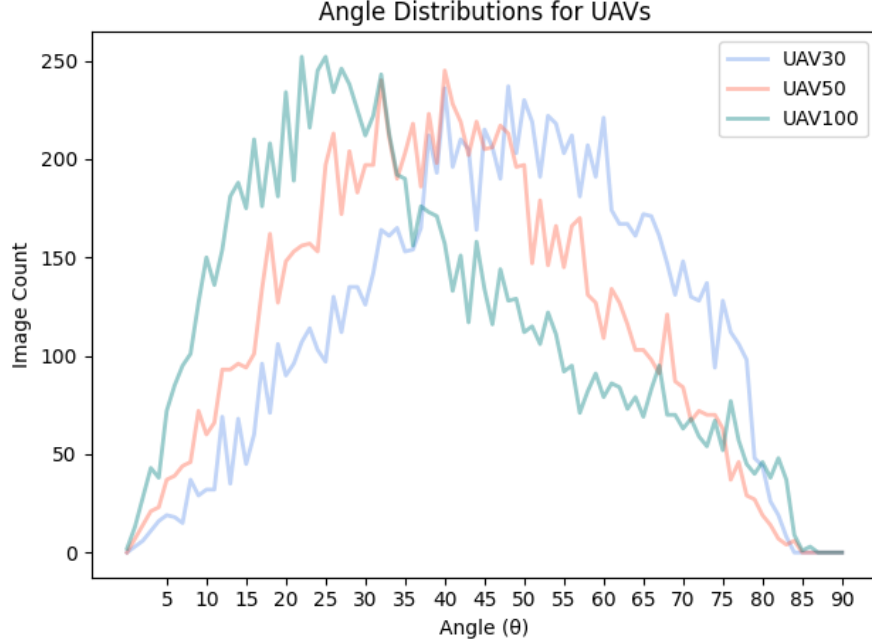


Figure 3.13: **Angle distribution** for each image-capturing device.

3.5.4 Bbox Size Distribution

Fig 3.14 shows the distribution of bounding box sizes over human instances captured by each device. The majority of bounding box sizes are small, which illustrates a common characteristic of aerial-view datasets. Interestingly, UAVs can capture human instances with larger bounding boxes than CCTVs. This could be due to the fact that, although UAVs are typically positioned at higher altitudes than CCTVs, there are more cases where UAVs get closer to real-time events and human instances, different from the fixed CCTVs.

3.5.5 Bbox Heatmap Distribution

Fig 3.15 presents the heatmap distribution of 2D bounding box data from various sensors in the SynPlay dataset. The heatmap visualization illustrates how

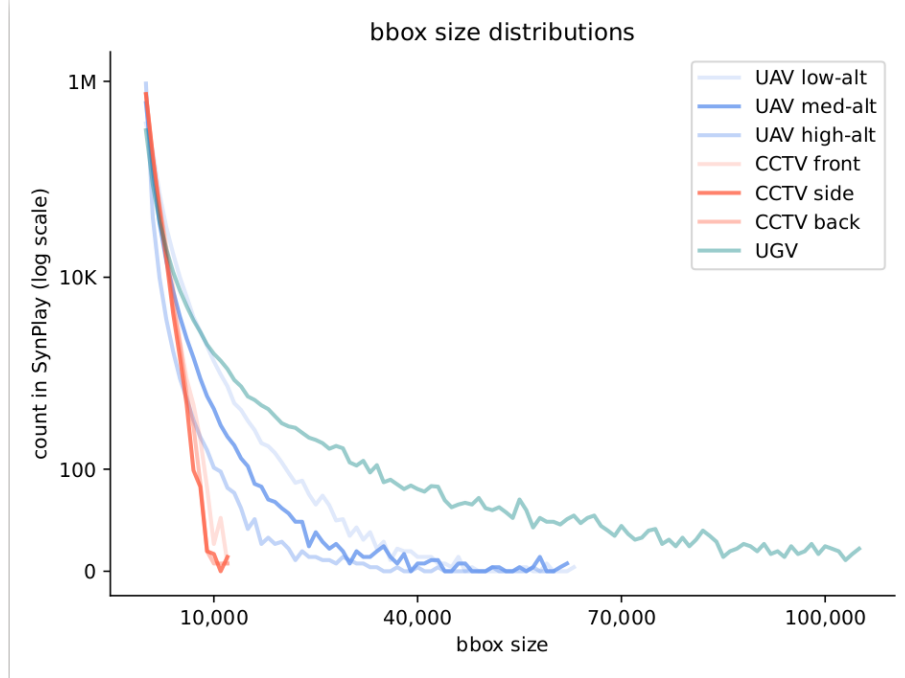


Figure 3.14: **BBox size distribution** for each image-capturing device.

frequently 2D bounding boxes cover different regions of the image frames. Typically, brighter colors indicate higher percentages of frames where the bounding boxes are located. Across most sensors, the central regions of the images exhibit brighter colors, suggesting a concentration of detected objects in these areas. Notably, excluding UGVs, the heatmap distributions tend to fade towards the edges of the images. This phenomenon suggests that while central areas may capture more frequent and prominent activity, the outer regions see fewer instances of bounding box detections. Understanding these spatial patterns is critical for optimizing surveillance tactics and improving the accuracy of object detection algorithms across diverse sensor types.

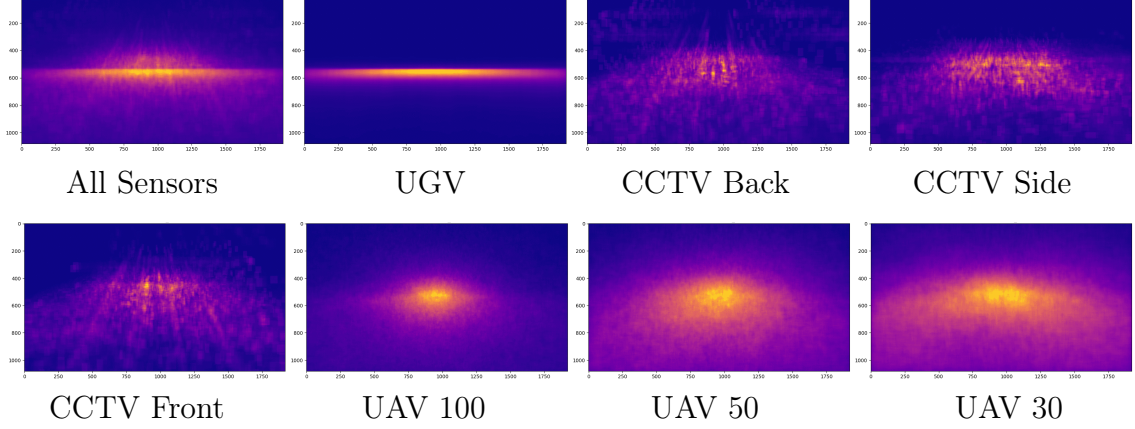


Figure 3.15: **Heatmap distribution from various sensors.** The X-axis and Y-axis represent the horizontal and vertical pixels of the image (1920 x 1080), respectively. The brightness of each pixel is determined by normalizing with the maximum value of each sensor.

3.5.6 Dataset Comparison

Table 3.1 provides a comparative analysis of various human datasets, categorized as real or synthetic and captured from either ground or aerial perspectives. Key observations from this comparison are outlined below:

1. Aerial-view sets, *thanks to their wide viewing angles*, generally have *more human instances per image* than ground-view sets, except for few cases that employ a fixed number of actors in a real set or design one instance per image in a synthetic set.
2. Aerial-view sets generally contain a wider range of viewpoints (mostly near~far).
3. For existing synthetic datasets, aerial-view sets typically feature *fewer motion variations* compared to ground-view sets. This is because *aerial-view datasets often prioritize leveraging a wide range of viewpoints over expanding the variety of human motions*.

dataset	domain	#inst	#img	#inst/img	natural motion	#motion	viewpoint
<i>ground-view</i>							
VOC 12 [17]	real	10K	11.5K	2.48	daily	infinite	near
COCO Dev17 [29]	real	649K	164K	9.72	daily	infinite	near
MPII Human Pose [2]	real	40K	24.9K	1.61	daily	20	near
Cityscapes [12]	real	21.4K	5K	7.85	daily	2	near
ADE20K [54]	real	30K	27.5K	4.36	daily	infinite	near
Human-Art [26]	real	123K	50K	2.46	art	infinite	near
GTA5 [39]	synth	1.4M	1.4M	1	X	20K _{pose}	near
SURREAL [51]	synth	6.5M	6.5M	1	detail+mocap	23	near
SOMASet [3]	synth	100K	100K	1	detail+mocap	250 _{pose}	near
PersonX [48]	synth	273K	273K	1	X	4 _{pose}	near
UnrealPerson [53]	synth	120K	120K	1	X	2	near
AGORA [36]	synth	.	19K	1~15	detail+mocap	4,240 _{pose}	near
BEDLAM [5]	synth	.	380K	1~10	detail+mocap	2,311 _{pose}	near
<i>aerial-view</i>							
Okutama-action [4]	real	.	77K	~9	detail	12	med
Semantic Drone [24]	real	1.5K	400	4.16	daily	unspecified	med
UAVid [31]	real	4.7K	420	20.06	daily	unspecified	med~far
VisDrone [55]	real	109K	40.0K	15.42	daily	unspecified	med
Archangel-real [45]	real	165.6K	41.4K	4	detail	3 _{pose}	near~far
Archangel-mannequin [45]	real	.	178.8K	6~7	detail	3 _{pose}	near~far
Archangel-synth [45]	synth	4.4M	4.4M	1	X	3 _{pose}	near~far
SynDrone [40]	synth	803K	72K	11.15	X	2	med~far
CARGO [52]	synth	108K	108K	1	X	2	near~far
SynPlay	synth	6.5M	73K	88.40	rule+mocap	infinite	near~far
* natural motion · daily: human motions engaged in daily activity · art: human motions shown in works of art · detail: human motions captured by ‘detail-guided design’ · rule: human motions captured by ‘rule-guided design’ · +mocap: human motions captured using a motion scanner							

Table 3.1: **Comparison of human datasets.** ‘#inst/img’ is acquired only on images that contain humans. ‘#motion’ indicates the number of unique motions depicted in the dataset, except the ones with the subscript ‘pose’ which indicate the number of static poses. Since a single motion can consist of multiple number of unique poses, #motion is generally smaller than the number of poses. For certain datasets, the test set without available labels is excluded from this comparison.

4. Rule-guided design can utilize *significantly larger range of human motions* compared to detail-guided design.

The comparison shown in the table also demonstrates that SynPlay successfully addresses the shortfall of aerial-view synthetic datasets (3rd observation), while maximizing the benefits of aerial-view datasets (1st and 2nd observations).

Moreover, the 4th observation supports that our proposed rule-guided design is successful in securing the diversity of human motions in the set. It is noteworthy that while SURREAL [51] (constructed with ‘detail+mocap’) contains a comparable

number (6.5M) of human instances as SynPlay, the number of motions manifested in the dataset is extremely limited when compared to SynPlay (23 vs. infinite).

3.6 SynPlay Sample Images

Fig 3.16 includes additional sample images from the SynPlay dataset. Various human appearances depending on human motion differently taken according to the game scenario, and camera viewpoints are observed. Various human appearances are observed that change depending on human motions taken differently according to the game scenario, and different camera viewpoints. In addition, various characters and backgrounds used for creating SynPlay are also visible.

(1) Red Light, Green Light



UGV



UAV med-alt



UAV low-alt



CCTV back



UAV high-alt

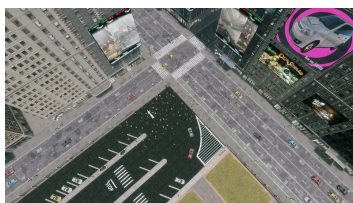


CCTV side

(2) Sugar Candy



CCTV side



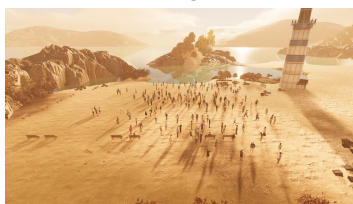
UAV high-alt



UGV



UAV low-alt



CCTV front

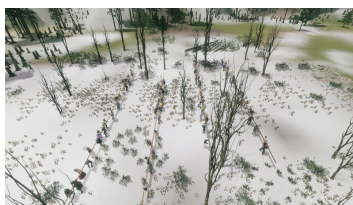


CCTV back

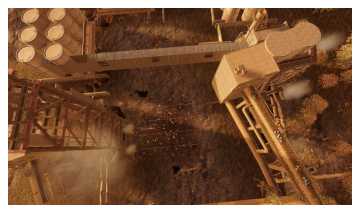
(3) Tug-of-War



UAV low-alt



CCTV front



UAV high-alt



CCTV back

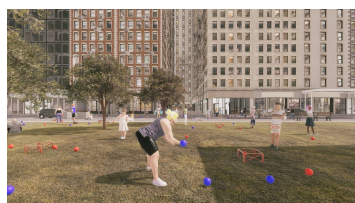


UAV low-alt

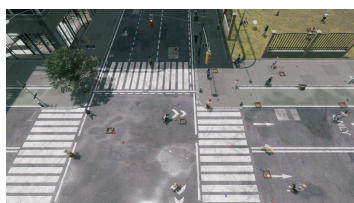


UGV

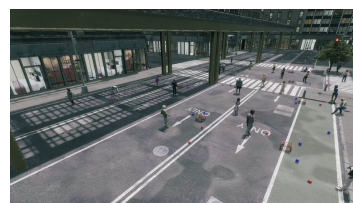
(4) Marbles



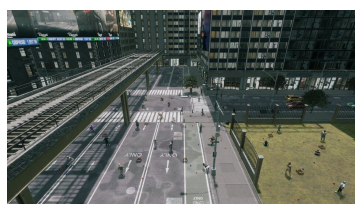
UGV



CCTV front



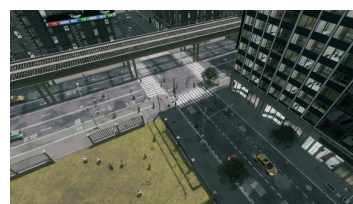
UAV med-alt



CCTV side

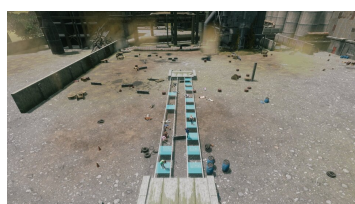


UAV med-alt

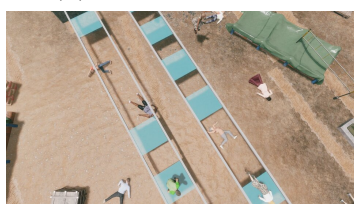


UAV low-alt

(5) Stepping Stones



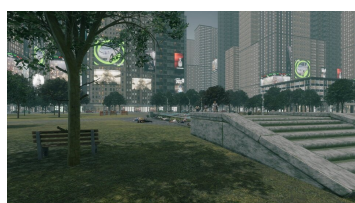
CCTV front



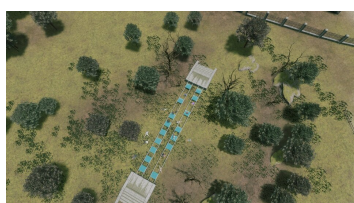
UAV med-alt



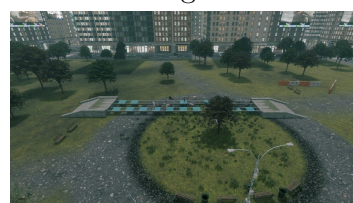
UAV high-alt



UGV



UAV med-alt



CCTV side

(6) Squid Game



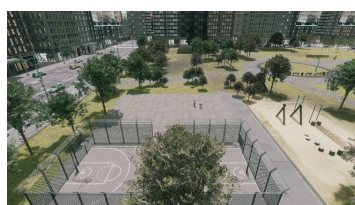
UGV



CCTV front



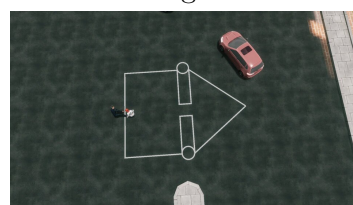
UAV high-alt



CCTV side



UGV



UAV low-alt

Figure 3.16: More example images from SynPlay are shown for all six Korean traditional games, each with various camera viewpoints.

Chapter 4

Task Evaluation

In line with the inherent purpose of synthetic data to serve as supplemental training data, we use the entire SynPlay dataset to train models for various computer vision tasks and evaluate its positive impact towards task performance. Our major counterpart models in evaluation are *trained-from-scratch*, which are trained only on real images (denoted as `real` in evaluation tables). We also validate the advantage of using SynPlay over other synthetic datasets.

In this experiment, we evaluate the effectiveness of the SynPlay dataset across two primary computer vision tasks: human detection and segmentation. These tasks are particularly challenging due to the need to identify diverse human appearances captured from varying distances in images.

4.1 General tasks: detection and segmentation

We evaluate the SynPlay dataset on two general vision tasks, human detection and segmentation. These tasks require the ability to identify diverse human appearances in images captured at a distance. To leverage synthetic data during training, we adopt a *pretrain-finetune* strategy, where a model is pre-trained on synthetic data and fine-tuned on target real-world data. The detectors used in the experiments are YOLO v8 models [25] with three different architecture sizes (small, medium, and

large). Mask2former [9] with the Swin-Base [30] backbone was used for segmentation. For evaluation metrics, we use the COCO-style APs [11] which are two bounding box APs (AP^{bb} and AP_{50}^{bb})¹ for human detection and Intersection-over-Union (IoU) for the segmentation.

The main tasks are conducted on aerial-view datasets, which feature a wider range of human appearances, making them ideal for validating the design philosophy behind the SynPlay dataset. We also conduct experiments on ground-view datasets to evaluate SynPlay on a more widely studied task in the community.

4.1.1 Experiment Setting

In our experiments, the goal is to explore SynPlay’s efficacy as supplementary training data across various tasks. We largely adhere to the original settings and implementations of the methods, with minimal modifications tailored to our specific experiments:

- **Architecture Modification:** Given that human detection and semantic segmentation can be considered as one-class problems, we adjust method architectures, particularly the dimensions of the last layer, accordingly.
- **Image Size in YOLOv8:** During training and inference of YOLOv8, we use an image size of 1280×1280 for most datasets, except COCO, which uses 640×640 . This decision is based on the original image sizes of the datasets, ensuring consistency in the range of human instances across datasets.

¹Detection accuracy for each model in the following tables are reported with two numbers in the form of AP^{bb}/AP_{50}^{bb} .

- **Training Mask2Former without the large-scale jittering (LSJ) augmentation [18].:** We did not use the default LSJ augmentation in training the Mask2Former segmentation models solely for performance reasons. In all cases, segmentation accuracy was found to be significantly lower when LSJ augmentation was used. LSJ augmentation, which greatly expands the range of image scaling, may not be suitable for aerial-view detection, which mainly includes small-sized human instances. This performance degradation with LSJ augmentation is also observed in [21], which is a reputable literature in the field of self-supervised learning.

By combining these experimental approaches, we aim to comprehensively assess the utility and performance of SynPlay in enhancing computer vision tasks across diverse datasets and scenarios.

4.1.2 Aerial-view tasks

Table 4.1 shows the results for aerial-view human detection and semantic segmentation tasks. Overall, for both tasks, using SynPlay in training provides remarkably better accuracy than all the compared cases, including *real* and all the other variations involving other synthetic data.

Notably, the results show that warming up the model with synthetic data before incorporating real data generally does not improve performance, except in the case of SynPlay. In other words, unless the synthetic dataset is properly designed and constructed, we cannot expect performance improvement simply from adding

Table 4.1: **Comparison with other synthetic datasets on aerial-view human detection and semantic segmentation.** The numbers in parentheses are the gaps from the model trained without synthetic data (‘real’). Positive and negative gaps are indicated in **green** and **red** fonts, respectively. The best accuracy for each setting is shown in **bold**. *Notations:* ‘+ real’ represents a model pre-trained with synthetic data and fine-tuned on a ‘real’ dataset, where ‘real’ is a training set derived from the dataset used for evaluation. ‘s’, ‘m’, and ‘l’ represent three YOLO v8 models with different architectures.

data in training	human detection									Semantic Drone [24]					
	VisDrone [55]			Okutama-action [4]											
	s	m	l	s	m	l	s	m	l	s	m	l	s	m	l
real	19.72/47.43	21.14/49.52	21.60/51.10	27.40/75.17	28.99/76.60	31.53/78.78	44.00/	77.20	44.52/	78.52	42.62/	79.87			
Archangel [45]	0.23/ 0.63	0.38/ 0.98	0.59/ 1.48	2.59/ 8.45	3.90/10.13	2.83/ 9.12	0.64/	1.59	2.42/	5.37	0.94/	1.62			
SynDrone [40]	0.31/ 0.81	0.36/ 0.84	0.71/ 1.89	0.00/ 0.00	0.00/ 0.01	0.00/ 0.00	0.00/	0.00	0.00/	0.00	0.00/	0.00			
SynPlay	5.29/11.75	4.31/ 9.12	2.79/ 5.87	12.74/40.86	8.19/25.43	8.15/25.23	7.02/	12.21	9.60/	15.51	15.71/	23.59			
Archangel + real	18.77/45.39	20.25/48.52	20.82/49.51	30.72/80.35	32.36 /80.63	31.71/79.63	46.60/	74.07	48.60/	75.86	44.62/	73.23			
	(-0.95/-2.04)	(-0.89/-1.00)	(-0.78/-1.59)	(+3.32/+5.18)	(+3.37/+4.03)	(+0.18/+0.85)	(+2.60/	-2.13)	(+4.08/	-2.66)	(+1.00/	-6.64)			
SynDrone + real	18.78/45.79	20.94/49.44	21.97/51.51	29.70/77.71	31.39/79.42	31.24/78.71	50.93/	82.28	53.71/	85.47	59.59/	85.02			
	(-0.94/-1.64)	(-0.20/-0.08)	(+0.37/+1.41)	(+2.30/+2.54)	(+2.40/+2.82)	(-0.29/-0.07)	(+6.93/	+5.08)	(+9.19/	+6.95)	(+16.97/	+5.15)			
SynPlay + real	20.88/49.31	22.34/52.12	22.98/52.93	32.47/81.60	31.96/ 81.13	33.17/82.52	66.52/	90.33	69.46/	91.35	68.82/	91.37			
	(+1.16/+1.88)	(+1.20/+2.60)	(+1.38/+1.83)	(+5.07/+6.43)	(+2.97/+4.53)	(+1.64/+3.74)	(+22.52/+13.13)	(+24.94/+12.83)	(+26.20/+11.50)						

data in training	semantic seg	
	Semantic Drone [24]	Aeroscapes [35]
real	0.66	22.25
Archangel [45]	0.74	0.04
SynDrone [40]	0.07	0.00
SynPlay	8.03	6.44
Archangel + real	9.28	20.61
	(+8.62)	(-1.64)
SynDrone + real	5.56	24.59
	(+4.90)	(+2.34)
SynPlay + real	23.32	32.19
	(+22.66)	(+9.94)

synthetic data to the training process. Moreover, among cases using synthetic data only in training, SynPlay presents unparalleled accuracy. In fact, the results using other sources of synthetic data are so poor that the other sources can be considered useless for this type of dataset utilization. Based on these two observations, *our design strategies for enhancing the diversity and realism of human appearance are shown to be highly effective in meeting expectations.*

4.1.3 Ground-view tasks

Table 4.2 explores the impact of using SynPlay for the general computer vision tasks of ground-view human detection and semantic segmentation. We also evaluate how models perform when only the subset with matching viewpoint (i.e., UGV images in SynPlay) is used in training. Overall, using the entire SynPlay yields the highest accuracy on both tasks, while using the UGV-subset still outperforms the model trained without SynPlay. These results demonstrate that our insight in ensuring diversity by varying the camera viewpoints is effective even in tasks that do not contain such multiple viewpoints. In addition, the greater improvement in semantic segmentation over object detection shows that ensuring diversity is more effective in tasks that require more detailed human representation models.

4.1.4 Combination with MS-COCO for pre-training dataset

The effect of using pre-training can be greater when applying two or more datasets with complementary properties. Here, we aim to investigate the potential synergy achieved by integrating MS COCO, a real dataset primarily comprising ground-view images, with SynPlay for the task of aerial-view human detection. Table 4.3, shows all combinations of SynPlay and MS COCO datasets when used for pre-training. The anticipated synergistic effect appears in all cases except in one case (AP_{50}^{bb} results on Okutama-action) when fine-tuned on the target dataset. Moreover, using *SynPlay only* provides comparable accuracy to using MS COCO when used indirectly through fine-tuning on the real dataset.

Table 4.2: **Impact of SynPlay on MS COCO** (person category). *Notation:* ‘SynPlay-UGV’ and ‘SynPlay-all’ are the UGV subset of SynPlay and the entire SynPlay, respectively.

data in training	(a) human detection			(b) sem.seg.
	s	m	l	
real	46.19/65.91	50.10/69.86	52.52/72.15	15.10
SynPlay-UGV	46.53/66.18	50.70/70.37	52.69/72.29	20.18
+real	(+0.34/+0.27)	(+0.60/+0.51)	(+0.17/+0.14)	(+5.08)
SynPlay-all	46.84/66.70	51.12/70.74	53.00/72.59	21.57
+real	(+0.65/+0.79)	(+1.02/+0.88)	(+0.48/+0.44)	(+6.47)

Table 4.3: **Synergy impact with MS COCO** on aerial-view human detection. YOLO v8 model with a medium size architecture.

data in training	VisDrone	Okutama-action	Semantic Drone
real	21.14/49.52	28.99/76.60	44.52/78.52
COCO	7.16/16.46	15.17/48.28	34.74/56.39
SynPlay	4.31/ 9.12	8.19/25.43	9.61/15.52
COCO + SynPlay	11.49/25.20	14.68/49.82	18.60/31.03
COCO + real	22.11/51.73	32.26/80.10	65.72/89.20
SynPlay + real	22.34/52.13	31.96/ 81.13	69.46/91.35
COCO + SynPlay + real	22.78/53.01	33.82/79.44	73.52/92.80

Interestingly, the results without fine-tuning show a different trend. On Okutama-action and Semantic Drone cases, using MS COCO performed better than other two baselines, with SynPlay specifically showing a much lower accuracy. We observe that synthetic data still lags behind real-world data in many respects, highlighting the need for further research to bridge the gap.

4.2 Data-scarce tasks: Few-shot and cross-domain learning

In this section, we compare SynPlay with other synthetic datasets on its ability to meet the demand for additional data in data-scarce tasks. For data-scarce tasks, we adopt few-shot and cross-domain learning tasks on aerial-view human detection,

which suffers more severely from a lack of training data than ground-view detection. Following the data-scarce task setups of [43], we train models with two few-shot regimes using 20 and 50 images of VisDrone (denoted by ‘Vis-20/50’). We test the models on VisDrone, Okutama-action, and Semantic Drone where the evaluations done on the last two datasets can be seen as ‘cross-domain’. To attenuate the potential random effects that may arise when selecting real training images, all reported numbers are average accuracy over three runs.

As baseline methods leveraging synthetic data in training, we use a pretrain-finetune strategy (PT-FT) and Progressive Transformation Learning (PTL) [43]. PTL is a progressive data augmentation approach that iteratively expands the training set by adding a subset of synthetic data, which is transformed to look real. In each PTL iteration, a subset of the synthetic data is selected, such that synthetic data that is closer to the real dataset is selected more often. For the data-scarce tasks experimented in [43], PTL was better than PT-FT while both outperformed the cases without synthetic data. We used RetinaNet [28] as the detector.²

4.2.1 Experiment Setting

- **Settings for PT-FT.** When using PT-FT in the general tasks, training specifications, including training epochs and learning rate, did not differ between pre-training and fine-tuning. In data-scarce tasks, we follow all the settings of [43] as outlined in PTL, while leaving out the progressive component.

²The currently available implementation of PTL is tailored to RetinaNet. For a fair comparison between PTL and PT-FT, we used RetinaNet instead of YOLO v8 for this experiment.

Table 4.4: **Few-shot and cross-domain learning accuracy (AP^{bb}/AP_{50}^{bb})** on aerial-view human detection. To clarify, the accuracy on Okutama-action and Semantic Drone refers to cross-domain learning performance. *Notation*: ‘Archangel*’ is an expanded ‘Archangel’ to be pose-diversified [44].

data in training	method	Vis-20						Vis-50					
		VisDrone		Okutama-action		Semantic Drone		VisDrone		Okutama-action		Semantic Drone	
real		0.58/ 2.27		3.64/ 14.54		0.62/ 1.89		0.76/ 3.30		7.82/ 28.66		1.30/ 5.65	
+ Archangel	PTL	2.07/ 6.72		7.90/ 31.53		8.81/ 33.71		2.92/ 9.26		11.49/ 42.51		8.98/ 33.21	
+ Archangel*		2.26/ 7.39		8.95/ 36.97		6.45/ 26.13		2.99/ 9.42		12.89/ 47.24		6.29/ 25.50	
+ SynPlay		3.08/ 9.03		14.39/ 49.53		6.94/ 24.22		3.71/11.20		15.67/ 52.06		7.74/ 26.99	
		(+2.50/+6.76)		(+10.75/+34.99)		(+6.32/+22.33)		(+2.95/+7.90)		(+7.85/+23.40)		(+6.44/+21.34)	
+ Archangel	PT-FT	0.76/ 2.48		4.24/ 17.17		6.53/ 23.67		1.29/ 3.76		5.32/ 20.96		7.10/ 27.95	
+ Archangel*		1.21/ 4.02		9.14/ 34.70		8.20/ 28.80		1.84/ 5.37		10.39/ 36.83		8.63/ 30.09	
+ SynPlay		2.94/ 9.38		12.19/ 40.88		11.32/ 37.30		3.72/11.87		13.66/ 44.06		12.76/ 41.66	
		(+2.36/+7.11)		(+8.55/+26.34)		(+10.70/+35.41)		(+2.96/+8.57)		(+5.84/+15.40)		(+11.46/+36.01)	

- **Settings for data-scarce tasks.** For all experiments performed for data-scarce tasks including the scaling behavior study, we follow all the settings and experimental environments of [44].

4.2.2 Comparison with other synthetic data

In Table 4.4, we compare the detection accuracy of the models trained with different synthetic datasets on the few-shot and cross-domain learning tasks. With PT-FT, SynPlay achieved significantly better accuracy than other synthetic datasets across all three datasets. With PTL, SynPlay performed the best on VisDrone and Okutama-action. These results were consistent for both Vis-20 and Vis-50 settings. Even on Semantic Drone, which shows an unusual performance trend, the best performance was achieved when SynPlay was used via PT-FT.

In addition, compared to SynPlay’s performance improvement on general tasks (in Table 4.1), the improvement achieved on data-scarce settings by SynPlay is much greater on VisDrone and Okutama-action on data-scarce tasks. This demonstrates that *SynPlay effectively meets the demand for additional data in data-*

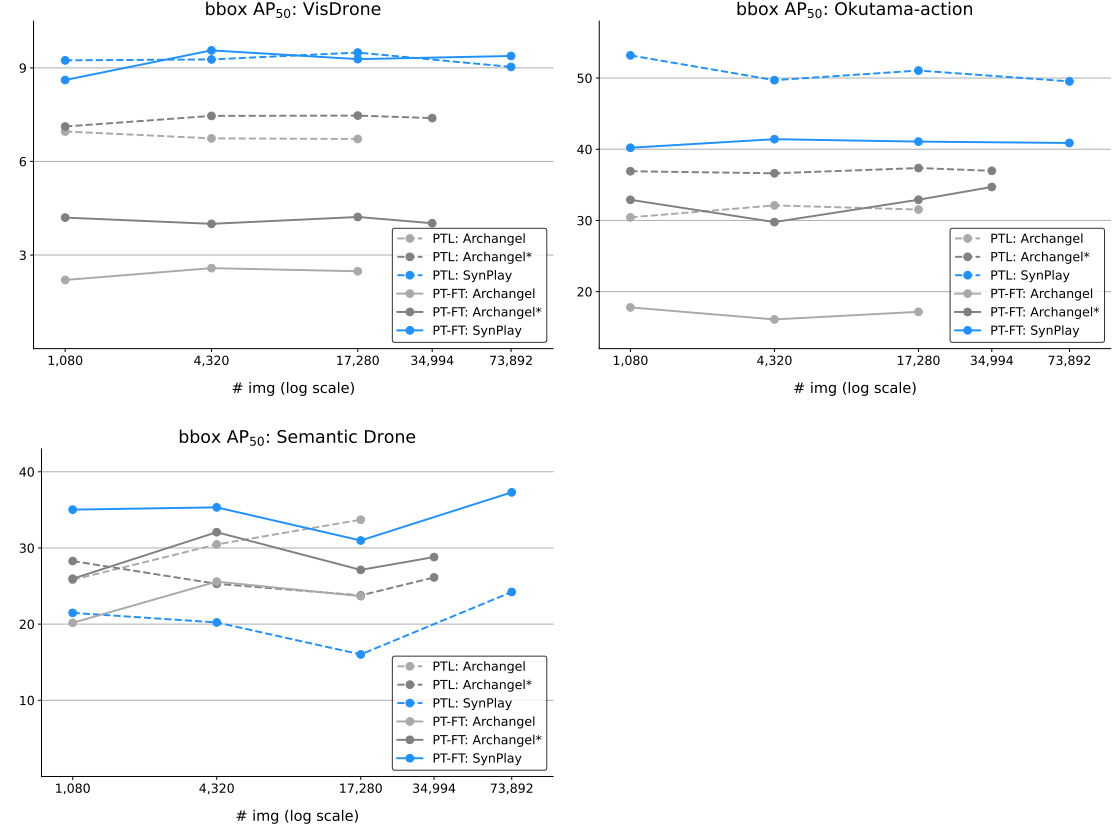


Figure 4.1: **Scaling behavior of synthetic datasets** under the Vis-20 setup (AP_{50}^{bb}). Scaling behavior of each dataset is compared by randomly sampled subsets of 1,080, 4,320, and 17,280 images, which correspond to 1/16th the size, 1/4th the size, and the size of Archangel. For reference, the sizes of Archangel* and SynPlay are 34,994 and 73,892, respectively.

scarce settings, which is greater than that in general tasks. We will discuss the unexpected performance trends on Semantic Drone in more detail in Chapter 5.

4.2.3 Scaling behaviors

To fairly validate the performance comparison without being affected by dataset size, we explore the scaling behavior of synthetic datasets. In Fig 4.1, we compare the detection accuracy of three synthetic datasets at multiple points where the datasets are randomly sampled to have the same size. On all three test sets, the

Table 4.5: **FID comparison.** In FID calculation, VisDrone serves as a reference representing real aerial-view human data.

COCO	Archangel	Archangel*	SynDrone	SynPlay
48.16	67.20	67.20	21.66	18.36

best performing models use SynPlay in training, *i.e.*, SynPlay + PTL on VisDrone and Okutama-action, SynPlay + PT-FT on Semantic Drone. *The performance gain achieved using SynPlay is not simply due to the large size of the dataset.*

4.3 Image Quality Evaluation

In Table 4.5, for all training datasets involved in our experiments, we calculate FID (Fréchet Inception Distance) [22] to assess their fidelity and diversity. We utilized the PyTorch implementation of FID in [42] with the default setup to assess the fidelity and diversity for all the training datasets involved in our experiments. We did not perform image scaling on the input for any dataset, and the final average pooling features were used to compute FID.

SynPlay presents the best score compared to other synthetic datasets, which is a result that aligns well with our task results. These results suggest that SynPlay’s superior task performance is achieved by better fidelity and diversity, which are our goals in designing SynPlay. Moreover, the better FID of SynPlay compared to MS COCO, which mainly includes ground-view images, is also reported, supporting the hypothesis that adopting multiple viewpoints effectively diversifies human appearances.

Chapter 5

Discussion and Conclusion

5.1 Discussion

Peculiar performance trend on Semantic Drone: The following conflicting phenomena were observed in experimental results (Table 4.4) when tested on Semantic Drone:

- *When using synthetic data in training via PTL for data-scarce tasks, involving SynPlay under-performed when compared to cases using other synthetic datasets.*
- *The performance gain acquired by incorporating synthetic data during training is remarkably large for Semantic Drone when compared to the other two cases (VisDrone, Okutama-action), with SynPlay’s incorporation showcasing the most substantial gain.*

We aim to find answers to the first conflicting phenomenon through the analysis of nadir-view instances. An instance with an elevation angle greater than 71.57° relative to the UAV is considered to be a nadir-view instance, representing the maximum elevation angle for Archangel [44]. To identify nadir-view instances for Archangel, we utilized the dataset metadata, i.e., UAV position. Similarly, for Archangel, we determined if an instance was a nadir-view instance based on the source instance, also using the dataset metadata. In the case of SynPlay, we computed

Table 5.1: **Proportion of nadir-view instances in synthetic datasets** used in data-scarce tasks. Instances with camera viewing angle from ground greater than 71.57° are considered as nadir-view instances.

Archangel	Archangel*	SynPlay
25.00%	12.82%	4.24%

the elevation angle for each instance using the absolute 3D coordinates of the instance and the UAV provided by SynPlay.

Most human instances in Semantic Drone are taken from nadir views, while VisDrone and Okutama-action have instances captured with fewer nadir-views. The portion of nadir-view instances in SynPlay is the smallest among the synthetic datasets used in data-scarce tasks (Table 5.1). As PTL continues to prioritize samples from synthetic data that closely resemble the seed data (*i.e.*, VisDrone) for training, the reduced selection of nadir-view instances from SynPlay may result in a lower gain (first phenomenon). On the other hand, the second phenomenon indicates that ensuring greater diversity via using supplemental synthetic data has greater impact on Semantic Drone, which lacks diversity due to its limited viewpoints.

Moreover, SynPlay, which is less similar to Semantic Drone while being more diverse in relation to the compared synthetic datasets, shows the largest impact, supporting our claim that improving the diversity is generally effective in constructing better synthetic data.

Broader impact. Utilizing real human datasets frequently entails inherent privacy concerns. We hope that our endeavors to enhance synthetic humans data, moving it one step closer to real-world fidelity, will contribute to alleviating these challenges.

Limitations. SynPlay was developed to provide richer representations of human appearance for tasks that involve localizing human in the scenes. We recognize the significance of incorporating distinctive features from diverse object categories. For future work, we are aiming to expand SynPlay to encompass a wider array of categories, thereby enriching its training capabilities.

5.2 Conclusion

What motion a human performs and *where a person is viewed from* are two crucial factors that make a difference in *how a human looks*. We create a synthetic human dataset called *SynPlay* with the aim of expanding the realism of human appearance by diversifying these factor. Enhancing the diversity allowed SynPlay to have a greater positive impact towards model training when compared to (train-from-scratch, using other synthetic data) on aerial-view/ground-view object detection and semantic segmentation. This positive impact of SynPlay becomes even greater in data-scarce tasks, where synthetic data is strongly desired as supplemental training data.

Bibliography

- [1] Adobe: Mixamo, <https://www.mixamo.com/#/>
- [2] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: Proc. CVPR (2014)
- [3] Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Comput. Vis. Image Underst.* **167**, 50–62 (Feb 2018)
- [4] Barekatain, M., Martí, M., Shih, H.F., Murray, S., Nakayama, K., Matsuo, Y., Prendinger, H.: Okutama-action: An aerial view video dataset for concurrent human action detection. In: Proc. CVPR Workshop (2017)
- [5] Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proc. CVPR (2023)
- [6] Blender Institute: Blender, <https://www.blender.org>
- [7] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proc. NeurIPS (2020)
- [8] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. ICML (2020)
- [9] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proc. CVPR (2022)
- [10] Cioppa, A., Giancola, S., AdrienDeli’ege, Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Droogenbroeck, M.V.: SoccerNet-Tracking: Multiple object tracking dataset and benchmark in soccer videos. In: Proc. CVPRW (2022)
- [11] COCO Consortium: COCO - common objects in context. <https://cocodataset.org/> (nd), accessed: July 18, 2024
- [12] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., ad Uwe Franke, R.B., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. CVPR (2016)
- [13] Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In: Proc. ICCV (2023)
- [14] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009)

- [15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. ICLR (2021)
- [16] Epic Games: Unreal engine, <https://www.unrealengine.com>
- [17] Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **111**(1), 98–136 (Jan 2015)
- [18] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proc. CVPR (2021)
- [19] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(1), 142–158 (Jan 2016)
- [20] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: Proc. NeurIPS (2020)
- [21] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proc. CVPR (2022)
- [22] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Proc. NeurIPS (2017)
- [23] Hwang, D.h.(Writer and Director):. Squid game (2021), <https://www.netflix.com/title/81040344?source=35>
- [24] Institute of Computer Graphics and Vision, Graz University of Technology: Aerial semantic segmentation drone dataset. <http://dronedataset.icg.tugraz.at>
- [25] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (2023), <https://github.com/ultralytics/ultralytics>
- [26] Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-Art: A versatile human-centric dataset bridging natural and artificial scenes. In: Proc. CVPR (2023)
- [27] Lee, H., Eum, S., Kwon, H.: ME R-CNN: Multi-expert R-CNN for object detection. *IEEE Trans. Image Process.* **29**, 1030–1044 (Sep 2019)
- [28] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. ICCV (2017)

- [29] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft COCO: Common objects in context. In: Proc. ECCV (2014)
- [30] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. ICCV (2021)
- [31] Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: UAVid: A semantic segmentation dataset for uav imagery. ISPRS J. Photogramm Remote Sens. (P&RS) **165**, 108–119 (Jul 2020)
- [32] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proc. ECCV (2018)
- [33] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proc. ICCV (2019)
- [34] MakeHuman Community: MakeHuman, <https://static.makehumancommunity.org/makehuman.html>
- [35] Nigam, I., Huang, C., Ramanan, D.: Ensemble knowledge transfer for semantic segmentation. In: Proc. WACV (2018)
- [36] Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: Proc. CVPR (2021)
- [37] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proc. CVPR (2019)
- [38] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (Jun 2016)
- [39] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Proc. ECCV (2016)
- [40] Rizzoli, G., Barbato, F., Caligiuri, M., Zanuttigh, P.: Syndrone-multi-modal UAV dataset for urban scenarios. In: Proc. ICCV Workshop (2023)
- [41] Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH: Aspatio-temporal maximum average correlation height filter for action recognition. In: Proc. CVPR (2008)
- [42] Seitzer, M.: pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid> (August 2020), version 0.3.0

- [43] Shen, Y.T., Lee, H., Kwon, H., Bhattacharyya, S.S.: Progressive transformation learning for leveraging virtual images in training. In: Proc. CVPR (2023)
- [44] Shen, Y.T., Lee, H., Kwon, H., Bhattacharyya, S.S.: Diversifying human pose in synthetic data for aerial-view human detection. arXiv:2405.15939 (2024), <https://arxiv.org/abs/2405.15939>
- [45] Shen, Y.T., Lee, Y., Kwon, H., Conover, D.M., Bhattacharyya, S.S., Vale, N., Gray, J.D., Leongs, G.J., Evensen, K., Skirlo, F.: Archangel: A hybrid UAV-based human detection benchmark with position and pose metadata. IEEE Access **11**, 80958–80972 (2023)
- [46] Stathopoulos, A., Han, L., Metaxas, D.: Score-guided diffusion for 3D human recovery. In: Proc. CVPR (2024)
- [47] Studio Chacre: The character creator, <https://charactercreator.org>
- [48] Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proc. CVPR (2019)
- [49] Unity Technologies: Unity, <https://unity.com/>
- [50] Unity Technologies: Animatorcontroller. <https://docs.unity3d.com/Manual/class-AnimatorController.html> (nd), accessed: July 18, 2024
- [51] Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proc. CVPR (2017)
- [52] Zhang, Q., Wang, L., Patel, V.M., Xie, X., Lai, J.: View-decoupled transformer for person re-identification under aerial-ground camera network. In: Proc. CVPR (2024)
- [53] Zhang, T., Xie, L., Wei, L., Zhuang, Z., Zhang, Y., Li, B., Tian, Q.: UnrealPerson: An adaptive pipeline towards costless person re-identification. In: Proc. CVPR (2021)
- [54] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proc. CVPR (2017)
- [55] Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 7380–7399 (Nov 2022)