ABSTRACT

Title of Dissertation: Spatial-Temporal Representation Learning For Aerial Perception

Ruiqi Xian
Doctor of Philosophy, 2026

Dissertation Directed by: Professor Dinesh Manocha
Department of Electrical and Computer Engineering

Aerial perception plays a central role in autonomous systems that rely on aerial, overhead, or elevated-view observations for tasks such as video understanding, localization, and multi-camera analysis. Compared with conventional ground-view perception, these settings introduce distinctive challenges, including severe viewpoint and scale variation, cluttered backgrounds, spatially sparse task-relevant foreground cues, temporal ambiguity, and substantial discrepancies across views, modalities, and camera networks. These characteristics make it difficult to directly apply standard visual learning methods developed for ground-view imagery and videos. In this dissertation, we address these challenges through the lens of representation learning for aerial perception, with a particular focus on how spatial structure, temporal dynamics, and cross-view consistency can be modeled more effectively.

Our central thesis is that effective aerial perception depends on representations that better capture action-relevant spatial cues, model temporal relationships more reliably, remain practical under real-world deployment constraints, and transfer across heterogeneous sensing conditions. Guided by this perspective, we develop a sequence of methods organized around three themes. First, we study

refined spatial-temporal representation learning for aerial video understanding, where we improve how aerial video representations capture spatially sparse action-relevant regions and temporally ambiguous motion under large background dominance and weak foreground visibility. In this theme, we develop methods for temporal feature alignment, informative frame sampling, and semantic guidance. Across representative aerial and video recognition benchmarks, our methods improve recognition accuracy by 2.2%–13.8% on UAV-Human, 6.8% on NEC-Drone, and 9.0% on Diving48.

Second, we study practical and scalable aerial perception under real-world constraints, where we develop aerial-specific methods that improve efficiency, reduce annotation dependence, and remain suitable for realistic deployment. In this theme, our deployment-oriented methods improve top-1 accuracy by 6.1%–7.4% on RoCoG-v2, outperform prior state of the art by 8.3%–10.4% on UAV-Human, and reach 95.9% accuracy on Drone Action, while also achieving 2$\times$ faster inference on an RB5 CPU. Complementing this, our self-supervised learning methods further improve top-1 accuracy by 2.9% on NEC-Drone and 5.8% on UAV-Human under limited annotation, while achieving 2$\times$–5$\times$ faster inference than supervised approaches with heavy test-time augmentation.

Third, we study transferable aerial perception across views, modalities, and camera networks, where we learn representations that preserve task-relevant consistency across aerial-ground correspondence and multi-camera settings. In this context, our methods enable more robust aerial-ground localization under cross-view, cross-modal, and scale discrepancies, and support calibration-free multi-camera association without requiring camera calibration or manual identity annotations.

Taken together, our contributions establish a unified perspective on aerial perception through representation learning that is spatially aware, temporally effective, practically deployable, and transferable across sensing conditions. We show that improving aerial perception is not only a matter of designing stronger task-specific predictors, but of learning representations that better reflect the spatial

structure, temporal dynamics, and sensing variability of aerial and elevated-view observations.

SPATIAL-TEMPORAL REPRESENTATION LEARNING FOR AERIAL PERCEPTION

by

Ruiqi Xian

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
2026

Advisory Committee:
Dr. Dinesh Manocha, Chair/Advisor
Dr. Mumu Xu, Dean's Representative
Dr. Shuvra S. Bhattacharyya
Dr. Pratap Tokekar
Dr. Michael Wilson Otte

Acknowledgments

I would like to express my heartfelt gratitude to everyone who has supported me throughout my Ph.D. journey at the University of Maryland.

First and foremost, I am deeply grateful to my advisor, Professor Dinesh Manocha, for his continuous guidance, support, and encouragement over the years. His mentorship has profoundly shaped my development as a researcher, and I sincerely appreciate the freedom, trust, and high standards he provided throughout my doctoral studies.

I would also like to thank my dissertation committee members for their time, valuable feedback, and thoughtful suggestions. Their questions and insights have helped strengthen this dissertation and have broadened my perspective on research.

I am grateful to my collaborators, lab mates, and colleagues for the many discussions, exchanges of ideas, and support along the way. I especially appreciate the positive and stimulating research environment that made this journey both rewarding and memorable.

I would also like to acknowledge my mentors and collaborators at NEC Laboratories America, especially Deep Patel and Iain Melvin, for their guidance and for providing me with valuable opportunities to learn and grow during my internship and collaborative research experience.

Beyond research, I am deeply thankful to my friends for their encouragement, understanding, and companionship throughout this journey. Their support helped me through many challenging moments.

Finally, I would like to thank my family and my girlfriend for their unconditional love, sacrifice, and unwavering belief in me. Their support has been the foundation of everything I have accomplished, and I am forever grateful to them.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background and Motivation

Recent advances in aerial and elevated sensing platforms, especially unmanned aerial vehicles (UAVs), have enabled a broad range of civilian and autonomous applications, including surveillance, traffic monitoring, disaster response, environmental inspection, infrastructure assessment, search and rescue, and navigation [7–9]. Compared with conventional ground-based sensing platforms, aerial systems provide wider spatial coverage, more flexible deployment, and observation viewpoints that are difficult to obtain from fixed cameras or vehicle-mounted sensors [10–12]. These advantages have made aerial and elevated-view perception increasingly important for large-scale real-world visual understanding and autonomous decision-making [13–15].

At the core of these applications lies *aerial perception*, which in this dissertation refers broadly to visual understanding from aerial, overhead, or elevated-view imagery and video [8, 13, 15]. This scope includes not only UAV-based sensing, but also related overhead and oblique-view settings that share common challenges such as large viewpoint variation, scale changes, sparse task-relevant foreground cues, cluttered backgrounds, and discrepancies across views or sensing conditions [5, 6, 16]. Depending on the application, aerial perception may involve action recognition, event understanding, localization, tracking, scene analysis, navigation, or correspondence learning across viewpoints, modalities, and camera networks [11, 14, 15]. Recent benchmarks and task-specific studies, including

Okutama-Action, Drone-Action, and UAV-Human, further highlight both the practical importance and the technical difficulty of visual understanding in aerial settings [5, 17, 18].

A useful perspective for studying these problems is *spatial-temporal representation learning*. In the broad sense, representation learning aims to learn features that capture the underlying structure of observations in a form that is useful for downstream tasks [19–21]. In visual perception, *spatial* representation learning concerns appearance, object structure, scene layout, and spatial relationships within an observation, while *temporal* representation learning concerns motion, temporal evolution, and dependencies across observations acquired over time [21–23]. Spatial-temporal representation learning therefore refers to learning feature representations that jointly encode informative structure across space and time, so that models can support tasks such as recognition, localization, prediction, tracking, and association in dynamic visual environments [21, 24, 25].

This representation-centric viewpoint has become increasingly important in modern visual understanding. Prior work has shown that strong performance in dynamic perception tasks depends heavily on the quality of learned spatial-temporal features, rather than only on task-specific prediction heads [2, 23, 26]. More recent architectures, including transformer-based models, further reinforce this trend by explicitly modeling relationships among spatial tokens and, when applicable, temporal tokens across observations [27–29]. Self-supervised and masked modeling approaches have also demonstrated that learning effective spatial-temporal representations is a central mechanism for improving generalization in visual understanding [25, 29, 30].

This perspective is particularly relevant to aerial perception because informative cues in aerial observations are often small, sparse, and unevenly distributed, while the underlying scene dynamics are affected by camera motion, viewpoint variation, occlusion, and large background regions [5, 16, 17]. As a result, successful aerial perception depends not only on recognizing what appears in individual

observations, but also on learning representations that capture where informative cues occur, how they are spatially organized, and how they evolve over time [21, 23, 24]. However, many existing visual learning methods have been developed primarily for conventional ground-view imagery and videos, and do not directly account for the distinctive visual characteristics and deployment constraints of aerial, overhead, and cross-view settings [6, 8, 15]. Recent surveys on drone vision and aerial action recognition similarly emphasize the need for learning methods that better address aerial-specific viewpoint changes, scale variation, sparse foreground signals, and real-world operational constraints [14–16].

Motivated by this gap, this dissertation studies *spatial-temporal representation learning for aerial perception*. The central premise is that effective aerial perception requires representations that are refined for aerial video understanding, practical and scalable under real-world constraints, and transferable across views, modalities, and camera networks. Guided by this premise, the dissertation is organized around three connected themes: refined spatial-temporal representation learning for aerial video understanding, practical and scalable aerial perception under real-world constraints, and transferable aerial perception across views, modalities, and camera networks.

1.2 Challenges in Aerial Perception

Although aerial and elevated sensing platforms provide important advantages for large-scale visual understanding, aerial perception remains substantially more challenging than conventional ground-view perception [8, 13, 15]. These challenges arise not only from the imaging geometry and motion characteristics of aerial or elevated viewpoints, but also from the practical conditions under which such data are collected, annotated, and deployed [11, 31, 32]. As a result, methods developed for standard

image recognition or ground-view video understanding often do not transfer effectively to aerial, overhead, or oblique-view scenarios [5, 6, 16, 33]. This section highlights the core challenges that motivate the study of spatial-temporal representation learning for aerial perception.

1.2.1 Spatial Challenges from Viewpoint, Scale, and Scene Composition

Aerial and elevated-view observations exhibit substantial spatial variability, meaning that the spatial appearance and arrangement of informative visual content can change significantly across observations due to viewpoint, altitude, camera orientation, scene coverage, and sensing configuration [8, 12, 34]. As a result, the same target or scene may appear drastically different across captures, even within the same task setting [5, 16, 33]. Targets of interest are often observed from overhead or oblique viewpoints, with visual patterns that differ substantially from the ground-view imagery on which many vision models are commonly developed or pretrained [6, 17, 35]. In addition, large scene coverage often causes foreground targets, such as people or vehicles, to occupy only a small portion of the image [5, 18, 34]. These difficulties are further amplified by cluttered backgrounds, weak local detail, and large intra-scene scale variation, all of which are widely recognized challenges in aerial and elevated-view vision [16, 35, 36].

These properties make aerial perception fundamentally different from standard visual recognition settings. Models must identify small, task-relevant regions within large and often cluttered scenes while remaining robust to changes in viewpoint, object scale, and scene composition [8, 16, 34]. Consequently, effective aerial perception requires representation learning strategies that preserve informative spatial structure and emphasize discriminative local cues rather than being dominated by global background appearance.

1.2.2 Temporal Complexity and Motion Ambiguity

Aerial perception also introduces substantial temporal challenges. In mobile aerial platforms such as UAVs, visual observations often contain strong camera ego-motion caused by platform movement, viewpoint adjustment, or trajectory changes [11, 14, 31]. Even in fixed elevated or multi-camera settings, temporal observations may still be affected by abrupt viewpoint changes, scale inconsistencies, occlusion, and background activity. These effects can obscure the underlying scene dynamics and make it difficult to separate target motion from camera-induced or viewpoint-induced appearance change [16, 37, 38]. Furthermore, aerial videos frequently contain long temporal spans with redundant or weakly informative observations, while action-relevant motion patterns may appear only briefly or in localized regions [5, 18, 39]. The temporal structure of aerial observations is therefore often noisy, sparse, and difficult to model consistently, especially when target motion, background motion, and camera effects are entangled [16, 31, 38].

These properties make temporal modeling particularly challenging, since effective understanding depends not only on what information is present, but also on when it is informative and how it evolves over time. Such challenges motivate spatial-temporal representations that emphasize informative temporal structure while suppressing noise introduced by camera motion and visual redundancy [3, 38, 39]. They also highlight the importance of learning mechanisms that focus computation on informative observations rather than treating all frames as equally useful [3, 39, 40].

1.2.3 Cross-View, Cross-Modal, and Cross-Camera Discrepancies

Many aerial perception problems extend beyond a single image or video stream. In practical settings, a perception system may need to align aerial observations with ground-view imagery, map-

based information, or other sensing modalities, or it may need to associate targets across multiple cameras with different viewpoints and observation conditions [15, 41, 42]. However, such scenarios introduce substantial discrepancies in appearance, geometry, scale, and semantic granularity. The same location may look dramatically different when observed from aerial and ground perspectives, and the same target may exhibit large appearance variation across camera networks with different viewpoints, heights, or scene layouts [6, 33, 42]. These cross-view, cross-modal, and cross-camera gaps make direct feature matching difficult and require learned representations that preserve task-relevant consistency despite changes in viewpoint, scale, modality, and sensing configuration.

As a result, transferability across observation conditions becomes a central requirement for aerial perception systems operating in realistic multi-view or multi-modal environments. Rather than being tied to a single sensing condition, representations should remain useful across viewpoints, sensing modalities, camera networks, and deployment environments, while still supporting robust correspondence, localization, and association.

1.2.4 Limited Supervision and Deployment Constraints

Aerial perception is also constrained by the cost and practicality of data collection, annotation, and deployment. Large-scale annotation of aerial imagery and video is expensive, especially for tasks involving long temporal sequences, rare events, dense correspondences, fine-grained action labels, or cross-camera identity association [5, 7, 18]. In many real-world settings, deployment cannot assume exhaustive manual labels, stable infrastructure, or computationally heavy inference-time preprocessing [10, 15, 32]. These constraints limit the applicability of methods that rely strongly on manual supervision or expensive processing pipelines [14, 15, 32]. As a result, aerial perception methods

must not only achieve strong accuracy, but also remain practical under limited supervision, restricted preprocessing capability, and real-world resource constraints.

These observations motivate learning strategies that reduce reliance on dense manual annotation and heavy preprocessing while maintaining strong representation quality. Such strategies are particularly important when aerial perception must be deployed at scale, adapted to new environments, or transferred across sensing conditions.

1.2.5 Implications for Representation Learning

Taken together, these challenges indicate that aerial perception is not simply a direct extension of conventional vision problems [8, 13, 15]. Instead, it requires representation learning methods that explicitly account for the distinctive spatial, temporal, and operational characteristics of aerial, overhead, and oblique-view observations. In particular, the challenges discussed above motivate three core directions for this dissertation.

First, aerial perception requires *refined spatial-temporal representation learning* that is better tailored to the spatial sparsity, background dominance, viewpoint variation, and temporal ambiguity of aerial observations. Effective models must better capture informative regions, emphasize useful temporal structure, and adapt learned features to aerial-specific visual conditions.

Second, aerial perception requires *practical and scalable learning under real-world constraints*, where annotation, computation, and preprocessing may all be limited. This calls for methods that remain effective under limited computation, restricted preprocessing capability, and limited supervision, and that can support deployment across larger datasets, longer observation streams, and more complex sensing environments.

Third, aerial perception requires *transferable representations across views, modalities, and camera networks*. Since many practical tasks involve aerial-ground correspondence, heterogeneous sensing conditions, or multi-camera association, learned representations must preserve task-relevant consistency even when the observation domain changes substantially.

These observations motivate the central research questions of this dissertation. The next section formalizes this perspective through the thesis statement and the main research questions that guide the remainder of this dissertation.

1.3 Thesis Statement and Research Questions

The discussions in the previous sections suggest that aerial perception should not be viewed merely as a collection of separate downstream tasks, but rather as a unified representation learning problem. Specifically, despite differences in task formulation, aerial video understanding, localization, multi-view correspondence, and multi-camera perception all depend on a common objective: learning feature representations from aerial, overhead, or elevated-view observations that preserve informative spatial structure, capture meaningful temporal dynamics, remain robust to viewpoint and sensing variation, and support reliable perception in real-world settings.

This shared perspective motivates the central thesis of this dissertation: effective aerial perception can be substantially improved by learning spatial-temporal representations that explicitly capture aerial-specific visual structure, remain practical under real-world computational and deployment constraints, and generalize across heterogeneous sensing conditions. Guided by this thesis, this dissertation investigates three research questions.

RQ1: How can we refine spatial-temporal representations for aerial video understanding? Aerial videos exhibit substantial viewpoint and scale variation, strong camera ego-motion, cluttered backgrounds, and small task-relevant foreground cues. These properties make it difficult for standard visual models to capture informative spatial structure and temporal dynamics effectively. This research question studies how aerial representations can be refined through improved temporal alignment, informative sampling, and semantically guided learning, so that they better support aerial video understanding.

RQ2: How can aerial perception remain practical and scalable under real-world constraints? In realistic deployment settings, aerial perception often faces limited computation, limited annotation, and restricted preprocessing capability. This research question investigates how to design learning strategies that remain effective under such constraints, including methods that improve efficiency, simplify deployment, and reduce reliance on dense manual supervision or computationally heavy processing.

RQ3: How can learned representations transfer across views, modalities, and camera networks? Many aerial perception systems must operate beyond a single observation stream, requiring correspondence across aerial and ground views, heterogeneous sensing modalities, or multiple cameras with different viewpoints and observation conditions. This research question studies how to learn representations that preserve task-relevant consistency despite changes in viewpoint, scale, modality, and sensing configuration.

Together, these research questions define the main scope of this dissertation. The next section provides an overview of how they are organized into the core themes of the dissertation.

1.4 Dissertation Overview

To address the research questions introduced above, this dissertation is organized around three complementary themes: *refined spatial-temporal representation learning for aerial video understanding*, *practical and scalable aerial perception under real-world constraints*, and *transferable aerial perception across views, modalities, and camera networks*(Figure 1.1). Rather than representing separate problem domains, these themes are developed under a unified perspective: aerial perception is studied throughout this dissertation as a spatial-temporal representation learning problem, where the goal is to learn feature representations that preserve informative spatial structure, capture meaningful temporal dynamics, remain practical for deployment, and generalize across sensing conditions.

Figure 1.1 provides a global overview of this dissertation. At the center of the dissertation is *spatial-temporal representation learning for aerial perception*, which serves as the common backbone connecting all chapters. Built on this foundation, the dissertation develops three thematic directions. The first theme focuses on improving the quality of learned representations for aerial video understanding. The second focuses on making aerial perception more practical and scalable under deployment constraints. The third focuses on improving representation transfer across views, modalities, and camera networks. Together, these themes organize the technical contributions of the dissertation into a coherent framework.

The first theme, *refined spatial-temporal representation learning for aerial video understanding*, focuses on improving how aerial videos are represented and understood. Compared with conventional ground-view videos, aerial observations often contain weak and spatially sparse action cues, substantial background clutter, strong viewpoint and scale variation, and noisy temporal structure [5, 16, 18]. These characteristics make effective aerial video understanding highly dependent on how informative

spatial and temporal cues are extracted, aligned, sampled, and semantically interpreted. This theme therefore studies methods that refine aerial video representations through stronger temporal alignment, more informative sampling, and improved semantic guidance.

The second theme, *practical and scalable aerial perception under real-world constraints*, focuses on making aerial perception more effective in realistic deployment settings. In practice, aerial and elevated-view perception systems often face limited computation, limited annotation, and constrained preprocessing capability [7, 15, 32]. These constraints require methods that are not only accurate, but also efficient, deployable, and scalable. This theme studies aerial-specific modeling strategies that improve perception quality while reducing reliance on excessive computation, dense manual supervision, or heavy processing pipelines.

The third theme, *transferable aerial perception across views, modalities, and camera networks*, focuses on how learned representations can remain useful when the sensing condition changes. Many aerial perception tasks extend beyond a single image or video stream and instead require correspondence across aerial and ground observations, heterogeneous sensing modalities, or multiple cameras with different viewpoints [15, 41, 42]. These settings introduce substantial discrepancies in appearance, geometry, scale, and semantic granularity, making direct matching or association difficult [6, 42]. This theme therefore studies representation learning strategies that preserve task-relevant consistency across diverse sensing conditions and support more reliable cross-view and cross-modal perception.

Although these themes emphasize different technical goals, they are unified by a shared representation learning viewpoint. The first theme strengthens how aerial visual information is represented within challenging video observations, the second addresses how such representations can be learned and deployed more practically, and the third examines how they can remain effective across changing sensing conditions. Together, they define the conceptual structure of the dissertation and motivate the

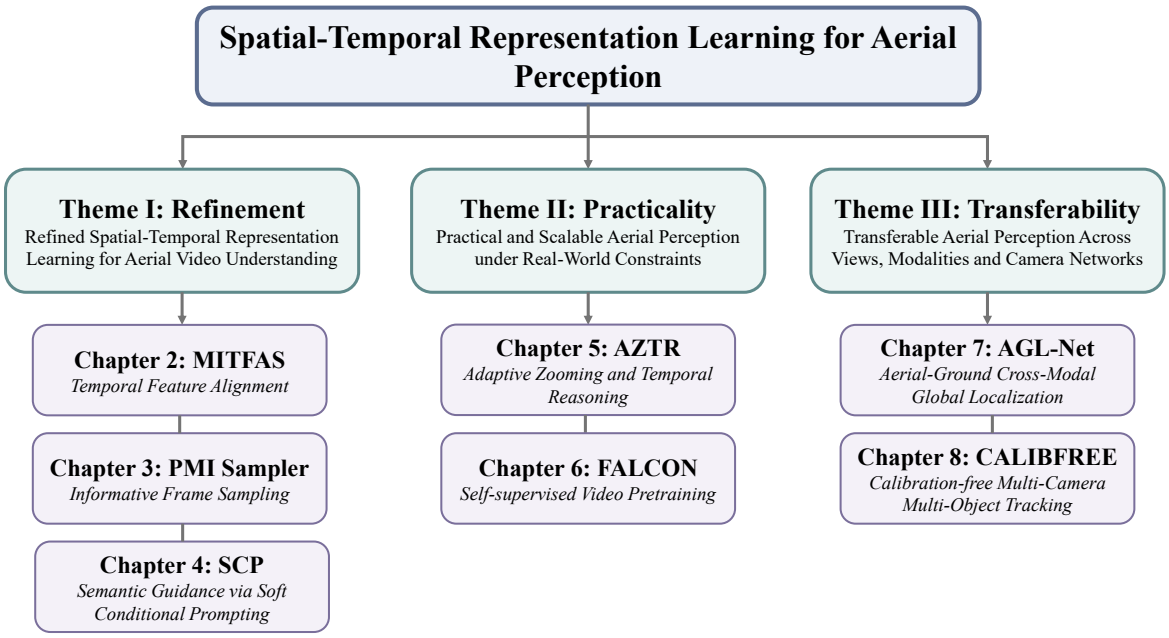


Figure 1.1: **Global overview of the dissertation.** The figure summarizes the dissertation structure under a unified spatial-temporal representation learning perspective, showing how the chapter-level contributions are organized into three complementary themes: refinement, practicality, and transferability.

chapter-level contributions presented in the following sections.

The next section summarizes the main contributions of this dissertation under these three themes, followed by an outline of the remaining chapters.

1.5 Summary of Contributions

The contributions of this dissertation are organized around three complementary themes: refined spatial-temporal representation learning for aerial video understanding, practical and scalable aerial perception under real-world constraints, and transferable aerial perception across views, modalities, and camera networks. Rather than constituting isolated task-specific advances, these contributions are connected by a common goal: learning representations that better capture aerial-specific spatial and

temporal structure, remain practical under deployment constraints, and transfer more effectively across heterogeneous sensing conditions.

1.5.1 Theme I: Refined Spatial-Temporal Representation Learning for Aerial Video Understanding

The first theme develops methods that improve the quality of aerial video representations. In particular, it studies how spatial-temporal features can be better aligned, sampled, and semantically guided so that aerial video understanding becomes more robust to motion ambiguity, viewpoint variation, cluttered backgrounds, and weak foreground cues.

- Chapter 2 develops a mutual-information-based temporal feature alignment framework for aerial action recognition. By explicitly encouraging more consistent temporal feature learning under motion ambiguity and visual variability, this chapter improves the model's ability to capture action-relevant temporal structure in aerial videos. Empirically, it yields up to 18.9% higher Top-1 accuracy than representative aerial action recognition baselines on UAV-Human, with additional gains of 7.3% on Drone-Action and 7.16% on NEC Drones, demonstrating the value of temporally aligned aerial video representations.

- Chapter 3 introduces a patch-similarity-guided informative frame sampling strategy for aerial videos. Instead of treating all frames as equally useful, it prioritizes observations that provide stronger spatial-temporal evidence, thereby reducing redundancy and improving representation quality. Empirically, this chapter yields relative gains of up to 13.8% over baseline sampling strategies on UAV-Human, together with improvements of 6.8% on NEC Drone and 9.0% on Diving48.

- Chapter 4 studies soft conditional prompt learning for aerial video action recognition. By improving the alignment between learned visual features and action semantics under aerial-specific viewing conditions, this chapter strengthens semantic discrimination in challenging aerial videos. Experimental results show improvements of up to 10.2% over existing aerial video recognition baselines on datasets such as Okutama and NECDrone, while also improving performance by up to 3.6% on SSV2, demonstrating both effectiveness on aerial data and generalization to ground-view video recognition.

Taken together, the contributions in this theme show that aerial video understanding benefits from more refined spatial-temporal representations that are better aligned in time, more selective in the observations they emphasize, and more effectively guided by aerial-specific semantics.

1.5.2 Theme II: Practical and Scalable Aerial Perception under Real-World Constraints

The second theme develops methods that make aerial perception more practical in realistic deployment settings. This includes improving inference efficiency, reducing reliance on manual annotation, and designing aerial-specific learning strategies that remain effective when computation, supervision, and preprocessing capability are limited.

- Chapter 5 develops an aerial video understanding framework based on adaptive zooming and temporal reasoning. By focusing computation more effectively on informative regions and motion patterns under challenging aerial viewpoints, this chapter improves perception quality while supporting more efficient deployment. Empirically, it improves Top-1 accuracy by 6.1%–7.4% over prior methods on RoCoG-v2, surpasses the previous state of the art by 8.3%–10.4% on

UAV-Human, and reaches 95.9% Top-1 accuracy on Drone-Action.

- Chapter 6 introduces a unified self-supervised video pretraining framework for aerial action recognition from raw RGB aerial videos. By combining object-aware masked autoencoding with object-centric dual-horizon future reconstruction, this chapter reduces reliance on manual annotation while learning stronger spatial-temporal representations for aerial video understanding. It improves downstream Top-1 accuracy by 2.9% on NEC-Drone and 5.8% on UAV-Human over strong pretrained baselines, while avoiding detector usage at inference and supporting a more practical deployment pipeline.

Taken together, these contributions show that aerial-specific modeling and self-supervised learning can improve not only perception quality, but also deployment efficiency, reduce annotation burden, and strengthen the practical scalability of aerial perception under real-world constraints.

1.5.3 Theme III: Transferable Aerial Perception Across Views, Modalities, and Camera Networks

The third theme develops methods for preserving task-relevant consistency when the sensing condition changes. It focuses on representation learning strategies that remain effective across aerial-ground correspondence, heterogeneous sensing modalities, and multi-camera environments with substantial viewpoint and scale discrepancies.

- Chapter 7 develops an aerial-ground cross-modal global localization framework that aligns local ground observations with global aerial context under varying scales. By improving representation transfer across views and modalities, this chapter enables more reliable localization under

strong appearance and scale discrepancies. Empirically, it reduces average localization error relative to prior cross-view localization baselines, improving performance on KITTI from 28.01 m to 18.02 m and on CARLA from 2.23 m / 19.15° to 1.83 m / 3.76°.

- Chapter 8 introduces a self-supervised calibration-free multi-camera multi-object tracking framework for cross-camera identity association without relying on camera calibration. By learning representations that preserve identity consistency across viewpoints and camera networks, this chapter improves cross-camera association robustness in realistic multi-view settings. Empirically, it improves overall accuracy by 3% and average F1 score by 7.5% over prior baselines on MMP-MvMHAT, while also demonstrating strong over-time tracking and cross-view association performance on the more diverse MvMHAT benchmark.

Taken together, these contributions establish a unified perspective on aerial perception through spatial-temporal representation learning. Across aerial video understanding, cross-modal localization, and multi-camera perception, this dissertation shows that effective aerial perception depends on representations that are refined for aerial-specific visual understanding, practical under real-world constraints, and transferable across diverse sensing conditions.

1.6 Organization of the Dissertation

The remainder of this dissertation is organized into three parts corresponding to the three main themes introduced in this chapter.

Part I: Refined Spatial-Temporal Representation Learning for Aerial Video Understanding. This part focuses on improving aerial video understanding through refined spatial-temporal representation learning. Chapter 2 studies temporal feature alignment for aerial action recognition. Chap-

ter 3 investigates informative frame sampling for aerial videos. Chapter 4 explores semantic guidance through soft conditional prompt learning for aerial video understanding.

Part II: Practical and Scalable Aerial Perception under Real-World Constraints. This part focuses on practical and scalable aerial perception under limited computation, annotation, and preprocessing capability. Chapter 5 studies adaptive zooming and temporal reasoning for efficient aerial video understanding. Chapter 6 explores self-supervised video pretraining for UAV action recognition under limited supervision.

Part III: Transferable Aerial Perception Across Views, Modalities, and Camera Networks. This part focuses on learning representations that remain effective across heterogeneous sensing conditions. Chapter 7 studies aerial-ground cross-modal global localization under varying scales. Chapter 8 investigates self-supervised calibration-free representation learning for multi-camera multi-object tracking.

Finally, Chapter 9 concludes the dissertation by synthesizing the main findings across all three parts, discussing the limitations of the current work, and outlining promising directions for future research in aerial perception and spatial-temporal representation learning.

Part I

Refined Spatial-Temporal Representation Learning for Aerial Video Understanding

Aerial video understanding is one of the most challenging settings in aerial perception because informative action cues are often spatially sparse, temporally ambiguous, and easily overwhelmed by large background regions. In this part, we study how spatial-temporal representations can be refined to better capture action-relevant information in aerial videos. In particular, we focus on three complementary directions: temporal feature alignment, informative frame sampling, and semantic guidance. Together, these methods improve how aerial videos are represented and interpreted under viewpoint variation, camera motion, background clutter, and weak foreground visibility.

Chapter 2: MITFAS: Mutual Information Based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition

2.1 Introduction

This chapter begins Part I, which focuses on refined spatial-temporal representation learning for aerial video understanding. As the first chapter in this part, we address a fundamental challenge in aerial action recognition: how can we extract spatially stable and temporally informative action representations from videos captured under severe viewpoint variation, camera motion, background clutter, and weak foreground visibility? To address this problem, we present **MITFAS** [38], short for *Mutual Information Based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition*.

Action recognition in aerial videos is substantially more challenging than in conventional ground-

view videos. Human actors typically occupy only a small portion of the frame, while the majority of pixels correspond to background regions [5, 16, 18]. In addition, because the video is captured from a moving aerial platform, the position, scale, and apparent orientation of the actor may vary significantly between adjacent frames. As a result, conventional spatial-temporal models may be forced to learn from unstable background changes and viewpoint variation rather than from action-relevant motion cues. These issues are further exacerbated by motion blur, partial occlusion, and the fact that not all frames in the sequence are equally informative for recognizing the action [8, 16].

Recent progress in video understanding has produced strong models for ground-view action recognition, including 3D ConvNet-based methods, efficient video architectures, and transformer-based temporal reasoning models [2, 23, 27, 28, 43]. However, these methods are not specifically designed for aerial videos, where weak actor visibility, strong background dominance, and platform motion create a more difficult spatial-temporal representation problem. More specifically, aerial videos present two coupled challenges. Spatially, the action-relevant foreground is small and unstable, so the model may attend to large but irrelevant background regions. Temporally, the apparent motion between frames is affected not only by the actor, but also by the motion of the aerial platform, which makes temporal learning less reliable.

MITFAS addresses these issues through two complementary ideas. First, we introduce *Temporal Feature Alignment (TFA)*, which uses mutual information to identify and align action-relevant regions across frames. Although motivated by temporal consistency, TFA plays an immediate spatial role: it stabilizes the spatial support of the actor across time by repeatedly locating the most relevant human-centered region in each frame. Second, we introduce *Mutual Information Sampling (MIS)*, which selects the most informative and distinctive frame sequence for recognition. By combining spatial support stabilization and temporal evidence selection, MITFAS improves both the spatial and temporal

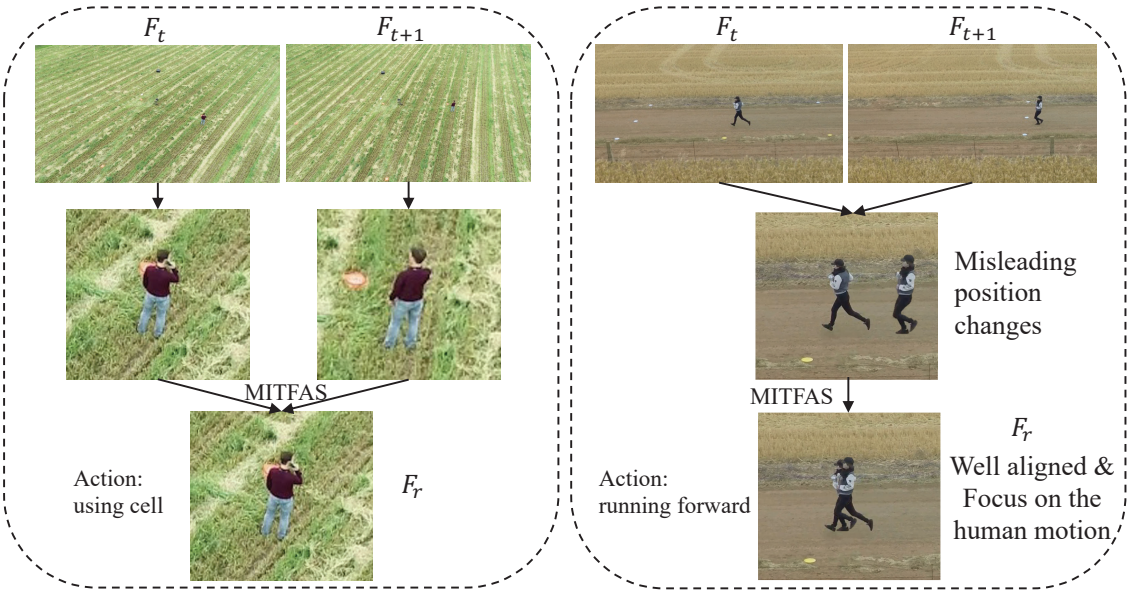
quality of aerial video representations.

Figure 2.1 illustrates the core challenge addressed in this chapter. In aerial videos, the human actor often occupies less than 10% of the frame, and its apparent position can shift significantly between adjacent frames due to platform motion. Under these conditions, raw temporal processing can easily be dominated by background change rather than by action-relevant motion. Figure 2.2 summarizes how MITFAS addresses this issue through spatial-temporal alignment followed by informative frame sampling.

From the perspective of this dissertation, MITFAS serves as the opening chapter of Part I by establishing a central idea that recurs throughout this theme: aerial video understanding can often be improved not simply by increasing model capacity, but by refining how spatial-temporal evidence is extracted and used. In particular, this chapter focuses on temporal feature alignment, while also showing that temporal alignment in aerial videos necessarily depends on stabilizing the spatial support of action-relevant regions.

The main contributions of this chapter are as follows:

- We develop a mutual-information-based spatial-temporal alignment method that identifies and aligns action-relevant regions across aerial video frames.
- We introduce a mutual-information-based frame sampling strategy that selects informative and distinctive temporal observations for aerial action recognition.
- We demonstrate strong quantitative improvements on multiple aerial action benchmarks. MITFAS improves Top-1 accuracy by **20.2%** over the baseline and **18.9%** over the prior state of the art on UAV-Human, improves the baseline by **16.6%** and the prior state of the art by **7.3%** on Drone Action, and achieves **78.62%** Top-1 accuracy on NEC-Drone, which is **7.18%** higher



(a) Target is too small, MITFAS will focus on the target. Target partially occluded, MITFAS will choose the more informative frame

(b) Misleading position changes, MITFAS will align the regions and focus on the human motion

Figure 2.1: **Illustration of the core challenges in aerial video action recognition.** The human actor in adjacent aerial frames often occupies less than 10% of the image and may shift significantly because of platform motion. MITFAS addresses this issue by identifying action-relevant regions, aligning them across time, and selecting more informative frames for recognition.

than the prior state of the art while using only half-resolution input frames [38].

2.2 Related Work

2.2.1 Video Action Recognition

Video action recognition has been extensively studied in conventional ground-view settings. Earlier deep learning approaches focused on two-stream architectures and 3D convolutional models that jointly capture appearance and motion [22, 23, 44]. More recent methods have emphasized efficient video architectures and transformer-based temporal reasoning, including X3D, MoViNet, ViViT, and TimeSformer [2, 27, 28, 43, 45]. These methods have achieved strong performance on large-scale

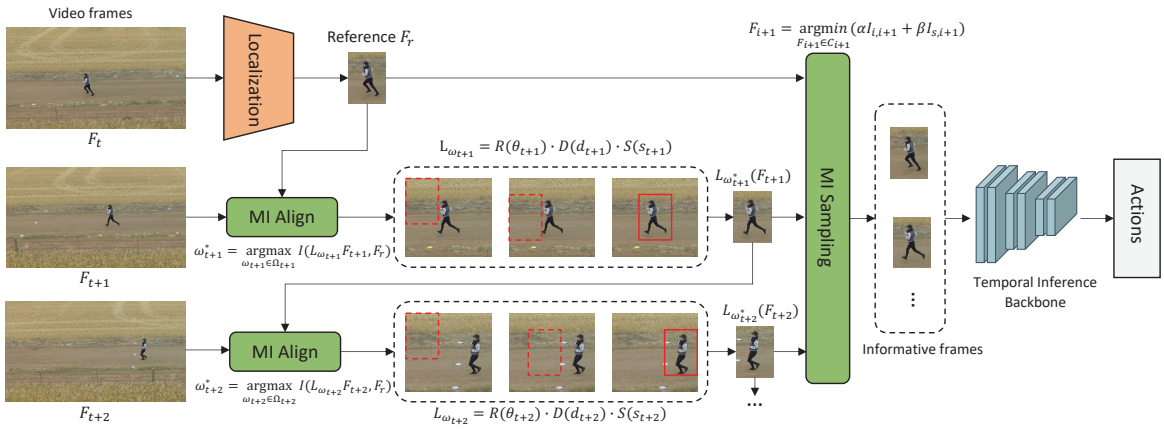


Figure 2.2: **Overview of MITFAS.** Given a starting frame F_t, we first localize the human action and crop a reference image F_r. For the next frame F_{t+1}, we estimate the optimal transformation parameter ω^*_{t+1} such that the mutual information between the transformed region $L_{\omega^*_{t+1}}(F_{t+1})$ and the reference image F_r is maximized. The aligned region is then used to update the reference for subsequent frames. After temporal alignment, we apply mutual-information-based sampling to select a distinctive and informative frame sequence, which is finally processed by a temporal recognition backbone such as X3D [2].

benchmarks, but they are primarily designed for scenarios in which actors occupy a substantial portion of the frame, viewpoints are relatively stable, and temporal variation is dominated by human motion rather than by platform motion.

In aerial videos, however, these assumptions often break down. The actor is typically small, the background dominates the frame, and camera motion can produce substantial apparent changes unrelated to the action itself. As a result, directly applying standard ground-view video models to aerial action recognition can lead to unstable temporal representations and poor use of the most informative frames.

2.2.2 Aerial Video Action Recognition

Aerial video action recognition has attracted increasing attention due to its relevance to surveillance, search and rescue, public safety, and robotic monitoring [7, 8, 16]. Public datasets such as

Okutama-Action, Drone Action, MultiViewPoint, NEC-Drone, and UAV-Human highlight the distinctive difficulties of aerial video understanding, including small actors, cluttered backgrounds, low-resolution humans, strong viewpoint change, and platform motion [5, 6, 17, 18, 33].

Prior methods have explored aerial-specific modeling, transfer learning, domain adaptation, and frequency-domain reasoning [6, 46–49]. While these methods improve robustness to aerial-view challenges, many still inherit temporal processing pipelines from conventional video recognition and do not explicitly address unstable temporal correspondence or uneven frame informativeness in aerial sequences.

2.2.3 Temporal Alignment and Correspondence

Temporal alignment is important whenever corresponding content shifts across frames due to viewpoint variation, motion, deformation, or partial occlusion. In conventional video settings, temporal correspondence is often handled implicitly through convolution, attention, or optical flow [23, 28, 45]. A related line of work studies feature alignment and correspondence more explicitly in image registration, tracking, and motion estimation, where the goal is to identify semantically or statistically consistent regions across observations [50–53]. These ideas are relevant because aerial videos also require identifying corresponding human-centered regions across time despite large apparent motion.

However, aerial videos introduce a more difficult alignment problem. The apparent motion between adjacent frames is influenced not only by the actor, but also by platform motion, scale change, background dominance, and viewing-angle variation. Under such conditions, dense correspondence or purely geometric alignment can be unreliable, especially when the actor occupies only a small region of the frame. MITFAS therefore differs from conventional temporal alignment approaches by using

mutual information to identify the most statistically consistent action-relevant region across frames. This makes the formulation more suitable for aerial videos, where exact geometric correspondence is difficult but statistical dependence between aligned regions remains informative.

2.3 Problem Formulation and Motivation

Consider an aerial video

$$V = \{F_1, F_2, \ldots, F_T\},$$

where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$ is captured by an aerial platform-mounted camera. The goal is to predict an action label

$$y \in \{1, \ldots, C\},$$

where C is the number of action categories.

In standard video recognition, the model learns a direct mapping from the input frame sequence to the action label. However, this direct formulation is problematic for aerial videos for two coupled reasons.

First, the spatial support of the action is weak and unstable. The actor often occupies only a small fraction of the frame, while the background dominates the image [5, 18]. Moreover, platform motion causes large changes in the apparent position, scale, and orientation of the actor between adjacent frames. This makes it difficult for the model to consistently focus on the correct spatial region.

Second, the temporal evidence itself is uneven. Not all frames are equally informative for recognition, because some frames may be blurred, partially occluded, or visually redundant. As a result, even if the actor is visible somewhere in the sequence, the model may still learn from weak or repeated

observations rather than from the most useful action cues.

These observations motivate a recognition strategy that explicitly improves both the spatial and temporal quality of the evidence supplied to the model. In particular, we seek a method that (1) aligns action-relevant content across time so that temporal learning becomes more spatially stable, and (2) selects the most informative temporal observations so that redundant or weak frames do not dominate recognition.

2.4 Method

2.4.1 Overview

MITFAS improves aerial action recognition by refining temporal evidence in two complementary ways. First, **Temporal Feature Alignment (TFA)** aligns action-relevant regions across frames so that the model observes more temporally consistent human-centered content despite viewpoint change and platform motion. Second, **Mutual Information Sampling (MIS)** selects the most informative and least redundant frame sequence for recognition. Together, these two components improve both the consistency and the informativeness of the temporal input before it is passed to the recognition backbone.

The overall pipeline is illustrated in Figure 2.2. Starting from an initial frame, we localize the human actor and construct a reference image. For each subsequent frame, TFA identifies the region that is most similar to the reference by maximizing mutual information. This aligned region is then used to update the temporal reference. After alignment, MIS selects a distinctive and representative frame sequence using mutual information and joint mutual information. The resulting sequence is finally processed by the action recognition backbone. The two stages serve different but complementary

roles: TFA refines *where* action-relevant evidence is extracted in each frame, while MIS refines *which* aligned observations are retained for recognition.

2.4.2 Temporal Feature Alignment

Temporal Feature Alignment is designed to improve temporal consistency in aerial videos by first stabilizing their spatial support. In this setting, human actors are often small and surrounded by large background regions. Due to platform motion and viewpoint change, the apparent position, scale, and orientation of the actor can vary substantially between adjacent frames. If temporal features are extracted directly from raw frames, the recognition model may be forced to learn from background variation rather than from action-relevant motion. Our goal is therefore to identify and align the dominant action-carrying region across time so that temporal learning becomes more spatially stable and semantically focused.

Let F_r denote a reference image containing a human-centered view of the actor. For each frame F_t, we seek a transformation parameter

$$\omega_t \in \Omega_t$$

such that the transformed region

$$L_{\omega_t}(F_t)$$

is best aligned with F_r. We model the transformation as a composition of rotation, translation, and scaling:

$$L_{\omega_t} = R(\theta_t) \cdot D(d_t) \cdot S(s_t), \tag{2.1}$$

where

$$\omega_t = (\theta_t, d_t, s_t).$$

The aligned region is obtained by maximizing mutual information:

$$\omega_t^* = \arg\max_{\omega_t \in \Omega_t} I(L_{\omega_t}(F_t); F_r), \tag{2.2}$$

where

$$I(L_{\omega_t}(F_t); F_r) = H(L_{\omega_t}(F_t)) + H(F_r) - H(L_{\omega_t}(F_t), F_r). \tag{2.3}$$

This objective encourages the selected region in the current frame to preserve the strongest statistical dependence with the reference image.

To compute mutual information between the transformed region and the reference image, we estimate their marginal and joint distributions using histograms. Let $v_{\omega_t}(p)$ denote the pixel value at position p in $L_{\omega_t}(F_t)$, and let $z_{\omega_t}(p)$ denote the intensity of the corresponding pixel in F_r. By binning all pixel pairs $(v_{\omega_t}(p), z_{\omega_t}(p))$, we obtain the joint histogram $h_{\omega_t}(v, z)$. The joint and marginal distributions are then approximated by

$$\begin{aligned} p_{VZ\omega_t}(v, z) &= \frac{h_{\omega_t}(v, z)}{\sum_{v,z} h_{\omega_t}(v, z)}, \\ p_{V\omega_t}(v) &= \sum_z p_{VZ\omega_t}(v, z), \\ p_{Z\omega_t}(z) &= \sum_v p_{VZ\omega_t}(v, z). \end{aligned} \tag{2.4}$$

The mutual information can then be written as

$$I(L_{\omega_t}(F_t); F_r) = \sum_{v,z} p_{VZ\omega_t}(v, z) \log \frac{p_{VZ\omega_t}(v, z)}{p_{V\omega_t}(v) p_{Z\omega_t}(z)}. \tag{2.5}$$

This formulation is particularly suitable for aerial videos because it measures statistical dependence rather than simple pixel similarity, making it more robust to appearance variation caused by viewpoint change, illumination change, and partial occlusion. Although the formulation is written at the image level, the same principle can also be interpreted at the feature level. If $M(\cdot)$ denotes a feature extractor, then TFA can be viewed as selecting the subset of features in the current frame that maximizes mutual information with the reference features:

$$M_s(F_t)^* = \arg \max_{M_s(F_t) \subset M(F_t)} I(M_s(F_t); M(F_r)). \tag{2.6}$$

This perspective helps explain why TFA improves downstream recognition: it does not merely crop the actor, but refines the spatial-temporal evidence presented to the recognition network.

2.4.3 Mutual Information Sampling

After alignment, not all frames are equally useful for recognition. In aerial videos, some frames are blurred, partially occluded, or nearly redundant due to high frame rate and weak motion change. Therefore, rather than using random or uniform sampling, we seek a frame sequence that is both informative and diverse.

Suppose we have already sampled

$$F_s = \{F_0, F_1, \ldots, F_i\},$$

and our goal is to select the next frame F_{i+1} from a candidate pool C_{i+1}. We choose the frame that is least redundant with respect to both the most recent frame and the set of previously selected frames:

$$F_{i+1} = \arg \min_{F_{i+1} \in C_{i+1}} \alpha I(F_i; F_{i+1}) + \beta I(F_s; F_{i+1}), \tag{2.7}$$

where the first term favors a frame that is distinctive relative to the current temporal state, and the second term reduces redundancy with the entire sampled set.

The second term can be decomposed through the chain rule:

$$I(F_s; F_{i+1}) = I(F_0, F_1, \ldots, F_i; F_{i+1}) = \sum_{j=0}^{i} I(F_j; F_{i+1} \mid F_{j-1}, \ldots, F_0). \tag{2.8}$$

In practice, conditional mutual information is difficult to compute exactly. We therefore adopt a low-dimensional approximation:

$$I(F_s; F_{i+1}) \approx \frac{1}{i+1} \sum_{j=0}^{i} I(F_j; F_{i+1}), \tag{2.9}$$

which leads to the tractable objective

$$F_{i+1} = \arg \min_{F_{i+1} \in C_{i+1}} \alpha I(F_i; F_{i+1}) + \frac{\beta}{i+1} \sum_{j=0}^{i} I(F_j; F_{i+1}). \tag{2.10}$$

This formulation balances two goals: selecting frames that introduce new information locally and maintaining diversity globally over the sampled sequence. To preserve randomness while still selecting informative frames, we first choose a random start frame F_0. Denoting its index by k_0, we define the candidate pool for the next frame using a randomly generated stride r_1, so that C_1 contains frames whose indices lie between k_0 and $k_0 + r_1$. The same strategy is repeated for subsequent frames. In this way, MIS does not deterministically jump to extreme temporal positions, but instead searches locally for the most informative continuation.

From a broader perspective, MIS complements TFA naturally. TFA stabilizes the spatial support of the actor across time, while MIS identifies which aligned observations carry the most useful temporal evidence. Together, they refine not only the consistency of the sequence, but also its informativeness.

2.4.4 MITFAS as a Full Recognition Pipeline

We now summarize how TFA and MIS are integrated into the full aerial action recognition pipeline. We first localize the actor in the start frame and crop a human-centered reference image F_r. To ensure that the action-relevant body region is fully preserved, we enlarge the actor region by 25% vertically and 10% horizontally before resizing it to the reference size. This reference image provides the anchor for subsequent alignment.

A naive sliding-window search over the full frame would be expensive. To improve efficiency, once we compute the optimal transformation ω_t^* at time t, we apply the same transformation to frame $t+1$ and expand the resulting region by 25% to define a smaller search area. Mutual information search is then performed only inside this local region rather than over the entire frame. We also occasionally

re-localize the actor to refresh the search area and prevent drift.

In our benchmarks, most videos are captured using platforms with anti-shake stabilization, so in practice we assume no rotation is required and set

$$R(\theta_t) = I.$$

For more general videos, the rotation matrix can still be written as

$$R(\theta_t) = \begin{bmatrix} \cos\theta_t & -\sin\theta_t \\ \sin\theta_t & \cos\theta_t \end{bmatrix}. \tag{2.11}$$

After obtaining all aligned frames, we apply MIS to generate a final sequence of 8 or 16 informative frames. This sampled sequence is passed to X3D-M [2] as the recognition backbone. Because TFA improves temporal consistency and MIS improves temporal informativeness, the backbone can focus more effectively on action-relevant motion patterns rather than on background change or redundant observations.

From the perspective of this dissertation, MITFAS can be understood as a spatial-temporal representation refinement method: TFA aligns *where* the action evidence appears across time, and MIS decides *which* temporal observations should be emphasized for recognition.

2.5 Experimental Evaluation

In this section, we evaluate MITFAS on multiple aerial action recognition benchmarks and analyze how its two key components contribute to performance. Since MITFAS is motivated by the weak spatial support and unstable temporal evidence of aerial videos, our experiments are designed not

only to measure recognition accuracy, but also to verify whether temporal alignment and informative sampling improve representation quality under aerial-specific challenges.

2.5.1 Datasets and Evaluation Setting

We evaluate MITFAS on three public aerial action recognition benchmarks that cover both indoor and outdoor scenes, as well as different degrees of viewpoint variation, background complexity, and actor visibility.

UAV-Human [5] is a large-scale aerial-view human behavior understanding benchmark containing diverse indoor and outdoor scenes with substantial variation in viewpoint, background, and motion. It is one of the most challenging datasets for aerial action recognition because the actors are often small, the scene context is complex, and camera motion introduces considerable temporal instability. For MITFAS, UAV-Human is particularly important because it directly tests whether improved spatial-temporal evidence extraction can help in a setting where foreground cues are weak and the viewing condition varies significantly across sequences.

NEC-Drone [6] is an indoor aerial action recognition dataset containing 5,250 videos and 16 action classes. Compared with UAV-Human, the environment is more constrained, but the actors remain relatively small in the frame and the viewpoint is still aerial. This dataset is useful for evaluating whether the proposed alignment and sampling strategy remains effective in a more controlled environment, and whether it can still improve recognition even under reduced image resolution.

Drone Action [18] is an outdoor aerial dataset containing 240 videos over 13 actions, captured using a free-flying aerial platform at relatively low altitude and speed. Although smaller than UAV-Human, it remains a valuable benchmark because it includes outdoor motion variation and realistic

Method	Backbone	Frames	Input Size	Initialization	Top-1 Acc (%) ↑
X3D-M	-	16	224×224	None	27.0
X3D-L	-	16	224×224	None	27.6
FAR	X3D-M	16	224×224	None	27.6
Ours (MITFAS)	X3D-M	16	224×224	None	**40.2**
FAR	X3D-M	8	540×540	None	28.8
Ours (MITFAS)	X3D-M	8	540×540	None	**38.4**
I3D	ResNet-101	8	540×960	Kinetics	21.1
FNet	I3D	8	540×960	Kinetics	24.3
FAR	I3D	8	540×960	Kinetics	29.2
FAR	X3D-M	8	620×620	Kinetics	39.1
Ours (MITFAS)	X3D-M	8	620×620	Kinetics	**46.6**
X3D-M	-	16	224×224	Kinetics	30.6
MViT	-	16	224×224	Kinetics	24.3
FAR	X3D-M	16	224×224	Kinetics	31.9
Ours (MITFAS)	X3D-M	16	224×224	Kinetics	**50.8**

Table 2.1: **Benchmarking on UAV-Human and comparison with prior methods.** MITFAS consistently outperforms both the X3D-M baseline and previous state-of-the-art methods across different resolutions, frame numbers, and initialization settings. The gains reach up to 13.2% over the baseline and 18.9% over the prior state of the art, demonstrating the effectiveness of the proposed temporal feature alignment strategy.

aerial camera behavior. It therefore provides an additional test of whether MITFAS can reliably select informative temporal evidence when the number of available training examples is limited.

Following the original MITFAS evaluation protocol, we use X3D-M [2] as the recognition backbone and compare against both baseline and prior state-of-the-art methods on each dataset. Our goal is not only to show that MITFAS improves final Top-1 accuracy, but also to analyze whether the gains are consistent across different aerial environments and whether they support the chapter's central claim that spatial-temporal evidence refinement is especially important for aerial video understanding.

2.5.2 Main Results

We compare MITFAS against prior state-of-the-art methods and corresponding baselines on all three datasets. The main results are shown in Table 2.1 and Table 2.2.

Method	Frames	Input Size	Init.	Top-1 Acc (%) ↑
X3D-M	8	960×540	Kinetics	66.1
FAR	8	960×540	Kinetics	71.4
Ours	8	540×540	Kinetics	**78.6**

(a) Results on NEC Drones. Our method shows an improvement of 12.5% on top-1 accuracy against the baseline X3D-M[2], 7.2% over current state-of-the-art FAR [48].

Method	Frames	Input Size	Init.	Top-1 Acc (%) ↑
HLPF	All	1920×1080	None	64.3
PCNN	-	1920×1080	None	75.9
X3D-M	16	224×224	Kinetics	83.4
FAR	16	224×224	Kinetics	92.7
Ours	16	224×224	Kinetics	**100.0**

(b) Results on Drone Action. Our method achieves 100% top-1 accuracy, 16.6% over the baseline method X3D-M[2], outperforming current state-of-the-art method FAR[48] by 7.3% under same configuration. (HLPF [54], PCNN [55])

Table 2.2: **Comparison of MITFAS with prior methods on NEC-Drone and Drone Action.**

Beyond the absolute accuracy gains, these results help clarify where MITFAS is most beneficial: namely, in aerial settings where action-relevant foreground cues are weak, spatial support is unstable, and useful temporal evidence is unevenly distributed across the sequence.

On **UAV-Human**, MITFAS improves Top-1 accuracy by **20.2%** over the baseline and by **18.9%** over the prior state of the art under the strongest configuration reported in the original work [38]. This large margin is particularly meaningful because UAV-Human contains substantial viewpoint variation, complex backgrounds, and noticeable changes in actor appearance and scale across time. The result suggests that when the actor is weakly visible and its spatial support changes significantly from frame to frame, explicitly aligning action-relevant regions before temporal recognition can substantially improve representation quality. In other words, the gain is not merely a consequence of using a stronger backbone, but of presenting the backbone with better spatial-temporal evidence.

On **NEC-Drone**, MITFAS achieves **78.6%** Top-1 accuracy, improving over the baseline by **12.5%** and over the prior state of the art by **7.2%**, while using only half-resolution input frames. This result is important for two reasons. First, it shows that the proposed alignment-and-sampling strategy remains effective even when the input resolution is reduced, which is relevant for practical aerial perception settings where bandwidth and computation may be constrained. Second, it indicates that

better spatial-temporal evidence extraction can compensate, at least in part, for reduced visual detail. This is consistent with the main thesis of the chapter: if the model is given more stable action-relevant regions and more informative temporal observations, recognition can improve even without increasing input fidelity.

On **Drone Action**, MITFAS reaches **100%** Top-1 accuracy, improving over the baseline by **16.6%** and over the prior state of the art by **7.3%**. Although this dataset is smaller than UAV-Human, the result still highlights an important property of MITFAS: when informative action cues are available but unevenly distributed over time, selecting a better aligned and more distinctive frame sequence can make recognition substantially more reliable. This is precisely the regime in which mutual-information-based sampling is most useful, since it reduces the chance that recognition will be dominated by redundant or weakly informative frames.

Taken together, these results show that MITFAS is effective across both indoor and outdoor aerial video settings, across different scene scales, and under different levels of action difficulty. More importantly, the consistent gains across all three datasets support the central claim of this chapter: aerial action recognition benefits not only from stronger recognition backbones, but also from explicitly refining the spatial-temporal evidence supplied to those backbones. In particular, TFA helps stabilize *where* action information is observed, while MIS improves *which* temporal observations are ultimately used for recognition.

2.5.3 Ablation Studies

We next analyze the effects of the two key components of MITFAS: Temporal Feature Alignment (TFA) and Mutual Information Sampling (MIS). The ablation results are summarized in Table 2.3.

Sampling Method	Top-1 Acc (%) ↑	Sampling Method	Top-1 Acc (%) ↑
Random	23.8	TFA + Random	39.8
Uniform	25.8	TFA + Uniform	42.2
MG Sampler	28.1	TFA + MG Sampler	45.5
MIS	**28.7**	TFA + MIS	**46.2**

(a) Temporal Feature Alignment (TFA) and Mutual Information Sampling (MIS) ablation studies on UAV-Human-Subset. The baseline is vanilla X3D with random [56] and uniform sampling [57], and we add our methods TFA and MIS step by step. From our experiments, TFA boost the accuracy by 16-17.5%. MIS outperforms the random sampling, uniform sampling, and MG Sampler[3].

Sampling Method	Alpha	Beta	UAV-Human Top-1 Acc (%) ↑	DroneAction Top-1 Acc (%) ↑
X3D + TFA + MIS	1.0	0.0	45.3	94.5
X3D + TFA + MIS	0.0	1.0	45.7	95.9
X3D + TFA + MIS	1.0	0.5	45.5	97.2
X3D + TFA + MIS	1.0	1.0	**46.2**	**100**

(b) Mutual Information Sampling (MIS) ablation studies on UAV-Human-subset and Drone Action. The baseline is vanilla X3D with TFA, we test the MITFAS Sampling in terms of two hyperparameters for mutual information and joint mutual information, α and β respectively. From our experiments, MITFAS obtains the best accuracy when $\alpha = 1.0$ and $\beta = 1.0$.

Method	UAV-Human Top-1 Acc (%) ↑	DroneAction Top-1 Acc (%) ↑
Bounding box tracking [58]	47.4	95.9
Spatial-temporal action detection [59]	47.9	95.9
TFA(ours)	**50.8**	**100.0**

(c) Comparison with other methods [58, 59].

Similarity Measure	Top-1 Acc (%) ↑
Euclidean Distance	42.1
Cosine Similarity	39.5
Peak Signal-to-Noise Ratio	43.4
Structural Similarity Index Measure	44.8
Mutual Information	**46.2**

(d) Comparison with other similarity measures on UAV-Human Subset. Compared to other similarity measures, mutual information achieves the best accuracy.

Table 2.3: **Ablation studies for MITFAS.**

The ablation results help explain why MITFAS works. Rather than contributing equally, TFA and MIS address different failure modes of aerial video understanding, and their effects are best understood as complementary.

Temporal Feature Alignment provides the largest gain. In our experiments, TFA improves Top-1 accuracy by approximately **16.0%–17.5%** when combined with X3D and different sampling strategies. This large improvement indicates that one of the main difficulties in aerial action recognition is not only weak temporal modeling, but unstable spatial support across frames. When the model is given a more consistent human-centered sequence, it can focus more effectively on action-relevant motion rather than background-dominated variation. From the thesis perspective, this result is especially important because it shows that temporal refinement in aerial videos is inseparable from spatial stabilization: improving the temporal sequence first requires aligning the spatial region from which

temporal evidence is extracted.

Mutual Information Sampling further improves recognition by selecting more informative temporal observations. Compared with random sampling, uniform sampling, and MGSampler [3], MIS improves Top-1 accuracy by **0.6%–6.4%**. Although the gain is smaller than that of TFA, it is still consistent and meaningful. This suggests that once the actor region has been better aligned, the next limiting factor becomes temporal redundancy. In other words, not all aligned frames are equally useful, and selecting a more distinctive subset still improves recognition quality. This supports the interpretation of MIS as a temporal evidence refinement mechanism rather than merely an efficiency heuristic.

The remaining ablations provide additional insight into the design choices of MITFAS. The sampling hyperparameter analysis shows that balancing informativeness and diversity is important for constructing an effective frame sequence, rather than simply selecting frames that are individually similar or individually distinct. The comparison with bounding-box tracking variants further suggests that explicit mutual-information-based alignment provides more useful support than relying on simpler tracking-style alternatives alone. Finally, the similarity-measure comparison shows that mutual information is more effective than alternative similarity measures in this setting, supporting our design choice to use statistical dependence rather than simpler pixel- or feature-level similarity alone.

Overall, the ablation results validate the complementary roles of the two components. TFA addresses the spatial-temporal instability of aerial videos by aligning action-relevant regions across time, while MIS improves the quality of temporal evidence by reducing redundancy among aligned observations. Their combination is therefore more effective than either component alone, and this complementarity is exactly what makes MITFAS a strong opening chapter for Part I.

2.6 Discussion

The results in this chapter suggest that one of the main difficulties in aerial action recognition is not simply limited model capacity, but the poor quality of the spatial-temporal evidence presented to the model. In aerial videos, action-relevant foreground cues are spatially sparse, unstable across frames, and easily overwhelmed by large background regions. At the same time, the temporal sequence itself is uneven in quality, since many frames are redundant, weakly informative, or distorted by platform motion. Under these conditions, directly applying a standard video backbone to raw frame sequences can lead to representations that are dominated by unstable background variation rather than by human motion.

MITFAS addresses this issue through a two-stage refinement strategy. The first stage, Temporal Feature Alignment, improves the *spatial stability* of the temporal input by repeatedly identifying the most action-relevant region in each frame. Although the method is framed as temporal alignment, its effect is fundamentally spatial-temporal: it stabilizes *where* the action evidence is extracted so that the sequence becomes more consistent for downstream modeling. The second stage, Mutual Information Sampling, improves the *temporal quality* of the aligned sequence by selecting frames that are informative and distinctive rather than redundant. Together, these two components refine both the spatial support and the temporal evidence available to the recognition model.

From the perspective of this dissertation, MITFAS establishes the first pillar of Part I. Its central lesson is that aerial video understanding can benefit substantially from refining the evidence presented to the model before attempting more sophisticated recognition. This is particularly important in aerial settings, where the foreground is weak and unstable and where background changes can easily dominate temporal reasoning. In this sense, MITFAS shows that temporal modeling in aerial videos cannot

be treated as purely temporal: it must also address the spatial instability of the action-bearing region.

The broader implication of this chapter is that representation refinement is a more suitable perspective for aerial video understanding than simply increasing backbone complexity. MITFAS improves performance not by introducing a larger recognition network, but by improving the quality of the sequence that the network receives. This idea naturally motivates the next two chapters in Part I. After establishing temporal feature alignment in this chapter, we next study informative frame selection more directly, and then investigate how semantic guidance can further improve aerial video understanding under weak and ambiguous visual evidence.

2.7 Limitations

Although MITFAS substantially improves aerial action recognition performance, it has several limitations.

First, the method relies on successful localization of the actor to initialize the reference image and guide subsequent alignment. If the actor is poorly localized or heavily occluded, the alignment process may deteriorate, which in turn can reduce the benefit of the downstream sampling stage.

Second, the current formulation is primarily designed for single-actor action recognition. In more complex aerial scenes containing multiple interacting people, overlapping actions, or ambiguous actor identities, a single reference-based alignment strategy may be insufficient. Extending the formulation to multi-agent aerial activity understanding would therefore require a more flexible mechanism for maintaining multiple action-relevant spatial supports over time.

Third, while MITFAS improves temporal alignment and frame informativeness, it still depends on a supervised downstream recognition model. In this sense, the method refines the quality of the

input evidence, but does not yet address the broader challenge of learning stronger aerial-specific representations under limited annotation. This limitation motivates later chapters in the dissertation that study more scalable learning strategies.

Finally, MITFAS is primarily designed for short-horizon action recognition rather than long-horizon activity understanding. More complex aerial video settings may require richer temporal reasoning over longer intervals, stronger semantic interpretation, or more explicit modeling of interactions between actors and context. These limitations motivate the later progression of Part I, where we move from alignment to frame selection and then to semantically guided representation learning.

2.8 Conclusion

In this chapter, we presented MITFAS [38], a mutual-information-based framework for temporal feature alignment and informative frame sampling in aerial video action recognition. By identifying action-relevant regions across frames and selecting more informative temporal observations, MITFAS substantially improves the quality of spatial-temporal representations for aerial action recognition. Experimental results on UAV-Human, NEC-Drone, and Drone Action demonstrate strong gains over both baselines and prior state-of-the-art methods.

More importantly, this chapter establishes a key thesis-level insight: in aerial videos, improving temporal modeling requires first stabilizing the spatial support of action-relevant evidence. MITFAS therefore serves not only as a strong aerial action recognition method, but also as the first concrete demonstration in this dissertation that aerial perception benefits from refining how evidence is extracted before it is interpreted.

As the opening chapter of Part I, MITFAS establishes the first pillar of the broader representation

refinement perspective developed in this dissertation. The next chapter builds on this idea by shifting the emphasis from temporal alignment to informative frame selection, studying how better temporal evidence can be selected even more directly for aerial video understanding.

Chapter 3: PMI Sampler: Patch Mutual Information Guided Frame Selection for Aerial Action Recognition

3.1 Introduction

This chapter continues Part I, which studies refined spatial-temporal representation learning for aerial video understanding. While Chapter 2 focused on improving aerial representations through temporal feature alignment, this chapter focuses on a complementary question: even if a recognizer is given a temporally ordered sequence, which frames should it actually observe? In aerial videos, this question is especially important because useful temporal evidence is often highly uneven across time.

Many existing video recognition methods rely on fixed temporal strides or uniform frame sampling. Although such strategies are simple and effective in many ground-view settings, they can be suboptimal for aerial videos. In practice, aerial videos frequently contain long intervals of redundant observations, while the most informative motion changes may only appear within a short temporal window. As a result, uniform sampling can easily over-represent motion-static regions while missing the most discriminative phase of the action.

This issue is particularly pronounced in aerial scenarios. Compared with conventional ground-view videos, aerial videos often exhibit smaller human actors, stronger viewpoint changes, and more severe background interference induced by camera motion. Consequently, simple frame-difference-

based criteria can become unreliable, since pixel-level changes may be dominated by background variation rather than by action-related motion. Frame selection is therefore not merely a computational convenience in aerial videos; it is a central representation problem.

Within the broader context of Part I, this chapter studies *temporal input refinement*. Unlike the previous chapter, which improves aerial representations by making temporal observations more spatially stable, this chapter improves the temporal observations presented to the recognition model in the first place. The key idea is that better temporal selection can itself serve as a lightweight but effective form of representation refinement: when the sampled clip contains more informative motion evidence, the downstream recognizer can learn a better action representation even without changing the backbone architecture.

To this end, we present **PMI Sampler** [39], a patch-similarity-guided frame selection strategy for aerial action recognition. Instead of relying on raw RGB differences, PMI Sampler measures the statistical dependence between adjacent frames using *patch mutual information* (PMI). The resulting motion salience sequence is then remapped by a shifted Leaky ReLU function and converted into a cumulative motion distribution, from which frames are adaptively sampled. In this way, the selected frame sequence better covers motion-salient temporal regions while avoiding excessive redundancy.

Figure 3.1 provides a qualitative motivation for the chapter. Compared with MGSampler [3], PMI Sampler allocates samples more consistently to motion-salient periods while remaining more robust to background changes induced by camera motion. This property is particularly important in aerial videos, where naive frame-difference-based sampling can easily be distracted by large but irrelevant background variation.

From the perspective of this dissertation, PMI Sampler establishes the second pillar of Part I. If Chapter 2 showed that aerial video understanding benefits from stabilizing *where* action evidence is

extracted, this chapter shows that it also benefits from selecting *when* the most informative evidence occurs. Together, these chapters argue that aerial video understanding improves when the temporal input is both more stable and more informative.

The main contributions of this chapter are as follows:

- We formulate adaptive frame selection for aerial videos as a problem of estimating the temporal distribution of informative motion.
- We introduce patch mutual information as a robust motion salience measure that is less sensitive to aerial-specific background noise than raw pixel differences.
- We present a simple yet effective motion-aware sampling strategy based on shifted Leaky ReLU remapping and cumulative motion segmentation, which consistently improves action recognition across multiple aerial and non-aerial benchmarks.

3.2 Related Work

Since Chapter 2 already reviewed the broader background of aerial action recognition and temporal modeling, here we focus on prior work most directly related to informative temporal selection and motion salience estimation.

3.2.1 Frame Selection and Adaptive Video Inference

Frame selection has been widely studied as a way to reduce temporal redundancy and improve the efficiency and effectiveness of video recognition pipelines [3, 40, 60–64]. A common observation in this literature is that not all frames contribute equally to recognition, so selecting a smaller but more

Figure 3.1: **Qualitative motivation for PMI Sampler.** Eight sampled frames from representative videos in Diving48 and UAV-Human are shown. Compared with MGSampler [3], PMI Sampler better reflects the temporal distribution of informative motion and more clearly separates motion-salient from motion-static periods. This behavior is particularly important in aerial videos, where background changes caused by camera motion can otherwise dominate naive sampling criteria.

informative subset can improve both computation efficiency and downstream accuracy [40, 60, 63, 64]. Existing approaches include adaptive frame policies learned from the input video [40], salient clip selection strategies [60], conditional early exiting and adaptive inference mechanisms [61], single-step clip compression methods [62], and motion-guided or importance-guided sampling schemes [3, 63].

Despite their effectiveness, many of these approaches are primarily developed for generic ground-view videos or broader video understanding settings [3, 40, 60–62]. In such settings, frame difference or learned selection signals can often serve as useful proxies for informativeness [3, 40, 63]. In aerial videos, however, this assumption is much weaker because camera motion, viewpoint change, and dominant background regions can produce large appearance changes that are unrelated to the target action [5, 16, 18, 33]. As a result, a frame may appear highly dynamic at the pixel level while still contributing little action-relevant information [5, 16, 18]. This makes informative frame selection a more difficult problem in aerial videos than in conventional video recognition [5, 6, 46].

PMI Sampler is motivated by this gap. Rather than learning a separate frame selection policy or relying directly on raw RGB difference, it estimates temporal salience through patch-level statistical dependence [39]. In this sense, the method is designed not only to reduce redundancy, but also to better match the temporal sampling density to the true distribution of informative motion in aerial videos [5, 18, 39].

3.2.2 Mutual Information for Temporal Salience Estimation

Mutual information has been widely used as a statistical measure of dependence in image analysis, registration, and representation learning [51–53, 65, 66]. In image registration, mutual information is particularly attractive because it can capture correspondence between observations without requiring

exact pixel-wise similarity, which makes it robust to appearance variation and modality shift [51–53]. In representation learning, mutual-information-based objectives have been used to measure shared information across views or transformations and to encourage more informative feature learning [65, 66]. More broadly, these lines of work suggest that mutual information is well suited for identifying meaningful dependence even when direct low-level comparison is unreliable [51, 65, 66].

This property is especially relevant for aerial videos. In aerial action recognition, raw frame differences are often dominated by platform motion, scale change, and background variation rather than by the action itself [5, 16, 18, 33]. Under such conditions, a dependence-based criterion can provide a more reliable signal for temporal salience estimation, because it measures how much genuinely new information one frame contributes relative to another [51, 52, 65]. PMI Sampler builds on this intuition by applying mutual information at the patch level rather than at the whole-frame level [39]. This design is important because global dependence measures may overlook localized spatial structure, whereas patch-level mutual information preserves local information that is more relevant to human action [39]. As a result, PMI Sampler combines two desirable properties: robustness to aerial-specific background noise and sensitivity to localized temporal change [5, 18, 39].

3.3 Problem Formulation and Motivation

Consider an input video

$$V = \{F_1, F_2, \ldots, F_T\},$$

where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$ and the task is to predict an action label

$$y \in \{1, \ldots, C\}.$$

A standard video recognition pipeline first samples a shorter frame sequence

$$\hat{V} = \{F_{t_1}, F_{t_2}, \ldots, F_{t_K}\}, \quad 1 \leq t_1 < t_2 < \cdots < t_K \leq T,$$

and then feeds $\hat{V}$ into a recognition backbone for classification.

In most existing pipelines, the indices $\{t_1, \ldots, t_K\}$ are obtained through uniform sampling or a fixed temporal stride. However, such a strategy can be suboptimal for aerial videos for three reasons.

First, aerial videos often contain long temporal intervals with weak or redundant motion. Uniform sampling may waste many sample slots on these motion-static periods.

Second, action-discriminative motion is often concentrated within a relatively short phase of the video. A fixed temporal schedule may miss these informative temporal transitions or sample them too sparsely.

Third, temporal salience is difficult to estimate reliably in aerial videos. Pixel-level changes are often affected by background motion, camera shake, and viewpoint change, so raw RGB differences do not necessarily reflect the true action-related motion.

These observations motivate a frame selection strategy that can identify informative temporal observations while remaining robust to aerial-specific nuisances. Rather than measuring motion through direct pixel difference, PMI Sampler estimates motion salience through patch-level statistical dependence between adjacent frames. In this sense, the method aims to refine the temporal input by better matching sampling density to the actual distribution of informative motion.

3.4 Method

3.4.1 Overview

Within this thesis, PMI Sampler can be understood as a form of *representation refinement through temporal selection*. Instead of modifying the recognition backbone itself, the method refines the temporal evidence presented to the backbone by constructing a sequence of more informative observations. This perspective makes the chapter naturally complementary to Chapter 2: MITFAS refines *how* temporal observations are aligned, whereas PMI Sampler refines *which* temporal observations are selected.

Given an input video V, the goal is to build a sampled sequence $\hat{V}$ that better captures the discriminative motion evolution of the action. PMI Sampler achieves this in three stages:

1. estimate motion salience between adjacent frames using a **patch mutual information (PMI) score**;

2. remap the salience sequence using a **shifted Leaky ReLU** to better separate motion-salient and motion-static periods;

3. sample frames according to the **cumulative motion information distribution** rather than a fixed temporal interval.

The overall pipeline of PMI Sampler is illustrated in Figure 3.2. Given an input video, we first estimate patch-mutual-information-based motion salience between adjacent frames, then remap the salience sequence to better separate motion-salient and motion-static periods, and finally sample frames according to the cumulative motion distribution. Conceptually, PMI Sampler first asks whether

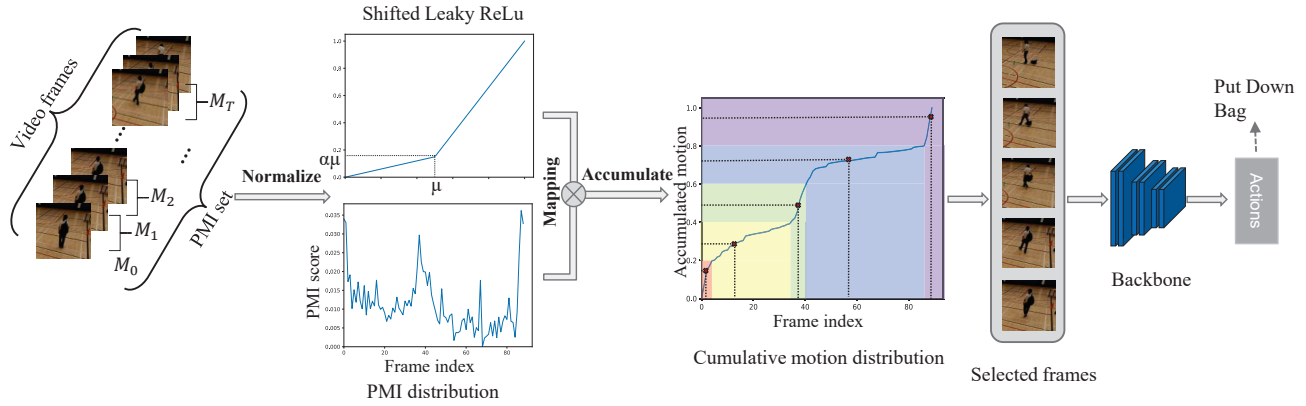


Figure 3.2: **Overview of PMI Sampler.** PMI Sampler first estimates motion salience between adjacent frames using patch mutual information. It then remaps the salience sequence to better distinguish motion-salient and motion-static periods, and accumulates the remapped scores into a cumulative motion distribution. Based on this distribution, the video is divided into N motion-balanced segments, from which representative frames are selected to form the final input sequence for action recognition.

a new frame provides genuinely new information relative to its predecessor, and then allocates sampling density according to how informative different parts of the video are.

3.4.2 Why Patch Mutual Information?

A key challenge in aerial videos is that motion cues are difficult to estimate robustly. In many cases, the human actor occupies only a small fraction of the image, while the majority of pixels belong to the background. When the UAV camera moves, the resulting background deviations can be much larger than the actual action motion. Therefore, pixel-wise RGB difference can be dominated by irrelevant background variation and fail to reveal the true temporal distribution of informative motion.

Mutual information provides a more robust alternative because it measures the statistical dependence between two signals rather than their exact pixel-level correspondence. In other words, it captures how much information one frame provides about another. If two adjacent frames are highly dependent, then the later frame contributes relatively little new motion information. If their depen-

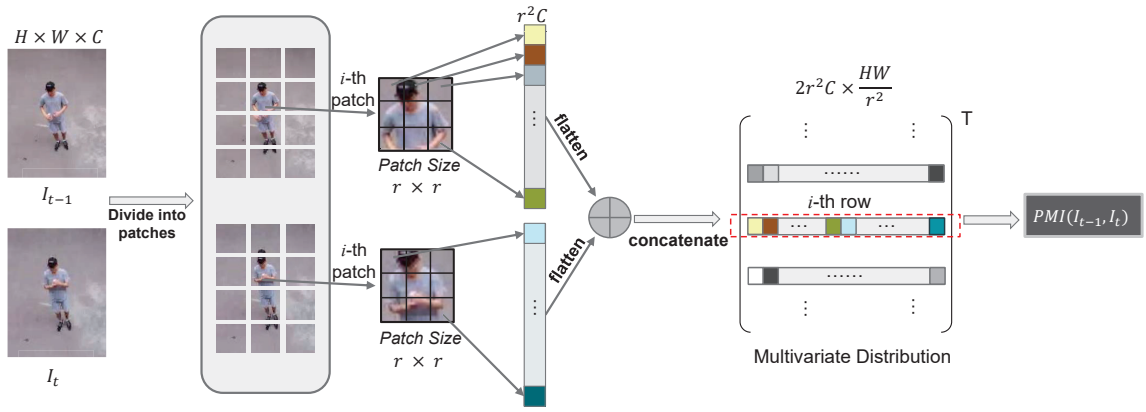


Figure 3.3: **Patch Mutual Information formulation.** For two adjacent frames I_{t-1} and I_t of size $H \times W \times C$, we partition each frame into non-overlapping patches of size $r \times r \times C$ and flatten each patch into a vector of dimension $d = r^2C$. Corresponding patches from the two frames are concatenated to form $2d$-dimensional samples, and all such samples are stacked to construct the multivariate point set for the frame pair. We then estimate its covariance matrix and use it to approximate the entropy terms required for patch mutual information. The resulting patch mutual information score is computed by Eq. 3.1, and is further converted into a motion salience estimate through the inversion and normalization steps in Eq. 3.3 and Eq. 3.4.

dence weakens, the later frame is more likely to contain a meaningful temporal change.

However, global mutual information alone ignores spatial structure, which is important for action recognition. PMI Sampler therefore computes mutual information at the *patch* level. This preserves local spatial structure while remaining more robust to noise than raw RGB difference. As a result, patch mutual information provides a better motion salience estimate for aerial videos than naive frame-difference measures, which are easily corrupted by camera-induced background motion. The patch-level construction used to estimate this dependence is illustrated in Figure 3.3.

3.4.3 Patch Mutual Information Score

Let two adjacent frames be denoted by I_{t-1} and I_t, each of size $H \times W \times C$. We divide each frame into non-overlapping square patches with side length r, and flatten each patch into a vector of

dimension

$$d = Cr^2.$$

For every corresponding patch pair between I_{t-1} and I_t, we concatenate the two flattened patch vectors into a single vector in $\mathbb{R}^{2d}$. For a pair of frames, the total number of patch correspondences is

$$N = \frac{HW}{r^2}.$$

By stacking all corresponding patch pairs, we obtain a multivariate point set

$$P_t = [p_1, p_2, \ldots, p_N],$$

where each $p_i \in \mathbb{R}^{2d}$.

Ideally, we would like to measure the mutual information between adjacent frames to quantify how much new information the current frame provides. However, direct estimation of mutual information in this setting is intractable, since the patch-level points are high-dimensional and their joint distribution is generally unknown and dependent.

To obtain a tractable formulation, we follow the general strategy of region-based mutual information [67] and adopt a covariance-based approximation. Specifically, we approximate the entropy of the point set using the entropy of a Gaussian distribution with the same covariance. This yields a lower-bound-style surrogate of mutual information that is efficient to compute in practice.

Let Ω denote the covariance matrix of P_t. Since each point is formed by concatenating a patch from I_{t-1} and one from I_t, Ω can be decomposed into marginal covariance blocks Ω_{t-1} and Ω_t, corresponding to the two frames. Under this approximation, we define the patch mutual information score

as

$$M_t = PMI(I_{t-1}, I_t) = H(\Omega_{t-1}) + H(\Omega_t) - H(\Omega), \tag{3.1}$$

where the entropy is approximated via

$$H(\Omega) \approx \frac{1}{2} \log \left((2\pi e)^d \det(\Omega) \right). \tag{3.2}$$

The resulting M_t should be interpreted as a covariance-based surrogate for mutual information rather than an exact information-theoretic quantity. A larger value of M_t indicates stronger dependency between adjacent frames, meaning that the current frame introduces less new information. Conversely, a smaller value suggests that the frame contributes more novel motion information.

To convert this dependence measure into a motion salience score, we invert it:

$$M'_t = \max_i(M_i) - M_t, \tag{3.3}$$

and normalize the sequence using an l_1 norm:

$$\hat{M}_t = \frac{M'_t}{\sum_i M'_i}. \tag{3.4}$$

The normalized score $\hat{M}_t$ serves as the motion salience estimate for frame I_t, where higher values indicate more informative temporal change.

We emphasize that this formulation does not assume that the patch-level points are strictly Gaussian. Instead, it uses a Gaussian entropy approximation to obtain a practical and differentiable objective. Empirically, this surrogate provides a robust estimate of motion salience in aerial videos, where

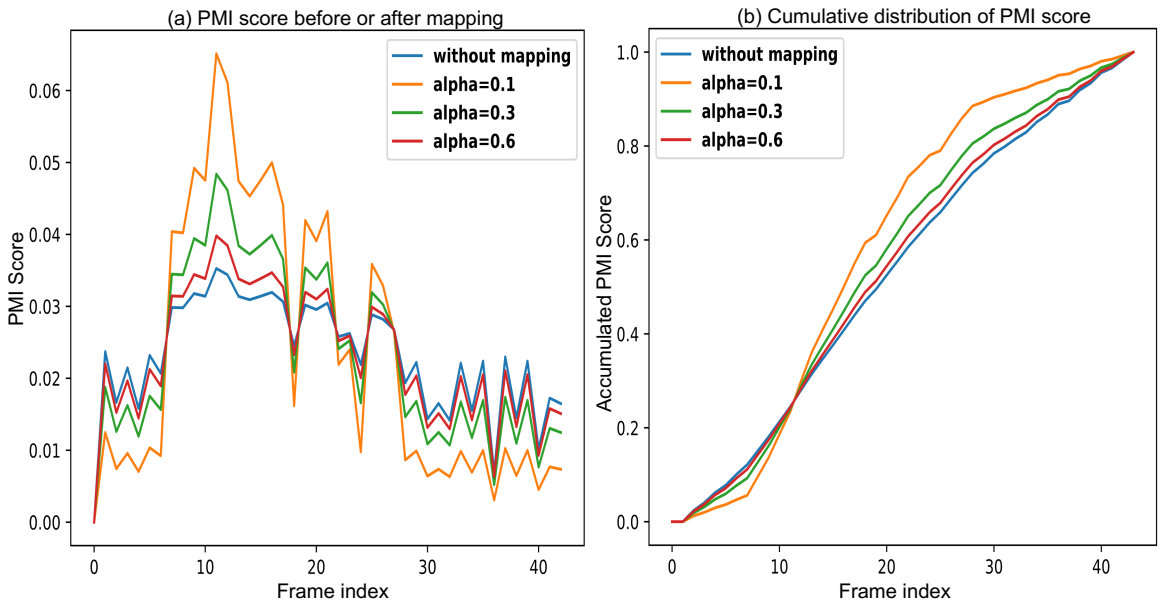


Figure 3.4: **Shifted Leaky ReLU remapping.** The normalized motion salience sequence $\hat{M}_t$ is remapped by the shifted Leaky ReLU function defined in Eq. 3.5. Compared with the original salience values, the remapped sequence assigns lower weight to motion-static intervals and higher contrast to motion-salient ones. The resulting scores are then accumulated into the cumulative motion distribution in Eq. 3.6, which is used to guide adaptive frame sampling.

direct pixel-level differences are often dominated by background motion.

3.4.4 Adaptive Sampling via Shifted Leaky ReLU and Cumulative Distribution

The sequence

$$S = \{\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_T\}$$

provides a temporal estimate of motion salience across the video. However, in aerial videos, the differences between motion-salient and motion-static regions can be subtle. To make these regions easier to distinguish, PMI Sampler further remaps the motion salience values using a shifted Leaky ReLU function. Figure 3.4 illustrates how this remapping increases the contrast between motion-static and motion-salient intervals before the cumulative distribution in Eq. 3.6 is constructed.

Let μ denote the mean of the normalized motion salience scores. For each $\hat{M}_t$, the remapped value is defined as

$$M_t^* = \begin{cases} \alpha \hat{M}_t, & \hat{M}_t \leq \mu, \\ \frac{1-\alpha\mu}{1-\mu}(\hat{M}_t - 1) + 1, & \hat{M}_t > \mu, \end{cases} \tag{3.5}$$

where α controls the smoothness of the remapping.

This transformation suppresses low-salience regions while preserving and emphasizing the relatively high-salience ones. As a result, the temporal contrast between motion-static and motion-salient intervals becomes more pronounced.

We then normalize the remapped sequence again and accumulate it into a cumulative distribution:

$$F(t) = \sum_{i \leq t} M_i^*. \tag{3.6}$$

Instead of uniformly dividing the video along the temporal axis, PMI Sampler divides the cumulative motion distribution into K equal portions. Each portion therefore corresponds to a temporal segment that contains approximately the same amount of motion information. From each segment, one representative frame is selected to form the final sampled sequence.

This design has two important advantages. First, it still preserves temporal coverage of the whole action sequence, since the final sample spans the full video. Second, it allocates more sampling density to motion-salient periods and fewer samples to motion-static intervals. In this way, the recognizer observes more discriminative temporal transitions without wasting sample slots on redundant observations.

3.5 Experimental Evaluation

In this section, we evaluate PMI Sampler on multiple aerial and non-aerial action recognition benchmarks and analyze why the proposed sampling strategy is effective. Since PMI Sampler is motivated by the uneven temporal distribution of informative motion in aerial videos, our experiments are designed not only to report final recognition accuracy, but also to verify whether patch-mutual-information-based sampling improves the quality of the temporal evidence supplied to downstream recognition backbones.

3.5.1 Datasets and Evaluation Setting

We evaluate PMI Sampler on both aerial and non-aerial action recognition benchmarks in order to assess not only its effectiveness in aerial videos, but also its generalization to broader motion-centric recognition settings.

For the two aerial benchmarks UAV-Human and NEC-Drone, we follow the same datasets introduced in Chapter 2. As discussed there, UAV-Human is a large-scale aerial-view human behavior understanding benchmark with substantial viewpoint, scale, and background variation [5], while NEC-Drone is an indoor aerial action recognition dataset containing 5,250 videos from 16 action classes [6]. In this chapter, these two datasets are mainly used to evaluate whether adaptive frame selection can better capture informative motion under aerial-specific challenges.

We further evaluate PMI Sampler on Diving48 [4], a fine-grained action recognition benchmark in which performance depends strongly on recognizing subtle temporal transitions rather than on static scene cues. This dataset is useful for testing whether the proposed sampling strategy generalizes beyond aerial videos to more general motion-centric action recognition scenarios.

In addition, we include Something-Something V2 [68], UCF101 [69], and HMDB51 [70] in additional experiments to evaluate transfer to more conventional ground-view action recognition settings. These datasets help clarify that PMI Sampler is aerial-specific in motivation, but not necessarily limited to aerial benchmarks in applicability.

Unless otherwise specified, the implementation follows the original MGSampler [3] setting. We compare against baseline sampling strategies and prior state-of-the-art methods using top-1 accuracy and top-5 accuracy, and we integrate PMI Sampler into multiple backbones in additional experiments to test whether the method should be viewed as a general temporal sampling module rather than as a design tied to a single recognizer. Some of the ablation studies are conducted on a randomly sampled subset of original dataset for time saving.

3.5.2 Implementation Details

Unless otherwise specified, all models are trained using SGD with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is set to 0.1 for training from scratch and 0.05 when initializing from Kinetics-pretrained weights. We use cosine or polynomial annealing for learning rate decay and optimize the models with multi-class cross entropy loss. During training, videos are decoded as a single clip and resized to 224×224. During testing, the shorter side is resized to 256, followed by three spatial crops of size 224×224, and the final prediction is obtained by averaging the scores over all crops. Unless otherwise stated, the rest of the setting follows the original paper.

Method	Frames	Input Size	Init.	Top-1 Acc (%) ↑
FAR	8	540×540	None	28.8
Ours	8	540×540	None	**39.7**
X3D-M	8	540×540	Kinetics	36.6
FAR	8	540×540	Kinetics	38.6
DiffFAR	8	540×540	Kinetics	41.9
Ours	8	540×540	Kinetics	**47.7**
FAR	8	620×620	Kinetics	39.1
Ours	8	620×620	Kinetics	**52.0**
X3D-M	16	224×224	Kinetics	30.6
FAR	16	224×224	Kinetics	31.9
AZTR	16	224×224	Kinetics	47.4
MG Sampler	16	224×224	Kinetics	53.8
Ours	16	224×224	Kinetics	**55.0**

Table 3.1: **Results on UAV-Human.** PMI Sampler consistently improves Top-1 accuracy under different frame counts, resolutions, and initialization settings.

3.5.3 Main Results

The main quantitative results evaluate PMI Sampler from three complementary perspectives. First, we test whether the method improves performance on a large-scale and highly challenging aerial benchmark. Second, we examine whether the benefit remains on a more controlled aerial dataset and on a fine-grained motion-centric benchmark beyond UAV videos. Third, we evaluate whether the proposed shifted Leaky ReLU remapping is an important part of the overall design rather than a minor implementation detail.

Table 3.1 shows that PMI Sampler consistently improves performance on UAV-Human across a wide range of settings, including different input resolutions, frame counts, and initialization schemes. This consistency is itself important. UAV-Human is a challenging aerial benchmark with substantial viewpoint variation, background clutter, and uneven temporal motion distribution, so a method that only works in one particular configuration would be far less convincing. Instead, PMI Sampler im-

Method	Frames	Backbone	Diving48 Top-1 Acc (%) ↑	NEC Drone Top-1 Acc (%) ↑
Random	16	X3D-M	71.1	52.0
Uniform	16	X3D-M	73.5	55.4
MG Sampler	16	X3D-M	74.6	58.5
K-centered	16	ViT	72.5	36.3
Ours	16	X3D-M	**81.3**	**62.5**

Table 3.2: **Results on Diving48 and NEC-Drone.** Our method relatively improves the top-1 accuracy by 9.0% on Diving48, by 6.8% on NEC-Drone.

proves over previous methods under every reported setting, with relative gains ranging from **2.2%** to **13.8%**. The largest gains appear in the weaker baseline settings, which suggests that adaptive temporal selection is especially helpful when the recognizer struggles to isolate useful motion evidence from noisy aerial observations.

The Kinetics-initialized 16-frame setting is particularly informative because it allows a direct comparison against recent aerial-specific methods under a stronger backbone setup. In this regime, PMI Sampler improves over MG Sampler from 53.8% to $\mathbf{55.0}\%$, while also surpassing AZTR and FAR. This indicates that even after the recognition backbone and initialization are fixed, the way temporal evidence is selected still matters. In other words, the gain does not come from changing the recognizer itself, but from improving the temporal observations presented to it.

Table 3.2 shows that the benefit of PMI Sampler is not limited to one aerial benchmark. On NEC-Drone, PMI Sampler improves over MG Sampler from 58.5% to $\mathbf{62.5}\%$, corresponding to a relative gain of **6.8%**. Although NEC-Drone is more controlled than UAV-Human, this result indicates that the proposed sampling strategy remains useful even when scene variation is reduced. The gain therefore seems to come from a more fundamental advantage in selecting informative temporal transitions, rather than from overfitting to one particular aerial benchmark.

The Diving48 result is even more revealing. PMI Sampler improves from 74.6% to $\mathbf{81.3}\%$, a

Mapping function	NEC Drone Top-1 Acc (%)	Diving48 Subset Top-1 Acc (%)
without Mapping	59.3	58.8
Quadratic	60.3	63.4
Sigmoid	61.4	65.7
Softmax	58.4	53.9
Tanh	60.9	56.6
ReLu	60.5	60.6
Shifted Leaky ReLu	**62.5**	**66.3**

Table 3.3: **Comparison between different mapping functions.** Shifted Leaky ReLU can better map the motion information distribution and make it easier to distinguish frames containing more motion information.

relative gain of **9.0%**. Since Diving48 is not an aerial dataset, this result shows that the proposed sampling strategy is aerial-specific in motivation but broader in applicability. More generally, it suggests that patch-mutual-information-based salience estimation is effective whenever fine-grained recognition depends on selecting subtle but informative motion transitions rather than simply observing evenly spaced frames.

Table 3.3 clarifies the importance of the proposed shifted Leaky ReLU remapping. Without any remapping, the motion salience distribution remains too weakly separated to support strong adaptive sampling. Standard nonlinear mappings such as quadratic, sigmoid, softmax, tanh, and ReLU all improve over the unmodified distribution in at least some settings, but none performs as consistently as the proposed shifted Leaky ReLU. In particular, it achieves the best Top-1 accuracy on both NEC-Drone and the Diving48 subset. This result supports the main design intuition of PMI Sampler: adaptive temporal selection depends not only on estimating temporal salience, but also on sharpening the distinction between motion-salient and motion-static intervals so that the final sampling distribution becomes more meaningful.

Taken together, the results in Tables 3.1–3.3 support the central claim of this chapter: temporal

input quality matters. Even without changing the recognition backbone, improving which frames are observed can substantially improve downstream action recognition. In this sense, PMI Sampler demonstrates that temporal selection is itself a useful form of representation refinement.

3.5.4 Robustness Across Backbone Models

To evaluate whether PMI Sampler is backbone-specific, we integrate it with multiple recognition architectures, including SlowOnly-R50, I3D, and X3D.

Method	Frames	Backbone	Top-1 Acc (%) ↑	Top-5 Acc (%) ↑
Uniform	8	I3D [23]	59.2	89.9
MG Sampler [3]	8	I3D [23]	55.2	87.6
Ours	8	I3D [23]	**61.8**	**91.7**
Uniform	8	SlowOnly-R50 [26]	60.0	90.3
MG Sampler [3]	8	SlowOnly-R50 [26]	57.1	88.4
Ours	8	SlowOnly-R50 [26]	**63.1**	**93.5**
Uniform	16	X3D [2]	73.5	95.1
MG Sampler [3]	16	X3D [2]	74.6	95.0
Ours	16	X3D [2]	**81.3**	**97.7**

Table 3.4: **Evaluation of our method with different recognition backbones on Diving48.** PMI Sampler can be incorporated with any recognition backbone models to improve the accuracy.

Table 3.4 shows that PMI Sampler consistently improves performance across all evaluated backbones. This is important because it indicates that the method should be understood as a general temporal sampling module rather than as a design tailored to a single recognizer. The gains are also nontrivial: PMI Sampler improves over uniform sampling by $+2.6\%$ on I3D, $+3.1\%$ on SlowOnly-R50, and $+7.8\%$ on X3D. The fact that the gains remain positive across architectures with different temporal modeling styles suggests that the benefit comes from better temporal evidence selection itself rather than from an accidental compatibility with one model family.

3.5.5 Dense Clip Sampling

We further evaluate PMI Sampler in a dense clip sampling setting.

Method	Frames	Backbone	Top-1 Acc (%) ↑	Top-5 Acc (%) ↑
Uniform	8	X3D-M [2]	74.0	95.3
MG Sampler [3]	8	X3D-M [2]	74.6	95.9
Ours	8	X3D-M [2]	**75.5**	**96.0**

Table 3.5: **Performance under dense clip sampling.** We evaluate PMI Sampler in a dense clip sampling setting, where clips are sampled more frequently to provide richer temporal coverage during inference. Compared with uniform sampling and the prior MG Sampler, our method achieves the best performance, improving Top-1 accuracy by 1.5% over uniform sampling and 0.9% over MG Sampler.

The gains remain consistent even when multiple clips are sampled from each video. This result matters because it shows that PMI Sampler is not merely exploiting a narrow single-clip evaluation protocol. Instead, it captures a more general principle: allocating temporal samples according to motion informativeness remains useful even when the overall evaluation already has broader temporal coverage.

3.5.6 Ablation Studies

We next analyze several key design choices in PMI Sampler, including the similarity measure used for temporal salience estimation, the shifted Leaky ReLU hyperparameter, the patch size used in PMI computation, the resulting speed-accuracy trade-off, and the transfer behavior to non-aerial datasets. Together, these studies help explain not only whether PMI Sampler works, but also why it works and under what conditions its design is most effective.

Table 3.6 directly validates one of the central design choices of PMI Sampler: the use of patch mutual information as the temporal salience measure. Simple similarity measures such as Euclidean

Similarity measure	UAV-Human subset Top-1 Acc (%)	Time cost per frame (ms)
Euclidean Distance	56.4	1.7
Cosine Similarity	55.0	2.5
Mutual Information	58.4	10.6
Regional Mutual Information (RMI)	59.2	20.2
Normalized Mutual Information (NMI)	58.8	32.8
Peak Signal-to-Noise Ratio (PSNR)	57.0	3.7
Structural Similarity Index Measure (SSIM)	57.7	79.9
Patch Mutual Information	**59.8**	**4.5**

Table 3.6: **Comparison between different similarity measures**. Our proposed Patch Mutual Information(PMI) outperforms other measures on UAV-Human.

	α	0	0.1	0.2	**0.3**	0.4	0.5	0.6	0.7
Diving48 subset	Top-1 (%) ↑	60.6	65.4	64.7	**66.3**	66.2	65.4	65.6	65.4
	Top-5 (%) ↑	92.6	92.8	94.4	**94.9**	94.6	93.3	94.5	94.6
UAV-Human subset	Top-1 (%) ↑	53.7	58.1	59.8	**59.8**	59.7	58.7	59.5	59.3
	Top-1 (%) ↑	83.8	85.3	**86.5**	86.3	85.3	86.2	85.5	86.0
NEC-Drone	Top-1 (%) ↑	54.1	57.9	58.0	58.2	60.7	61.2	**62.5**	61.7
	Top-1 (%) ↑	85.4	89.5	88.6	89.3	89.5	**90.0**	89.6	89.5

Table 3.7: **Impact of the hyper-parameter α.** The dynamic camera movement in UAV-Human and Diving48 videos necessitates a lower $\alpha(\sim 0.3)$ to differentiate real motion from background noise. In contrast, NEC-Drone videos, captured by a hovering UAV, require a higher $\alpha(\sim 0.6)$ to preserve the motion info distribution.

distance and cosine similarity are computationally cheap, but their recognition accuracy is clearly lower. Full-image statistical measures such as mutual information, regional mutual information, and normalized mutual information improve accuracy, but at a substantially higher computational cost. Patch Mutual Information achieves the best overall result, reaching the highest Top-1 accuracy while remaining considerably more efficient than RMI, NMI, and SSIM. This result is important because it shows that the benefit of PMI Sampler does not come from using a more expensive similarity function in general, but from using a similarity function that preserves local spatial structure while remaining robust to aerial-specific noise.

Table 3.7 shows that the hyperparameter α controls how strongly the shifted Leaky ReLU sep-

Size of Patch ($r \times r$)	UAV-Human Top-1 Acc (%)	Time cost per frame (ms)
3×3	53.8	5.4
5×5	54.6	6.8
7×7	**55.0**	**9.2**
11×11	52.3	17.9
21×21	51.9	67.8

Table 3.8: **Impact of the patch size.** A better tradeoff between accuracy and time cost is achieved with patch size 7×7.

arates motion-salient from motion-static intervals, and that the preferred value varies across datasets. This is an important observation rather than a nuisance implementation detail. On UAV-Human and Diving48, smaller α values are more effective, which is consistent with the fact that these datasets contain stronger camera or scene variation and therefore benefit from a sharper suppression of weak salience responses. NEC-Drone, in contrast, performs best with a larger α, suggesting that in more stable aerial videos it is preferable to preserve a smoother motion-information distribution. This pattern supports the interpretation that the remapping function is not arbitrary: it adjusts temporal contrast according to the motion characteristics of the dataset.

Table 3.8 analyzes the patch size used in PMI computation. The results show that 7×7 provides the best trade-off between recognition accuracy and computation cost. Very small patches are less effective because they capture less stable local statistics, while very large patches are significantly more expensive and lose the local structure that makes patch-based mutual information useful in the first place. This result directly supports the core design of PMI Sampler: mutual information should be computed at a scale that preserves local motion structure while remaining robust to aerial-specific noise.

Table 3.9 is particularly important from a thesis perspective because it shows that PMI Sampler is not only accurate, but also useful for improving the speed-accuracy trade-off. Even when fewer frames

Method	Frames	Input Resolution	Training Time per Epoch (s) ↓	Training Time per Video (ms) ↓	Diving48 Top-1(%) ↑	Diving48 Top-5(%) ↑
Random	16	224 × 224	196.6	13.1	71.1	94.8
Uniform	16	224 × 224	191.4	12.7	73.5	95.1
MG Sampler	16	224 × 224	287.4	19.2	74.6	95.0
PMI Sampler **(Ours)**	**8**	224 × 224	248.3 **(-39.1)**	16.5 **(-2.7)**	75.0 **(+0.4)**	95.6 **(+0.6)**
PMI Sampler **(Ours)**	**12**	224 × 224	276.8 **(-10.6)**	18.4 **(-0.8)**	77.7 **(+3.1)**	97.5 **(+2.5)**
PMI Sampler **(Ours)**	16	**172 × 172**	294.8 **(+7.4)**	19.6 **(+0.4)**	79.0 **(+4.4)**	97.7 **(+2.7)**
PMI Sampler **(Ours)**	16	224 × 224	295.1 **(+7.7)**	19.6 **(+0.4)**	81.3 **(+6.7)**	97.7 **(+2.7)**

Table 3.9: **Training efficiency.** PMI Sampler achieves higher accuracy with less spatial/temporal information, leading to better speed-accuracy tradeoffs. We show results for fewer frames (8 and 12) and smaller input sizes (172×172).

Dataset	**SthSthV2**		**UCF101**		**HMDB51**	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
Uniform	57.5	84.3	94.3	99.3	64.6	88.1
MG Sampler	59.2	85.2	94.6	99.2	64.9	87.9
Ours (Patch 7×7)	59.7 **(+0.5)**	85.2 **(+0)**	94.4 **(-0.2)**	99.6 **(+0.4)**	64.5 **(-0.4)**	87.2 **(-0.7)**
Ours (Patch 3×3)	60.3 **(+1.1)**	85.8 **(+0.6)**	95.1 **(+0.5)**	99.5 **(+0.3)**	65.5 **(+0.6)**	88.1 **(+0.2)**

Table 3.10: **Results on SthSthV2, UCF101, HMDB51**. Given that the majority of videos in these datasets feature fixed cameras capturing close-range scenes with prominent human actors and minimal background changes, utilizing smaller patches (3×3) can effectively capture finer details of the actions, leading to improved accuracy.

are used, or when the spatial resolution is reduced, PMI Sampler still matches or exceeds stronger baselines. For example, with only 8 frames it already outperforms MG Sampler while reducing training time, and with reduced spatial resolution it still yields substantial accuracy gains. This suggests that the method does not merely rearrange information more cleverly; it effectively concentrates the available temporal budget on higher-value observations.

Finally, Table 3.10 shows that PMI Sampler generalizes beyond aerial datasets, but also reveals an important nuance. On Something-Something V2, UCF101, and HMDB51, smaller patches can be more effective than the 7×7 configuration preferred on aerial data. This is consistent with the visual characteristics of these benchmarks: their videos are typically captured by fixed or near-fixed cameras, the human actor is larger in the frame, and the background is less disruptive. Under such conditions, finer local structure becomes more informative. This result is valuable because it highlights both the

generality and the dataset dependence of the method: PMI Sampler is broadly applicable, but the optimal patch granularity depends on the visual scale and motion statistics of the domain.

Overall, the ablation studies indicate that the success of PMI Sampler depends on two complementary ingredients: first, a motion salience estimate that is robust to aerial-specific noise, and second, a remapping strategy that sharpens the distinction between motion-salient and motion-static intervals. Together, these findings reinforce the interpretation of PMI Sampler as a lightweight but general temporal refinement module.

3.5.7 Qualitative Analysis

A key strength of PMI Sampler is its robustness to background noise and its ability to allocate samples according to the true temporal distribution of useful motion information. In aerial videos, camera motion can produce large pixel-level changes even when the action itself is temporally static. Methods based on raw RGB difference can therefore be misled by background fluctuations and incorrectly assign high salience to uninformative frames. In contrast, PMI Sampler estimates temporal salience through patch-level mutual information, which is more robust to such noise and better reflects action-related temporal change. This property is particularly important for aerial videos, where the actor often occupies less than 10% of the frame and the background dominates the visual signal.

Figures 3.5–3.7 compare PMI Sampler with MGSampler [3] on representative videos from Diving48, UAV-Human, and NEC-Drone. These examples illustrate two recurring observations. First, PMI Sampler is less sensitive to background-induced variation and therefore produces a motion distribution that better matches the actual progression of the action. Second, PMI Sampler allocates more samples to motion-salient intervals while avoiding oversampling temporally redundant regions.

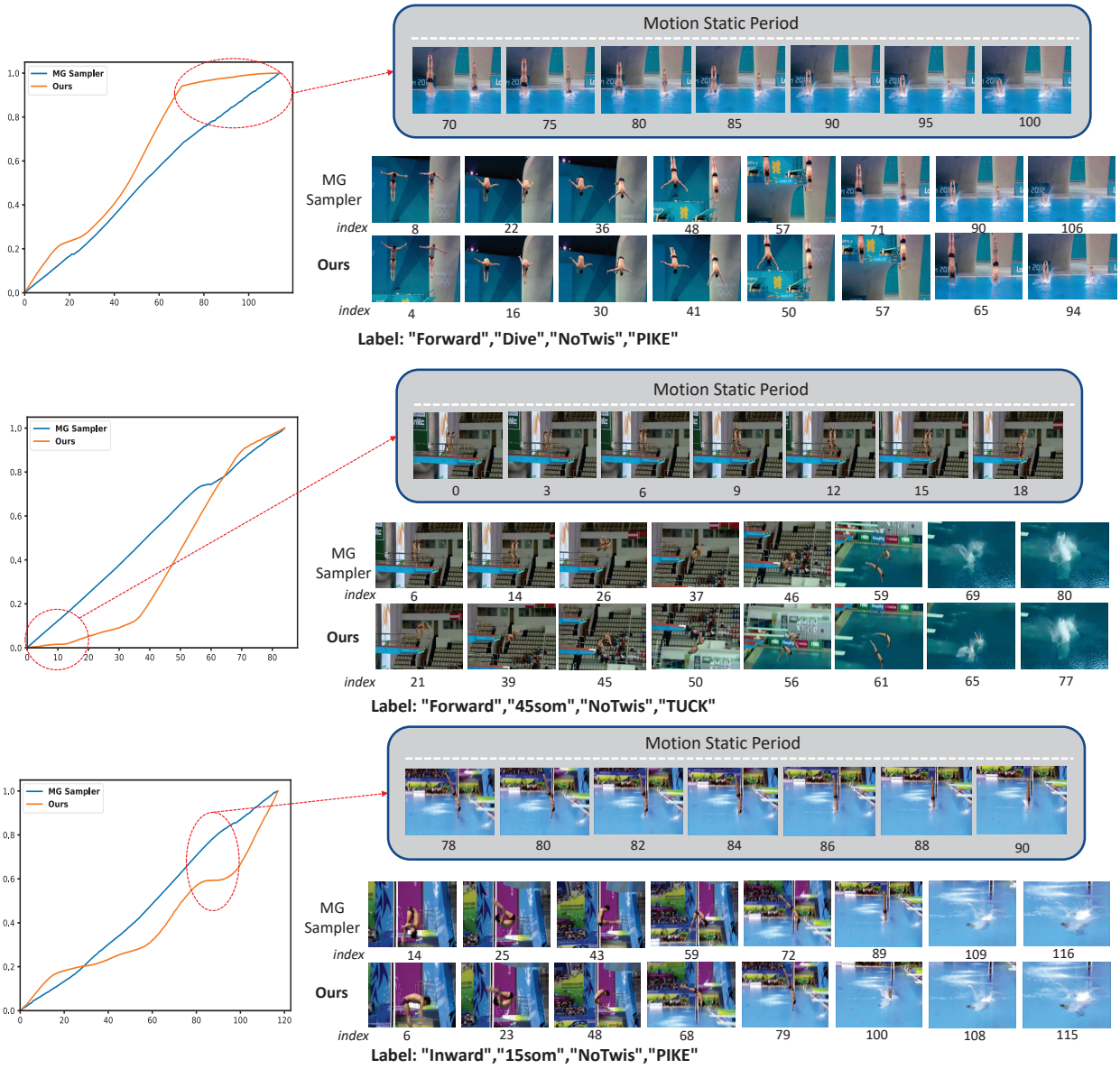


Figure 3.5: **Comparison with MGSampler on Diving48.** Representative examples from Diving48 [4] show that PMI Sampler more accurately reflects the temporal distribution of useful motion than MGSampler [3]. Even outside aerial videos, the proposed patch-mutual-information-based criterion is better aligned with the actual action progression and is less affected by irrelevant visual variation.

Figure 3.5 first shows that the advantage of PMI Sampler is not limited to aerial videos. On Diving48, which is a fine-grained and motion-centric benchmark, PMI Sampler better identifies when informative temporal change actually occurs. Compared with MGSampler, it avoids spending too many samples on visually changing but less informative intervals and instead concentrates more effectively on the discriminative phase of the action. This supports the interpretation that PMI-based salience estimation captures useful temporal structure more faithfully than simpler difference-based criteria.

The difference becomes even more pronounced in aerial videos. Figure 3.6 shows that PMI Sampler is better at recognizing motion-static intervals, even when the camera is moving and the background is highly dynamic. This is a difficult regime for RGB-difference-based methods, because large background change can create the false impression that the frame is highly informative. By relying on patch-level mutual information instead, PMI Sampler is less affected by these nuisance changes and is therefore more likely to preserve sample budget for genuinely useful motion transitions.

Figure 3.7 further shows the complementary effect on motion-salient periods. Once background noise is better suppressed, PMI Sampler can allocate more samples to the intervals that actually contain discriminative action motion. This is precisely the behavior desired for adaptive temporal selection: fewer samples should be spent on redundant or weakly informative segments, and more should be concentrated on the parts of the video that are most useful for recognition. In this sense, the qualitative evidence directly supports the chapter's main claim that adaptive sampling should follow the distribution of informative motion rather than the distribution of raw visual change.

Overall, these visualizations reinforce the same conclusion as the quantitative experiments. The main benefit of PMI Sampler is not simply that it samples fewer or different frames, but that it samples frames in a way that better matches the true temporal distribution of useful motion information. This

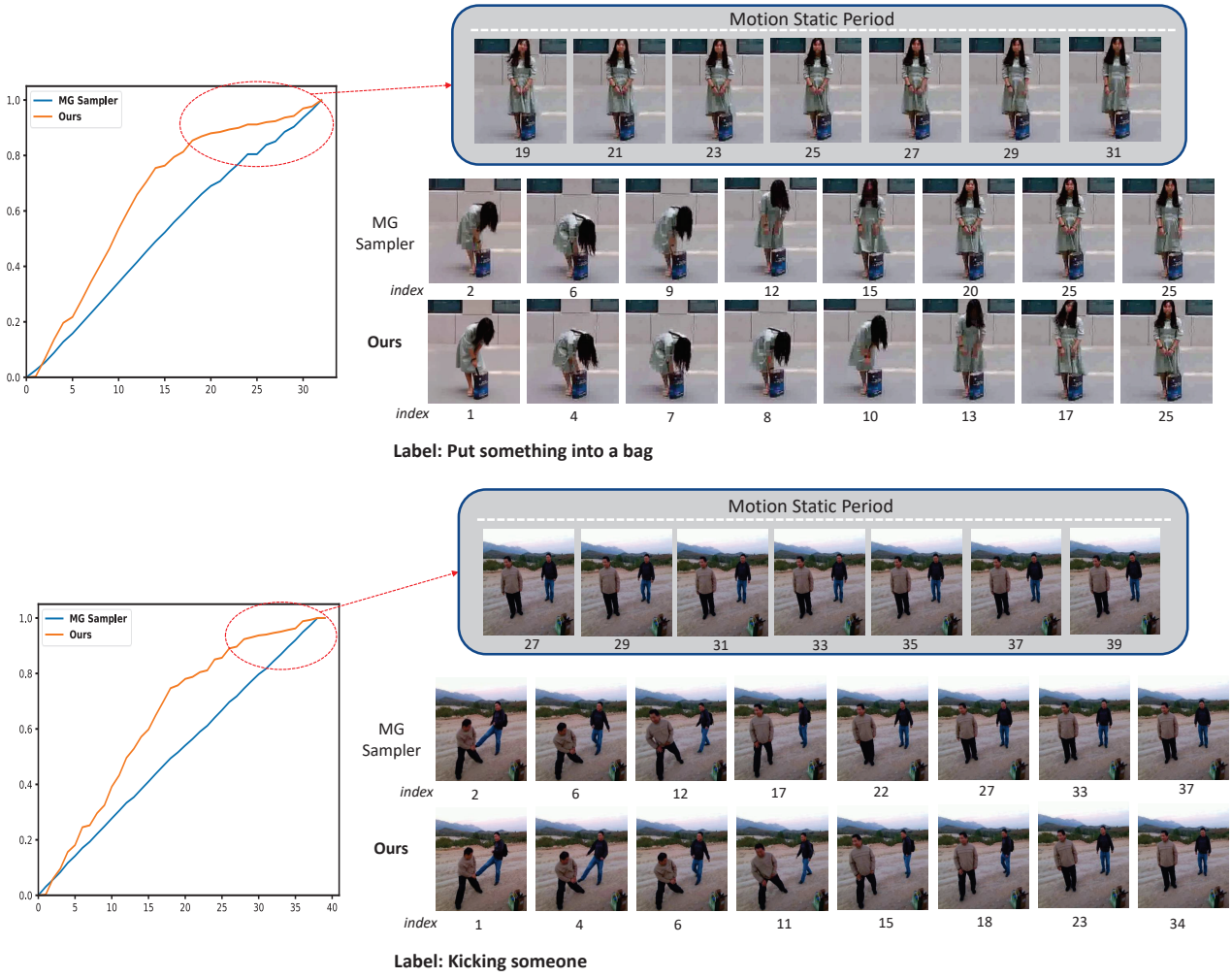


Figure 3.6: **Comparison with MGSampler on motion-static intervals in aerial videos.** Representative examples from UAV-Human [5] and NEC-Drone [6] show that PMI Sampler is more robust to UAV-induced background noise and more accurately identifies motion-static periods than MGSampler [3].

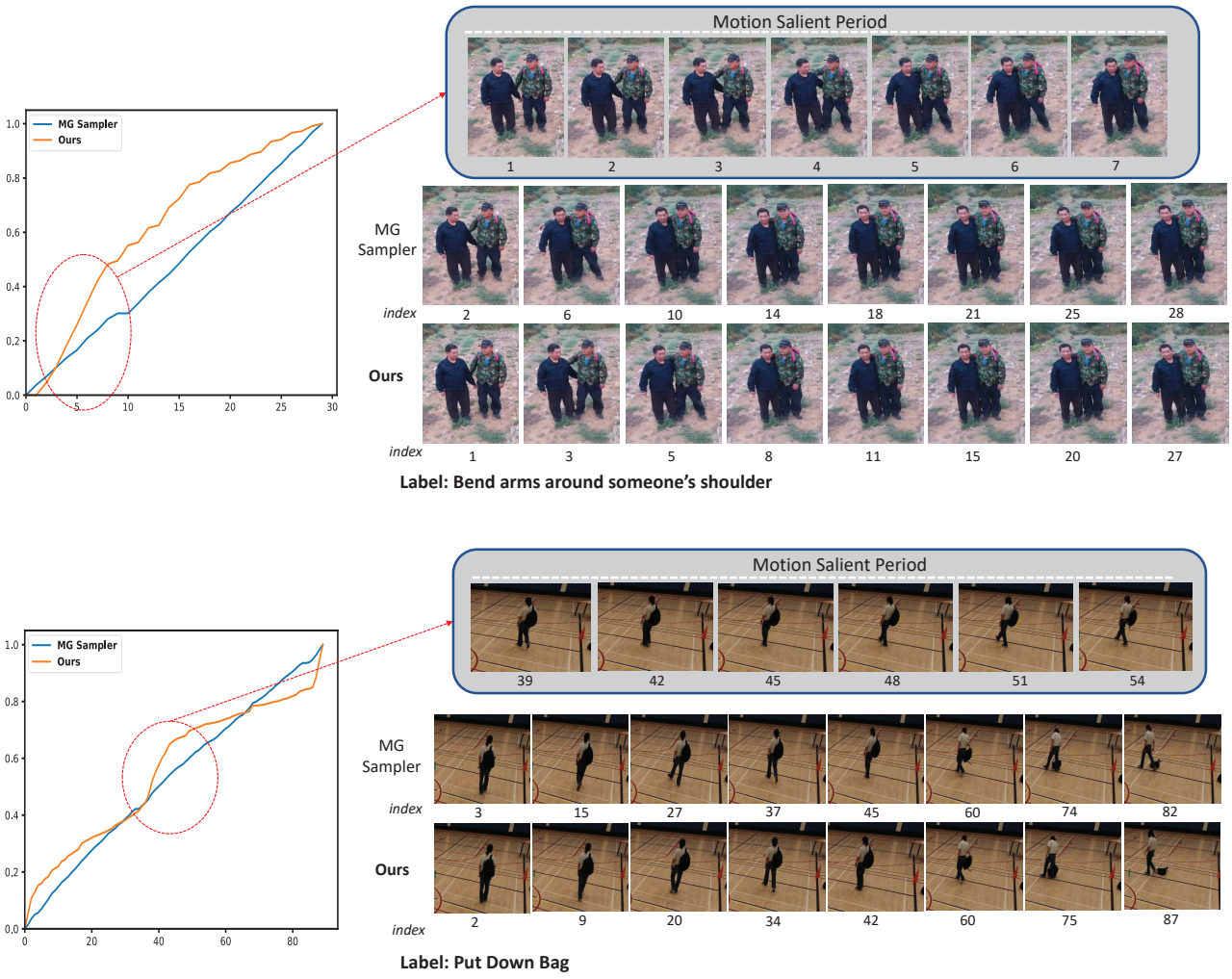


Figure 3.7: **Comparison with MGSampler on motion-salient intervals in aerial videos.** Additional examples from UAV-Human [5] and NEC-Drone [6] show that PMI Sampler allocates more samples to motion-salient periods and fewer to motion-static ones, producing a temporal allocation that is more consistent with actual action progression.

is why the method improves recognition not only on aerial datasets such as UAV-Human and NEC-Drone, but also on a non-aerial fine-grained benchmark such as Diving48.

3.6 Discussion

PMI Sampler can be understood as a lightweight temporal refinement module for aerial video understanding. Its main significance is therefore not only that it improves action recognition accuracy, but that it shows how temporal input quality itself can become a meaningful axis of representation improvement. Rather than changing the recognition backbone, the method improves the temporal observations supplied to the backbone, allowing the downstream model to focus more consistently on motion-salient and action-relevant periods.

This perspective is especially important in aerial videos. Compared with ground-view action recognition, aerial videos often contain weaker foreground signals, stronger background interference, and more severe camera-induced motion. Under these conditions, the temporal distribution of useful information becomes much more uneven, and uniform sampling is more likely to waste observations on redundant or misleading intervals. The results of this chapter therefore suggest that temporal sampling in aerial videos should not be treated merely as a preprocessing detail or an efficiency heuristic. Instead, it should be regarded as part of the representation design itself.

From the perspective of Part I, PMI Sampler complements Chapter 2 in a particularly direct way. MITFAS improves aerial understanding by making temporal observations more stable across time, whereas PMI Sampler improves it by making those observations more informative in the first place. Together, these chapters show that refined aerial video understanding depends on both *where* temporal evidence is aligned and *when* informative evidence is selected.

More broadly, the experiments also suggest that the benefit of PMI Sampler is not restricted to one aerial benchmark or one backbone architecture. Its gains on NEC-Drone, UAV-Human, Diving48, and additional backbones indicate that the core idea is more general: when informative motion is unevenly distributed over time, sampling according to motion salience can improve the quality of the temporal evidence used for recognition. In this sense, PMI Sampler supports a broader thesis claim that better aerial video understanding can be achieved not only through stronger models, but also through better-structured inputs.

3.7 Limitations

PMI Sampler has several limitations. First, when the motion is smooth and consistent throughout the video, the method may behave similarly to uniform sampling. In such cases, the temporal distribution of informative motion is already relatively even, so adaptive sampling provides less advantage.

Second, when the action label depends strongly on a static pose or gesture rather than on a motion-salient transition, sampling fewer frames from motion-static periods may reduce effectiveness. This reflects an inherent limitation of motion-oriented sampling criteria: they are strongest when recognition depends on temporal change rather than on isolated static evidence.

Third, although PMI Sampler avoids training a dedicated selection policy, it still relies on a hand-designed statistical formulation whose effectiveness depends on the chosen remapping function and patch size. Future work may benefit from combining statistical frame salience with stronger semantic or learned guidance.

Figure 3.8 illustrates two representative failure modes of PMI Sampler. When motion is smooth and uniformly distributed throughout the video, adaptive sampling behaves similarly to uniform sam-

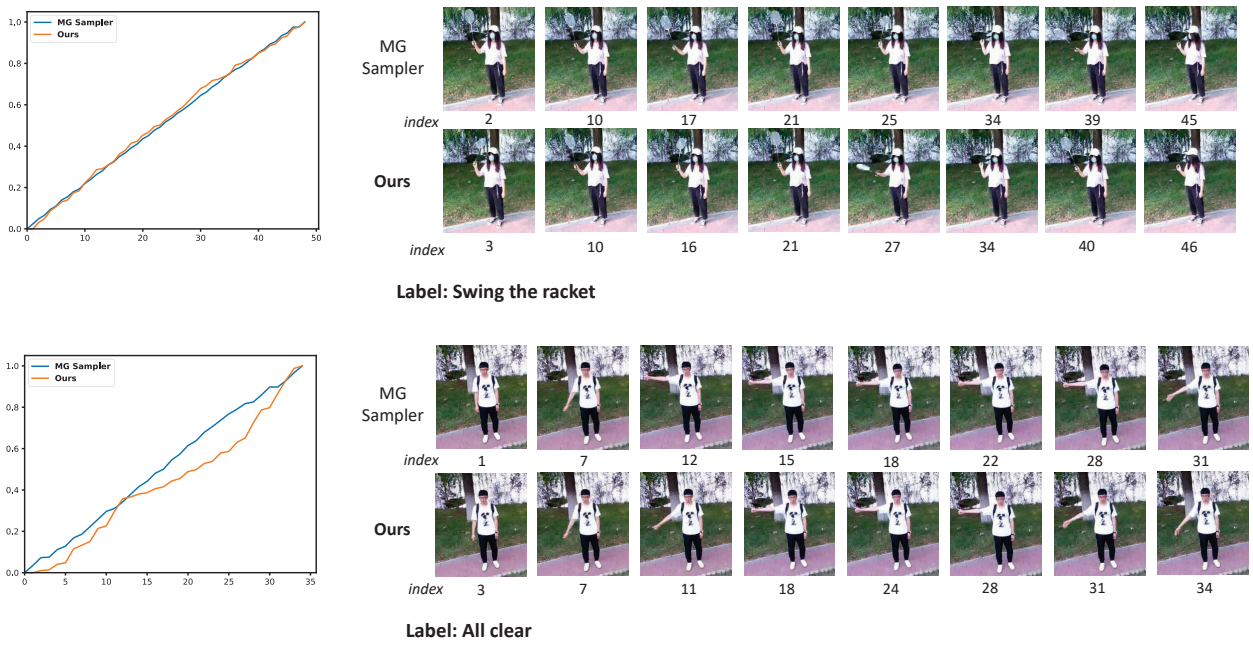


Figure 3.8: **Representative limitations of PMI Sampler.** The first example shows that when motion is smooth and consistently distributed across the whole video, PMI Sampler behaves similarly to uniform sampling. The second example shows that when the action label depends strongly on a static gesture during a motion-static interval, assigning fewer samples to that interval may reduce effectiveness.

pling. In addition, when the action label depends primarily on a static gesture rather than on a motion-salient transition, under-sampling motion-static periods may reduce effectiveness.

Finally, PMI Sampler improves temporal informativeness but does not explicitly address semantic ambiguity in aerial videos. This motivates the next chapter, which studies how semantic guidance can further refine aerial video representations once temporal evidence has been better selected.

3.8 Conclusion

This chapter presented PMI Sampler [39], a patch-similarity-guided frame selection strategy for aerial action recognition. By using patch mutual information to quantify discriminative motion information and by adaptively selecting frames according to the cumulative motion distribution, PMI

Sampler constructs more informative temporal inputs for aerial video understanding. Experimental results across UAV-Human, NEC-Drone, Diving48, and additional benchmarks demonstrate that the proposed sampling strategy consistently improves action recognition performance.

Within the broader context of Part I, this chapter complements Chapter 2 by showing that aerial video representations can be improved not only through temporal alignment, but also through better temporal selection. Whereas MITFAS focuses on making temporal observations more consistent, PMI Sampler focuses on making them more informative. Together, these results reinforce the thesis theme that refined spatial-temporal processing is essential for robust aerial video understanding.

The next chapter builds on this progression by moving from temporal selection to semantic guidance, studying how aerial video understanding can be improved further when weak visual evidence is interpreted with stronger semantic support.

Chapter 4: SCP: Soft Conditional Prompt Learning for Semantic Guidance in Aerial Video Understanding

4.1 Introduction

This chapter continues Part I, which studies refined spatial-temporal representation learning for aerial video understanding. While Chapter 2 improves aerial representations through temporal feature alignment and Chapter 3 improves them through informative frame selection, the present chapter addresses a complementary question: *how can a recognition model be better guided to interpret weak and noisy aerial visual evidence?* In other words, even if the temporal observations have been better aligned and better selected, how can the model more reliably associate those observations with the correct action semantics?

Aerial action recognition remains considerably more challenging than conventional ground-view video understanding. In aerial videos, human actors are often small, motion cues are weak, and discriminative regions may occupy only a small fraction of the frame. In addition, aerial videos commonly exhibit substantial viewpoint variation, background clutter, and camera motion, all of which can obscure the semantic cues required for action recognition. As a result, a model may fail not because informative evidence is completely absent, but because the available evidence is semantically ambiguous, weakly emphasized, or difficult to interpret under aerial-specific observation conditions.

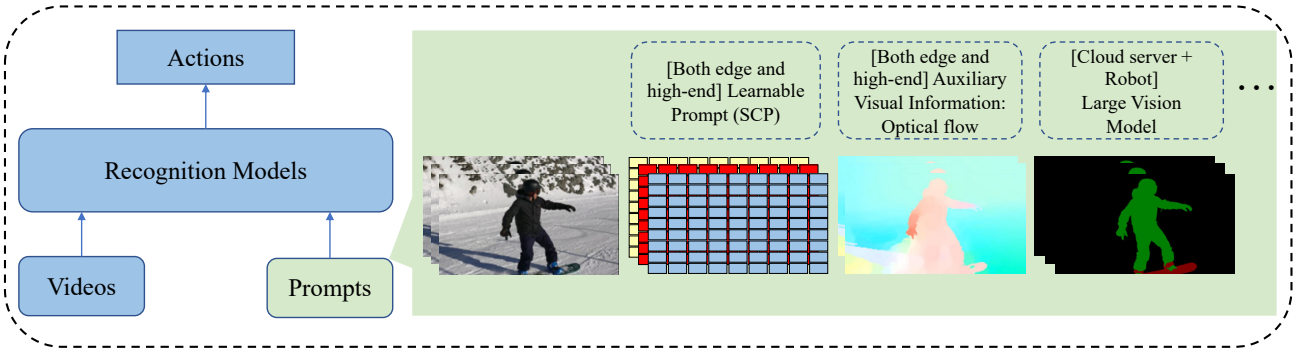


Figure 4.1: **Task Overview.** We use prompt learning for action recognition. Our method leverages the strengths of prompt learning to guide the learning process by helping models better focus on the descriptions or instructions associated with actions in the input videos. We explore various prompts, including optical flow, large vision models, and proposed SCP to improve recognition performance. The recognition models can be CNNs or Transformers.

These challenges motivate a representation learning strategy that goes beyond improving temporal evidence alone. In addition to selecting better frames or aligning temporal features, the recognition model should also be guided toward action-relevant semantic cues under aerial-specific observation conditions. Prompt learning provides a natural mechanism for such guidance. Rather than substantially redesigning the recognition backbone, prompts can provide lightweight conditioning signals that steer the model toward informative spatial-temporal patterns and improve how those patterns are semantically interpreted.

To this end, this chapter presents **Soft Conditional Prompting (SCP)** [71], a prompt-based learning framework for aerial video understanding. The key idea is to refine semantic guidance through a combination of *input-invariant prompt knowledge* and *input-adaptive prompt conditioning*. Specifically, SCP maintains a pool of learnable prompt experts that capture shared task-level information and dynamically combines them according to the input video to produce customized prompts. In this way, the model can preserve common action semantics while adapting its guidance signal to different scenes, viewpoints, and action instances.

Beyond learnable prompts, this chapter also studies auxiliary prompt forms derived from optical flow and large vision models, showing that multiple prompt sources can provide complementary guidance for action recognition. To further strengthen temporal reasoning, the prompt-enhanced visual features are processed by an auto-regressive temporal module that models sequential dependencies across the video. Figure 4.1 illustrates the overall task perspective of prompt-guided aerial action recognition, while Figure 4.2 summarizes how SCP and auxiliary prompts can be incorporated into a practical robotic perception system.

From the perspective of this dissertation, SCP establishes the third pillar of Part I. If MITFAS shows that aerial video understanding benefits from stabilizing *where* action evidence is extracted, and PMI Sampler shows that it benefits from selecting *when* informative evidence is observed, SCP shows that it also benefits from improving *how* that evidence is semantically interpreted. Together, these three chapters argue that robust aerial video understanding requires not only better temporal evidence, but also stronger semantic guidance under aerial-specific observation conditions.

The main contributions of this chapter are as follows:

- We introduce a prompt-based semantic guidance framework for aerial video understanding, showing that lightweight prompt conditioning can improve action recognition under challenging aerial observation conditions.
- We propose **Soft Conditional Prompting (SCP)**, which learns a pool of input-invariant prompt experts and dynamically generates input-adaptive prompts for different videos.
- We show that auxiliary prompt forms, including optical flow and large vision model outputs, can also provide useful semantic guidance signals for action recognition.
- We demonstrate that the proposed framework improves recognition performance across both

aerial and ground-view benchmarks, including NEC-Drone, Okutama-Action, and Something-Something V2 [71].

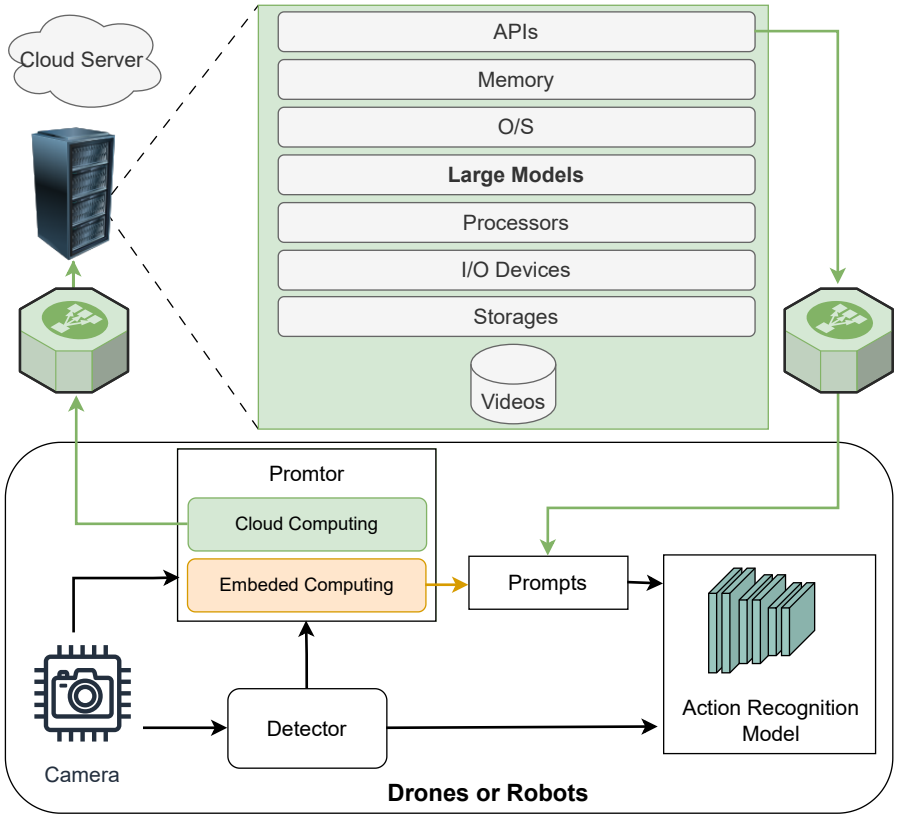


Figure 4.2: **Overall Architecture of SCP.** Our action recognition method is designed to run on edge devices (on mobile robots) and cloud servers. This includes lightweight prompts (embedded), which can be easily embedded in any action recognition model without much extra computational cost. For large vision models, we perform these computations on cloud server and use low-latency communication with the robots.

4.2 Related Work

Since Chapter 2 and Chapter 3 have already reviewed the broader background of aerial action recognition and temporal modeling, here we focus on prior work most directly related to semantic guidance and prompt-based conditioning for action recognition.

4.2.1 Prompt Learning for Visual Recognition

Prompt learning has emerged as an effective way to adapt large pretrained models to downstream tasks without fully redesigning or retraining the backbone [72–76]. Early work such as CLIP demonstrated that large-scale vision-language pretraining can produce representations that transfer well across tasks when guided by text prompts [72]. Subsequent methods showed that prompts can be optimized rather than hand-crafted, enabling more effective adaptation to downstream recognition tasks [73]. Conditional prompt learning further introduced the idea that prompts should depend on the input instance rather than remain fixed across the entire dataset [74]. Related work has also explored prompt distributions or prompt pools to improve flexibility and generalization [75, 76].

These developments are especially relevant to aerial video understanding because they suggest that prompt learning can provide lightweight semantic conditioning without substantially modifying the recognition backbone. However, most prior prompt-learning work has been developed for image recognition or vision-language adaptation rather than for aerial video understanding, where the visual evidence is weak, noisy, and temporally structured. SCP builds on this literature but focuses on prompt learning as a form of semantic guidance for aerial video action recognition.

4.2.2 Prompt-Based Action and Video Recognition

Prompt learning has also begun to influence action recognition and video understanding. Recent work has explored prompt-based adaptation for action recognition, including prompt learning for video backbones, vision-language alignment for action semantics, and prompt-based transfer of pretrained large models to temporal recognition tasks [77–79]. These works suggest that prompt-based conditioning can improve how models associate visual observations with action labels, especially when large

pretrained models already contain useful semantic priors.

At the same time, video understanding differs from image recognition in an important way: useful semantics must be inferred not only from appearance, but also from temporal evolution. This means that prompts for video understanding should ideally support both semantic discrimination and temporal reasoning. Existing prompt-based approaches do not directly address the specific challenges of aerial videos, where the actor is small, the background is dominant, and semantic cues are often ambiguous under unusual viewpoints [5, 16, 18]. SCP is motivated by this gap. Rather than using a single static prompt, it learns a pool of prompt experts and dynamically combines them according to the input video, allowing semantic guidance to adapt across different aerial scenes and action instances.

4.2.3 Semantic Guidance for Action Recognition

A broader line of work has explored how external semantic cues can improve action recognition. These cues may come from object information, motion priors, auxiliary modalities, or large pretrained models that provide richer supervisory signals [80–83]. The common intuition behind these methods is that action recognition becomes easier when the model is encouraged to focus on semantically meaningful structure rather than relying only on low-level visual evidence.

This perspective is especially important in aerial videos. In many aerial settings, the visual evidence for an action is weak and can be difficult to interpret directly. Semantic guidance can therefore serve as a complementary mechanism that improves how the model interprets whatever spatial-temporal evidence is available. SCP follows this perspective, but differs from prior work by casting semantic guidance in a prompt-learning framework. In particular, it combines input-invariant prompt knowledge with input-adaptive prompt conditioning, and further studies auxiliary prompt sources such

as optical flow and large vision model outputs. This makes SCP a natural semantic counterpart to the alignment and sampling strategies developed in the previous two chapters.

4.3 Problem Formulation and Motivation

Consider an input video

$$V = \{F_1, F_2, \ldots, F_T\},$$

where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$, and the task is to predict an action label

$$y \in \{1, \ldots, C\}.$$

In a standard video recognition pipeline, the model learns a direct mapping

$$\hat{y} = f(V),$$

where f is typically implemented using a visual backbone followed by a temporal reasoning module.

This formulation is often insufficient for aerial video understanding. In aerial videos, action-relevant evidence is frequently weak, spatially sparse, and difficult to interpret reliably. The actor may occupy only a small fraction of the frame, appear under unusual viewpoints, or be partially obscured by clutter and camera-induced motion. As a result, the model may fail not because the evidence is entirely missing, but because the available evidence is semantically ambiguous and difficult to associate with the correct action concept.

To address this issue, we introduce a prompt-conditioned recognition formulation

$$\hat{y} = f([V, P]),$$

where P denotes a guidance signal that conditions the recognition model. The role of the prompt is not to replace visual representation learning, but to steer it by emphasizing action-relevant spatial-temporal evidence and encouraging the model to interpret that evidence through a more discriminative semantic prior.

A useful prompt mechanism for aerial videos should satisfy two properties. First, it should encode shared task-level semantics that remain useful across videos. Second, it should remain adaptive to the specific input, since different scenes, viewpoints, and action instances may require different forms of guidance. These observations motivate the design of **Soft Conditional Prompting (SCP)**, which maintains a pool of prompt experts and dynamically combines them according to the input video.

4.3.1 Prompt Definition in Our Setting

Before introducing the proposed method, we clarify what *prompts* mean in our setting. In conventional vision–language models, prompts are usually textual templates or learned context tokens that adapt a pretrained model toward downstream classification. In this chapter, however, we use the term more broadly.

Specifically, we define prompts as *auxiliary conditioning signals that guide spatial-temporal representation learning for aerial videos*. Under this view, prompts are not limited to fixed text templates. Instead, they may encode semantic guidance, spatial structure, or temporal/motion-related cues

that help the model focus on more informative evidence in aerial observations.

This distinction is important because aerial videos differ substantially from conventional ground-view videos. Informative action cues are often weak, spatially sparse, and easily dominated by cluttered background regions. As a result, effective prompting in aerial video understanding should not only adapt class semantics, but also guide the model toward more relevant spatial-temporal patterns.

Based on this perspective, SCP introduces *soft conditional prompts* that are input-dependent and directly influence representation learning. Therefore, SCP should be understood not simply as prompt tuning in the text space, but as a semantic-aware prompting framework for improving spatial-temporal representations in aerial video understanding.

4.4 Method

4.4.1 Overview

SCP follows a prompt-conditioned recognition framework for aerial video understanding. Given an input video X, the model predicts the action label through

$$f = f_a \circ f_e([X, P]), \tag{4.1}$$

where f_e denotes the prompt-guided input encoder, P denotes the prompt, and f_a denotes the temporal reasoning module.

The framework contains two main components. The first is a *prompt-guided input encoder*, which injects semantic guidance into visual feature extraction. The second is an *auto-regressive temporal reasoning module*, which models temporal dependencies over the prompt-enhanced features.

As shown in Figure 4.3, the overall framework supports multiple prompt forms, including learnable prompts, optical-flow-based prompts, and large-vision-model-based prompts. Among these, the proposed **Soft Conditional Prompting (SCP)** serves as the central learnable prompt mechanism in this chapter.

The rest of this section is organized as follows. We first describe prompt-guided input encoding, then present the proposed SCP mechanism, followed by auxiliary prompt forms, temporal reasoning, and learning objectives.

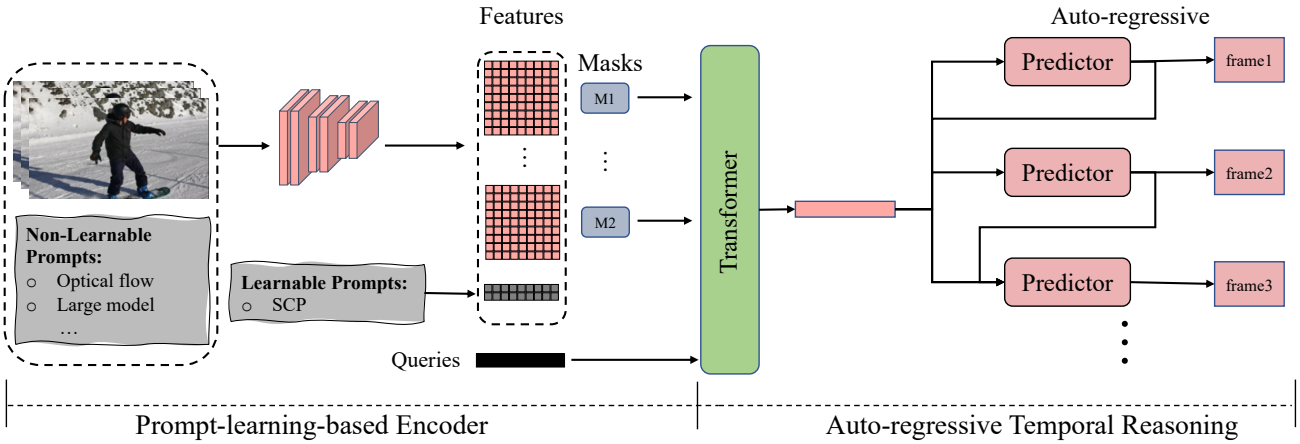


Figure 4.3: **Overview of the action recognition framework.** We use transformer-based action recognition methods as an example. We designed a prompt-learning-based encoder to help better extract the feature and use our auto-regressive temporal reasoning algorithm for recognition models for enhanced inference ability.

4.4.2 Prompt-Guided Input Encoding

Prompt learning aims to improve representation learning by providing additional task-relevant guidance to the model input. In aerial video understanding, such guidance is particularly useful because the actor is often small and action-relevant cues can be easily overwhelmed by background variation, camera motion, and viewpoint change. By conditioning the encoder on prompt signals, the model can more effectively focus on action-relevant regions and capture more discriminative spatial-

temporal patterns.

Formally, given an input video X, the prompt-guided encoder takes as input the combination $[X, P]$, where P may be either predefined or learnable. The output of the encoder is a prompt-enhanced feature sequence

$$X^p = \{x_1^p, x_2^p, \ldots, x_m^p\}.$$

These features are then processed by the temporal reasoning module for action prediction. In this chapter, prompts are not treated as generic add-ons, but as semantic guidance signals that help the model better associate weak aerial evidence with the corresponding action semantics.

4.4.3 Soft Conditional Prompting

A single static prompt is often insufficient for aerial video understanding. Different videos may involve different action categories, scene layouts, camera viewpoints, motion patterns, and levels of ambiguity. A prompt that is useful for one input may therefore be suboptimal for another. This motivates a prompt mechanism that can preserve shared task-level information while still adapting to the specific content of each video.

To address this, we propose **Soft Conditional Prompting (SCP)**, which dynamically generates prompts from a pool of prompt experts. As illustrated in Figure 4.4, the key idea is to combine two forms of prompt knowledge: *input-invariant prompt experts*, which capture shared task information across the dataset, and *input-adaptive prompt composition*, which allows the prompt to change with the input video.

4.4.3.1 Prompt Expert Pool

SCP maintains a pool of learnable prompt experts

$$P = \{P_1, P_2, \ldots, P_l\}, \tag{4.2}$$

where each P_k denotes a learnable prompt token and l is the number of experts. These prompt experts are shared across all training samples and are updated through the training process. Their role is to encode prompt knowledge that remains broadly useful across videos, such as shared action semantics or recurrent spatial-temporal patterns.

Unlike a single learned prompt, the expert-pool design allows the model to represent prompt knowledge in a more flexible way. Different experts may capture different aspects of the task, while the final prompt can be composed differently for different videos.

4.4.3.2 Input-Adaptive Prompt Generation

For a specific input video X^*, SCP computes dynamic weights that determine how the prompt experts should be combined:

$$P^* = \mathrm{Matmul}(\sigma(\mathrm{FC}(X^*)), P), \tag{4.3}$$

where $\mathrm{FC}(\cdot)$ is a fully connected projection and $\sigma(\cdot)$ denotes the sigmoid activation. The resulting weight vector is input-dependent and allows the model to generate a customized prompt P^* for the current video.

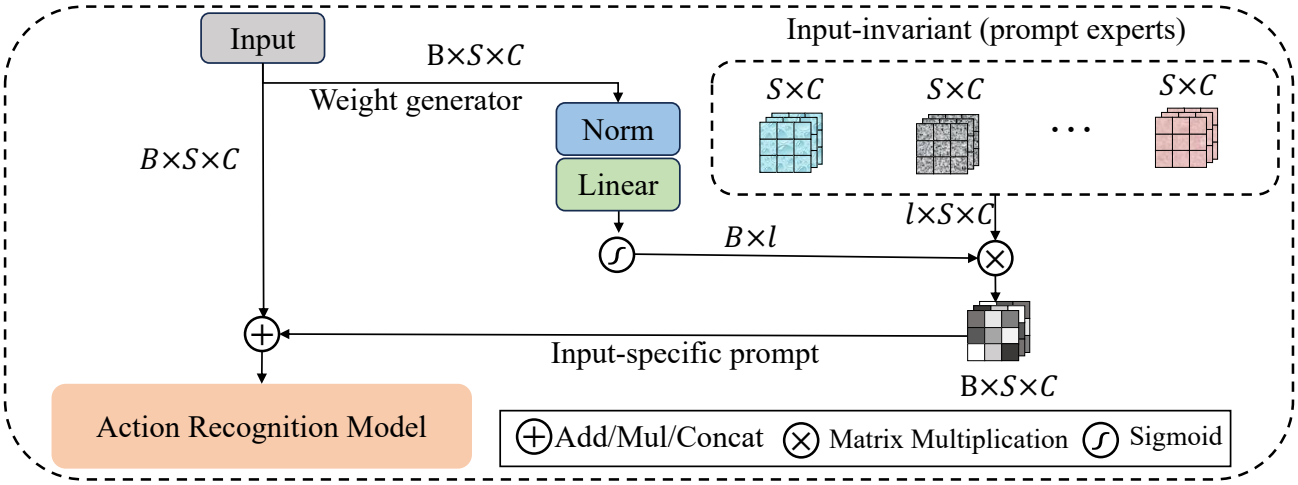


Figure 4.4: **Soft Conditional Prompt Learning (SCP):** Learning input-invariant (prompt experts) and input-specific (data dependent) prompt. The input-invariant prompts will be updated from all the inputs, which contain task information, and we use a dynamic mechanism to generate input-specific prompts for different inputs. Add/Mul means element-wise operations. $B \times S \times C$ is the input features' shape, and l is the expert's number in the prompt pool.

The prompt-enhanced feature for frame x_i is then written as

$$x_i^p = f_e([x_i, p_i]), \qquad x_i \in X^*,\ p_i \in P^*, \tag{4.4}$$

where f_e denotes the prompt-guided encoder. In this design, the prompt experts capture shared semantic priors, while the dynamic composition mechanism adapts the prompt to the current video. This allows SCP to jointly model *input-invariant prompt knowledge* and *input-specific semantic guidance*.

This formulation is especially suitable for aerial video understanding. Because aerial videos vary substantially in viewpoint, motion strength, and scene context, a fixed prompt may not be expressive enough to provide useful guidance across all inputs. SCP instead provides a soft and adaptive form of semantic conditioning, allowing the model to preserve common action knowledge while tailoring the prompt to the specific visual evidence available in each video.

4.4.4 Auxiliary Prompt Forms

Although SCP is the primary learnable prompt mechanism in this chapter, the framework also supports non-learnable prompt forms that provide auxiliary semantic cues. These additional prompts are useful for understanding the broader design space of prompt-guided aerial video understanding and for evaluating whether external sources of guidance can complement learnable prompts.

4.4.4.1 Optical Flow Prompt

Optical flow provides explicit motion information between frames and can serve as a prompt that highlights dynamic regions. For adjacent clips or selected frames, the optical flow is computed as

$$o_i = O(x_i, x_j), \tag{4.5}$$

where $O(\cdot, \cdot)$ denotes the optical flow operator and x_i, x_j are selected frames from neighboring clips. The resulting prompt-conditioned input is written as

$$[X, P] = \{x_k * o_i \mid x_k \in clip_i\}, \tag{4.6}$$

where $*$ denotes element-wise fusion.

The optical flow prompt is useful because it explicitly emphasizes motion information, which can guide the recognition model toward action-relevant temporal change. At the same time, unlike SCP, it does not learn semantic prompt knowledge from data and therefore serves as a complementary but less adaptive guidance mechanism.

4.4.4.2 Large Vision Model Prompt

Large vision models can also provide useful high-level guidance without requiring task-specific retraining. In this chapter, we use the Segment Anything Model (SAM) to generate mask-based prompts from image prompts such as points or bounding boxes. For frame x_i, the generated prompt is

$$p_i = SAM(x_i, \text{boxes/points}), \tag{4.7}$$

and the prompt-conditioned input becomes

$$[X, P] = \{x_i * p_i \mid i \in [1, m]\}. \tag{4.8}$$

These mask-based prompts can guide the model toward semantically important regions in the frame and can therefore serve as another form of external semantic prior. In the context of this chapter, they help illustrate that prompt-guided aerial action recognition is not restricted to one specific prompt type, but instead supports a broader family of semantic guidance signals.

4.4.5 Auto-Regressive Temporal Reasoning

Prompt-enhanced features still require temporal reasoning in order to support action recognition. To model temporal dependencies more effectively, SCP uses an auto-regressive temporal reasoning module over the feature sequence

$$X^p = \{x_1^p, x_2^p, \ldots, x_m^p\}.$$

Let x_i^p denote the prompt-enhanced feature at time step i. The model predicts the subsequent

temporal representation through

$$\hat{x}_{i+1}^{p} = f_a\left(\prod_{j<i+1} f_a(x_j^p) + x_{i+1}^p\right), \tag{4.9}$$

where f_a denotes the auto-regressive temporal reasoning module. The role of this component is to accumulate temporal context over the sequence and improve the model's ability to reason about action progression after the visual evidence has already been semantically conditioned by prompts.

In the thesis context, this temporal component should not be viewed as the central novelty of the chapter. Rather, it is a supporting module that complements the prompt-guided semantic conditioning by ensuring that the semantically enhanced features are also temporally integrated.

4.4.6 Learning Objectives

The framework supports both single-agent and multi-agent action recognition.

For single-agent action recognition, we use the standard cross-entropy loss:

$$L_n = -\sum_{c=1}^{C} y_{n,c} \log \frac{\exp(\hat{x}_{n,c}^p)}{\sum_{i=1}^{C} \exp(\hat{x}_{n,i}^p)}, \tag{4.10}$$

where C is the number of classes, $\hat{x}_{n,c}^p$ is the predicted logit for class c, and $y_{n,c}$ is the ground-truth label indicator.

For multi-agent action recognition, where multiple action labels may be present, we use binary cross-entropy with logits:

$$L_{n,c} = -\left[y_{n,c} \log \sigma(\hat{x}_{n,c}^p) + (1 - y_{n,c}) \log(1 - \sigma(\hat{x}_{n,c}^p))\right], \tag{4.11}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

These objectives allow the same prompt-guided framework to operate across both single-agent and multi-agent settings, making SCP applicable to a broader range of aerial and robotic video understanding problems.

4.5 Experimental Evaluation

In this section, we evaluate SCP on both aerial and ground-view action recognition benchmarks in order to assess two questions. First, does prompt-based semantic guidance improve action recognition under challenging aerial observation conditions? Second, do the benefits of SCP generalize beyond aerial datasets to broader video understanding settings?

4.5.1 Datasets and Evaluation Setting

We evaluate SCP on three representative benchmarks that cover controlled aerial recognition, more challenging multi-agent aerial recognition, and large-scale ground-view video understanding.

For NEC-Drone, we follow the same dataset introduced in Chapters 2 and 3. In this chapter, NEC-Drone mainly serves as a controlled aerial benchmark for evaluating whether prompt-based semantic guidance improves recognition when actors remain small and action cues are visually weak [6].

For Okutama-Action, we evaluate SCP in a more challenging multi-agent aerial setting. Compared with NEC-Drone, Okutama-Action involves more complex aerial scenes, multiple people, and stronger ambiguity in the spatial distribution of action-relevant evidence, making it especially suitable for evaluating semantic guidance under realistic aerial observation conditions [17].

For Something-Something V2, we follow the same benchmark setting introduced earlier in Chapter 3. In this chapter, it is used to evaluate whether the benefits of SCP extend beyond aerial datasets to a broader video recognition setting in which semantic interpretation of object interactions and temporal change remains important [68].

We compare against baseline and prior state-of-the-art methods on top-1 and top-5 accuracy, and evaluate SCP in both aerial and non-aerial domains to test whether prompt-based semantic conditioning should be understood as a general recognition enhancement rather than as a dataset-specific design.

4.5.2 Implementation Details

Unless otherwise specified, experiments follow the original paper settings. For NEC-Drone, training uses SGD with momentum 0.9 and weight decay 5×10^{-4}, with the initial learning rate set to 0.1 for training from scratch and 0.05 for Kinetics-pretrained initialization. Videos are resized to 224×224 during training and evaluated with multi-crop testing after resizing the shorter side to 256. For Okutama-Action, ROI-aligned features are extracted and optimized with the corresponding multi-label objective. For Something-Something V2, we follow the MViTv2 fine-tuning protocol used in the original SCP implementation. Unless otherwise stated, the rest of the setup follows the original SCP paper.

4.5.3 Main Results

The main quantitative results are shown in Tables 4.1, 4.2, and 4.3. Together, these results evaluate SCP under three complementary perspectives: performance on a controlled aerial benchmark,

performance on a more challenging multi-agent aerial benchmark, and transfer to a large-scale non-aerial video benchmark.

Method	pretrain	Top-1 Acc (%) ↑
X3D-random [2]	Kinetics400	52.0
X3D-uniform [2]	Kinetics400	55.4
K-centered [84]	N/A	36.3
SCP-uniform (Ours)	Kinetics400	59.4

Table 4.1: **Comparison with existing methods on NEC-Drone.** Our SCP improves 4.0-7.4% over X3D and 23.1% over K-centered.

Table 4.1 shows that SCP improves performance on NEC-Drone over baseline sampling-based alternatives. In particular, SCP reaches 59.4% Top-1 accuracy, improving over X3D-uniform by **4.0%** and over K-centered by **23.1%**. This result is meaningful because NEC-Drone is already a relatively controlled aerial benchmark. The gain therefore suggests that SCP is not only helpful in highly cluttered aerial scenes, but also beneficial in settings where the main challenge lies in weak actor visibility and semantic ambiguity rather than in large scene complexity alone. In other words, even when temporal sampling and backbone choice are fixed, better semantic guidance still improves aerial recognition.

Table 4.2 further shows that the benefits of SCP become even more pronounced on Okutama-Action, which is substantially more challenging than NEC-Drone due to multi-agent activity, cluttered aerial scenes, and stronger ambiguity in action-relevant regions. Without bounding boxes, SCP reaches 71.54%, outperforming DroneAttention by **3.17%**. With bounding-box guidance, SCP reaches 75.93%, exceeding the corresponding prior method by **10.20%**. These gains are especially important because they suggest that semantic guidance is most valuable when the visual scene is more complex and when multiple candidate actors or activities compete for attention. In such settings, prompt conditioning appears to help the model better associate the available aerial evidence with the correct action semantics.

Method	Frame size	Top-1 Acc (%) ↑
AARN [85, 86]	crops	33.75
Lite ECO [86, 87]	crops	36.25
I3D(RGB)[23, 86]	crops	38.12
3DCapsNet-DR[86, 88]	crops	39.37
3DCapsNet-EM[86, 88]	crops	41.87
DroneCaps[86]	crops	47.50
DroneAttention without bbox[89]	720×420	61.34
SCP without bbox (Ours)	224×224	71.54
DroneAttention with bbox [89]	720×420	72.76
SCP with bbox (Ours)	224×224	75.93

Table 4.2: **Comparison with the state-of-the-art results on the Okutama dataset.** With bbox information, we achieved 10.20% improvement over the SOTA method. Without bbox information, we outperformed the SOTA by 3.17%.

Method	pretrain	Top-1 Acc (%) ↑	Top-5 Acc (%) ↑
TEA [90]	ImageNet 1k	65.1	89.9
MoViNet-A3 [43]	N/A	64.1	88.8
ViT-B-TimeSformer [28]	ImageNet 21k	62.5	/
SlowFast R101, 8×8 [26]	Kinetics400	63.1	87.6
MViTv1-B, 16×4 [45]	Kinetics400	64.7	89.2
MViTv2-S, 16×4 [91]	Kinetics400	67.3	91.0
SCP (Ours)	Kinetics400	68.3	91.4

Table 4.3: **Comparison with the state-of-the-art results on the Something Something V2.** Our SCP improves 3.6% over MViTv1 and 1.0% over strong SOTA MViTv2.

A second important observation from Table 4.2 is that SCP remains strong both with and without explicit bounding-box input. This indicates that the benefit of the method does not depend solely on precise external localization. Instead, the learned prompt mechanism itself contributes meaningful semantic guidance, which is consistent with the main thesis of this chapter.

Finally, Table 4.3 shows that the benefit of SCP generalizes beyond aerial video understanding. On Something-Something V2, SCP improves over strong recent baselines, including MViTv1 and MViTv2. This result is important because it shows that prompt-based semantic guidance is not only useful under aerial-specific conditions, but also beneficial in broader video recognition settings where

semantic interpretation of temporal interactions matters. At the same time, the improvement is smaller than on Okutama-Action, which is consistent with the intuition that semantic guidance becomes especially valuable when the visual evidence is weaker or more ambiguous, as is often the case in aerial videos.

Taken together, the results in Tables 4.1–4.3 support the central claim of this chapter: even after temporal evidence has been better aligned and better selected, recognition can still benefit substantially from improved semantic guidance. SCP therefore complements the previous chapters by refining not *where* or *when* evidence is observed, but *how* that evidence is interpreted.

4.5.4 Comparison Across Prompt Forms

A key question is whether the gains observed in the previous section come specifically from SCP or more generally from the use of prompt guidance. To study this, we compare several prompt forms, including optical flow prompts, large vision model prompts, and the proposed learnable SCP prompts. The results are shown in Table 4.4.

Method	Frame size	Top-1 Acc (%) ↑
Baseline	224×224	71.54
Baseline + Optical Flow	224×224	72.13
Baseline + Large Vision Model (SAM)	224×224	74.68
Baseline + SCP	224×224	75.93

Table 4.4: **Ablation study in terms of different prompts on the Okutama dataset.** We evaluated various prompts, including optical flow, a large vision model(SAM [1]), and SCP. The large vision model and SCP achieved better accuracy.

Table 4.4 shows that prompt guidance is beneficial more broadly, but that different prompt forms are not equally effective. Compared with the baseline, optical flow prompts provide a small improvement, suggesting that explicit motion cues can help guide the model toward action-relevant temporal

change. Large vision model prompts yield a larger gain, which indicates that stronger region-level semantic priors are particularly helpful in aerial videos, where the actor is small and difficult to isolate directly from raw RGB frames. SCP achieves the best result overall, improving from 71.54% to 75.93%. This result is important because it suggests that learnable and input-adaptive semantic guidance is more effective than fixed auxiliary prompts alone. In other words, prompt guidance is useful in general, but the adaptive and expert-based formulation of SCP provides the strongest form of guidance among the tested prompt types.

4.5.5 Component Analysis

To understand the contribution of different parts of the full framework, we next perform a component-wise analysis on Okutama-Action. The results are shown in Table 4.5.

Component	Frame size	Top-1 Acc (%) ↑
Baseline	224x224	71.54
Baseline + ROI	224x224	73.61
Baseline + ROI + Large Vision Model (SAM)	224x224	74.68
Baseline + ROI + SCP	224x224	76.34

Table 4.5: **Ablation study in terms of the effect of different components in our method on the Okutama dataset.** We evaluated ROI, Large Vision Model (SAM), and SCP. The experiments showed the effectiveness of our proposed methods.

Table 4.5 shows that each component contributes positively, but their effects are not identical. Adding ROI alignment already improves over the baseline, which is expected because it helps the model focus on the most relevant region of the frame. Adding the large vision model prompt yields a further gain, indicating that external semantic priors provide additional useful guidance beyond spatial focusing alone. The best result is achieved by replacing that auxiliary prompt with SCP, which increases accuracy to 76.34%. This is especially informative because it shows that the main benefit

does not come merely from providing any extra guidance signal, but from using a learnable and input-adaptive semantic conditioning mechanism. The component analysis therefore supports the core claim of this chapter: semantic guidance is most effective when it is both structured and adaptive.

4.5.6 Prompt Expert Analysis

We also analyze the number of prompt experts used in the SCP prompt pool. This experiment is important because SCP relies on a balance between shared prompt knowledge and input-adaptive flexibility. The corresponding results are shown in Table 4.6.

Method	Frame size	Top-1 Acc (%) ↑
Baseline	224x224	71.54
Baseline + SCP with 4 Experts	224x224	76.16
Baseline + SCP with 8 Experts	224x224	76.34
Baseline + SCP with 16 Experts	224x224	75.93
Baseline + SCP with 32 Experts	224x224	73.70

Table 4.6: **Ablation study in terms of different experts number on the Okutama dataset.** We evaluated various experts number including 4, 8, 16, and 32. From our experiment, with the expert number of 8 achieved better accuracy.

Table 4.6 shows that a moderate number of prompt experts provides the best trade-off between flexibility and optimization stability. Increasing the expert number from 4 to 8 yields a small improvement, suggesting that a richer prompt pool helps the model encode more diverse shared task knowledge. However, performance begins to decrease once the number of experts becomes too large. This trend indicates that an excessively large prompt pool may make the conditional combination problem harder to optimize and may weaken the stability of the learned semantic guidance. The best result is achieved with 8 experts, which supports the design choice adopted in the main experiments. More broadly, this experiment reinforces the idea that SCP works by combining *shared prompt structure* with *input-adaptive prompt generation*; both under-capacity and over-capacity can weaken this

balance.

4.5.7 Large Vision Model Prompt Analysis

To better understand the behavior of large vision model prompts, we further compare different input forms used to generate prompts from SAM. The results are shown in Table 4.7.

Method	Frame size	Top-1 Acc (%) ↑
Large Vision Model (SAM) with 1 point	224x224	58.12
Large Vision Model (SAM) with 2 points	224x224	58.02
Large Vision Model (SAM) with 4 points (Line)	224x224	66.30
Large Vision Model (SAM) with bbox point	224x224	74.68

Table 4.7: **Ablation study in terms of different inputs for large vision model on the Okutama dataset.** We evaluated various inputs including single point, two points, four points, and bbox. From our experiment, the large vision model with bbox achieved better accuracy.

Table 4.7 shows that the quality of externally generated prompts depends strongly on the prompt input format. Sparse point-based prompts lead to relatively weak performance, while more structured inputs such as line prompts and especially bounding-box prompts produce much stronger results. This trend is meaningful because it suggests that large vision model prompts are only as useful as the spatial guidance used to generate them. Bounding-box-based prompts perform best because they provide more stable and complete coverage of the actor region, which is especially important in aerial videos where the foreground is small and ambiguous. This analysis also clarifies the role of large vision model prompts in this chapter: they provide a valuable source of auxiliary semantic guidance, but their effectiveness still depends on the quality and stability of the prompting signal.

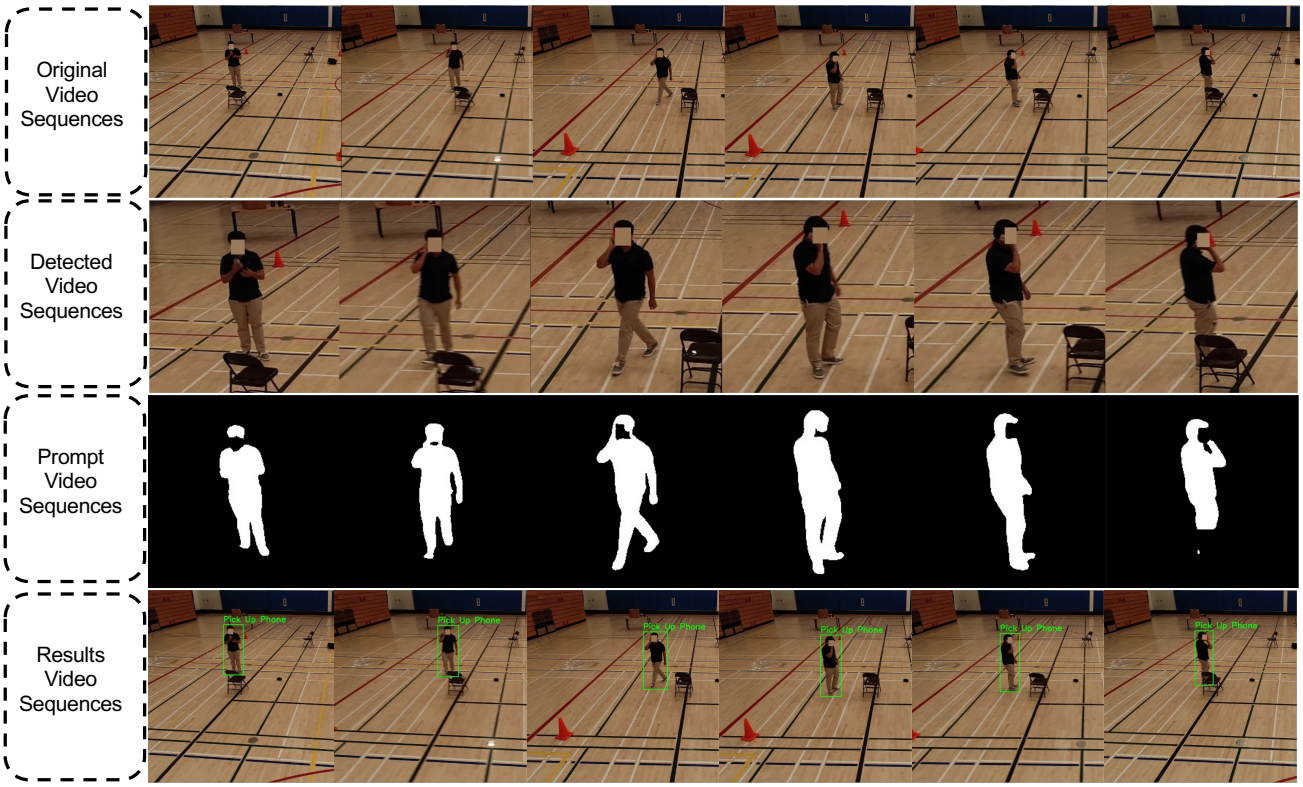


Figure 4.5: **Visualization.** We first detect the interested target and generate the prompts, then predict the action.

4.5.8 Qualitative Analysis

To further understand how prompt guidance affects recognition, we visualize both the prompt-guided action recognition process and the outputs produced by large vision model prompts. These qualitative results are shown in Figures 4.5 and 4.6.

Figure 4.5 illustrates the overall behavior of prompt-guided recognition. The visualization shows that the framework first identifies the target region, then conditions recognition on the corresponding prompt, and finally predicts the action. This process helps clarify the role of semantic guidance in the chapter: the prompt is not simply an auxiliary feature, but an explicit mechanism for directing the model toward action-relevant evidence before temporal reasoning is applied.

Figure 4.6 further illustrates the behavior of prompts generated from a large vision model. Dif-

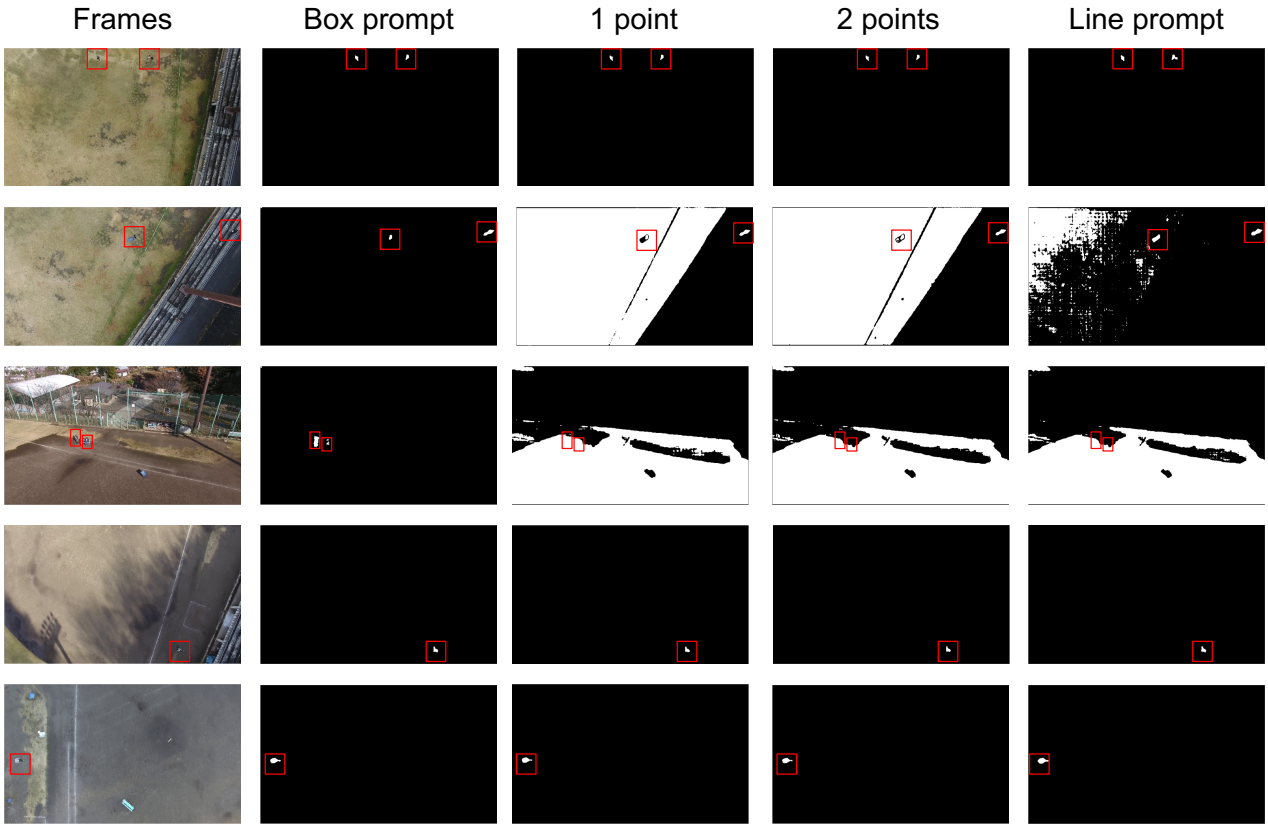


Figure 4.6: **Large Vision Model Prompts from the large vision model, no supervision needed.** We visualize the outputs in terms of different prompts, including bbox, line, and different points. bbox and line have more stable outputs, which means better prompts result in better outputs.

ferent input prompt types lead to noticeably different outputs, and the more structured prompts, such as bounding boxes and line-based inputs, produce more stable and complete masks than sparse point prompts. This is consistent with the quantitative results in Table 4.7. Together, these visualizations suggest that prompt guidance helps the model focus on action-relevant targets and regions, especially under challenging aerial viewpoints where the actor is small, visually weak, or partially ambiguous.

Taken together, the analyses in this section provide a more detailed picture of why SCP is effective. Prompt guidance is useful more generally, but SCP provides the strongest gains because it learns a structured pool of prompt experts and generates prompts adaptively for each input video. At the same time, auxiliary prompts derived from optical flow and large vision models reveal a broader

design space for semantic guidance, suggesting that future work may benefit from combining learned and externally generated prompts in a more unified framework.

4.6 Discussion

SCP can be understood as a lightweight semantic guidance framework for aerial video understanding. Its main significance is therefore not only that it improves recognition accuracy on NEC-Drone, Okutama-Action, and Something-Something V2, but that it shows how weak and noisy aerial evidence can be interpreted more reliably when the model is guided by adaptive semantic priors. In this setting, the challenge is not merely to extract stronger visual features, but to help the model associate limited and ambiguous visual observations with the correct action semantics.

This perspective helps explain why SCP is a natural complement to the previous two chapters. Chapter 2 showed that aerial video understanding benefits from making temporal features more stable. Chapter 3 showed that it also benefits from making temporal inputs more informative. SCP addresses the remaining step: even when evidence has been better aligned and better selected, the model must still interpret that evidence correctly. The role of prompt-guided conditioning is precisely to improve this semantic interpretation process.

The experimental results support this view. The strongest gains appear on more challenging aerial settings, especially Okutama-Action, where multiple agents, cluttered scenes, and ambiguous action-relevant regions make semantic interpretation particularly difficult. This suggests that semantic guidance is most valuable not when evidence is already clean and obvious, but when the recognition problem is limited by ambiguity in how the available evidence should be understood.

From the perspective of Part I, SCP completes the progression established across the first three

chapters. MITFAS improves *where* evidence is aligned, PMI Sampler improves *when* informative evidence is selected, and SCP improves *how* that evidence is semantically interpreted. Together, these chapters show that robust aerial video understanding requires not only stronger spatial-temporal evidence, but also stronger guidance for interpreting that evidence under aerial-specific observation conditions.

More broadly, this chapter suggests that prompt learning can play a useful role beyond standard vision-language adaptation. In aerial video understanding, prompts act less as generic transfer tools and more as lightweight semantic controllers that steer the model toward task-relevant interpretation. This makes SCP a meaningful example of how semantic guidance can become part of representation refinement rather than simply an auxiliary add-on.

4.7 Limitations

Although SCP improves action recognition across both aerial and ground-view benchmarks, it still has several limitations.

First, the quality of prompt guidance depends on the quality of the underlying visual evidence. When the actor is extremely small, heavily occluded, or poorly localized, the prompt may not provide sufficient semantic support to fully resolve the ambiguity. In such cases, prompt conditioning can improve interpretation only to the extent that useful visual evidence is still present in the input.

Second, SCP introduces additional design choices that affect performance, including the number of prompt experts, the prompt type, and the form of auxiliary guidance. As shown in the additional analysis, the effectiveness of prompt-based conditioning depends on how these components are configured. This means that semantic guidance is helpful, but not entirely plug-and-play; it still requires a

suitable prompt design for the target setting.

Third, although the framework supports auxiliary prompt forms such as optical flow and large vision model outputs, these auxiliary prompts are themselves dependent on external preprocessing or upstream models. Their usefulness therefore depends not only on the recognition task, but also on the quality and stability of the generated prompt signal. In particular, large vision model prompts can vary substantially depending on the prompt input format, and may not always provide equally reliable guidance.

Finally, while SCP improves how the model semantically interprets aerial observations, it does not directly solve all remaining challenges in aerial video understanding. In particular, it does not explicitly address longer-horizon activity structure, richer interaction modeling, or cross-domain transfer beyond the evaluated benchmarks. These limitations suggest that semantic guidance is an important part of aerial representation refinement, but not the final step.

4.8 Conclusion

In this chapter, we presented SCP [71], a prompt-based semantic guidance framework for aerial video understanding. By conditioning the recognition model with prompts that encode shared task knowledge while remaining adaptive to the input video, SCP improves how weak and ambiguous aerial visual evidence is interpreted. The framework further supports auxiliary prompt forms derived from optical flow and large vision models, showing that prompt-guided recognition provides a flexible and effective mechanism for semantic conditioning.

Experimental results on NEC-Drone, Okutama-Action, and Something-Something V2 show that SCP consistently improves recognition performance across both aerial and non-aerial settings. The

gains are especially strong in more challenging aerial scenarios, which supports the central claim of this chapter: after temporal evidence has been better aligned and better selected, recognition can still benefit substantially from stronger semantic guidance.

From the perspective of Part I, SCP completes the progression developed across the first three chapters. MITFAS improves temporal consistency, PMI Sampler improves temporal informativeness, and SCP improves semantic interpretability. Together, these chapters establish the central argument of Part I: robust aerial video understanding depends on refining not only the spatial-temporal evidence available to the model, but also how that evidence is semantically interpreted.

With this chapter, Part I concludes by showing that improved aerial video understanding can be achieved through three complementary forms of representation refinement: alignment, selection, and semantic guidance.

Part II

Practical and Scalable Aerial Perception under Real-World Constraints

Practical aerial perception requires more than strong recognition performance. In real-world UAV settings, perception models must often operate under limited computation, restricted onboard resources, and scarce labeled data, making both efficiency and scalability essential. The chapters in this part study how aerial perception can be designed to remain effective under such constraints. In particular, we focus on two complementary directions: computation-aware inference and supervision-efficient learning. Together, these methods improve how aerial perception systems are deployed and trained in realistic environments where both hardware and annotation budgets are limited.

Chapter 5: AZTR: Aerial Video Action Recognition with Auto Zoom and Temporal Reasoning

5.1 Introduction

This chapter begins Part II, which studies *practical and scalable aerial perception under real-world constraints*. While Part I focused on improving representation quality for aerial video understanding, the emphasis of this chapter is different: it asks how aerial perception can remain effective when computation, preprocessing capability, and deployment resources are limited. In this sense, the goal is not only to recognize actions accurately, but also to do so in a way that remains compatible with realistic aerial platforms.

Aerial video action recognition is an important problem in aerial perception, with applications in surveillance, public safety, search and rescue, human-robot interaction, and autonomous monitoring

[7, 9, 13]. Compared with conventional ground-view video understanding, aerial action recognition is substantially more challenging because the actor often occupies only a small fraction of the frame, the background dominates the visual field, actor scale changes with altitude and viewpoint, and camera ego-motion introduces substantial appearance variation across time [5, 16, 18, 33]. These properties make it difficult for standard action recognition models to extract robust and discriminative features directly from full-frame aerial videos.

The problem becomes even more difficult in practical deployment settings. Many strong video recognition models are designed for desktop GPUs and assume abundant memory, computation, and support for heavy spatio-temporal processing [2, 23, 27, 28]. In practical aerial systems, however, inference may need to run on mobile, embedded, or edge hardware with much tighter resource constraints. Under such conditions, simply increasing model size or input resolution is often inefficient, since most of the additional computation is still spent on irrelevant background rather than on action-relevant foreground content.

These observations motivate a design that is both *aerial-specific* and *deployment-aware*. Rather than processing the entire frame uniformly, it is more effective to first focus computation on the target actor and then perform temporal reasoning on the resulting target-centered representation in a way that can adapt to different hardware budgets.

To this end, this chapter presents **AZTR** [37], short for *Auto Zoom and Temporal Reasoning*. AZTR addresses aerial action recognition through two complementary ideas. First, it introduces an **auto zoom** mechanism that localizes the target actor, dynamically selects an appropriate crop size, and rescales the target-centered region to increase the proportion of action-relevant visual information. Second, it introduces a **temporal reasoning** module that aggregates target-centered features over time using computation-aware designs suitable for both edge devices and desktop GPUs. In this way, AZTR

improves not only recognition accuracy but also deployment practicality.

Figure 5.1 illustrates the main challenge addressed in this chapter. In aerial videos, the actor often occupies only a small portion of the frame, while background clutter, scale variation, and camera motion make direct full-frame recognition inefficient. AZTR addresses this challenge through target-centered spatial adaptation and efficient temporal reasoning. From the perspective of this dissertation, AZTR serves as an early example of practical aerial perception: it shows that aerial-specific spatial adaptation and hardware-aware temporal modeling together form an effective foundation for real-world deployment.

The main contributions of this chapter are as follows:

- We introduce an **auto zoom** strategy for aerial action recognition that localizes the target actor, dynamically rescales the crop region, and reduces the dominance of background content.
- We develop a **computation-aware temporal reasoning** framework that supports different temporal modeling choices on edge devices and desktop GPUs.
- We demonstrate that combining target-centered spatial adaptation with efficient temporal reasoning yields strong gains on multiple aerial action recognition benchmarks while preserving favorable speed-accuracy trade-offs for practical deployment [37].

5.2 Related Work

Since the earlier chapters of this dissertation have already reviewed the broader literature on conventional action recognition and aerial video understanding, here we focus more narrowly on prior work most relevant to practical aerial perception under real-world constraints. In particular, this chap-

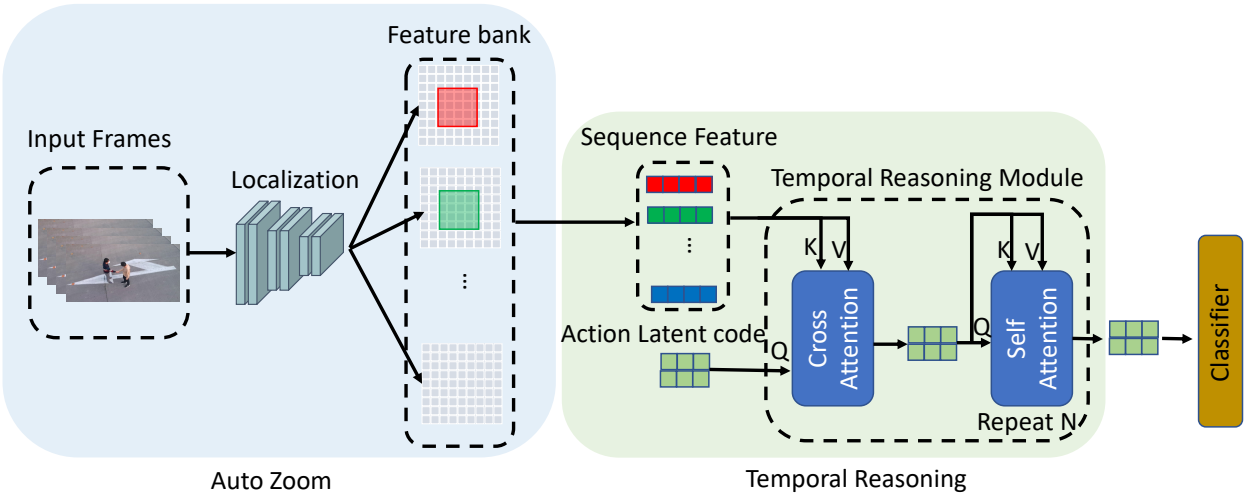


Figure 5.1: **Overview of the aerial action recognition challenges addressed by AZTR.** In aerial videos, the actor often occupies only a small portion of the frame, while background clutter, scale variation, and camera motion make direct full-frame recognition inefficient. AZTR addresses this problem through target-centered auto zoom and efficient temporal reasoning.

ter is most closely connected to research on aerial action recognition under deployment limitations, as well as adaptive spatial focus and efficient temporal modeling.

5.2.1 Aerial Action Recognition under Practical Constraints

Aerial action recognition has attracted increasing attention because of its relevance to surveillance, search and rescue, public safety, and autonomous monitoring [7, 8, 16]. Benchmarks such as Okutama-Action, Drone Action, MultiViewPoint, and UAV-Human highlight the distinctive difficulties of aerial action understanding, including small actors, cluttered backgrounds, viewpoint variation, and camera motion [5, 17, 18, 33]. Prior methods have explored aerial-specific modeling, transfer learning, domain adaptation, and frequency-domain reasoning [6, 46–48].

While these methods improve recognition under aerial viewpoints, many still inherit the basic processing assumptions of conventional video recognition and do not explicitly address practical de-

ployment constraints. In particular, they often process the full frame uniformly even though most pixels correspond to irrelevant background, and they do not directly study how aerial action recognition can remain effective on resource-constrained hardware. AZTR is motivated by this gap and focuses explicitly on both recognition accuracy and deployment efficiency.

5.2.2 Adaptive Spatial Focus and Efficient Temporal Modeling

The design of AZTR is also related to a broader line of work on adaptive computation and efficient temporal reasoning in video understanding. Prior methods have studied how to reduce video redundancy through frame selection, dynamic inference, or lightweight temporal aggregation [2, 3, 43]. These ideas are relevant because aerial videos also contain substantial redundancy, and not all frames contribute equally to recognition.

However, aerial action recognition introduces an additional spatial inefficiency that is less pronounced in many conventional video settings: the action-relevant actor often occupies only a small region of the frame, while the rest of the image is dominated by background [5, 18, 34]. This means that practical aerial perception requires more than efficient temporal modeling alone. It also requires a way to allocate spatial computation more effectively so that the recognition model focuses on informative target-centered content rather than uniformly processing the entire frame.

AZTR differs from general video efficiency methods by explicitly combining *spatial adaptation* through auto zoom with *temporal reasoning* under a controllable computational budget [37]. In this sense, the method is not simply an efficient inference strategy, but an aerial-specific recognition framework that addresses both spatial inefficiency and temporal modeling under practical deployment constraints.

5.3 Problem Formulation and Motivation

Consider an input aerial video

$$V = \{F_1, F_2, \ldots, F_T\},$$

where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$, and the task is to predict an action label

$$y \in \{1, \ldots, C\},$$

where C is the number of action categories. In a standard video recognition pipeline, the model learns a direct mapping

$$\hat{y} = f(V),$$

where f is typically implemented using a spatial feature extractor together with a temporal aggregation module.

This formulation is often inefficient for aerial action recognition. In aerial videos, the actor frequently occupies only a small fraction of the frame, while the rest of the image is dominated by background. As a result, applying full-frame recognition uniformly across all spatial locations spends substantial computation on irrelevant content rather than on the action-relevant foreground. At the same time, aerial videos still require temporal reasoning, since many actions can only be distinguished through motion evolution across frames rather than from a single appearance cue.

These two factors create a practical tension. On the one hand, the model should focus more computation on the target-centered region in order to improve spatial efficiency and increase the proportion

of useful visual evidence. On the other hand, it must still preserve sufficient temporal context to support reliable action recognition. This tension becomes especially important in real-world deployment, where inference may need to run on hardware with limited memory, latency, or compute budget.

These observations motivate a recognition framework that is both *aerial-specific* and *deployment-aware*. Rather than processing the full frame uniformly, the model should first adapt its spatial input to the target actor and then perform temporal reasoning on the resulting target-centered representation under a controllable computational budget.

This is the motivation behind **AZTR**. The framework addresses practical aerial action recognition through two complementary components: *auto zoom*, which dynamically selects a target-centered crop and rescales it to emphasize action-relevant content, and *temporal reasoning*, which aggregates the resulting features over time using hardware-aware designs suitable for different deployment settings. In this sense, AZTR refines not only what is recognized, but also how computation is allocated during recognition.

5.4 Method

5.4.1 Overview

Within this dissertation, AZTR is studied as a practical aerial perception framework for action recognition under real-world computational constraints. The central goal is not only to improve recognition accuracy, but also to allocate computation more effectively so that aerial action recognition remains suitable for deployment on both edge devices and desktop GPUs. Unlike the chapters in Part I, which primarily refine representation quality for aerial video understanding, this chapter focuses on how spatial-temporal processing itself can be made more deployment-aware.

Given an input aerial video

$$V = \{I_1, I_2, \ldots, I_T\},$$

AZTR first transforms the full-frame video into a target-centered sequence through an *auto zoom* mechanism, and then applies a *temporal reasoning* module to infer the action label. This overall structure can be written as

$$\hat{y} = f_{\text{temp}}(f_{\text{zoom}}(V)),$$

where $f_{\text{zoom}}(\cdot)$ denotes the target-centered spatial adaptation stage and $f_{\text{temp}}(\cdot)$ denotes the temporal reasoning stage.

The framework contains two main components. The first is **auto zoom**, which localizes the target actor, dynamically selects a crop size, and rescales the resulting target-centered region so that the model receives a higher proportion of action-relevant visual information. The second is **computation-aware temporal reasoning**, which aggregates target-centered features over time using designs that can be adjusted to different hardware settings. Figure 5.2 illustrates the overall AZTR pipeline, while Figures 5.3 and 5.4 detail the two main components.

5.4.2 Auto Zoom

A central difficulty in aerial action recognition is that the target actor often occupies only a very small portion of the frame. As a result, full-frame processing wastes a large amount of computation on irrelevant background while providing only a limited amount of actor-centered evidence to the recognition model. To address this issue, AZTR introduces an **auto zoom** mechanism that automatically localizes the target actor, crops the relevant region, and rescales it before recognition.

Let b_t denote the bounding box associated with the target actor in frame I_t. Based on this

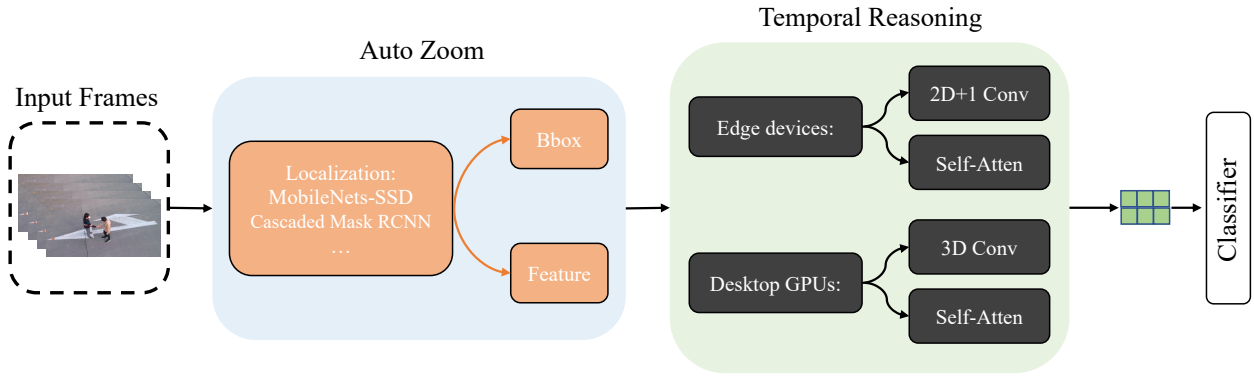


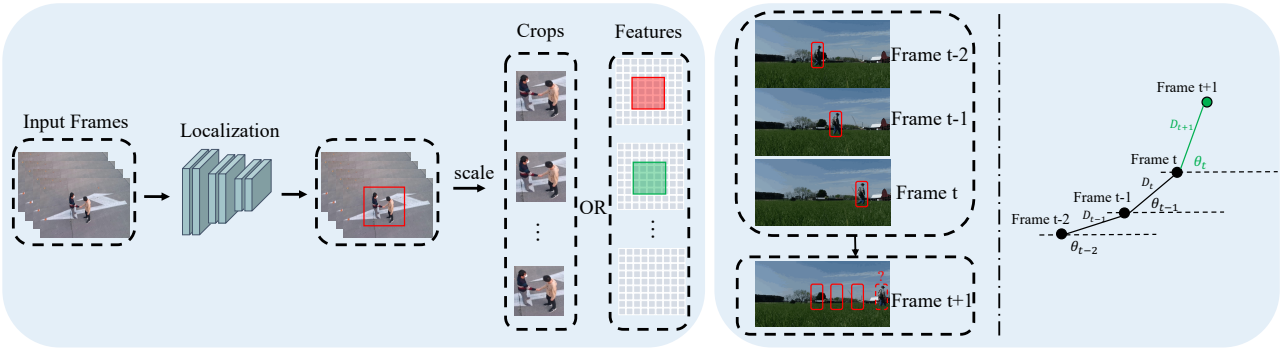
Figure 5.2: **Overall pipeline of AZTR.** The input aerial video is first processed by the auto zoom module to produce target-centered spatial crops. These target-centered observations are then passed to a temporal reasoning module whose form can be adapted to different deployment settings, enabling practical aerial action recognition under varying computational budgets.

estimate, AZTR constructs a target-centered region

$$\tilde{I}_t = \mathcal{Z}(I_t, b_t),$$

where $\mathcal{Z}(\cdot)$ denotes the zooming, cropping, and resizing operation. Unlike a fixed crop, AZTR chooses the crop size dynamically from a discrete set of candidate resolutions, such as $[480, 640, 720, 960]$, based on both the raw video resolution and the actor size. The crop is selected so that the actor occupies approximately 15% to 20% of the cropped region, which provides a good balance between local detail and surrounding context [37]. After cropping, the selected region is resized to the backbone input resolution.

This design provides two important benefits. First, it increases the proportion of actor pixels in the model input, allowing the backbone to focus more directly on the action. For example, the original study reports that on UAV-Human, a human actor may occupy only about 2% to 5% of the full frame, whereas after auto zoom the proportion increases to roughly 16% to 22% [37]. Second, because the



(a) Auto Zoom with Crops or Features

(b) Bbox shifting

Figure 5.3: **Auto zoom module.** AZTR localizes the target actor, selects an appropriate crop region based on actor scale and frame resolution, and resizes the resulting target-centered crop for feature extraction. This increases the proportion of action-relevant content while suppressing irrelevant background.

crop follows the target across time, the resulting view partially compensates for UAV motion and keeps the actor more spatially centered, thereby reducing background-dominated temporal noise.

5.4.3 Sparse Detection and Bounding Box Propagation

To make target-centered processing practical for real deployment, AZTR does not run the detector on every frame. Instead, it computes detections only on a sparse set of *key frames*, typically corresponding to only 10% or 20% of all frames [37]. These key frames serve as anchors for target localization, while the remaining frames use propagated or interpolated boxes.

For initialization, key frames are selected uniformly with a fixed stride. Since the detector may occasionally produce low-confidence or inaccurate boxes, AZTR first retains only detections with confidence above a threshold, and then predicts the next key-frame location based on the previous target trajectory:

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} x_t + (D_t + \delta_{D_t}) \cos(\theta_{t-1} + \delta_\theta) \\ y_t + (D_t + \delta_{D_t}) \sin(\theta_{t-1} + \delta_\theta) \end{pmatrix}, \tag{5.1}$$

where

$$D_t = \text{distance}\left(\begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}, \begin{pmatrix} x_t \\ y_t \end{pmatrix}\right), \qquad \delta_{D_t} = D_t - D_{t-1},$$

and

$$\theta_{t-1} = \arctan\left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}\right), \qquad \delta_\theta = \theta_{t-1} - \theta_{t-2}.$$

The predicted location is then compared with the detector output. If the detector output is missing or too far from the predicted location, the prediction is used instead. For normal frames between key frames, the bounding boxes are obtained by linear interpolation. This sparse-detection design substantially reduces localization cost while preserving sufficiently accurate target-centered crops for downstream recognition.

5.4.4 Target-Centered Feature Extraction

After auto zoom, each transformed frame $\tilde{I}_t$ is processed by a visual backbone:

$$x_t = \phi(\tilde{I}_t),$$

where $\phi(\cdot)$ denotes the backbone network. For edge deployment, AZTR uses lightweight backbones such as MoViNet, whereas for higher-capacity desktop settings it uses stronger backbones such as X3D-M [2, 37, 43]. The resulting feature sequence

$$X = \{x_1, x_2, \ldots, x_T\}$$

serves as the input to temporal reasoning.

Because the features are extracted from target-centered regions rather than full frames, the representation is less dominated by background appearance and more concentrated on the actor and its local context. This makes the subsequent temporal modeling both more informative and more efficient.

5.4.5 Computation-Aware Temporal Reasoning

Although auto zoom improves the spatial support of the representation, action recognition still depends on modeling temporal evolution. AZTR therefore introduces a temporal reasoning module whose specific form can be adjusted according to the hardware setting.

For edge devices, AZTR uses lightweight temporal reasoning based on $(2\mathrm{D} + 1\mathrm{D})$ convolution and shallow attention mechanisms. The main idea is to extract spatial features frame by frame and then fuse them temporally with inexpensive 1D temporal operations. This design is suitable for platforms that do not efficiently support heavy 3D convolution.

For desktop GPUs, AZTR can use stronger temporal modules, including 3D convolution or deeper attention-based reasoning, to improve recognition performance when computational resources are less restrictive.

The most flexible formulation in AZTR is the attention-based temporal reasoning module, which uses both **cross-attention** and **self-attention**. Let

$$X_Q \in \mathbb{R}^{N \times M}, \qquad X_{KV} \in \mathbb{R}^{T \times D}$$

denote the query sequence and the input feature sequence, respectively, where T is the number of frames, D is the frame embedding dimension, N is the number of learned query tokens, and M is the

query dimension. The projected query, key, and value features are written as

$$F_Q, F_K, F_V = p(X_Q), p(X_{KV}), p(X_{KV}), \tag{5.2}$$

where the projection maps the key and value to a controllable dimension S.

The attention weights are then computed as

$$A = \text{softmax}\left(\frac{F_Q F_K^\top}{\sqrt{S}}\right), \tag{5.3}$$

and the output is obtained by

$$Z = AV. \tag{5.4}$$

Cross-attention is used to map the input frame sequence to a new sequence of controlled size, while self-attention further models dependencies within the resulting sequence. An important property of this design is that its complexity is linear in the input temporal dimension T and the number of self-attention layers L, while the projection dimension S provides an explicit handle to trade accuracy for efficiency [37]. This makes the temporal reasoning stage particularly suitable for practical deployment, where different platforms may require different computational budgets.

5.5 Experimental Evaluation

In this section, we evaluate AZTR on multiple aerial action recognition benchmarks and analyze both its recognition performance and deployment efficiency. Since AZTR is motivated by the practical challenge of running aerial perception under limited computational resources, our evaluation is designed to answer two questions. First, does target-centered spatial adaptation improve aerial ac-

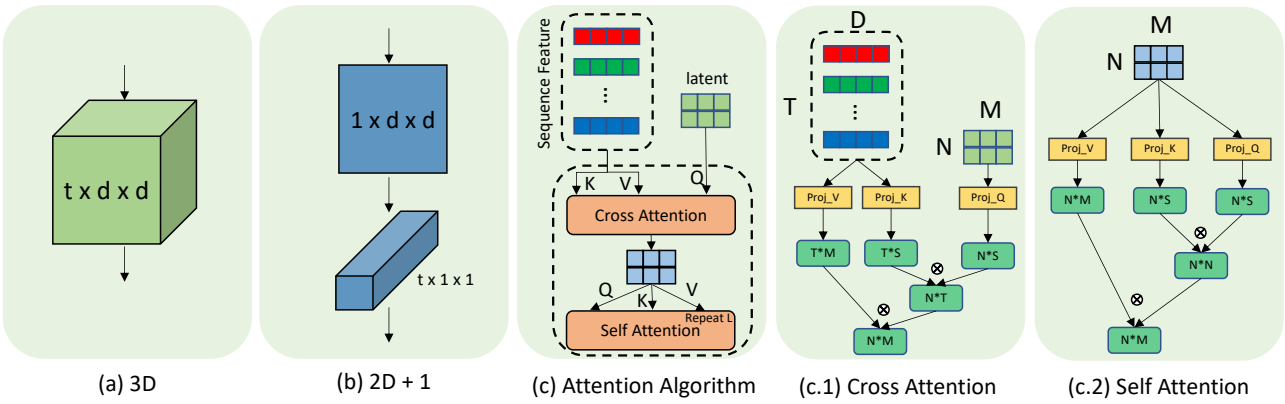


Figure 5.4: **Computation-aware temporal reasoning in AZTR.** After extracting target-centered per-frame features, the model performs efficient temporal aggregation using hardware-aware reasoning modules, enabling different accuracy-efficiency trade-offs for edge and desktop settings.

tion recognition across different datasets and difficulty levels? Second, can AZTR preserve favorable speed-accuracy trade-offs under realistic deployment constraints?

5.5.1 Datasets

AZTR is evaluated on three aerial action recognition benchmarks that together cover both practical edge deployment and larger-scale aerial generalization.

RoCoG-v2 contains 99 long aerial videos with 17 action categories. The videos are trimmed into 5,828 short clips of duration 1–5 seconds, and the original study further obtained 70,124 human bounding-box annotations in a semi-automatic manner to support localization [37]. In this chapter, RoCoG-v2 mainly serves as the practical edge-deployment benchmark, since it is paired with lightweight models evaluated on the Qualcomm Robotics RB5 platform.

For UAV-Human and Drone Action, we follow the same datasets introduced earlier in this dissertation. As discussed in previous chapters, UAV-Human is a large-scale aerial-view human behavior understanding benchmark with substantial viewpoint, scale, and background variation [5], while Drone

Action is an outdoor aerial action dataset with realistic camera motion and large viewpoint differences [18]. In this chapter, these two datasets are mainly used to evaluate whether the practical design of AZTR also remains effective on larger-scale and more diverse aerial recognition benchmarks beyond the edge-deployment setting.

Taken together, these datasets allow us to evaluate AZTR under both constrained and large-scale aerial settings, and to assess not only recognition performance but also the practical value of target-centered computation.

5.5.2 Implementation Settings

AZTR is evaluated under two deployment regimes.

Edge deployment. A lightweight implementation is deployed on the Qualcomm Robotics RB5 platform with a Kryo 585 CPU and Adreno 650 GPU. In this setting, MoViNet serves as the backbone and the temporal reasoning module uses efficient streaming or $(2\mathrm{D} + 1\mathrm{D})$ temporal operations [37]. This setting is particularly important because it reflects the real deployment constraints that motivate this chapter.

Desktop GPU setting. A higher-capacity implementation is evaluated on NVIDIA RTX A5000 GPUs using X3D-M and stronger temporal reasoning modules. This setting allows us to study whether the same target-centered design remains beneficial even when the computational budget is less restrictive.

This dual-setting evaluation is central to the chapter, since AZTR is intended not only to improve recognition accuracy but also to preserve a favorable speed-accuracy trade-off under realistic deployment constraints.

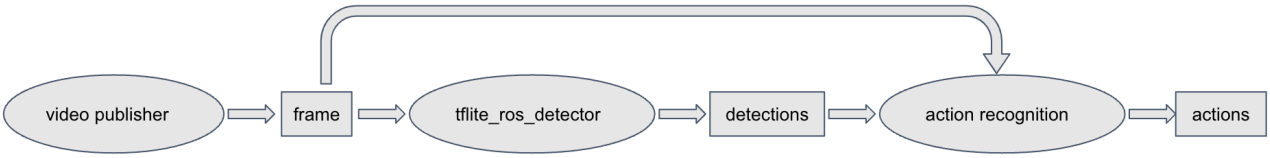


Figure 5.5: **Practical deployment setup of AZTR on the Qualcomm Robotics RB5 platform.** The system highlights the real-world inference pipeline used for edge deployment and illustrates that AZTR is designed not only for recognition accuracy but also for practical on-device operation.

Method	Frames	Input Size	Init.	Top-1 Acc (%) ↑
MoViNet A0 [43]	8	172×172	None	17.8
MoViNet A0 [43]	8	172×172	Kinectics	23.4
MoViNet A2 [43]	8	224×224	Kinectics	28.1
MoViNet A3 [43]	8	256×256	Kinectics	29.0
AZTR (Ours)	8	172×172	Kinectics	**29.5**
MoViNet A0 [43]	20	172×172	None	27.5
MoViNet A0 [43]	20	172×172	Kinectics	32.8
MoViNet A2 [43]	20	224×224	Kinectics	34.1
AZTR (Ours)	20	172×172	Kinectics	**40.2**

Table 5.1: **Results on RoCoG-v2.** We demonstrate that our approach can improve the top-1 accuracy by 6.1%-7.4%, outperforms all SOTA methods that can be deployed on the RB5 platform.

5.5.3 Main Results

The main quantitative results are shown in Tables 5.1, 5.2, and 5.3. Together, these results evaluate AZTR under three complementary perspectives: practical edge deployment on RoCoG-v2, large-scale aerial generalization on UAV-Human, and high-accuracy outdoor aerial recognition on Drone Action.

Table 5.1 shows the practical edge-deployment behavior of AZTR on RoCoG-v2. AZTR consistently outperforms lightweight MoViNet baselines in both the 8-frame and 20-frame settings, despite using only 172×172 target-centered inputs. In the 8-frame case, AZTR improves over MoViNet A3 from 29.0% to $\mathbf{29.5}\%$, even though A3 uses a larger input resolution. In the 20-frame case, the gain be-

Method	Backbone	Frames	Input Size	Init.	Top-1 Acc (%) ↑
X3D-S [2]	-	16	224×224	None	21.5
X3D-M [2]	-	16	224×224	None	27.0
X3D-L [2]	-	16	224×224	None	27.6
FAR [49]	X3D-M	16	224×224	None	27.6
FAR [49]	X3D-M	8	540×540	None	28.8
AZTR (Ours)	X3D-M	16	224×224	None	**39.2**
I3D [23]	ResNet-101	8	540×960	Kinetics	21.06
X3D-M [2]	-	16	224×224	Kinetics	30.6
FNet [92]	I3D	8	540×960	Kinetics	24.39
FAR [49]	I3D	8	540×960	Kinetics	29.21
FAR [49]	X3D-M	16	224×224	Kinetics	31.9
FAR [49]	X3D-M	8	620×620	Kinetics	39.1
AZTR (Ours)	X3D-M	16	224×224	Kinetics	**47.4**

Table 5.2: **Benchmarking UAV Human and comparisons with prior arts.** We compared with the state-of-the-art methods, which demonstrates an improvement of $8.3\% - 10.4\%$ over SOTA methods. Trained on high-end desktop GPUs.

comes much larger: AZTR reaches $\mathbf{40.2}\%$, improving over MoViNet A2 by **6.1%** and over MoViNet A0 by **7.4%**. This is an important result for the chapter because it shows that better spatial allocation can compensate for, and even outperform, stronger full-frame backbones under strict deployment constraints. In other words, AZTR does not simply add complexity; it reallocates computation toward the actor-centered region in a way that improves recognition efficiency.

Table 5.2 shows that the benefits of AZTR extend well beyond the edge setting. On UAV-Human, AZTR substantially improves over prior aerial action recognition methods under both training-from-scratch and Kinetics-initialized settings. Without pretraining, AZTR improves over FAR from 28.8% to $\mathbf{39.2}\%$, which is a very large gain on a difficult benchmark. With Kinetics initialization, AZTR reaches $\mathbf{47.4}\%$, outperforming FAR by **8.3%** in its strongest reported setting and improving over the weaker aerial baselines by as much as **10.4%**. These gains are especially meaningful because UAV-Human contains substantial viewpoint variation, background clutter, and scene diversity. The result therefore suggests that target-centered spatial adaptation is useful not only for deployment efficiency,

Method	Frames	Input Size	Init.	Top-1 Acc (%) ↑
HLPF [54]	All	1920×1080	None	64.36
PCNN [55]	-	1920×1080	None	75.92
X3D-M [2]	16	224×224	Kinetics	83.4
FAR [49]	16	224×224	Kinetics	92.7
AZTR (Ours)	16	224×224	Kinetics	**95.9**

Table 5.3: **Results on Drone Action.** We demonstrate that AZTR improves the state-of-the-art accuracy by 3.2%, reaching 95.9% on Drone Action. Trained on high-end desktop GPUs.

but also for improving recognition robustness at scale.

Table 5.3 shows that AZTR also achieves very strong performance on Drone Action, reaching $\mathbf{95.9}\%$ Top-1 accuracy and improving over FAR by $\mathbf{3.2\%}$. Since Drone Action is an outdoor aerial benchmark with relatively small actor regions and significant viewpoint differences, this result further supports the chapter's central claim: focusing computation on the action-relevant target while preserving efficient temporal reasoning improves aerial action recognition across different environments and dataset scales.

Taken together, the results in Tables 5.1–5.3 show that AZTR is effective across both practical edge deployment and stronger desktop-GPU settings. More importantly, the gains are consistent across datasets with different scales and difficulties, which suggests that AZTR captures a general principle of practical aerial perception: spatially reallocating computation toward the actor is often more valuable than uniformly scaling up full-frame processing.

5.5.4 Deployment Efficiency and Platform Analysis

Since AZTR is designed for practical aerial perception under real-world constraints, it is also important to evaluate its deployment efficiency across different platforms. Tables 5.4 and 5.5 summarize the platform support analysis and the measured inference speed on RB5 CPU.

Platforms	C3D	MP3D	AP3D	DC3D
TL (CPU)[93]	✓	×	×	×
TL + NNAPI (CPU/GPU/DSP)[93]	×	×	×	×
Tensorflow + MNN (CPU/GPU)[93][94]	✓	✓	✓	×
TL + MNN[93][94]	×	×	×	×

Table 5.4: **Supported operations on different platforms** 3D operators are not well supported on most edge devices or processors, as highlighted here. Therefore, we use 2D+1 conv and an efficient attention mechanism on the RB5 platform. TL: Tensorflow Lite, C3D: Conv3D, MP3D: MaxPooling 3D, AP3D: AveragePooling 3D, DC3D: Depthwise Conv3D

Method	Input Size	Inference Time per frame
MoViNet A0 [43]	172×172	33.2 ms
MoViNet A2 [43]	224×224	106.4 ms
MoViNet A3 [43]	256×256	124.0 ms
AZTR (Ours)	172×172	**56.5 ms**

Table 5.5: **Inference Time on RB5 CPU.** Our method takes 56.5 ms for inference on one frame (on average) which is 2 times faster than MoViNet A3 on the RB5, and also results in improvement on top-1 accuracy, see Table 5.1.

Table 5.4 highlights an important systems-level motivation for AZTR. Many 3D video operators are not well supported on typical edge-device software stacks, especially when deployment involves mobile processors and lightweight runtimes. This limitation means that strong conventional video models cannot always be transferred directly to practical aerial systems. AZTR addresses this issue by using more deployment-friendly temporal operations, such as $2\mathrm{D} + 1\mathrm{D}$ convolution and efficient attention, rather than relying on heavy 3D operators that may be poorly supported in practice.

Table 5.5 further shows that AZTR preserves a favorable speed-accuracy trade-off on the RB5 platform. Although AZTR is slower than the smallest MoViNet A0 baseline, it remains approximately **2× faster** than MoViNet A3 while also achieving higher Top-1 accuracy, as shown in Table 5.1. This comparison is especially important because it shows that AZTR is not simply accurate or simply efficient in isolation; rather, it occupies a more favorable operating point in the deployment trade-off

Figure 5.6: **Qualitative visualization of AZTR.** Compared with full-frame processing, the proposed auto zoom strategy allocates more visual emphasis to the actor and provides a cleaner target-centered representation for downstream temporal reasoning.

space. In practical terms, this means that AZTR is better aligned with real-world aerial perception needs, where recognition quality must be balanced against limited on-device computation.

Figure 5.5 complements these results by illustrating the actual ROS2-based deployment pipeline on RB5. Together, the platform analysis and the measured inference speed show that AZTR is not merely a recognition method evaluated on offline benchmarks, but a framework explicitly designed for realistic edge deployment.

5.5.5 Qualitative Analysis

Figure 5.6 illustrates the qualitative behavior of AZTR. Compared with full-frame processing, the auto zoom strategy produces a target-centered representation that better preserves action-relevant

visual evidence while suppressing irrelevant background clutter. This visual difference is important because it explains why AZTR improves both accuracy and efficiency: the model spends more of its representational and computational budget on the actor rather than on the surrounding scene. The qualitative evidence therefore supports the same conclusion as the quantitative results, namely that practical aerial perception benefits substantially from target-centered spatial adaptation.

5.6 Discussion

AZTR can be understood as a target-centered and deployment-aware spatial-temporal framework for aerial video understanding. Its main significance is therefore not only that it improves action recognition accuracy, but that it shows how aerial perception can be redesigned so that computation is allocated more effectively under real-world deployment constraints. In this setting, the challenge is not simply to build a stronger recognizer, but to ensure that the limited computational budget available on practical aerial platforms is spent on the most informative visual content.

This perspective helps explain why AZTR differs from the methods in Part I. The earlier chapters focused primarily on improving representation quality: MITFAS improved temporal consistency, PMI Sampler improved temporal informativeness, and SCP improved semantic interpretability. By contrast, AZTR addresses a different but equally important question: how should spatial-temporal computation itself be organized when aerial perception must run under realistic hardware constraints? The answer proposed in this chapter is that practical aerial recognition should first reallocate spatial computation toward the target actor and then perform temporal reasoning in a form that can adapt to different deployment platforms.

The experimental results support this interpretation. On RoCoG-v2, AZTR improves recognition

while remaining practical for RB5 deployment, showing that target-centered spatial adaptation can outperform stronger full-frame baselines even under strict edge-device constraints. On UAV-Human and Drone Action, the same design also yields substantial gains at larger scale, which suggests that the benefit of target-centered computation is not limited to one deployment setting or one dataset. In this sense, AZTR captures a broader principle of practical aerial perception: spatially focusing computation on the actor is often more valuable than uniformly increasing model capacity over the whole frame.

From the perspective of Part II, this chapter plays an important opening role. It marks the shift from improving aerial representations under idealized training settings to designing aerial perception systems that remain effective under limited memory, latency, and compute. AZTR therefore establishes a central theme of the second part of the dissertation: robust aerial perception in the real world depends not only on recognition quality, but also on computation-aware design choices that make deployment feasible.

More broadly, the chapter suggests that practical aerial perception requires jointly considering *what* information is most relevant and *where* computation should be spent. In aerial videos, these two questions are tightly coupled because the actor is small and the background dominates the frame. AZTR shows that explicitly addressing this coupling leads to better operating points in the accuracy-efficiency trade-off space, which is likely to remain important for future UAV perception systems.

5.7 Limitations

Although AZTR improves aerial action recognition under practical deployment constraints, it still has several limitations.

First, the framework depends on successful target localization. If the actor is poorly detected,

heavily occluded, or ambiguously separated from nearby people, the auto zoom stage may provide an inaccurate target-centered crop, which can reduce the benefit of the downstream temporal reasoning module.

Second, AZTR is primarily designed for target-centered single-actor recognition. In more complex aerial scenes with multiple interacting people or overlapping activities, a single target-centered crop may be insufficient to capture all action-relevant context.

Third, while AZTR improves the speed-accuracy trade-off, it does not eliminate the broader challenge of adapting aerial perception models across different hardware, sensing conditions, and deployment domains. In particular, the best temporal reasoning configuration may still vary across platforms and resource budgets.

These limitations suggest that practical aerial perception benefits from target-centered spatial adaptation, but that more flexible multi-actor reasoning and broader deployment generalization remain important directions for future work.

5.8 Conclusion

In this chapter, we presented AZTR [37], a practical aerial action recognition framework that combines target-centered auto zoom with computation-aware temporal reasoning. By reallocating computation toward the actor and reducing the dominance of irrelevant background, AZTR improves both recognition accuracy and deployment efficiency across multiple aerial benchmarks.

Experimental results on RoCoG-v2, UAV-Human, and Drone Action show that AZTR achieves strong gains under both edge-device and desktop-GPU settings. These results support the central claim of this chapter: practical aerial perception is not only a matter of making models smaller or faster, but

of redesigning spatial-temporal processing so that computation is directed toward the most informative evidence under real-world constraints.

From the perspective of Part II, AZTR serves as the opening chapter by establishing a key principle for practical and scalable aerial perception: aerial-specific spatial adaptation and hardware-aware temporal modeling are both necessary for effective real-world deployment. This chapter therefore marks the transition from representation refinement in Part I to deployment-aware aerial perception under realistic operating conditions.

Chapter 6: FALCON: Future-Aware Object-Centric Self-Supervised Learning for Aerial Action Recognition

6.1 Introduction

This chapter continues Part II, which studies practical and scalable aerial perception under real-world constraints. While Chapter 5 addressed the inference side of this problem by improving deployment efficiency through target-centered spatial adaptation and hardware-aware temporal reasoning, this chapter addresses a complementary challenge on the learning side: *how can we learn effective aerial action representations when labeled aerial videos are limited and expensive to curate?* [5, 16, 18, 37]

Aerial action recognition is a fundamental capability for aerial perception, with applications in surveillance, search and rescue, intent estimation, and human–robot collaboration [7–9]. Compared with conventional ground-view videos, aerial footage presents a substantially more difficult learning problem [5, 16, 18, 33]. Human actors and manipulated objects are often extremely small, large cluttered backgrounds dominate the field of view, and camera ego-motion introduces substantial appearance variation across time [5, 18, 33]. As a result, action-relevant evidence is both weak and spatially sparse, making it easy for a model to overfit to background statistics while under-emphasizing the subtle human- and object-centric cues that actually determine the action label [5, 16, 34].

The supervision problem further amplifies this difficulty. Large-scale labeled aerial datasets remain much smaller than their ground-view counterparts, and obtaining dense annotations for aerial videos is costly [5, 16, 18]. This makes self-supervised learning especially attractive for aerial action recognition [25, 29, 95–97]. Among existing self-supervised paradigms, masked autoencoding has emerged as a strong pretraining strategy for videos because it learns spatial-temporal representations by reconstructing masked content from visible context [29, 97–101]. However, standard masked autoencoding is not well aligned with the visual structure of aerial videos [5, 18, 102]. Under severe background dominance, random masking and uniform reconstruction often allow the pretraining objective to be solved primarily through background appearance and scene layout, while the most action-relevant human and object regions remain under-represented in both the visible set and the reconstruction target [102–106].

A second mismatch arises in the temporal domain. Conventional masked reconstruction mainly focuses on recovering missing content within the observed clip, which can often be solved through short-range appearance continuity [29, 97, 99]. For aerial action recognition, however, discriminative cues often depend on how human- and object-centric regions evolve over time [5, 18, 107, 108]. Naively extending reconstruction to entire future frames is not an ideal solution, since full-frame future prediction can again be dominated by ego-motion and cluttered background change rather than by action-relevant motion [5, 33, 102, 109].

These observations suggest that effective self-supervised learning for aerial action recognition should satisfy three requirements [102]. First, it should preserve sufficient visibility of action-carrying regions under masking [102–104]. Second, it should allocate reconstruction supervision toward those regions rather than uniformly across the frame [102, 105, 106]. Third, it should provide stronger temporal supervision by modeling how object-centric motion evolves into the future, while avoiding

background-dominated future prediction [102, 107–109].

To this end, this chapter presents **FALCON**, an aerial-tailored self-supervised pretraining objective that is both *object-aware* and *future-aware* [102]. FALCON uses off-the-shelf object cues only during pretraining to mitigate background-dominated learning and introduces an object-centric dual-horizon future reconstruction objective to encourage anticipatory motion understanding [102]. Importantly, these object cues are used only to instantiate pretraining priors: no detector input, bounding boxes, or additional preprocessing are required during downstream fine-tuning or inference [102]. This makes FALCON practical not only as an aerial-specific pretraining strategy, but also as a deployment-friendly action recognition framework [37, 102].

Figure 6.1 illustrates the central challenge addressed in this chapter. In aerial videos, action-relevant humans and objects occupy only a small portion of the frame, while standard reconstruction-based pretraining can be dominated by background content [5, 18, 102]. FALCON addresses this problem by using object-aware masking and loss design together with future-aware supervision that is focused on object-centric motion evolution [102].

From the perspective of this dissertation, FALCON complements AZTR by addressing a different aspect of practical aerial perception. Whereas AZTR improves *how inference-time computation is allocated*, FALCON improves *how representations are learned when dense supervision is limited* [37, 102]. Together, these chapters show that practical aerial perception depends on both efficient inference and supervision-efficient representation learning.

The main contributions of this chapter are as follows:

- We identify two key aerial-specific objective mismatches in masked video pretraining: *background-dominated reconstruction learning* and *weak supervision for object-centric motion evolution*

[102].

- We introduce **object-aware masked reconstruction**, which uses objectness priors to balance token visibility and reallocate reconstruction supervision toward action-relevant regions in observed clips [102].
- We propose an **object-centric dual-horizon future reconstruction** objective that models both short- and long-horizon motion evolution using future supervision restricted to object-centric regions [102].
- We show that FALCON consistently improves aerial action recognition on NEC-Drone and UAV-Human, while also generalizing to standard transfer benchmarks and preserving efficient RGB-only inference without detector overhead at test time [102].

6.2 Related Work

Since the earlier chapters of this dissertation have already reviewed the broader literature on action recognition and aerial video understanding, here we focus more narrowly on prior work most directly related to self-supervised video pretraining and supervision designs that emphasize action-relevant spatial-temporal structure.

6.2.1 Self-Supervised Video Pretraining

Self-supervised learning has become an important direction for video representation learning because it reduces dependence on large-scale manual annotation while still producing transferable spatial-temporal features. Early work explored contrastive and predictive objectives for learning video

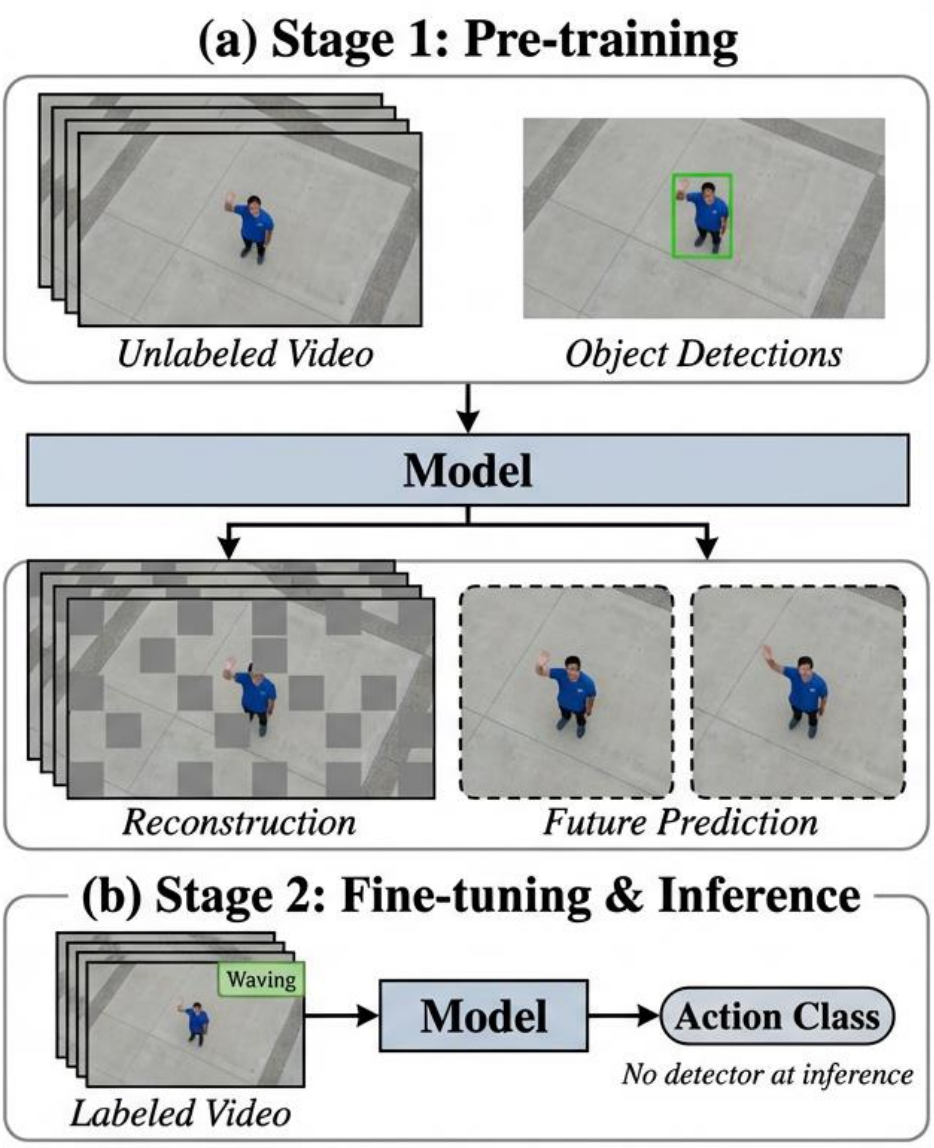


Figure 6.1: **Overview of FALCON. Stage 1 (Pre-training):** Given unlabeled UAV videos, we use off-the-shelf detections *only during pretraining* to instantiate objectness cues and optimize an object-aware masked reconstruction objective together with object-centric dual-horizon future reconstruction. **Stage 2 (Fine-tuning & Inference):** The pre-trained model is transferred to UAV action recognition using labeled videos, and inference is performed end-to-end from raw RGB without detectors or region processing.

representations from unlabeled data [107, 108]. More recently, masked pretraining has emerged as a particularly strong paradigm for visual and video representation learning [29, 97–101]. The central idea is to reconstruct masked content or predict masked features from visible context, thereby encouraging the encoder to learn informative spatial-temporal representations directly from video.

These methods have shown strong performance on conventional video benchmarks, but most of them treat reconstruction supervision in a relatively uniform manner across the frame [29, 97, 100]. In many standard video settings, this is reasonable because foreground objects occupy a substantial fraction of the image and the visual signal is relatively balanced. In aerial videos, however, the actor and manipulated objects are often spatially sparse, while large cluttered backgrounds dominate the frame [5, 16, 18, 33]. Under such conditions, standard masked reconstruction can be solved too easily through background appearance and scene layout, weakening the learning signal associated with the human- and object-centric regions that actually determine the action label. FALCON is motivated by this mismatch and studies how masked video pretraining should be modified to better match the structure of aerial footage.

6.2.2 Object-Aware and Future-Aware Learning

A related line of work has explored how auxiliary structure, such as object-centric supervision, region-aware learning, or future prediction, can improve visual representation learning. Object-aware approaches encourage the model to focus on semantically meaningful regions rather than treating all patches equally [104–106, 109]. This idea is especially relevant to aerial videos, where action-relevant humans and objects occupy only a small part of the frame and can otherwise be overwhelmed by background-dominated learning.

Future-aware learning provides another important source of supervision, since many actions can only be understood by modeling how motion evolves over time rather than by reconstructing appearance within the observed clip alone [107, 108]. However, in aerial settings, naively predicting the future of the whole frame is problematic because full-frame future change is often dominated by ego-motion and cluttered background variation rather than by object-centric action evolution [5, 18, 33]. This means that future supervision must also be structured in a more targeted way.

FALCON differs from prior self-supervised video learning methods by combining these two ideas in an aerial-specific manner [102]. It uses objectness priors only during pretraining to guide both masking and reconstruction supervision, and it introduces an object-centric dual-horizon future reconstruction objective to strengthen temporal learning without requiring detector input during downstream fine-tuning or inference. In this sense, FALCON is not simply another masked video pretraining method, but a self-supervised learning framework designed to better match the spatial and temporal structure of aerial videos.

6.3 Problem Formulation and Motivation

Consider an input aerial video

$$V = \{F_1, F_2, \ldots, F_T\},$$

where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$, and the task is to predict an action label

$$y \in \{1, \ldots, C\}.$$

In a standard video recognition pipeline, a model learns a mapping

$$\hat{y} = f(V),$$

where f is typically initialized either from supervised pretraining or from a generic self-supervised video representation learner.

In this chapter, we are particularly interested in the self-supervised setting, where the encoder is first pretrained by reconstructing masked video content and is then fine-tuned for downstream action recognition. Let the observed clip be denoted by

$$V^o = \{F_1, F_2, \ldots, F_{T_o}\},$$

and let the future clip be

$$V^f = \{F_{T_o+1}, \ldots, F_T\}.$$

A standard masked video pretraining objective learns a representation by reconstructing masked content in the observed video, and in some cases may be extended to future prediction. Although this formulation is effective for conventional videos, it is not well aligned with the structure of aerial action data.

The first mismatch is **background-dominated reconstruction**. In aerial videos, the actor and manipulated objects often occupy only a very small portion of the frame, while most visible tokens correspond to background. Under random masking and uniform reconstruction, the pretraining objective can often be solved primarily through background appearance and scene layout, even though these regions contribute relatively little to the action semantics. As a result, the learned representation may

under-emphasize the small human- and object-centric regions that are most relevant for downstream aerial action recognition.

The second mismatch is **weak supervision for object-centric motion evolution**. Conventional masked reconstruction mainly focuses on recovering missing content within the observed clip, which can often be solved through local appearance continuity. However, aerial action recognition frequently depends on how human- and object-centric regions evolve over time. A naive future-prediction objective over the whole frame is not an ideal solution, because future change in aerial videos is often dominated by ego-motion and cluttered background variation rather than by the action itself. What is needed is not simply more future prediction, but future supervision that is concentrated on the regions most likely to carry action-relevant motion.

These observations suggest that effective self-supervised learning for aerial action recognition should satisfy two requirements. First, it should preserve and emphasize action-relevant human and object regions during masked reconstruction. Second, it should provide stronger temporal supervision by modeling future motion evolution in an object-centric manner rather than through full-frame future reconstruction.

These requirements motivate an aerial-tailored pretraining framework that is both object-aware and future-aware. Rather than treating all spatial regions and temporal targets uniformly, such a framework should emphasize action-relevant human and object regions during reconstruction and strengthen temporal supervision through object-centric motion evolution. This is the central motivation behind FALCON, whose design is presented in the next section.

6.4 Method

6.4.1 Overview

Within this dissertation, FALCON is studied as a supervision-efficient pretraining framework for aerial action recognition. The central goal is to improve how representations are learned from aerial videos when dense manual annotation is limited and expensive to obtain. Unlike Chapter 5, which focuses on practical inference through target-centered spatial adaptation and hardware-aware temporal reasoning, this chapter focuses on improving the *pretraining objective* itself so that self-supervised learning is better matched to the structure of aerial videos.

FALCON follows an object-aware and future-aware self-supervised learning framework. As illustrated in Figure 6.2, the method first performs *object-aware masked reconstruction* on the observed clip, where objectness cues are used both to balance token visibility under masking and to allocate reconstruction supervision toward action-relevant regions. It then introduces an *object-centric dual-horizon future reconstruction* objective on the future clip to strengthen temporal learning beyond the observed frames.

The rest of this section is organized as follows. We first describe the overall pretraining formulation, then present object-aware reconstruction on the observed clip, followed by the proposed object-centric dual-horizon future reconstruction objective.

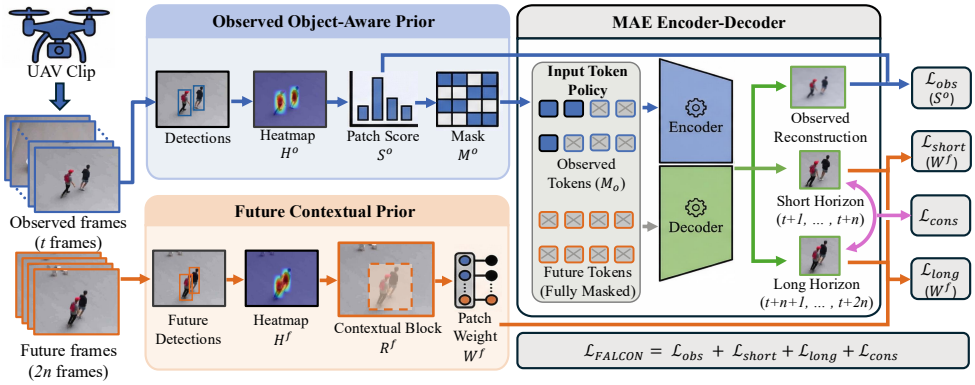


Figure 6.2: **FALCON pipeline.** Given a UAV clip, we split it into observed and future frames. Pretraining-time detections instantiate an *observed object-aware prior* ($H^o \rightarrow S^o \rightarrow M^o$) to enable stratified masking and object-centric supervision for observed reconstruction, and a *future contextual prior* ($H^f \rightarrow \mathcal{R}^f \rightarrow W^f$) that restricts *where* dual-horizon future reconstruction is supervised. An MAE encoder–decoder reconstructs observed tokens and fully masked future tokens with short/long losses and a horizon-consistency regularizer, optimized by $\mathcal{L}_{\text{FALCON}} = \mathcal{L}_{obs} + \mathcal{L}_{short} + \mathcal{L}_{long} + \mathcal{L}_{cons}$.

6.4.2 Overall Pretraining Formulation

Let the input clip be decomposed as

$$V = \{V^o, V^f\},$$

where $V^o \in \mathbb{R}^{T_o \times C \times H \times W}$ denotes the observed clip and $V^f \in \mathbb{R}^{T_f \times C \times H \times W}$ denotes the future clip. FALCON uses an asymmetric encoder–decoder architecture for self-supervised pretraining. The pretrained encoder is then transferred to downstream action recognition.

The pretraining objective contains two complementary parts. On the observed clip, FALCON performs *object-aware masked reconstruction*, using objectness cues to address the spatial imbalance between small action-relevant regions and dominant background content. On the future clip, it performs *object-centric dual-horizon future reconstruction*, using future supervision restricted to a contextual object-motion region. Importantly, the object cues are used only during pretraining to in-

stantiate masking and supervision priors; no detector input is required during downstream fine-tuning or inference.

6.4.3 Object-Aware Reconstruction on Observed Frames

FALCON pretrains on the observed clip V^o with an object-aware masked reconstruction objective. The key idea is to couple *token visibility* under masking with *supervision allocation* during reconstruction under a shared objectness prior. This is essential in aerial videos, where foreground humans and objects are extremely sparse relative to the background.

6.4.3.1 Objectness Prior

Given off-the-shelf detections on observed frames, denoted as

$$\mathbb{B}^o = \{\{b_{t,i}\}_{i=1}^{n_t}\}_{t=1}^{T_o},$$

we construct a pixel-level objectness heatmap by Gaussian aggregation:

$$H^o(x,y) = \frac{1}{T_o}\sum_{t=1}^{T_o}\sum_{i=1}^{n_t}\exp\left(-\frac{(x-x_{t,i}^c)^2+(y-y_{t,i}^c)^2}{2\sigma^2}\right), \tag{6.1}$$

where $(x_{t,i}^c, y_{t,i}^c)$ is the center of bounding box $b_{t,i}$. Projecting H^o to the patch grid yields patch-wise objectness scores

$$S^o = \{S_j^o\}_{j=1}^{N_s}, \qquad N_s = \frac{H}{h}\frac{W}{w}. \tag{6.2}$$

This objectness prior provides a soft estimate of which spatial regions are more likely to contain action-relevant human or object evidence.

6.4.3.2 Balanced Object-Aware Masking

Aerial videos suffer from severe foreground–background imbalance. If masked autoencoding uses semantics-agnostic random masking, then the visible token set can easily be dominated by background patches, leaving insufficient action-relevant evidence for the encoder.

To address this issue, FALCON introduces a *balanced object-aware masking* strategy based on stratified token visibility. Specifically, patches are first sorted by their objectness score S_j^o and partitioned into equal-length bins corresponding to objectness quantiles. Visible tokens are then selected so as to cover the full score range by sampling one visible patch from each bin. The resulting spatial mask pattern is replicated across time to form the observed mask M^o.

This design has two desirable effects. First, it prevents small human- and object-centric regions from being systematically excluded from the visible set. Second, it still preserves contextual background cues rather than over-collapsing visibility onto foreground only. In this way, the encoder receives a more balanced mixture of foreground evidence and context, which is especially important under the extreme spatial imbalance of aerial videos. Figure 6.3 illustrates this masking process.

6.4.3.3 Object-Centric Reconstruction Loss

Balancing visibility alone is not sufficient if the reconstruction loss is still dominated by background regions. FALCON therefore also reallocates reconstruction supervision according to the same objectness prior.

Let $K^{inv,o}$ denote the masked observed tokens and let $\hat{V}^o$ be the decoder predictions. The

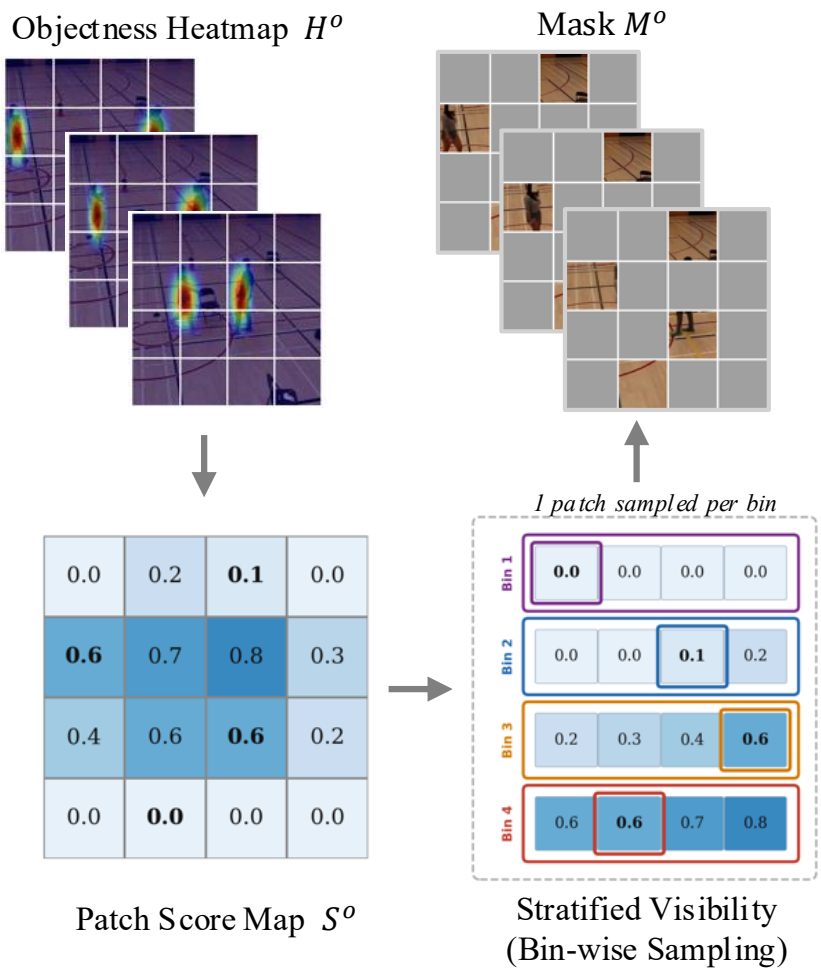


Figure 6.3: **Object-aware masking via stratified bin sampling.** Detections are aggregated into an objectness heatmap H^o and projected to a patch score map S^o. Patches are sorted by S^o and partitioned into equal-length bins; we sample one visible patch per bin to form a balanced mask M^o, ensuring coverage of both action-relevant (high-score) and contextual (low-score) regions.

observed-frame reconstruction loss is defined as

$$\mathcal{L}_{obs} = \sum_{i \in K^{inv,o}} \tilde{w}_i^o \, \|V_i^o - \hat{V}_i^o\|_2^2, \tag{6.3}$$

$$\tilde{w}_i^o = \frac{S_i^o + \mu}{\sum_{j \in K^{inv,o}} (S_j^o + \mu)}, \tag{6.4}$$

where μ is the mean objectness score.

The additive μ provides a background floor that avoids zero-weight tokens and stabilizes optimization, while still prioritizing action-relevant regions through the normalized objectness scores. Overall, the observed-clip branch mitigates UAV background dominance by jointly shaping the *input evidence* through balanced masking and the *learning signal* through object-centric supervision allocation.

6.4.4 Object-Centric Dual-Horizon Future Reconstruction

Beyond observed-clip reconstruction, FALCON introduces a future-aware objective to explicitly learn motion evolution. Let

$$V^f = \{V^s, V^l\},$$

where V^s and V^l denote short- and long-horizon future clips, respectively. All future tokens are masked and reconstructed from observed context, while supervision is restricted to an object-centric region in order to reduce ego-motion- and background-dominated learning.

6.4.4.1 Future Supervision Region

We compute a future objectness heatmap H^f in the same manner as Eq. (6.1), using pretraining-time detections on future frames. From this heatmap, we form a high-response support set

$$\Omega = \{(x, y) \mid H^f(x, y) \geq \gamma \max(H^f)\}, \tag{6.5}$$

compute its bounding rectangle, and dilate it by a fixed margin to obtain a contextual block:

$$\mathcal{R}^f = \mathrm{Dilate}(\mathrm{BBox}(\Omega), m), \tag{6.6}$$

where $\mathrm{Dilate}(\cdot, m)$ expands the rectangle by m patch units along each spatial direction. Projecting $\mathcal{R}^f$ to the patch grid yields a future-region weight map

$$W^f = \{w_j^f\}_{j=1}^{N_s}. \tag{6.7}$$

Compared with the token-level objectness scores S^o used on the observed clip, the future supervision region $\mathcal{R}^f$ defines a broader contextual object-motion block. This is important because future motion often depends not only on the object itself, but also on its immediate local context.

6.4.4.2 Dual-Horizon Reconstruction Objective

Let $K^{inv,s}$ and $K^{inv,l}$ denote the masked token sets for the short- and long-horizon future clips, respectively. Using normalized weights derived from W^f, FALCON defines two horizon-specific

losses:

$$\mathcal{L}_{short} = \sum_{i \in K^{inv,s}} \tilde{w}_i^f \, \|V_i - \hat{V}_i\|_2^2, \tag{6.8}$$

$$\mathcal{L}_{long} = \sum_{i \in K^{inv,l}} \tilde{w}_i^f \, \|V_i - \hat{V}_i\|_2^2, \tag{6.9}$$

where $\tilde{w}_i^f$ is normalized from W^f.

To encourage temporal coherence across horizons, FALCON further applies a consistency loss:

$$\mathcal{L}_{cons} = \|\bar{z}^s - \bar{z}^l\|_2^2, \tag{6.10}$$

where $\bar{z}^s$ and $\bar{z}^l$ denote the mean decoder-predicted token features over the short- and long-horizon futures, respectively, averaged over spatial tokens within the future supervision region.

This dual-horizon design is important because the two horizons capture complementary aspects of action dynamics. The short horizon emphasizes local motion continuity, whereas the long horizon encourages more anticipatory temporal reasoning. By restricting both objectives to object-centric regions, FALCON strengthens temporal supervision without allowing future prediction to be dominated by full-frame background change.

6.4.5 Unified Objective

The full pretraining objective is

$$\mathcal{L} = \mathcal{L}_{obs} + \mathcal{L}_{short} + \mathcal{L}_{long} + \mathcal{L}_{cons}. \tag{6.11}$$

This formulation directly addresses the two mismatches identified in the previous section. The observed-clip branch mitigates background-dominated reconstruction learning, while the future branch strengthens supervision for object-centric motion evolution. Importantly, object cues are required only during pretraining to instantiate the masking and supervision priors. During downstream fine-tuning and inference, the model operates on RGB video alone.

6.5 Experimental Evaluation

In this section, we evaluate FALCON on both aerial action recognition benchmarks and standard transfer benchmarks in order to assess three questions. First, does aerial-tailored self-supervised pretraining improve downstream aerial action recognition? Second, do the learned representations transfer across datasets with different aerial biases and capture conditions? Third, does FALCON preserve practical inference efficiency by avoiding detector overhead at test time?

6.5.1 Datasets and Evaluation Setting

We evaluate FALCON on two aerial action recognition benchmarks and two standard transfer benchmarks.

For NEC-Drone and UAV-Human, we follow the same datasets introduced earlier in this dissertation. As discussed in previous chapters, NEC-Drone is a controlled aerial action benchmark with relatively small actors and indoor aerial viewpoints, while UAV-Human is a much larger and more diverse aerial-view human behavior understanding benchmark with substantial viewpoint, background, and illumination variation [5, 6]. In this chapter, these two datasets are used to evaluate whether aerial-tailored self-supervised pretraining improves downstream aerial action recognition under both

controlled and large-scale conditions.

To assess broader transferability, we further evaluate FALCON on UCF101 [69] and HMDB51 [70]. These two standard video action benchmarks are used here not as the primary target domain, but as transfer datasets for evaluating whether the representations learned by FALCON remain useful beyond aerial footage.

Unless otherwise specified, FALCON follows the original pretraining setting in which object cues are used only during pretraining to instantiate masking and supervision priors, while downstream fine-tuning and inference use only RGB inputs. This distinction is important because it allows us to evaluate not only recognition accuracy, but also whether the learned representations remain practical for deployment.

6.5.2 Implementation Details

Following the original FALCON setting, we pretrain the model with masked video reconstruction on aerial data and then fine-tune on downstream action recognition tasks. Unless otherwise specified, the default configuration uses a ViT-B or ViT-L backbone with 16-frame inputs at resolution 224×224, initialized from Kinetics-400 pretraining. The observed and future segments are jointly used during self-supervised pretraining, while only RGB inputs are required during downstream fine-tuning and inference. Object cues are used only to construct pretraining priors and are not required at test time.

6.5.3 Main Results on Aerial Benchmarks

The main aerial action recognition results are shown in Table 6.1.

Method	Backbone	Extra data	Input Size	Frames	GFLOPs	Params.	NEC-Drone Top-1 Acc (%) ↑	UAV-Human Top-1 Acc (%) ↑
			supervised					
FAR [48]	X3D-M	K400	540 × 960	8	65	4M	71.4	38.6
DiffFAR [49]	X3D-M	K400	540 × 960	8	130	4M	80.7	41.9
AZTR [37]	X3D-M	K400	224 × 224	16	7	4M	-	47.4
MITFAS[38]	X3D-M	K400	224 × 224	16	7	4M	78.6	50.8
PMI Sampler [39]	X3D-M	K400	224 × 224	16	7	4M	62.5	55.0
MViT v1 [45]	MViT-B	K400	224 × 224	16	71	37M	34.6	24.3
ViViT FE [27]	ViT-B	IN-21K	224 × 224	16	284	116M	38.4	34.1
TimesFormer [28]	ViT-B	K400	224 × 224	8	196	131M	40.5	38.4
MotionFormer [110]	ViT-B	IN21K + K400	224 × 224	8	370	109M	73.6	50.4
			self-supervised					
ST-MAE [30]	ViT-B	K400	224 × 224	16	180	87M	-	45.1
VideoMAE [29]	ViT-B	K400	224 × 224	16	180	87M	82.5	62.1
MVD [111]	ViT-B	IN21K + K400	224 × 224	16	180	87M	77.6	60.5
SiamMAE [112]	ViT-B	K400	224 × 224	16	180	87M	76.4	60.1
FALCON(Ours)	ViT-B	K400	224 × 224	16	180	87M	**85.4**	**67.9**
VideoMAE [29]	ViT-L	K400	224 × 224	16	597	305M	88.4	71.5
VideoMAE v2 [83]	ViT-G	K400	224 × 224	16	5100	632M	82.2	61.1
FALCON(Ours)	ViT-L	K400	224 × 224	16	597	305M	**91.0**	**77.4**

Table 6.1: **Results on NEC-Drone and UAV-Human.** FALCON sets new state of the art, improving over VideoMAE by +2.9%/+5.8% (ViT-B) and +2.6%/+5.9% (ViT-L) top-1 on NEC-Drone/UAV-Human.

Table 6.1 shows that FALCON consistently improves over prior self-supervised video pretraining methods on both NEC-Drone and UAV-Human. With a ViT-B backbone, FALCON improves over VideoMAE by **+2.9%** on NEC-Drone and **+5.8%** on UAV-Human. With a ViT-L backbone, the gains remain strong, reaching **+2.6%** on NEC-Drone and **+5.9%** on UAV-Human. These results are especially important because they show that aerial-tailored pretraining objectives remain beneficial even when starting from strong masked autoencoding baselines.

The larger gain on UAV-Human is particularly meaningful. Compared with NEC-Drone, UAV-Human is more diverse and contains stronger variation in background, illumination, and scene scale. The fact that FALCON yields even larger improvements there suggests that its object-aware and future-aware design is especially helpful when the foreground is weaker and the action signal is more easily overwhelmed by nuisance variation. This directly supports the main claim of the chapter: self-supervised pretraining for aerial videos should not treat foreground and background uniformly, since standard reconstruction objectives otherwise under-emphasize the most action-relevant regions.

Method	Backbone	Extra data	Param (M)	UCF101 Top-1 Acc (%) ↑	HMDB51 Top-1 Acc (%) ↑
Vi^{2}CLR [95]	S3D	K400	9	89.1	55.7
CORP [113]	Slow-R50	K400	32	93.5	68.0
CVRL [24]	Slow-R50	K400	32	92.9	67.9
CVRL [24]	Slow-R152	K600	328	94.4	70.6
ρBYOL [25]	Slow-R50	K400	32	94.2	72.1
VIMPAC [114]	ViT-L	HowTo100M	307	92.7	65.9
MVD [111]	ViT-B	IN-1K + K400	87	97.0	76.4
VideoMAE [29]	ViT-B	K400	87	96.1	73.3
FALCON (Ours)	ViT-B	K400	87	**97.0**	**76.6**

Table 6.2: **Results on UCF101 and HMDB51.** FALCON improves over VideoMAE by +0.9% on UCF101 and +3.3% on HMDB51.

6.5.4 Transfer to Standard Action Recognition Benchmarks

To evaluate whether the learned representation remains useful beyond aerial action recognition, we next report transfer results on UCF101 and HMDB51, shown in Table 6.2.

Table 6.2 shows that FALCON also improves over the corresponding VideoMAE baseline on standard transfer benchmarks. The gains are modest on UCF101 and clearer on HMDB51, which is consistent with the fact that HMDB51 is generally more challenging and benefits more from stronger pretrained representations. These results indicate that FALCON does not simply overfit to aerial-specific data statistics. Instead, object-aware and future-aware pretraining appears to improve the quality of the learned visual-temporal representation in a way that remains transferable to broader action recognition settings.

This transfer behavior is important from a thesis perspective. Although FALCON is motivated by aerial action recognition, its benefits are not restricted to one narrow aerial domain. Rather, the method demonstrates that pretraining objectives designed around foreground relevance and motion evolution can improve downstream performance even when the target benchmark is more conventional.

Method	Backbone	Extra data	NEC-Drone → UAV-Human Top-1 Acc (%) ↑	UAV-Human → NEC-Drone Top-1 Acc (%) ↑
VideoMAE [29]	ViT-B	K400	59.4	76.6
MVD [111]	ViT-B	IN-1K + K400	58.7	76.1
SiamMAE [112]	ViT-B	K400	57.2	75.2
FALCON (Ours)	ViT-B	K400	**64.1**	**80.6**
VideoMAE [29]	ViT-L	K400	61.6	79.1
FALCON (Ours)	ViT-L	K400	**66.0**	**82.7**

Table 6.3: **Transfer Learning Ability.** Cross-dataset transfer between NEC-Drone and UAV-Human.

6.5.5 Cross-Dataset Transfer Between Aerial Benchmarks

A central motivation for self-supervised pretraining is improved transferability across datasets with different biases and capture conditions. To evaluate this property more directly in the aerial domain, we report cross-dataset transfer results between NEC-Drone and UAV-Human in Table 6.3.

Table 6.3 shows that FALCON achieves the strongest transfer performance in both directions. With a ViT-B backbone, it improves NEC-Drone $\rightarrow$ UAV-Human transfer from 59.4% to $\mathbf{64.1}\%$, and UAV-Human $\rightarrow$ NEC-Drone transfer from 76.6% to $\mathbf{80.6}\%$. With a ViT-L backbone, the gains remain similarly strong. This result is particularly important because it suggests that FALCON improves not only in-domain recognition, but also cross-dataset robustness. By centering pretraining supervision on human and object regions rather than on background-dominated reconstruction, the learned encoder becomes less sensitive to dataset-specific nuisance factors and transfers more effectively across aerial benchmarks.

Method	Data Aug.	Backbone	NEC Drone Top-1 Acc (%) ↑	UAV-Human Top-1 Acc (%) ↑	Inference time /video (ms)
AZTR [37]	✓	X3D-M	-	47.4	37.1
MITFAS [38]	✓	X3D-M	78.6	50.8	92.4
FALCON (Ours)	×	ViT-B	**85.4**	**67.9**	**18.7**

Table 6.4: **Inference time comparison.** On an RTX A5000 GPU, FALCON runs at 18.7 ms/video, achieving ∼2× and ∼5× speedups over AZTR and MITFAS while improving accuracy.

6.5.6 Inference Efficiency

Since object cues are used only during pretraining, FALCON incurs no detector overhead during downstream evaluation. Table 6.4 compares inference latency with prior aerial-specific methods.

Table 6.4 shows that FALCON achieves the lowest per-video latency among the compared methods while also improving recognition accuracy. This is an important practical advantage. Unlike prior methods that require online alignment, detector input, or additional object-aware processing at test time, FALCON retains the simplicity of RGB-only inference. In this sense, the method improves supervision efficiency during learning without increasing inference complexity during deployment. This property makes FALCON especially relevant to Part II, where practical and scalable aerial perception is a primary concern.

6.5.7 Ablation Studies

To understand why FALCON works, we perform a series of ablation studies on UAV-Human, as summarized in Table 6.5. These ablations evaluate the contribution of the core components, the design of the future reconstruction objective, the masking strategy, the observed–future clip split, the masking ratio, the amount of available object cues, and the sensitivity to detector quality.

The core component ablation in Table 6.5(a) shows the contribution of the main losses and

Model Variant	$\mathcal{L}_{\text{obs}}$	OAM	$\mathcal{L}_{\text{short}}$	$\mathcal{L}_{\text{long}}$	$\mathcal{L}_{\text{cons}}$	Top-1 Acc (%) ↑
Baseline						62.1
+ OAM		✓				64.0
+ $\mathcal{L}_{\text{obs}}$	✓	✓				66.4
+ $\mathcal{L}_{\text{short}}$	✓	✓	✓			66.8
+ $\mathcal{L}_{\text{long}}$	✓	✓	✓	✓		67.1
Full FALCON	✓	✓	✓	✓	✓	**67.9**

(a) Core Component Ablation. Each component independently contributes.

Future Design	Top-1 Acc (%) ↑
No future prediction	66.4
Single horizon (uniform)	67.1
Short-horizon only	66.8
Long-horizon only	66.5
Bidirectional (past + future)	67.0
Dual-horizon (ours)	**67.9**

(b) Future Horizon Design.

Supervision Region	Top-1 Acc (%) ↑
Full frame	66.1
Tight bounding box	67.5
Heatmap weights	66.5
Expanded $\mathcal{R}^f$ (ours)	**67.9**

(c) Future Supervision Region.

Masking	Top-1 Acc (%) ↑	
	w/o Future	w/ Future
Random	63.7	65.5
Tube [29]	64.1	66.3
Block	62.1	61.8
Ours	**66.4**	**67.9**

(d) Masking Strategy.

#Observed	#Future	Top-1 Acc (%) ↑
12	4	**67.9**
10	6	67.5
8	8	67.2

(f) Observed vs. Future Split.

Mask Ratio	50%	60%	70%	80%	90%	95%
Top-1 Acc (%) ↑	67.8	67.9	**67.9**	67.3	66.2	64.7

(e) Masking Ratio.

Videos with boxes	25%	50%	75%	100%
Top-1 Acc (%) ↑	61.8	65.4	67.1	**67.9**

(g) Partial Bboxes.

Detector	Backbone	CrowdHuman mAP ↑	UAV-Human Top-1 Acc (%) ↑
None (Vanilla VideoMAE)	–	–	62.1
Cascade Mask R-CNN	MobileNet	65.7	63.9
Faster R-CNN	HRNet	72.3	66.1
Cascade Mask R-CNN	HRNet	84.1	**67.9**

(h) Detector Quality. Consistent gains even with weaker detections.

Table 6.5: **Ablation Studies.** OAM: object-aware masking. Results on UAV-Human unless noted.

the object-aware masking strategy. Starting from the baseline VideoMAE-style pretraining objective, object-aware masking (OAM) alone already improves Top-1 accuracy from 62.1% to 64.0%, indicating that balancing token visibility is beneficial even before changing the loss. Adding the object-aware observed-frame reconstruction loss further improves performance to 66.4%, which confirms that reweighting supervision toward action-relevant regions is also important. Incorporating short-horizon and long-horizon future objectives yields additional gains, and the full model reaches 67.9%. This progressive improvement shows that the gains are cumulative rather than coming from a single dominant component.

The future horizon comparison in Table 6.5(b) evaluates different future prediction designs. Future prediction is clearly useful compared with using no future objective, but not all designs are equally effective. A single-horizon future objective improves performance, while the proposed dual-horizon design performs best at 67.9%. This result suggests that short-range and long-range motion evolution provide complementary supervisory signals. In aerial videos, where action semantics often depend on both immediate motion continuity and broader temporal progression, combining the two horizons produces a stronger pretraining objective than either one alone.

The supervision-region comparison in Table 6.5(c) further studies how future supervision should be spatially allocated. Reconstructing the full frame performs worse than restricting supervision to object-centric regions, which is consistent with the observation that full-frame future prediction can be dominated by ego-motion and background change. A tight bounding-box region is already effective, but the proposed expanded contextual future region $\mathcal{R}^f$ performs best. This indicates that future supervision should remain object-centric while still preserving a modest amount of surrounding context, since local context is often useful for modeling motion evolution.

The masking strategy comparison in Table 6.5(d) shows that random masking and tube masking

improve over the baseline, but both are inferior to the proposed object-aware masking strategy. The gap is especially clear when future supervision is included, where the proposed masking reaches 67.9%. This result directly supports one of the central arguments of the chapter: in aerial videos, masking should not treat all patches uniformly, because the small action-relevant foreground can otherwise be easily overwhelmed by background-dominated visibility patterns.

The masking ratio study in Table 6.5(e) shows that performance is strongest around 60% to 70%, with 70% giving the best result. Lower masking ratios do not force the model to learn sufficiently strong representations from sparse observations, while very high masking ratios make the pretraining task too difficult. This behavior is broadly consistent with masked autoencoding literature, but the result here also shows that FALCON remains stable over a reasonably wide range of ratios.

The observed–future split in Table 6.5(f) shows that the best result is obtained with 12 observed frames and 4 future frames, while allocating more frames to the future branch slightly reduces performance. This suggests that future supervision is beneficial, but only when the model still retains sufficient observed context for stable reconstruction and representation learning. In other words, future prediction should complement rather than replace strong observed-frame supervision.

The partial-bbox study in Table 6.5(g) evaluates the effect of partial object annotations during pretraining. As the percentage of videos with object cues increases, downstream performance improves steadily, reaching the best result when all videos are equipped with boxes. Importantly, the model still retains a substantial gain even when only a fraction of training videos provide object cues. This suggests that FALCON is not overly brittle with respect to the availability of pretraining objectness priors, which is encouraging for practical settings where object annotations may be incomplete or only partially available.

Finally, the detector-quality analysis in Table 6.5(h) shows that stronger detectors improve down-

stream performance, but even weaker detectors still provide gains over vanilla VideoMAE. This result is important because it shows that FALCON does not require perfect detections in order to be effective. Instead, the method benefits from objectness priors in a graded manner, which makes it more practical for real aerial datasets where detector quality may vary across scenes and domains.

Taken together, these ablations show that FALCON works because it addresses the two key objective mismatches identified earlier in the chapter: spatial imbalance in observed-frame reconstruction and weak supervision for object-centric motion evolution. The gains therefore reflect a better alignment between the pretraining objective and the structure of aerial videos, rather than simply a stronger backbone or a larger pretraining recipe.

6.5.8 Qualitative Analysis

We finally provide qualitative analysis to better understand the behavior of FALCON. Figure 6.4 visualizes attention maps on UAV-Human scenes under both in-domain and cross-domain settings. Specifically, we compare *in-domain* pretraining on UAV-Human followed by fine-tuning on UAV-Human, and *cross-domain* pretraining on NEC-Drone followed by fine-tuning on UAV-Human. Compared with the baseline, FALCON consistently produces more concentrated attention on the human regions while suppressing irrelevant background responses. This difference is especially pronounced in the cross-domain setting, where the baseline tends to attend more diffusely to background content, whereas FALCON remains more focused on the action-relevant foreground. These qualitative observations are consistent with the quantitative transfer results in Table 6.3, and further support the claim that object-aware and future-aware pretraining helps the model learn representations that are less sensitive to dataset-specific background bias and more robust across aerial domains.

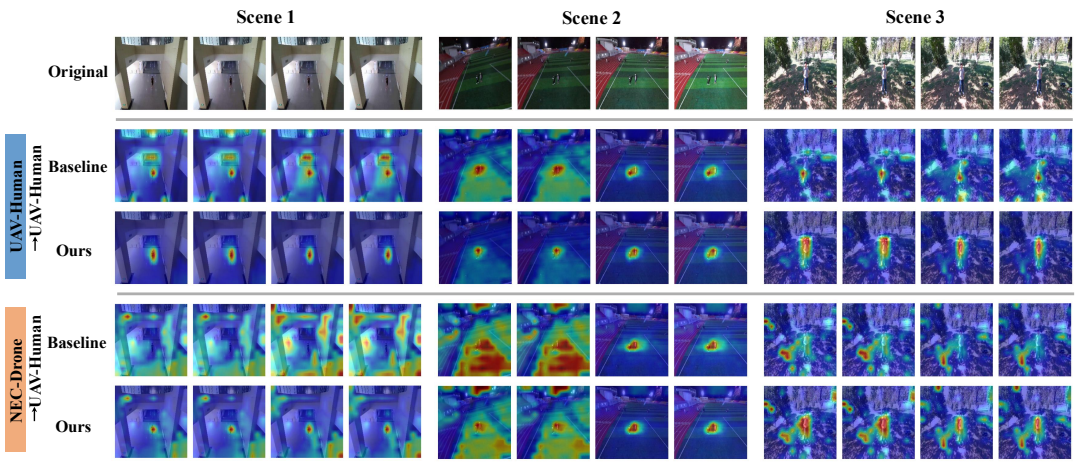


Figure 6.4: **Attention visualization under in-domain and cross-domain transfer.** We visualize attention maps on UAV-Human scenes under two settings: *in-domain* pretraining on UAV-Human followed by fine-tuning on UAV-Human, and *cross-domain* pretraining on NEC-Drone followed by fine-tuning on UAV-Human. Compared with the baseline, FALCON produces more concentrated attention on human regions while suppressing background responses. The difference is more pronounced in the cross-domain setting, qualitatively supporting the improved transfer robustness shown in Table 6.3.

6.6 Discussion

FALCON can be understood as an aerial-tailored self-supervised pretraining framework that reallocates learning emphasis toward action-relevant spatial-temporal structure. Its central significance is therefore not only that it improves downstream aerial action recognition, but that it shows how pretraining objectives themselves should be redesigned when the visual structure of the target domain is highly imbalanced. In aerial videos, both the visible token distribution and the reconstruction target can easily be dominated by background content. FALCON addresses this issue by shifting pretraining emphasis toward human- and object-centric regions and by strengthening supervision for object-centric motion evolution.

The experimental results support this interpretation. The gains are not explained by a larger model or a more generic pretraining recipe alone, but by better alignment between the pretraining objective and the spatial-temporal structure of aerial videos. In particular, the stronger improvements

on UAV-Human and the cross-dataset transfer gains suggest that object-aware and future-aware supervision helps the encoder become less dependent on dataset-specific nuisance factors and more sensitive to the motion patterns that actually determine aerial actions. This indicates that supervision-efficient aerial perception depends not only on reducing the need for labels, but also on allocating self-supervised learning signals in a way that better matches aerial data.

From the perspective of the overall dissertation, FALCON complements AZTR by addressing the learning side of practical and scalable aerial perception. While AZTR improves how computation is allocated during inference, FALCON improves how supervision is allocated during representation learning. Together, these chapters show that practical aerial perception requires both deployment-aware inference and supervision-efficient pretraining. In this sense, FALCON helps establish a broader thesis point of Part II: aerial perception under real-world constraints is improved not only by making inference more efficient, but also by making representation learning better matched to weakly supervised aerial data.

6.7 Limitations

Although FALCON improves self-supervised aerial action pretraining, it still has several limitations.

First, the method relies on objectness priors during pretraining. Although these priors are not required at downstream inference time, the quality and coverage of the pretraining object cues still affect how effectively the masking and supervision are allocated.

Second, FALCON is designed primarily for action recognition settings in which the most relevant signal is concentrated on a small set of human- and object-centric regions. In scenes with highly

distributed activity patterns or more complex multi-agent interactions, object-centric priors alone may be insufficient to capture all relevant semantics.

Third, while the proposed dual-horizon future objective improves temporal supervision, it still focuses on relatively local future evolution. Longer-horizon activity structure, richer interaction dynamics, and broader cross-domain generalization remain important open directions for future work.

These limitations suggest that object-aware and future-aware pretraining is an important step toward supervision-efficient aerial perception, but not a complete solution on its own.

6.8 Conclusion

In this chapter, we presented FALCON [102], an aerial-tailored self-supervised pretraining framework for action recognition. By combining object-aware masking, object-aware reconstruction supervision, and object-centric dual-horizon future reconstruction, FALCON improves how representations are learned from aerial videos under weak and spatially sparse action evidence.

Experimental results on NEC-Drone, UAV-Human, and standard transfer benchmarks show that FALCON consistently improves downstream recognition performance while preserving efficient RGB-only inference without detector overhead at test time. These results support the central claim of this chapter: practical aerial perception depends not only on efficient inference, but also on pretraining objectives that are better aligned with the spatial and temporal structure of aerial videos.

From the perspective of Part II, FALCON complements AZTR by addressing the learning side of practical and scalable aerial perception. Whereas AZTR improves how computation is allocated during inference, FALCON improves how supervision is allocated during representation learning. Together, these two chapters show that pratical and scalable aerial perception requires both deployment-aware

inference and supervision-efficient pretraining.

Part III

Transferable Aerial Perception Across Views, Modalities, and Camera Networks

Aerial perception often extends beyond a single viewpoint or sensing stream. Many practical systems require correspondence across aerial and ground observations, heterogeneous sensing modalities, or multiple cameras with different viewpoints and scales. The chapters in this part study how learned representations can preserve task-relevant consistency when the sensing condition changes. In particular, we focus on two complementary directions: transferable cross-view localization and transferable cross-camera identity association. Together, these methods improve the robustness of aerial perception across views, modalities, and camera networks.

Chapter 7: AGL-Net: Transferable Aerial-Ground Cross-Modal Localization under Scale Variation

7.1 Introduction

This chapter begins Part III, which studies transferable aerial perception across views, modalities, and camera networks. Unlike the previous parts of this dissertation, which focused on aerial video understanding and practical aerial perception under deployment constraints, the emphasis here is on *transferable correspondence across sensing conditions* [37, 102, 115]. In many real-world robotic systems, aerial perception does not operate in isolation within a single sensing regime. Instead, aerial observations often serve as global references that must be related to local ground-level sensing captured from different viewpoints, modalities, and spatial scales [116–119]. This chapter studies such transfer in the context of global localization, where a ground robot must localize itself within a large-

scale aerial map [120, 121].

Global localization is a fundamental capability for long-range navigation in autonomous driving, field robotics, and GPS-denied environments [32, 120, 121]. A robot must estimate its global position and orientation in order to plan over large areas, maintain consistent situational awareness, and recover from drift in local odometry [11, 32, 121]. While short-range motion can often be handled using onboard sensing alone, reliable global localization remains difficult when GPS is unavailable, inaccurate, or unreliable [32, 120, 121]. In such settings, the robot must infer its pose by matching local observations against a global reference map [120, 122, 123].

This problem becomes substantially more challenging when the local observation and the global map come from *different viewpoints and sensing modalities* [116, 118, 119, 124]. In this chapter, the local query is a ground LiDAR scan, while the global reference is an aerial RGB map. This creates a severe representation gap [120, 125]. The LiDAR scan captures sparse local 3D geometry from a ground viewpoint, whereas the aerial map provides dense visual appearance from an overhead perspective [120, 125, 126]. Establishing correspondence across these inputs is difficult because the two modalities encode the scene in fundamentally different ways and share only limited direct appearance similarity [119, 120, 124, 127].

A further challenge arises from *scale variation*. The LiDAR scan is naturally measured in metric coordinates, but the aerial map may have uncertain or varying resolution, especially in practical deployment scenarios where map crop scale and coverage are not guaranteed to be fixed in advance [128–130]. As a result, even if the correct local structure is present in the aerial map, direct matching can still fail when the local and global observations are misaligned in scale [128, 131, 132]. This issue is especially important for metric localization, where the goal is not merely coarse place recognition, but accurate estimation of position and heading in the global map frame [118, 120, 131].

These observations suggest that effective aerial-ground localization should satisfy two requirements. First, it should learn transferable representations that remain robust across large viewpoint and modality changes [116, 124, 125, 127]. Second, it should explicitly account for scale inconsistency between the local ground observation and the global aerial reference [128, 129, 131]. Methods that rely only on same-modality registration, fixed-scale assumptions, or appearance-level similarity are therefore not well suited to this setting [120, 122, 124].

To address these challenges, this chapter presents **AGL-Net**, an **Aerial-Ground Localization Network** for cross-modal metric global localization [128]. AGL-Net uses a unified two-stage matching design. In the first stage, it extracts neural representations from the ground LiDAR scan and the aerial RGB map and performs coarse cross-modal matching [128]. In the second stage, it extracts structural skeleton features from both modalities and performs refined matching with explicit scale alignment [128]. In addition, the model is trained with pose, scale, and skeleton supervision so that the learned representation becomes not only correspondence-aware, but also scale-aware and structurally meaningful [128]. This design allows AGL-Net to localize under unknown map scales without requiring preprocessed aerial imagery at a fixed metric resolution [128].

Figure 7.1 illustrates the central challenge addressed in this chapter. A ground robot must localize itself using a local LiDAR scan and a global aerial RGB map despite substantial differences in viewpoint, sensing modality, and spatial scale. AGL-Net addresses this problem through transferable cross-modal matching with explicit scale alignment.

From the perspective of this dissertation, AGL-Net serves as the first chapter in Part III by studying *transferable correspondence across sensing regimes*. Whereas the earlier parts primarily focused on understanding aerial videos themselves, this chapter studies how aerial perception can function as a global reference for ground-level robotic observations under large changes in viewpoint,

modality, and scale [37, 102, 115, 128]. In this sense, the chapter extends aerial perception from *understanding within one sensing regime* to *correspondence across sensing regimes*.

The main contributions of this chapter are as follows:

- We formulate aerial-ground metric localization as a transferable cross-modal matching problem between local ground LiDAR scans and global aerial RGB maps.
- We propose **AGL-Net**, a two-stage matching architecture that combines neural feature matching and skeleton-based structural refinement for robust pose estimation [128].
- We introduce an explicit **scale alignment** module together with scale-aware supervision to handle varying and unknown aerial map scales without requiring pre-aligned imagery [128].
- We construct a CARLA-based dataset for precise metric localization training and evaluation, and show that AGL-Net outperforms strong baselines on both CARLA and KITTI [128].

7.2 Related Work

Since the earlier chapters of this dissertation have already reviewed the broader literature on aerial perception and robotic visual understanding, here we focus more narrowly on prior work most directly related to aerial-ground localization across viewpoints, modalities, and scales. In particular, this chapter is most closely connected to three lines of research: *cross-view geo-localization*, *LiDAR- and map-based global localization*, and *cross-modal metric localization under structural and scale mismatch*.

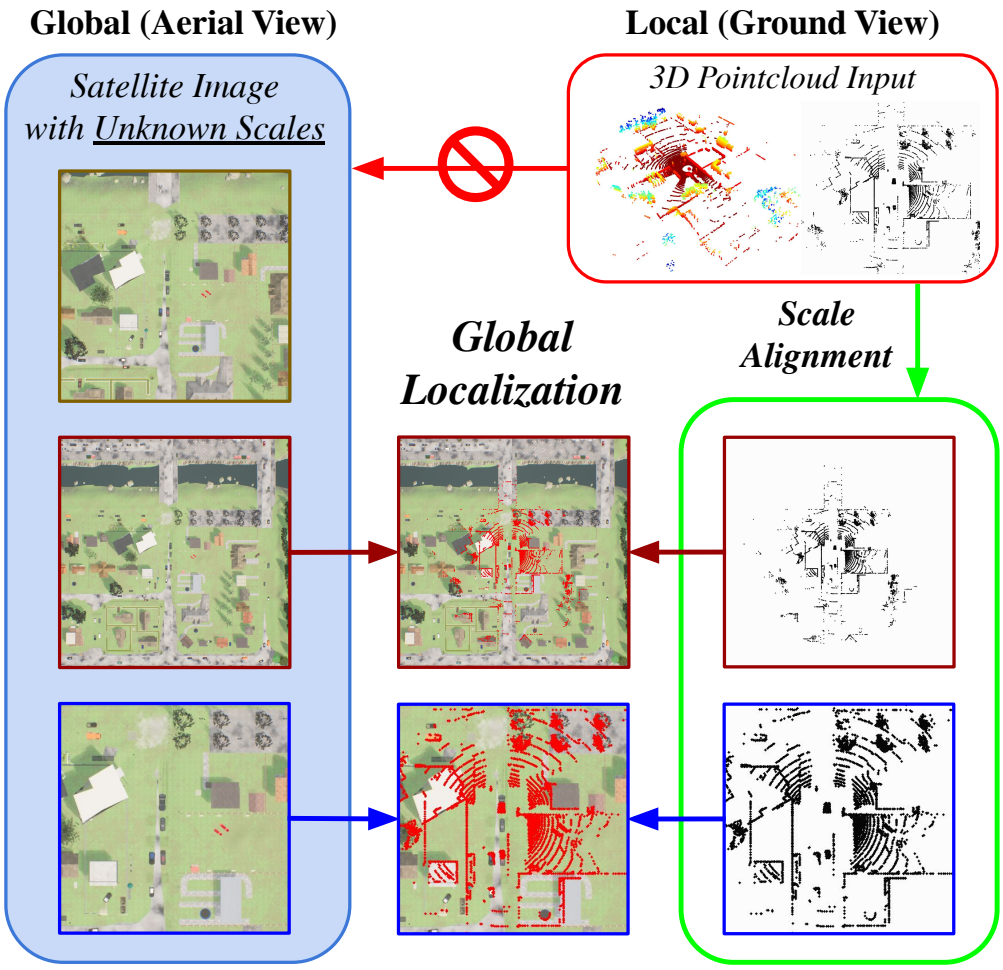


Figure 7.1: **Overview of aerial-ground cross-modal metric localization.** A ground robot must localize itself using a local LiDAR scan and a global aerial RGB map under substantial viewpoint, modality, and scale differences. AGL-Net addresses this problem through transferable cross-modal matching with explicit scale alignment.

7.2.1 Cross-View and Aerial-Ground Geo-Localization

Cross-view geo-localization studies how to localize a ground-level observation using an overhead aerial or satellite reference. Early work formulated this problem as matching between street-view and aerial imagery, showing that direct appearance matching fails under drastic viewpoint changes and that cross-view representation learning is necessary [116–118]. More recent learning-based methods substantially improved retrieval performance by using siamese metric learning, global feature aggregation, and increasingly large-scale benchmarks [124, 127, 133–136]. These methods demonstrate that cross-view correspondence can be learned despite strong viewpoint variation, and they establish aerial-ground matching as an important tool for GPS-denied localization and robotic navigation.

A related direction has moved beyond coarse retrieval toward *fine-grained* or *metric* cross-view localization. Rather than asking only which aerial tile best matches a ground query, these methods aim to estimate more precise position and, in some cases, orientation within the aerial map frame [127, 132, 137, 138]. This shift is important because practical robotic localization often requires metric pose estimation rather than approximate place recognition alone. However, most existing cross-view geo-localization methods still focus on *image-to-image* matching, typically between ground RGB images and aerial or satellite RGB imagery. As a result, they do not directly address the more severe cross-modal gap considered in this chapter, where the query is a local ground LiDAR observation and the reference is a global aerial RGB map.

7.2.2 LiDAR- and Map-Based Global Localization

Global localization has also been studied extensively from the perspective of LiDAR and map-based robotic localization. A large body of work localizes a robot by aligning LiDAR scans to a prior

map, often using geometric registration, learned place recognition, or learned localization features [121–123]. These approaches are effective when a dense prior LiDAR map is available and when the reference and query are relatively compatible in sensing modality. However, maintaining a prior LiDAR map is costly, and the reference map may not always be captured by the same sensing platform as the robot.

To reduce this dependence on prior LiDAR maps, several works have studied localization against overhead imagery or public map sources. FLAG explored air-ground localization using feature-based alignment between onboard sensing and aerial references [126]. More recently, *Get to the Point* showed that a robot can perform both place recognition and metric localization from LiDAR observations using overhead imagery as a map proxy [120]. Related work further investigated localization against OpenStreetMap priors and point-based localization between LiDAR and overhead imagery [129, 139]. These methods are especially relevant to this chapter because they move toward the same general goal of localizing ground sensing against overhead map references without requiring a prior LiDAR map. Nevertheless, many of them either rely on handcrafted structural alignment, focus primarily on place recognition before localization refinement, or do not explicitly model the *scale inconsistency* between local ground observations and aerial reference maps that arises in practical deployment.

7.2.3 Cross-Modal Localization under Structural and Scale Mismatch

The problem addressed in this chapter lies at the intersection of the two lines above, but introduces an additional difficulty: the query and reference differ not only in viewpoint, but also in sensing modality and spatial scale. A ground LiDAR scan captures sparse local geometry in metric coordinates, whereas an aerial RGB map provides dense visual appearance from an overhead view.

Establishing correspondence across such inputs is harder than conventional image-based cross-view matching, since direct texture or appearance similarity is weak and only higher-level structural cues are consistently shared.

A few recent works have begun to explore this broader cross-source setting. In particular, Cross-Loc3D studies aerial-ground cross-source 3D place recognition and shows that transferable aerial-ground correspondence can be learned even across different sensing regimes [125]. At a broader level, recent surveys on cross-view geo-localization and aerial visual place recognition also emphasize the growing importance of localization across heterogeneous viewpoints and sensing conditions [119, 130]. However, existing methods still leave two important gaps. First, many focus on *retrieval-style* localization rather than direct metric pose estimation. Second, even when structural matching is considered, *scale variation* is often assumed away through fixed-resolution aerial tiles or pre-aligned references.

This chapter addresses these gaps through **AGL-Net** [128]. Unlike standard cross-view image geo-localization methods, AGL-Net performs aerial-ground *cross-modal* localization between ground LiDAR and aerial RGB maps. Unlike approaches that depend primarily on a fixed reference scale, AGL-Net explicitly models *scale alignment* and combines coarse neural matching with refined skeleton-based structural matching. In this way, the method is designed not only to recognize aerial-ground correspondence, but also to localize robustly under unknown and varying aerial map scales.

7.3 Problem Formulation and Motivation

We consider aerial-ground metric localization, where a ground robot must estimate its pose with respect to a global aerial reference map. Let the local observation be a ground LiDAR scan

$$L \in \mathbb{R}^{N\times 3},$$

and let the global reference be an aerial RGB map

$$A \in \mathbb{R}^{H\times W\times 3}.$$

The goal is to estimate the robot pose

$$p = (x, y, \theta),$$

where (x, y) denotes the ground-plane position in the aerial map coordinate frame and θ denotes the heading angle. Unlike coarse place recognition, this task requires *metric localization*, namely, recovering both the matched region and the corresponding position and orientation with sufficient spatial precision for downstream navigation [118, 120, 131]. Figures 7.1 and 7.2 illustrate this setting.

A straightforward solution would be to learn or optimize direct correspondence between the local observation and the aerial reference. However, this is difficult in the present setting for three reasons. First, the query and reference come from different viewpoints and sensing modalities, making direct appearance-level matching unreliable [116, 124, 125, 127]. Second, the LiDAR query covers only a limited local region, so coarse similarity alone may be insufficient for precise pose recovery in large environments [120, 131, 138]. Third, the aerial reference may be observed at varying or

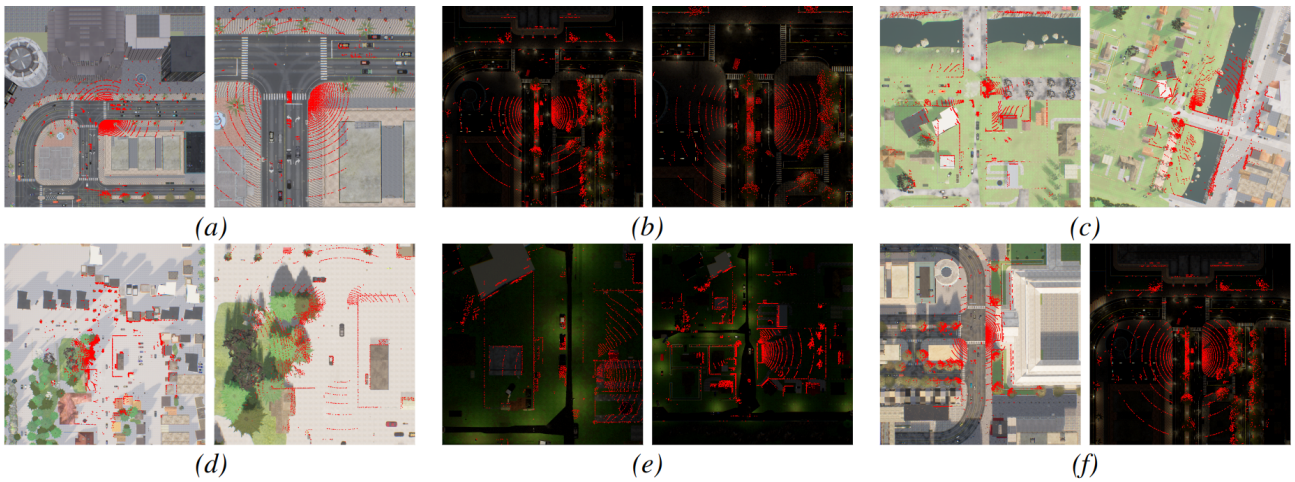


Figure 7.2: **Examples of the aerial-ground localization setting.** The local observation is a ground LiDAR scan, while the global reference is an aerial RGB map. This creates substantial differences in viewpoint, modality, and scale, motivating a transferable localization framework.

unknown scales, which further complicates fine-grained correspondence and metric pose estimation [128, 129, 132].

These challenges imply that effective aerial-ground localization should satisfy three requirements: it should learn transferable representations across viewpoint and modality gaps, preserve structural information for fine-grained matching, and explicitly account for scale inconsistency between the local and global observations [125, 128, 131]. These requirements motivate a localization framework that combines coarse cross-modal correspondence with structure-aware refinement and explicit scale alignment. The next section presents **AGL-Net**, which is designed around these principles.

7.4 Method

7.4.1 Overview

Within this dissertation, AGL-Net is studied as a transferable aerial-ground localization framework that connects aerial perception with ground-level robotic sensing. Unlike the earlier chapters, which primarily focused on understanding aerial videos themselves, this chapter studies how aerial observations can serve as global references for localization across viewpoint, modality, and scale changes. The central goal is to estimate the pose of a ground robot by matching a local LiDAR observation against a global aerial RGB map.

As illustrated in Figure 7.3, AGL-Net follows a two-stage localization pipeline. In the first stage, the model extracts neural representations from the ground LiDAR scan and the aerial RGB map and performs coarse cross-modal matching to estimate an initial pose hypothesis. In the second stage, it extracts structural skeleton features from both modalities and performs refined matching with explicit scale alignment. This combination is important because coarse neural matching provides transferable cross-modal correspondence, while structure-aware refinement improves localization precision under local ambiguity and scale mismatch.

To support this design, AGL-Net is trained with supervision on pose estimation, scale prediction, and skeleton consistency. The resulting representation is therefore encouraged to be not only correspondence-aware across sensing modalities, but also structurally meaningful and robust to scale variation. In this way, the framework addresses the three requirements identified in the previous section: cross-modal robustness, fine-grained structural matching, and explicit scale awareness.

The rest of this section is organized as follows. We first describe the feature extraction process

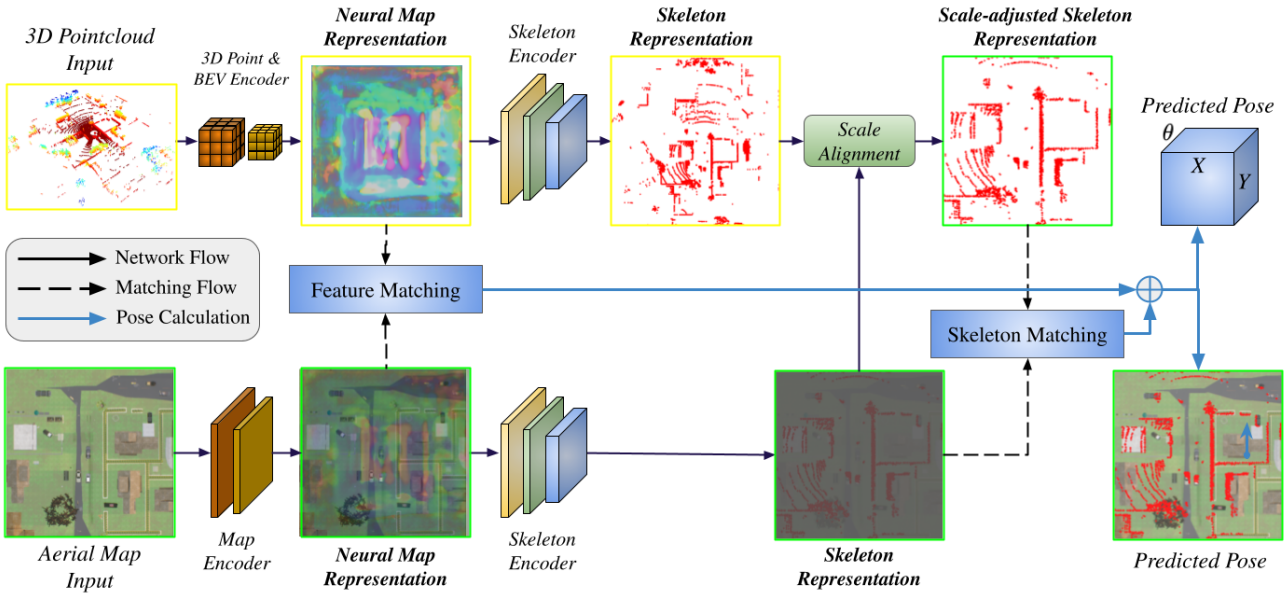


Figure 7.3: **Overview of AGL-Net.** Given a ground LiDAR scan and an aerial RGB map patch, AGL-Net extracts neural and skeleton representations, performs coarse neural feature matching, and then refines localization through skeleton-based matching with explicit scale alignment.

for the ground LiDAR scan and the aerial RGB map. We then introduce the scale alignment module and the proposed two-stage matching strategy. Finally, we present the training objectives used to optimize the full localization framework.

7.4.2 Ground LiDAR Encoding

Given a ground LiDAR scan $\mathcal{L}_g$, our goal is to produce a local bird's-eye-view (BEV) representation with consistent spatial resolution. This is important because, unlike the aerial map whose image resolution may vary, LiDAR measurements are naturally defined in metric space. We therefore encode the LiDAR scan into a BEV feature map $f_{bev} \in \mathbb{R}^{H \times W \times C}$, which serves as the local query representation for subsequent localization.

We first discretize the raw point cloud into equally spaced 3D voxels $V \in \mathbb{R}^{H \times W \times L}$, where each voxel stores the points that fall into its spatial cell. For each voxel, the contained points are either

downsampled or zero-padded to a fixed size. A PointNet encoder [140] is then applied within each voxel to extract local point features, and Transformer blocks [141] are used to model the relations among voxels. Following the general idea of PointPillars [142], the voxel features are scattered back to their original spatial locations and compressed along the height dimension using 2D convolutions, yielding the BEV pseudo-image f_{bev}. This representation reduces the viewpoint gap between the ground LiDAR input and the aerial overhead map while preserving spatial structure such as road boundaries, intersections, and building contours.

In addition to the dense BEV feature map, AGL-Net extracts a binary skeleton representation $f_{bev}^{s} \in \mathbb{R}^{H \times W \times 2}$. We use a CNN Φ_{bev}^{s} with 2D deformable convolution layers to generate this structure-aware feature. The goal of this branch is to emphasize the geometric layout of the local scene while suppressing redundant feature variation, so that the resulting representation can support more reliable structural refinement in the later matching stage.

Overall, the LiDAR encoder produces two complementary outputs: the dense neural feature map f_{bev} for coarse cross-modal correspondence and the structural representation f_{bev}^{s} for geometry-aware refinement.

7.4.3 Aerial Map Encoding

Let $\mathcal{I}_{map}$ denote an RGB patch cropped from the aerial overhead map M. The purpose of the map encoder is to transform this global visual reference into a dense spatial representation that can be matched against the local BEV LiDAR query.

We employ a pre-trained ResNet-50 [143] followed by a VGG19 network [144] to extract a dense map feature $f_{map} \in \mathbb{R}^{H \times W \times C}$. Unlike the LiDAR branch, which begins from sparse geometric

measurements, the aerial branch starts from dense RGB appearance. Its role is therefore not merely to preserve texture, but to produce a feature map whose spatial organization remains useful for aerial-ground matching despite the severe modality gap.

As in the LiDAR branch, we additionally derive a structural skeleton representation $f^s_{map} \in \mathbb{R}^{H\times W\times 2}$ from the neural map feature. This branch is implemented using a CNN Φ^s_{map} and is trained to emphasize edges, lines, and layout cues that are more stable across modalities than raw color or texture. In aerial imagery, structures such as road boundaries, geometric outlines, and topological patterns provide a more reliable bridge to the LiDAR-side BEV structure than appearance alone.

The aerial encoder therefore also produces two complementary outputs: the dense neural feature map f_{map} for coarse matching and the skeleton feature map f^s_{map} for refined structural alignment. This design allows the two modalities to interact first at the level of learned correspondence and later at the level of explicit geometry.

7.4.4 Scale Alignment

A central challenge in our setting is that the local LiDAR scan and the aerial RGB map may not be aligned in scale. The LiDAR input is always represented in metric coordinates, whereas the aerial map can be provided at uncertain or varying spatial resolutions depending on the crop size, map source, or deployment configuration. Even when the correct aerial region is identified, direct structural matching may still fail if the local and global observations correspond to different scales.

To address this issue, we introduce a CNN-based scale classifier Φ_{scale} that operates on the map skeleton feature f^s_{map} rather than on the full neural map feature f_{map}. This choice is motivated by the observation that structural cues, such as road widths, intersection shapes, and building outlines,

provide a cleaner basis for scale reasoning than raw RGB appearance. Given a predefined range of discrete scale bins, Φ_{scale} predicts a softmax-normalized weight for each bin, and the final continuous scale factor $\mathbb{S}$ is computed as the weighted sum over these bins.

The predicted scale factor $\mathbb{S}$ is then used to transform the LiDAR-side skeleton feature f_{bev}^{s} before refined matching. We denote the scaled feature by $f_{bev}^{\mathbb{S}}$ and use nearest-neighbor interpolation so that the local skeleton can be resized while preserving its binary structural pattern.

When $\mathbb{S} > 1$, the BEV skeleton corresponds to a smaller support region in the aerial map, so we shrink the effective spatial support of f_{bev}^{s} and place the result back into a zero-padded feature map of the original size. When $\mathbb{S} < 1$, we instead enlarge the relevant central region to align with a larger support in the map. In both cases, the transformed feature retains the same tensor size as the original representation, which simplifies the subsequent matching pipeline.

This scale alignment step is important because it explicitly compensates for spatial-resolution mismatch before structural refinement. Without such correction, even accurate structural cues may fail to align, which can significantly degrade metric pose estimation.

7.4.5 Two-Stage Template Matching

Given the neural and skeleton features from both branches, AGL-Net estimates the camera pose $\eta^* = (u^*, v^*, \theta^*)$ through a two-stage matching process. The key idea is to separate coarse cross-modal correspondence from fine-grained structural refinement. These two tasks are related but not identical: identifying the correct neighborhood in a large aerial map requires robust high-level matching, whereas accurate pose recovery requires precise local alignment.

In the first stage, we perform exhaustive matching between the neural map feature f_{map} and

the BEV feature f_{bev} across candidate poses η. This produces a score volume Ω, where each score measures how well the transformed BEV feature aligns with the aerial map feature under the candidate pose. Formally,

$$\Omega(\eta) = \frac{1}{XY} \sum_{p \in X \times Y} f_{map}(\eta(p))^\top f_{bev}(p),$$

where $\eta(p)$ transforms a 2D point p from the BEV coordinate system into the aerial map frame. In practice, the matching is implemented efficiently by rotating f_{bev} only N_r times and performing the convolutions using batched multiplication in the Fourier domain. This stage provides a coarse localization prior that is robust to the large modality gap, but it may still be ambiguous in environments with repeated local structure.

In the second stage, we refine the localization hypothesis using the structural skeleton features. After applying the predicted scale alignment, the model compares the LiDAR-side structural map $f^{\mathbb{S}}_{bev}$ with the aerial skeleton feature f^{s}_{map} and computes a refined matching score

$$\Psi(\eta) = \frac{1}{XY} \sum_{p \in X \times Y} f^{s}_{map}(\eta(p))^\top f^{\mathbb{S}}_{bev}(p).$$

Compared with Ω, this term is less sensitive to modality-specific appearance differences and focuses more directly on structural consistency between the two views.

The final pose likelihood combines both matching terms. We construct a discretized probability volume over candidate poses by applying a softmax to the summed scores,

$$P = \mathrm{softmax}(\Omega + \Psi),$$

and obtain the predicted pose by maximum likelihood estimation:

$$\eta^* = \arg\max_{\eta} P\big(\eta \mid f_{bev}, f_{map}, f_{bev}^{\mathbb{S}}, f_{map}^{s}\big).$$

This two-stage design is important because the two matching terms are complementary. The neural feature term Ω captures broad cross-modal compatibility and helps identify plausible candidate regions, while the structural term Ψ improves metric precision by enforcing geometry-aware local alignment after scale correction.

7.4.6 Training Objectives

To optimize the full localization framework, we use a composite training loss

$$\mathbb{L} = L_{uv\theta} + L_{\mathbb{S}} + L_{skeleton},$$

which jointly supervises pose estimation, scale prediction, and skeleton extraction.

The first term, $L_{uv\theta}$, supervises the estimated camera pose. Since AGL-Net predicts a discretized probability distribution over candidate poses, we train the model using the negative log-likelihood of the ground-truth pose:

$$L_{uv\theta} = -\log P(\eta^*).$$

This objective encourages the matching pipeline to assign high probability to the correct position and heading, and therefore directly optimizes the final localization behavior rather than a surrogate retrieval score.

To guide the scale alignment module, we introduce a scale supervision term that penalizes the

discrepancy between the predicted scale factor $\mathbb{S}$ and the ground-truth scale $\mathbb{S}_{gt}$. We use a mean squared error loss:

$$L_{\mathbb{S}} = \frac{1}{B} \sum_{i=1}^{B} \left(\mathbb{S} - \mathbb{S}_{gt} \right)^2 ,$$

where B is the batch size. This term is important because learning scale purely from the final pose objective can be unstable under large scale variation.

Finally, we supervise the predicted skeleton map using a binary cross-entropy loss. Let $y_{gt}(p)$ denote the ground-truth skeleton label at spatial location p. Then

$$L_{skeleton} = -\frac{1}{X \times Y} \sum_{p \in X \times Y} \Big[y_{gt}(p) \log f^{s}_{map}(p) + \big(1 - y_{gt}(p)\big) \log\big(1 - f^{s}_{map}(p)\big) \Big].$$

The ground-truth skeleton mask is obtained by projecting the LiDAR points onto the aerial map, as illustrated in Figure 7.2. During inference, the model predicts the skeleton of the full map patch, while the loss is computed only over the region covered by the LiDAR observation.

These three objectives act on different but complementary parts of the model. The pose loss drives accurate global localization, the scale loss improves robustness to varying aerial resolution, and the skeleton loss encourages meaningful structural representations for refined cross-modal alignment.

7.5 Experimental Evaluation

In this section, we evaluate AGL-Net on both simulated and real-world data to answer three questions. First, can aerial-ground cross-modal localization be performed accurately under large viewpoint and modality discrepancies? Second, does explicit scale alignment improve metric localization when the aerial reference is observed at varying scales? Third, do the results support the broader claim of this

chapter—namely, that aerial perception can serve as a transferable global reference for ground-level robotic sensing rather than being limited to aerial-only scene understanding?

7.5.1 Datasets and Evaluation Setting

We evaluate AGL-Net on both CARLA [145] and KITTI [146]. These two datasets serve complementary roles in this chapter. CARLA provides a controlled environment for training and evaluation with accurate pose supervision and systematically varied aerial map scales, while KITTI provides a real-world benchmark for assessing whether the learned localization framework transfers beyond simulation.

CARLA is used as the primary benchmark for both training and quantitative evaluation. It provides paired ground LiDAR observations and aerial reference maps together with precise pose annotations, making it well suited for studying metric localization under controlled scale variation. Since the main challenge of this chapter lies in aerial-ground localization under viewpoint, modality, and scale differences, CARLA allows us to isolate these factors and evaluate the proposed scale-aware design in a clean setting.

We also report results on KITTI [146] for comparison in a real autonomous driving scenario. Compared with CARLA, KITTI introduces additional challenges such as sensing noise, map mismatch, and greater appearance variation, making aerial-ground localization substantially more difficult. Evaluating on both datasets therefore allows us to study AGL-Net in both a controlled metric localization setting and a more realistic deployment setting. From the perspective of this dissertation, this distinction is important: CARLA tests whether aerial-ground correspondence can be learned under clean supervision, while KITTI tests whether such correspondence remains useful when aerial

observations are used as transferable references in real-world sensing conditions.

We report both recall-based and error-based metrics. Specifically, *Recall@X m/* denotes the percentage of predictions whose position and orientation errors are within a specified threshold. We report position recall at 1, 3, and 5 meters and orientation recall at 1, 3, and 5 degrees. In addition, we report the average position error in meters and the average heading error in degrees. Together, these metrics capture both strict localization success and overall pose estimation quality.

7.5.2 Implementation Details

We first convert each LiDAR scan into a BEV representation by voxelizing the point cloud into pillars of size $[2, 2, 30]$. The point ranges along the x, y, and z axes are set to $[-100, 100]$, $[-100, 100]$, and $[-10, 20]$, respectively, and the maximum number of points per voxel is 128. For the aerial reference, we crop overhead map patches of size 1024×1024 and 256×256 at different random scales so that the resulting patches cover different physical areas while still containing the ground-truth pose.

For the LiDAR branch, we set the PointNet feature dimension to 16 and use two attention blocks with four heads and feature dimension 16. After four Conv2D layers with kernel size 1, the matching dimension of f_{bev} is reduced to 8. The skeleton extraction networks Φ^s_{bev} and Φ^s_{map} share the same architecture, consisting of three deformable Conv2D layers with ReLU activations followed by a linear layer. The scale classifier Φ_{scale} consists of three deformable Conv2D layers with ReLU activations followed by two linear layers, with 33 scale bins covering the range $[0.5, 10]$.

We train all models on four NVIDIA RTX A6000 GPUs for 200k iterations using the Adam optimizer with a learning rate of 10^{-4}. The batch size is 48 for map patches of size 1024 and 4 for map

Map	Ground Modality	Method	Map Size (pixels)	Avg. Loc. Error (m) ↓	Avg. Ori. Error (°) ↓	Lat. Recall (%) @ X m ↑			Long. Recall (%) @ X m ↑			Ori. Recall (%) @ X° ↑		
						1m	3m	5m	1m	3m	5m	1°	3°	5°
OSM (KITTI)	Image	retrieval	256×256	–	–	37.47	66.24	72.89	5.94	16.88	26.97	2.97	12.32	23.27
	Image	refinement	256×256	–	–	50.83	78.10	82.22	17.75	40.32	52.40	31.03	66.76	76.07
	Image	OrienterNet	256×256	–	–	51.26	84.77	91.81	22.39	46.79	57.81	20.41	52.24	73.53
	LiDAR	OrienterNet*	256×256	28.01	**8.30**	2.86	8.80	15.37	4.03	11.47	18.48	8.62	24.11	36.38
	LiDAR	**AGL-Net**	256×256	**18.02**	8.59	6.55	18.88	31.17	5.16	15.00	24.54	6.25	17.94	31.87
Satellite (CARLA)	LiDAR	OrienterNet*	256×256	2.23	19.15	66.87	87.67	98.61	44.38	81.51	97.69	55.78	62.71	66.26
	LiDAR	**AGL-Net**	256×256	**1.83**	**3.76**	76.58	92.14	100.0	44.07	88.44	99.54	57.47	68.41	74.42
	LiDAR	OrienterNet*	1024×1024	18.75	82.71	33.44	40.99	56.55	13.41	18.64	22.50	30.35	31.43	32.51
	LiDAR	**AGL-Net**	1024×1024	**14.96**	**57.25**	50.54	57.94	69.49	18.64	24.19	29.89	42.06	44.22	46.22

Table 7.1: **Results on KITTI and CARLA.** Compared with the adapted OrienterNet baseline, AGL-Net yields lower localization error and stronger recall on most evaluation metrics, demonstrating more robust aerial-ground localization across different environments and map scales. Average localization and orientation errors are reported in meters and degrees, respectively. All recall values are reported as percentages. * indicates that we make small modifications to the original method to take LiDAR modality input.

patches of size 256.

7.5.3 Main Results on CARLA and KITTI

The main quantitative comparisons are shown in Table 7.1. We compare AGL-Net with OrienterNet [147], which is originally designed for image-based localization. To provide a fair comparison in our setting, we adapt OrienterNet to use LiDAR input instead of ground RGB images.

On CARLA, AGL-Net consistently outperforms OrienterNet across both position and orientation metrics. For example, with a map size of 224×224, AGL-Net reduces the average position error by 0.4 meters and the average orientation error by 15.39. The advantage remains clear when the map size is increased to 1024×1024, indicating that the proposed localization framework remains effective even as the global search space becomes larger. These results support the central design of the method: coarse neural matching identifies plausible aerial-ground correspondences, while the second-stage skeleton refinement further improves pose accuracy under local ambiguity and scale variation.

More importantly, the CARLA results show that aerial observations can indeed function as structured global references for local ground sensing. In other words, the model is not merely learning a

benchmark-specific matching shortcut; it is learning a form of transferable aerial-ground correspondence that remains useful for metric pose recovery even when the two inputs differ substantially in viewpoint, modality, and scale. This directly supports the broader theme of Part III, where aerial perception is studied not only as a source of semantic understanding, but also as a source of transferable spatial reference.

On KITTI, the localization problem is substantially harder. Compared with image-map geo-localization, matching LiDAR scans to overhead RGB maps introduces a stronger cross-modal gap, and the real-world data also contains additional noise from sensor calibration, map mismatch, and localization uncertainty. Even under these conditions, AGL-Net achieves encouraging performance and improves the average position error over OrienterNet by 9.99 meters. The orientation improvement is less consistent than on CARLA, which likely reflects the greater difficulty of precise structural alignment on real-world data. Nevertheless, the KITTI results still demonstrate that the proposed framework transfers beyond simulation and remains effective in realistic driving environments.

This real-world behavior is particularly important from a dissertation perspective. If the gains were limited to CARLA alone, the chapter would mainly show that aerial-ground localization is feasible in a controlled setting. By contrast, the KITTI results suggest that aerial representations can remain useful even when the target sensing platform, environmental conditions, and data quality differ substantially from the training setup. This moves the contribution beyond a single benchmark and supports the broader claim that aerial perception can be extended from aerial-only understanding to transferable cross-sensing spatial reasoning.

Taken together, these results support a broader point of Part III. The value of aerial perception is not restricted to understanding aerial observations themselves; it can also provide a transferable global reference for ground-level robotic sensing. The gains on both CARLA and KITTI therefore suggest

Feature Match.	Skeleton Match.	Scale Align.	Scale Aug.	Avg. Loc. Error (m) ↓	Avg. Ori. Error (°) ↓	Loc. Recall (%) @ X m ↑		Ori. Recall (%) @ X° ↑	
						1m	3m	1°	3°
✓	✗	✗	✗	10.12	41.86	9.24	29.58	32.36	35.90
✓	✗	✗	✓	2.23	19.15	30.97	68.88	55.78	62.71
✗	✓	✗	✗	3.53	10.61	7.55	37.29	10.32	13.56
✗	✓	✗	✓	3.40	10.26	8.47	43.76	12.17	16.18
✓	✓	✗	✗	9.14	33.25	9.71	30.66	50.23	54.24
✓	✓	✗	✓	2.33	17.53	31.28	69.34	54.24	62.10
✓	✓	✓	✗	9.50	40.41	10.94	33.59	32.67	36.52
✓	✓	✓	✓	**1.83**	**3.76**	**34.05**	**78.74**	**57.47**	**68.41**

Table 7.2: **Ablation studies on different components.** Combining feature matching, skeleton matching, scale alignment, and scale augmentation yields the best overall localization performance, demonstrating that structural refinement and explicit scale handling are both important for robust aerial-ground localization.

that aerial-ground correspondence can be learned despite large viewpoint, modality, and scale discrepancies, extending aerial perception from scene understanding to cross-sensing spatial reasoning.

7.5.4 Ablation Studies

We next evaluate the contribution of the main components of AGL-Net using ablation studies on map patches of size 256×256. The results are summarized in Tables 7.2 and 7.3.

Table 7.2 evaluates the contribution of different matching components. The results show that scale augmentation already improves performance by exposing the model to aerial maps with varying spatial resolutions during training. This helps the network learn more robust cross-modal correspondences and reduces sensitivity to fixed-scale assumptions. The table also shows that skeleton matching is more robust to scale variation than feature matching alone, since the structural representation emphasizes edges and layout cues that are less sensitive to raw appearance or local detail changes. When scale alignment is introduced by explicitly predicting a scale factor and resizing the BEV skeleton accordingly, the model achieves the best performance. This confirms that accurate scale compensation is a key ingredient of metric aerial-ground localization.

NLL Loss	Scale Loss	Skeleton Loss	Avg. Loc. Error (m) ↓	Avg. Ori. Error (°) ↓	Loc. Recall (%) @ X m ↑		Ori. Recall (%) @ X° ↑	
					1m	3m	1°	3°
✓	✗	✗	2.23	19.15	30.97	68.88	55.78	62.71
✓	✓	✗	3.04	26.07	30.35	67.18	**59.48**	64.56
✓	✗	✓	2.67	22.38	31.28	67.80	59.17	64.10
✓	✓	✓	**1.83**	**3.76**	**34.05**	**78.74**	57.47	**68.41**

Table 7.3: **Ablation studies on different loss terms.** Combining both the scale loss and the skeleton loss with the NLL objective yields the best overall localization performance, giving the lowest localization and orientation errors together with the strongest recall on most metrics.

This result is also important at the thesis level. Once aerial observations are used as references for ground sensing, generic cross-modal similarity alone is no longer sufficient. The model must additionally reason about how local structure transfers across sensing modalities and how that structure should be spatially aligned under different map scales. The gains from scale alignment therefore support a broader conclusion of this chapter: transferable aerial perception requires explicit geometric reasoning once it is used for cross-sensing localization rather than aerial-only recognition.

Table 7.3 studies the role of the training objectives. Using both the scale loss $L_{\mathbb{S}}$ and the skeleton loss $L_{skeleton}$ yields better performance than using either loss alone. This result is important because the two losses supervise different but interdependent parts of the model. The scale loss encourages the LiDAR-side structural representation to align with the aerial map at the correct resolution, but its effectiveness depends on the quality of the learned structural cues. Conversely, the skeleton loss improves the quality of the structure-aware representation, but by itself does not guarantee correct scale reasoning. Their combination therefore leads to more stable and accurate pose estimation.

From the perspective of this dissertation, this finding is important because it shows that transferable aerial perception requires more than generic feature matching alone. Once aerial observations are used as global references for ground sensing, structural alignment and scale reasoning become essential for reliable cross-sensing localization. The ablation results therefore reinforce the broader theme

of Part III: transferring aerial perception across sensing regimes requires representations that are not only semantically meaningful, but also structurally and geometrically compatible.

Overall, the ablation results support the main design choices of AGL-Net. The gains do not come from a single isolated modification, but from the interaction between cross-modal feature matching, structure-aware refinement, and explicit scale alignment. This is consistent with the problem formulation in the previous sections: robust aerial-ground localization requires not only transferable correspondence across sensing modalities, but also fine-grained structural matching under varying spatial scales.

7.5.5 Qualitative Results

Figure 7.4 presents qualitative examples of the predicted localization results on CARLA and KITTI. These examples illustrate several important behaviors of AGL-Net. First, the method is able to identify the correct aerial region despite the substantial difference between a sparse local LiDAR observation and a dense aerial RGB map. Second, the refined matching stage improves local alignment in scenes with repeated patterns or partial structural ambiguity. Third, the method remains reasonably robust even when the aerial reference is observed under different scales.

The qualitative results are consistent with the quantitative findings. On CARLA, the predicted poses tend to align closely with the ground truth, reflecting the effectiveness of the full coarse-to-fine matching pipeline under accurate supervision. On KITTI, although the predictions are noisier, AGL-Net still captures meaningful aerial-ground correspondence and often localizes near the correct map region. These visualizations therefore support the main claim of this chapter: aerial observations can function as transferable global references for ground-level localization when cross-modal matching is

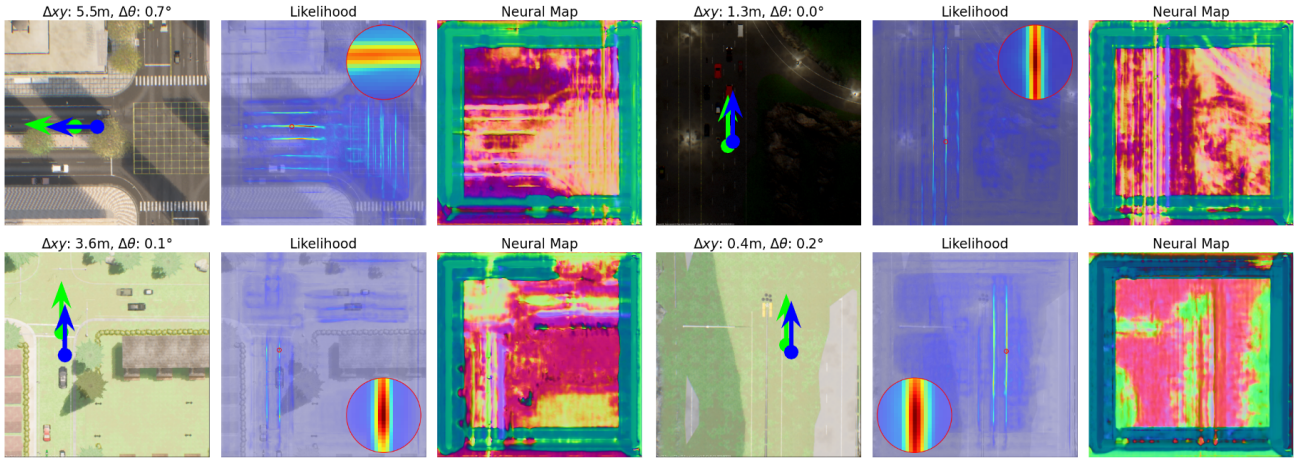


Figure 7.4: **Qualitative localization results on CARLA and KITTI.** AGL-Net is able to recover meaningful aerial-ground correspondence despite substantial viewpoint, modality, and scale differences. The examples also show that the proposed refinement stage improves local structural alignment in challenging scenes.

combined with structure-aware refinement and explicit scale alignment.

7.6 Discussion

AGL-Net can be understood as a step from *aerial perception for aerial-only understanding* toward *aerial perception for cross-sensing spatial reasoning*. The central contribution of this chapter is therefore not only that it improves localization accuracy on CARLA and KITTI, but that it shows how aerial observations can function as transferable global references for ground-level robotic sensing. In this setting, the aerial map is no longer just a visual context source; it becomes a spatial prior against which local ground observations must be aligned despite large discrepancies in viewpoint, modality, and scale.

This perspective helps explain why the proposed design differs from standard cross-view retrieval or same-modality registration. Once aerial imagery is used as a global reference for ground LiDAR localization, generic feature similarity is not sufficient. The model must first establish broad

aerial-ground correspondence and then resolve fine-grained structural alignment under uncertain scale. The two-stage design of AGL-Net reflects exactly this requirement: coarse neural matching captures transferable cross-modal compatibility, while skeleton-based refinement and explicit scale alignment improve metric precision by focusing on geometry that is more stable across sensing regimes.

From the perspective of Part III, this chapter also plays an important bridging role. The earlier parts of this dissertation focused primarily on learning better representations for aerial videos and improving practical aerial perception under weak supervision or deployment constraints. By contrast, this chapter shows that aerial perception can also be useful when transferred beyond the aerial sensing regime itself. In particular, it suggests that aerial observations can provide globally organized spatial information that complements the local but metrically grounded structure captured by ground LiDAR. This extends the role of aerial perception from *understanding what is observed from above* to *reasoning across sensing platforms using aerial information as a shared reference frame*.

More broadly, the results suggest that transferable aerial perception depends on both semantic correspondence and geometric compatibility. The semantic side is needed to bridge the large appearance and modality gap between aerial RGB and ground LiDAR, while the geometric side is needed to support precise localization once a plausible region has been identified. This dual requirement is likely to remain important in other aerial-ground tasks beyond localization, including cross-platform mapping, multi-robot coordination, and large-scale spatial reasoning.

7.7 Limitations

Although AGL-Net improves aerial-ground metric localization, it still has several limitations.

First, the current formulation assumes that the aerial map already provides sufficient structural

cues for localization. In highly texture-poor environments or regions with weak geometric layout, the skeleton-based refinement stage may become less reliable. Similarly, when large portions of the local scene are occluded or poorly observed in the LiDAR scan, the structural correspondence between the local and global observations may degrade.

Second, the current scale alignment module predicts a single global scale factor for the local query. This is effective for the map variation studied in this chapter, but it may be insufficient when scale mismatch is spatially nonuniform or when aerial reference quality varies significantly across regions. More flexible alignment mechanisms may therefore be needed for broader deployment settings.

Third, while the KITTI results show promising transfer beyond simulation, the real-world evaluation remains limited in scope. A broader study across more cities, sensing platforms, and map sources would be needed to fully characterize how robust the learned aerial-ground correspondence is under stronger domain shift. In particular, seasonal change, map aging, sensor noise, and large geographic variation may all affect localization performance.

Finally, this chapter focuses on single-query localization against a static aerial reference. In practical robotic systems, localization is often performed jointly with temporal filtering, mapping, or multi-sensor fusion. Integrating AGL-Net into such broader systems remains an important direction for future work.

These limitations suggest that AGL-Net should be viewed as an initial step toward transferable aerial-ground localization rather than a complete solution. Nevertheless, the results of this chapter indicate that explicit structure-aware and scale-aware matching is a promising direction for extending aerial perception across sensing regimes.

7.8 Conclusion

In this chapter, we presented AGL-Net [128], a transferable aerial-ground localization framework for cross-modal metric pose estimation. By combining BEV LiDAR encoding, aerial map representation learning, explicit scale alignment, and two-stage neural-plus-structural matching, AGL-Net addresses the central challenge of aligning local ground LiDAR observations with global aerial RGB references under large differences in viewpoint, modality, and scale.

Experimental results on CARLA and KITTI showed that AGL-Net consistently improves over strong baselines and that its gains arise from the interaction between coarse cross-modal correspondence, structure-aware refinement, and scale reasoning. These results support the main claim of this chapter: accurate aerial-ground localization requires more than generic cross-view matching alone. Once aerial observations are used as global references for ground sensing, structural compatibility and explicit scale alignment become essential for reliable metric localization.

From the perspective of this dissertation, this chapter establishes the first part of a broader argument developed in Part III. Aerial perception is valuable not only for understanding aerial observations themselves, but also for serving as a transferable spatial reference across sensing regimes. In this sense, AGL-Net extends aerial perception from aerial-only scene understanding to cross-sensing spatial reasoning, showing how aerial information can support ground-level robotic localization even under substantial modality and viewpoint mismatch.

This chapter also sets up the next stage of the dissertation. Whereas AGL-Net studies transfer across aerial and ground sensing modalities, the following chapter investigates transfer across camera networks, where correspondence must be established across viewpoints within multi-camera environments. Together, these chapters broaden the role of aerial and visual perception from isolated

understanding within a single sensing regime to transferable reasoning across sensing platforms and viewpoints.

Chapter 8: CALIBFREE: Self-Supervised Calibration-Free Multi-Camera Multi-Object Tracking

8.1 Introduction

This chapter continues Part III, which studies transferable aerial perception across views, modalities, and camera networks. While Chapter 7 focused on transferable correspondence between aerial and ground sensing for localization, this chapter addresses a different but closely related problem: *how can identity representations remain consistent across multiple cameras when camera calibration is unavailable, unreliable, or impractical to maintain?* [128, 148, 149]

Multi-object tracking (MOT) has made substantial progress in recent years through improvements in detection quality, temporal association, and learned appearance modeling [150–153]. As camera networks become increasingly common in surveillance, transportation, and smart-space environments, *multi-camera multi-object tracking* (MCMOT) has become an important extension of this problem [148, 149]. Compared with single-camera MOT, MCMOT must preserve identity not only over time within each view, but also across cameras with different viewpoints, overlap patterns, and camera-specific appearance bias [148, 154, 155]. This broader setting is especially important in practical deployments because multiple cameras can reduce blind spots, improve robustness to occlusion, and extend tracking beyond the field of view of any individual camera [148, 149, 154].

Despite this promise, robust MCMOT in real-world camera networks remains difficult [149, 156]. Cross-camera observations often differ substantially due to viewpoint change, scale variation, partial occlusion, illumination shift, and camera-dependent appearance bias [148, 149, 155]. As a result, an identity representation that is stable within one camera may transfer poorly across other views [148, 155]. Many existing MCMOT methods address this problem by relying on camera calibration, geometric constraints, overlap priors, or scene-specific structural assumptions to establish cross-view correspondence [149, 154, 157]. While these assumptions can be effective in controlled environments, they are often difficult to maintain in practical deployments where cameras are reoriented, replaced, drift over time, or undergo changes in overlap coverage [149, 156].

These limitations motivate a different direction. Rather than depending on camera geometry, the model should learn cross-camera identity correspondence directly from raw visual observations [155, 158]. In partially overlapping camera networks, synchronized detections across views naturally provide a self-supervisory signal: detections that co-occur in time are likely to share useful cross-view information even without manual identity labels [155, 158]. This suggests that a purely RGB-driven, self-supervised representation may provide a stronger foundation for truly calibration-free MCMOT [155, 158].

To this end, this chapter presents **CALIBFREE**, a self-supervised representation learning framework for multi-camera tracking without camera calibration, geometric constraints, or manual identity annotations [115]. The key idea is to learn two complementary detection-level embeddings with distinct roles: a *view-agnostic* feature for cross-camera association and a *view-specific* feature for within-camera temporal tracking [115]. These two representations are learned through a joint distillation-and-reconstruction scheme. Single-view distillation stabilizes identity-sensitive features within each camera, while cross-view reconstruction aligns identity-relevant cues across cameras using synchronized

visual overlap alone [115]. A feature-separation regularizer further encourages the two embeddings to specialize rather than collapse into a single mixed representation [115].

From the perspective of this dissertation, CALIBFREE extends transferable aerial perception from cross-modal localization to *cross-camera identity consistency*. Whereas AGL-Net learns correspondence across viewpoint, modality, and scale for localization, CALIBFREE learns transferable identity representations across camera networks for tracking [115, 128]. Together, these chapters show that perception becomes more useful when it preserves task-relevant consistency under changing observation conditions rather than remaining confined to a single sensing setup [115, 128].

The main contributions of this chapter are as follows:

- We introduce a fully self-supervised MCMOT framework that requires neither camera calibration nor manual identity annotations, enabling robust deployment in dynamic camera networks [115].
- We propose a representation learning strategy that separates **view-agnostic** and **view-specific** embeddings, allowing cross-camera and within-camera associations to rely on features specialized for their respective roles [115].
- We develop a joint **single-view distillation** and **cross-view reconstruction** training scheme that improves identity consistency across time and across cameras without geometric constraints [115].
- We demonstrate strong and robust performance on challenging MCMOT benchmarks, showing that CALIBFREE improves both over-time and cross-view tracking while remaining resilient to detector noise and camera-dependent appearance variation [115].

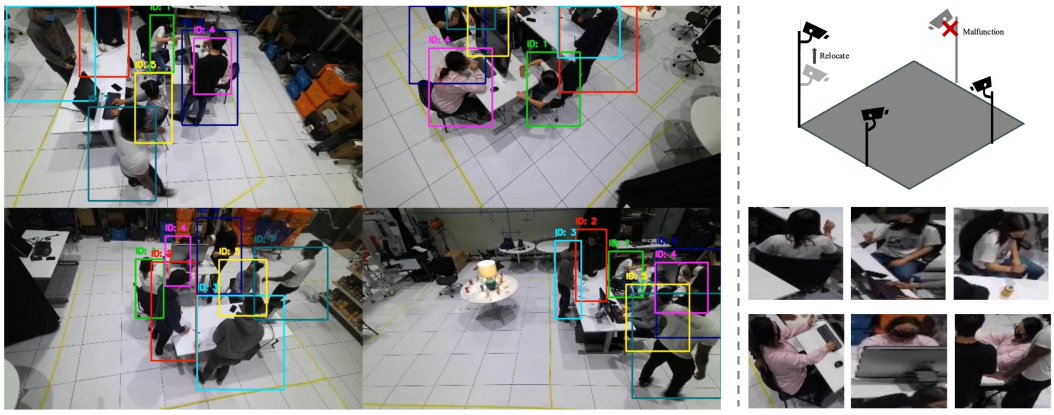


Figure 8.1: **Multi-Camera Multi-Object Tracking (MCMOT) setup.** *Left:* Multi-view scenes with individuals tracked across overlapping camera views, each assigned a unique color-coded bounding box. *Top right:* Flexible camera configurations illustrating variable camera numbers and placements across scenarios, or due to factors like malfunction or relocation. *Bottom right:* Examples of appearance variations for individuals across different viewpoints, highlighting the challenge of maintaining consistent identity association in multi-view tracking.

8.2 Related Work

Since the earlier chapters of this dissertation have already reviewed the broader literature on aerial perception and transferable representation learning, here we focus more narrowly on prior work most directly related to multi-object tracking across camera networks. In particular, this chapter is most closely connected to four lines of research: *single-camera multi-object tracking*, *multi-camera multi-object tracking and its benchmarks*, *geometry- or calibration-dependent cross-camera association*, and *self-supervised or unsupervised representation learning for cross-camera tracking*.

8.2.1 Single-Camera Multi-Object Tracking

Single-camera MOT has advanced rapidly through improvements in detection quality, temporal association, and learned appearance modeling. Early online tracking methods such as SORT and DeepSORT showed that strong detection together with motion prediction and appearance affinity can

already provide effective within-camera tracking [159, 160]. Subsequent methods further improved performance by integrating detection and tracking more tightly, for example through joint detection-tracking pipelines, point-based tracking formulations, and transformer-based association frameworks [150–153, 161–163]. These methods demonstrate that stable identity association within a single view can be learned effectively when temporal continuity and appearance consistency are both available.

However, single-camera MOT methods are fundamentally limited to within-view tracking. They do not address the additional challenge of preserving identity across different cameras, where appearance, viewpoint, scale, and overlap conditions may vary substantially. As a result, features that work well for temporal association within one camera do not necessarily transfer reliably to cross-camera identity matching.

8.2.2 Multi-Camera Multi-Object Tracking and Benchmarks

Multi-camera multi-object tracking extends the tracking problem across distributed camera networks by requiring both temporal continuity within each view and identity consistency across views. Early work established this setting through benchmark design, performance measures, and graph-based or optimization-based formulations that connect observations across cameras [148, 164, 165]. A number of datasets and benchmarks have since played an important role in this area, including DukeMTMC, WILDTRACK, CityFlow, and more recent multi-view human association datasets such as MvMHAT [155, 166–169]. These benchmarks collectively show that MCMOT is substantially more challenging than single-camera MOT because association must remain reliable despite viewpoint change, partial overlap, and camera-dependent visual bias.

Recent methods have continued to improve MCMOT performance using graph modeling, global

association strategies, and stronger identity features [170–173]. At the same time, recent surveys emphasize that practical deployment remains difficult because many methods still rely on assumptions that become brittle in dynamic real-world camera networks [149, 156]. This motivates approaches that reduce dependence on fixed camera geometry or manual scene engineering.

8.2.3 Geometry- and Calibration-Dependent Cross-Camera Association

A major line of prior work addresses cross-camera association through explicit spatial or temporal constraints. Earlier approaches modeled multi-camera tracking using space-time consistency, inter-camera transition patterns, or graph structures defined over calibrated views [154, 165, 174, 175]. More recent methods also exploit geometric relationships or structured scene priors to narrow the association search space and improve tracking consistency across cameras [157, 172, 173].

These methods can be highly effective when calibration is accurate and camera layouts remain stable. However, such assumptions are often difficult to maintain in practical deployments. Camera networks may evolve over time due to camera replacement, reorientation, drift, or changing overlap coverage, making calibration-dependent solutions increasingly costly to maintain at scale [149, 156]. This limitation is especially important for the problem studied in this chapter, where the goal is not merely to improve association in a fixed network, but to learn identity representations that remain useful even when explicit camera geometry is unavailable or unreliable.

8.2.4 Self-Supervised and Unsupervised Representation Learning for Cross-Camera Tracking

A complementary direction seeks to reduce annotation and calibration dependence by learning representations directly from the visual data itself. In person re-identification, unsupervised and tracklet-based learning methods have shown that useful identity structure can emerge without full manual labels [176–178]. Related work has also explored self-supervised association across multiple views, showing that co-occurring observations from different cameras can provide a meaningful supervisory signal for cross-view learning [179].

This idea is especially relevant to MCMOT, where synchronized observations naturally provide cross-view information even without explicit identity annotations. Recent work such as self-supervised multi-view multi-human association demonstrates that cross-camera correspondence can indeed be learned from synchronized RGB observations alone [155, 158]. However, existing methods still face an important challenge: the same feature representation is often asked to support both within-camera temporal tracking and cross-camera association, even though these two tasks emphasize different types of information. Within-view tracking benefits from retaining camera-specific appearance details, while cross-view matching requires suppressing such bias in favor of more stable identity-preserving cues.

This chapter addresses that gap through **CALIBFREE** [115]. Unlike calibration-dependent MCMOT methods, CALIBFREE learns identity correspondence without camera parameters or geometric constraints. Unlike approaches that rely on a single shared embedding, it explicitly learns two complementary representations: a *view-agnostic* feature for cross-camera matching and a *view-specific* feature for within-camera temporal association. These representations are learned jointly through single-view distillation, cross-view reconstruction, and feature-separation regularization, enabling a

fully self-supervised and calibration-free solution to multi-camera tracking.

8.3 Problem Formulation and Motivation

We consider multi-camera multi-object tracking (MCMOT) over synchronized video streams from V cameras. At time step t, let

$$D_t^v = \{D_i^v \mid i = 1, 2, \ldots, N_t^v\}$$

denote the set of person detections in camera v, where N_t^v is the number of detections in that view. The full multi-camera detection set at time t is therefore

$$\bar{D}_t = \{D_t^1, D_t^2, \ldots, D_t^V\}.$$

The goal of MCMOT is to assign detections to identities consistently across both time and cameras. This requires solving two coupled association problems. The first is *intra-camera tracking*, where detections must be associated over time within each camera to form temporally stable tracklets. The second is *cross-view matching*, where detections from different cameras must be associated when they correspond to the same subject. In contrast to single-camera MOT, MCMOT therefore requires a representation that remains useful not only under temporal variation within a view, but also under cross-camera viewpoint and appearance changes [148–153].

A straightforward approach would be to learn a single detection-level embedding and use it for both within-camera and cross-camera association. However, this is problematic in the calibration-free setting considered in this chapter. Temporal tracking within a camera often benefits from preserving

local appearance details, including pose, clothing texture, and camera-dependent imaging characteristics. By contrast, cross-camera matching requires suppressing such view-specific bias and emphasizing identity-preserving cues that remain stable across viewpoints [148, 155, 158]. A single shared representation is therefore asked to satisfy two partially conflicting requirements.

A second challenge arises from the absence of camera calibration and manual identity annotations. Many existing MCMOT approaches use geometric constraints, overlap priors, or scene-specific calibration to reduce cross-camera ambiguity [149, 154, 157]. Without these signals, the model must infer cross-view identity consistency directly from raw RGB detections alone. This makes the learning problem substantially harder, since the supervision must come from synchronized observation structure rather than explicit labels or geometry.

These observations motivate learning two complementary detection-level embeddings with different association roles:

- a **view-agnostic feature** f_a, intended to capture identity-preserving cues that remain consistent across cameras; and
- a **view-specific feature** f_s, intended to preserve appearance details that are useful for temporal tracking within a camera.

The central question is therefore not only how to learn a strong tracking feature, but how to learn *specialized* features that support different forms of association without labels or calibration. In particular, an effective calibration-free MCMOT framework should satisfy three requirements. First, it should learn temporally stable within-view representations from masked or partial detections. Second, it should exploit synchronized cross-view observations to improve identity consistency across cameras. Third, it should encourage functional specialization between the two learned embeddings rather than

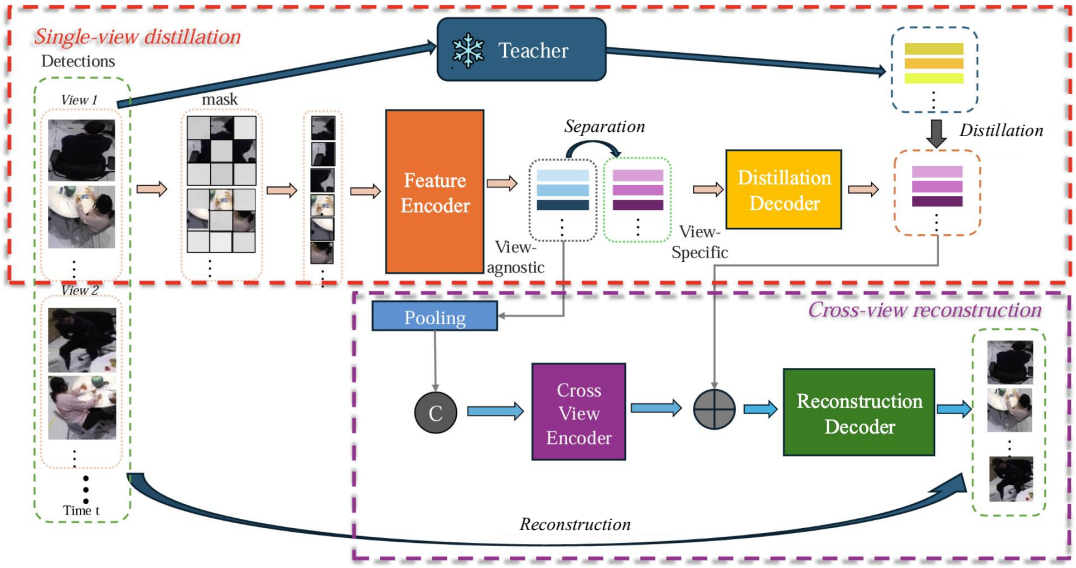


Figure 8.2: **Overview of CALIBFREE.** The method includes single-view distillation, feature separation, and cross-view reconstruction. In single-view distillation (red box), masked detections are encoded, and feature quality is supervised by a teacher model using a distillation loss. A separation regularizer encourages the learned features to specialize into view-agnostic and view-specific components. For cross-view reconstruction (purple box), pooled view-agnostic features are processed to reconstruct masked patches across views, optimized with a reconstruction loss.

allowing them to collapse into a single mixed representation [115, 155, 158].

These requirements motivate the design of **CALIBFREE**. Rather than relying on fixed camera geometry, CALIBFREE learns representation specialization directly from synchronized RGB observations. As shown in Figure 8.1 and Figure 8.2, the framework combines single-view distillation, cross-view reconstruction, and feature-separation regularization to learn one feature for within-camera temporal association and another for cross-camera identity matching. The next section presents this formulation in detail.

8.4 Method

8.4.1 Overview

Within Part III, CALIBFREE studies transferable identity representation learning across camera networks. Unlike Chapter 7, which focused on transferable correspondence across sensing modalities for localization, this chapter considers how to preserve task-relevant identity consistency across views when camera calibration is unavailable.

As illustrated in Figure 8.2, CALIBFREE consists of three main components. First, a shared feature encoder extracts detection-level representations and splits them into a view-agnostic branch and a view-specific branch. Second, single-view distillation improves feature stability within each camera by training a masked student to match an unmasked teacher. Third, cross-view reconstruction uses synchronized observations from overlapping cameras to align identity-relevant information across views. A feature-separation regularizer is further applied to encourage the two branches to specialize for different association roles.

The rest of this section first describes feature extraction and representation specialization, then introduces single-view distillation and cross-view reconstruction, and finally presents the separation loss, training objective, and inference procedure for MCMOT.

8.4.2 Feature Extraction and Representation Specialization

At each synchronized time step t, person detections are first extracted independently from all camera views. Let D_i^v denote the i-th detection in camera v. Each detected region is cropped, resized to a fixed spatial resolution, and partitioned into non-overlapping patches. After patch embedding,

every detection is represented as a sequence of patch tokens that serves as the input to the shared feature encoder.

To improve robustness to partial visibility and encourage contextual reasoning, CALIBFREE applies random masking to the patch tokens before feature extraction. Specifically, for each detection we randomly mask a subset of tokens with masking ratio ρ. Since cross-view reconstruction will later use synchronized detections from multiple cameras, the same mask pattern is applied across all views at the same time step. This preserves spatial correspondence across observations and ensures that the reconstruction objective focuses on differences induced by viewpoint rather than arbitrary masking inconsistency.

The visible tokens are then processed by a shared ViT-based encoder. Let the resulting patch-level feature for a detection be denoted by

$$f \in \mathbb{R}^{M^{\text{vis}} \times E},$$

where M^{vis} is the number of visible tokens and E is the embedding dimension. Instead of using this feature as a single mixed representation, CALIBFREE splits it channel-wise into two equally sized components:

$$f_a \in \mathbb{R}^{M^{\text{vis}} \times E/2}, \qquad f_s \in \mathbb{R}^{M^{\text{vis}} \times E/2}.$$

Here, f_a is the *view-agnostic* feature, intended to encode identity-preserving information that transfers across cameras, while f_s is the *view-specific* feature, intended to retain appearance details that are useful for temporal tracking within a single view.

This separation is motivated by the observation that cross-camera association and within-camera

tracking are related but not identical tasks. Cross-camera identity matching should suppress camera-dependent bias and emphasize stable identity cues, whereas temporal tracking within a camera can still benefit from local appearance information specific to that view. A single shared embedding must therefore balance two partially conflicting objectives. By contrast, CALIBFREE explicitly allocates representational capacity to each role.

In practice, the two branches are not assumed to be strictly disentangled in a strong theoretical sense. Rather, the goal is *functional specialization*: the view-agnostic branch should become more suitable for cross-view association, while the view-specific branch should become more effective for within-view temporal tracking. The following two subsections describe the two supervisory signals that drive this specialization.

8.4.3 Single-View Distillation

Single-view distillation is introduced to improve the stability of the learned detection representation within each camera. At a high level, this module trains a masked student encoder to recover the representation produced by a stronger unmasked teacher on the same detection crop. This encourages the student to preserve identity-relevant information even when only partial visual evidence is available.

Formally, let the teacher process the full detection crop and output a patch-level representation f^t, while the student processes the masked version and outputs f^s. The student is then trained to match the teacher through a distillation objective

$$\mathcal{L}_{\text{distill}}.$$

In practice, this loss is applied at the patch-feature level so that the student learns to reconstruct stable

local identity structure rather than only a global pooled descriptor.

This objective is especially useful for within-camera tracking. In a single view, detections may vary over time because of pose change, occlusion, motion blur, or partial visibility. By forcing the masked student to approximate the unmasked teacher, CALIBFREE encourages both f_a and f_s to remain informative under incomplete observations. As a result, the learned representation becomes more stable for temporal association within each camera.

However, single-view distillation alone is not sufficient for calibration-free MCMOT. It improves robustness within a view, but it does not explicitly encourage identity consistency across views. This motivates the second self-supervisory signal, cross-view reconstruction.

8.4.4 Cross-View Reconstruction

Cross-view reconstruction provides the main mechanism by which CALIBFREE learns cross-camera consistency without camera calibration or identity labels. The key observation is that synchronized detections from different cameras may contain complementary views of the same subject or scene region. Even without explicit correspondence labels, these co-occurring observations provide a meaningful supervisory signal for learning transferable identity features.

CALIBFREE uses the view-agnostic branch f_a for this purpose. For each synchronized time step, detection-level features from different cameras are first pooled and aggregated, and the resulting multi-view representation is passed through a cross-view encoder–decoder pathway. The decoder is trained to predict the masked patches of each target detection using information available from other views. The associated reconstruction objective is denoted by

$$\mathcal{L}_{\text{recon}}.$$

Unlike single-view masked modeling, this objective explicitly encourages the model to use information from other cameras. As a result, the learned view-agnostic representation is pushed to encode identity-relevant structure that remains informative under viewpoint change. This is precisely the property needed for cross-camera association in a calibration-free setting.

Another important advantage of this design is that it does not require geometry, epipolar constraints, or camera parameters. The only supervision comes from synchronized co-occurrence in RGB observations. In this sense, cross-view reconstruction is the core mechanism that enables CALIBFREE to learn cross-camera correspondence directly from raw multi-view video.

8.4.5 Feature Separation Regularization

Although single-view distillation and cross-view reconstruction provide complementary supervision, the two learned representations can still overlap substantially if no additional constraint is imposed. In particular, without regularization, the view-agnostic and view-specific branches may collapse into similar mixed features, weakening the intended specialization of the model.

To address this issue, CALIBFREE applies a separation loss

$$\mathcal{L}_{\text{sep}},$$

which encourages f_a and f_s to capture different aspects of the detection representation. The purpose of this term is not to claim strict disentanglement in a strong information-theoretic sense. Instead, the goal is functional specialization: f_a should become less sensitive to camera-dependent variation and more suitable for cross-view association, while f_s should retain appearance cues that remain useful for temporal association within each camera.

This regularization plays an important supporting role in the overall framework. Single-view distillation tends to preserve stable within-view information, while cross-view reconstruction encourages cross-camera consistency. The separation loss complements these two objectives by allocating representational capacity more explicitly across the two branches. Later experiments show that this leads to more meaningful specialization and improved tracking behavior across intra-camera and cross-camera settings.

8.4.6 Training Objective

The final training objective combines the three components described above:

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{sep}}. \tag{8.1}$$

These terms serve distinct but complementary roles. The distillation loss improves the stability and quality of masked detection features within each camera. The reconstruction loss aligns identity-relevant structure across synchronized views and provides the key self-supervisory signal for cross-camera consistency. The separation loss encourages the two feature branches to specialize rather than collapse into redundant representations.

Together, these objectives allow CALIBFREE to learn calibration-free identity representations from synchronized RGB detections alone. Importantly, no manual identity labels, camera calibration parameters, or explicit geometric constraints are required during training. The representation is instead shaped entirely by temporal consistency within a view, synchronized co-occurrence across views, and the pressure to specialize for different association roles.

8.4.7 Inference for MCMOT

At inference time, each detection is encoded once and produces both f_a and f_s. These two features are then used differently during tracking.

The view-specific feature f_s is used primarily for within-camera temporal association. Since this branch preserves appearance details that remain informative within a single view, it is more suitable for linking detections over time inside each camera. By contrast, the view-agnostic feature f_a is used primarily for cross-camera association, since it is encouraged during training to suppress camera-dependent bias and retain more stable identity-preserving cues across viewpoints.

This division of roles is central to CALIBFREE. Instead of forcing a single shared embedding to solve both within-view and cross-view association, the framework uses two specialized embeddings whose behavior is learned self-supervised during training. In this way, inference remains simple—each detection is encoded once—but the learned representation is better aligned with the distinct demands of MCMOT in dynamic camera networks.

8.5 Experimental Evaluation

In this section, we evaluate CALIBFREE on two challenging multi-camera benchmarks in order to answer three questions. First, can calibration-free identity representations learned from synchronized RGB observations support both within-camera and cross-camera association? Second, does the proposed specialization into view-agnostic and view-specific features improve tracking performance over strong baselines and simplified variants? Third, do the results support the broader claim of this chapter—namely, that task-relevant identity consistency can be learned and transferred across camera networks without camera calibration or manual identity annotations?

8.5.1 Datasets and Evaluation Protocol

We evaluate CALIBFREE on two challenging multi-camera benchmarks: MMP-MvMHAT and MvMHAT. These datasets serve complementary roles in this chapter. MMP-MvMHAT provides an indoor setting with relatively dense view overlap and substantial identity ambiguity caused by cluttered office-like layouts, while MvMHAT includes both indoor and outdoor scenes with greater variation in camera layout, overlap pattern, and appearance conditions. Together, they allow us to test whether CALIBFREE can remain effective across different levels of cross-view difficulty and deployment complexity [155, 157, 158].

Following prior work, we evaluate performance along three axes. The first is *over-time tracking*, which measures temporal identity consistency within each camera using metrics such as IDP, IDR, IDF1, MOTA, and HOTA. The second is *cross-view tracking*, which measures cross-camera association quality using AIDP, AIDR, AIDF1, and MHAA. The third is *overall performance*, summarized by combined accuracy and F1-style measures. This protocol is important because calibration-free MC-MOT must succeed not only in temporal tracking within each view, but also in maintaining identity consistency across views [155, 157].

From the perspective of this dissertation, this evaluation setup is especially meaningful because it tests whether transferable identity consistency can be learned across camera networks under changing observation conditions. In other words, the goal is not merely to track well within one camera, but to preserve task-relevant identity information when the view itself changes.

8.5.2 Main Results on MMP-MvMHAT

The main comparison on MMP-MvMHAT is shown in Table 8.1.

Methods	*Over-time Tracking*					*Cross-view Tracking*				*Overall*	
	IDP	**IDR**	**IDF1**	**MOTA**	**HOTA**	**AIDP**	**AIDR**	**AIDF1**	**MHAA**	***A***	***F***
Supervised											
Tracktor++[150]	67.0	56.0	61.0	67.2	46.2	62.0	23.2	33.8	19.1	43.1	47.4
CenterTrack[151]	35.9	24.1	28.8	48.2	27.1	29.3	3.9	6.8	3.1	25.7	17.8
TraDeS[152]	59.7	50.1	54.5	66.1	42.7	54.5	17.3	26.2	13.3	39.7	40.4
TrackFormer[153]	41.8	28.6	34.0	46.7	30.2	39.9	5.3	9.4	3.2	25.0	21.7
DeepCC[148]	51.6	52.5	52.1	92.5	59.3	42.7	23.4	30.2	19.7	56.1	41.2
SVT[180]	63.1	63.4	63.3	96.7	68.8	53.8	33.4	41.2	29.8	63.1	52.3
Self-supervised											
MvMHAT (YOLOX)[155]	51.1	53.2	52.7	82.1	47.2	30.4	17.1	23.2	14.1	48.1	37.9
MvMHAT (GT)[155]	58.6	58.8	58.7	93.7	65.0	35.4	21.2	26.5	20.3	57.0	42.6
MvMHAT++ (YOLOX)[157]	59.3	60.5	60.1	82.2	52.1	45.7	34.1	40.5	28.9	55.5	50.3
MvMHAT++ (GT)[157]	67.1	67.6	67.3	95.0	70.2	**62.1**	42.5	50.4	40.6	67.8	58.9
CALIBFREE (YOLOX)	81.3	77.1	79.1	82.5	59.2	52.3	48.4	50.3	34.4	58.4	64.7
CALIBFREE (GT)	**82.2**	**78.0**	**80.0**	**97.6**	**75.7**	55.0	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 8.1: **Results on MMP-MvMHAT.** CALIBFREE achieves the strongest overall performance and consistently improves identity tracking quality across both over-time and cross-view evaluation settings. All metrics are reported as percentages.

CALIBFREE achieves the strongest overall performance among both supervised and self-supervised baselines. In over-time tracking, it reaches the best IDF1, MOTA, and HOTA, showing that the learned features remain highly stable within each camera. In cross-view tracking, it also improves AIDR, AIDF1, and MHAA over prior methods, indicating stronger identity consistency across camera viewpoints.

A particularly important practical result is that CALIBFREE remains more robust to imperfect detection boxes than prior self-supervised approaches. This matters because detector noise is unavoidable in real deployments, and a useful calibration-free system must remain effective even when bounding boxes are not ideal. The gains on both temporal and cross-view metrics therefore suggest that the proposed representation specialization improves not only benchmark performance, but also robustness under realistic tracking conditions.

More broadly, these results support the central claim of this chapter. CALIBFREE is not simply learning a stronger generic embedding; rather, it is learning identity features that remain useful under changing camera viewpoints without relying on calibration. This directly supports the broader theme of Part III: perception becomes more useful when it preserves task-relevant consistency across

Methods	*Over-time Tracking*					*Cross-view Tracking*				*Overall*	
	IDP	**IDR**	**IDF1**	**MOTA**	**HOTA**	**AIDP**	**AIDR**	**AIDF1**	**MHAA**	***A***	***F***
Supervised											
Tracktor++[150]	54.2	40.1	46.1	66.5	42.8	34.3	14.6	20.5	37.1	51.8	33.3
CenterTrack[151]	44.3	33.5	38.1	63.5	37.8	29.7	9.1	13.9	34.1	48.8	26.0
TraDeS[152]	46.7	43.2	44.9	69.5	42.9	32.4	14.0	19.6	36.0	52.8	32.2
TrackFormer[153]	52.3	47.2	49.6	**70.4**	47.3	47.8	23.2	31.3	40.2	55.3	40.4
DeepCC[148]	44.7	44.2	44.4	63.9	41.1	57.9	34.8	43.4	43.8	53.9	43.9
SVT[180]	47.9	47.2	47.6	65.4	43.1	61.7	45.7	52.5	50.4	56.9	50.0
Self-supervised											
MvMHAT[155]	53.1	52.0	52.5	64.7	47.9	53.0	46.4	49.5	51.7	58.2	51.0
MvMHAT++[157]	58.5	57.4	57.9	66.3	51.8	**63.8**	56.0	**59.6**	**59.7**	**63.0**	**58.8**
CALIBFREE	**59.1**	**58.4**	**58.7**	60.4	**52.0**	58.9	**56.2**	57.4	57.0	58.4	58.1

Table 8.2: **Results on MvMHAT.** CALIBFREE achieves best or second-best performance across most key metrics, showing strong tracking robustness in a more diverse multi-camera benchmark. All metrics are reported as percentages.

observation regimes instead of remaining tied to a single view.

8.5.3 Results on MvMHAT

Table 8.2 reports results on the more diverse MvMHAT benchmark.

On this dataset, CALIBFREE again achieves best or second-best performance across most key metrics. In particular, it attains the strongest over-time identity stability and remains highly competitive for cross-view association. Although MvMHAT++ achieves slightly stronger precision-oriented cross-view metrics in some settings, CALIBFREE delivers a more unified annotation-free framework with a strong overall balance between within-view and cross-view performance.

This result is especially meaningful because MvMHAT includes more challenging deployment conditions, including sparser overlap and greater camera diversity. The fact that CALIBFREE remains competitive in this setting suggests that the learned specialization of view-agnostic and view-specific features is not tied to a narrowly controlled benchmark. Instead, it remains useful when the camera network becomes more heterogeneous and the assumptions needed by calibration-dependent methods become less reliable.

From a dissertation perspective, this point is important. If the gains were limited to the denser and more controlled benchmark, the chapter would mainly show that self-supervised calibration-free tracking is feasible in favorable settings. By contrast, the MvMHAT results suggest that transferable identity consistency can remain useful even when the camera network becomes more diverse, which better supports the broader argument of Part III.

8.5.4 Ablation Studies

We next analyze the main design components of CALIBFREE through three complementary ablation experiments. These experiments are designed to answer different questions. First, does the full framework outperform simpler variants that remove one or more training signals? Second, how sensitive is the learned representation to the detector used during pretraining and inference? Third, do the two learned feature branches indeed specialize into different functional roles, as intended by the design of CALIBFREE?

Table 8.3 evaluates the effectiveness of the main training components by comparing four model variants: a teacher-only feature baseline, a distillation-only variant, a reconstruction-only variant, and the full CALIBFREE model. This experiment is designed to test whether the three proposed ingredients—single-view distillation, cross-view reconstruction, and feature separation—play complementary roles or whether the final gains can already be obtained from a simpler training objective.

The teacher-only baseline already provides reasonably strong over-time tracking, which suggests that detection-level visual features contain useful identity information within each camera. Adding single-view distillation further improves temporal stability, indicating that masked student learning helps preserve identity-relevant cues under partial visibility and appearance variation. By contrast, the

Methods	*Over-time Tracking*					*Cross-view Tracking*				*Overall*	
	IDP	**IDR**	**IDF1**	**MOTA**	**HOTA**	**AIDP**	**AIDR**	**AIDF1**	**MHAA**	***A***	***F***
ViT-L (teacher)	82.1	77.8	79.9	97.5	75.6	47.9	44.4	46.1	36.7	67.1	63.0
Distillation only	78.4	67.8	72.7	97.2	68.4	47.5	40.1	43.5	34.4	65.8	58.1
Reconstruction only	81.2	77.1	79.1	97.5	75.2	52.3	48.4	50.3	41.8	69.7	64.7
CALIBFREE	**82.2**	**78.0**	**80.0**	**97.6**	**75.7**	**55.0**	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 8.3: **Ablation studies on CALIBFREE variants.** The full model achieves the best overall performance, showing that distillation, cross-view reconstruction, and feature specialization provide complementary benefits for calibration-free multi-camera tracking. All metrics are reported as percentages.

Pretraining Detector	**Inference Detector**	*Over-time Tracking*					*Cross-view Tracking*				*Overall*	
		IDP	**IDR**	**IDF1**	**MOTA**	**HOTA**	**AIDP**	**AIDR**	**AIDF1**	**MHAA**	***A***	***F***
YOLOv7	YOLOv7	59.9	57.5	58.6	44.9	47.7	47.2	44.3	45.7	16.4	30.7	52.2
Ground Truth	YOLOv7	59.8	57.4	58.5	44.9	47.7	49.3	46.1	47.6	18.3	31.6	53.1
YOLOX	YOLOX	81.3	77.1	79.1	82.5	59.2	52.3	48.4	50.3	34.4	58.4	64.7
Ground Truth	YOLOX	81.0	76.8	78.9	82.5	59.0	54.0	50.1	52.0	37.5	60.0	65.5
Ground Truth	**Ground Truth**	**82.2**	**78.0**	**80.0**	**97.6**	**75.7**	**55.0**	**50.9**	**52.8**	**44.2**	**70.8**	**66.4**

Table 8.4: **Ablation study on detector choice during pretraining and inference.** CALIBFREE remains relatively robust to the detector used during pretraining, while detector quality at inference has a larger effect on final tracking performance. All metrics are reported as percentages.

reconstruction-only variant is more beneficial for cross-view matching, since it explicitly encourages the model to use synchronized observations from other cameras to recover masked content. The full CALIBFREE model achieves the strongest overall balance, showing that temporal stabilization, cross-view alignment, and representation specialization are all needed for robust calibration-free MCMOT. This experiment therefore confirms that the gains do not come from a single stronger feature extractor, but from the interaction between the three proposed training signals.

Table 8.4 studies the effect of detector choice during pretraining and inference. In this experiment, different detectors are used to generate the person bounding boxes for representation learning and for the final tracking stage. The purpose is to evaluate how strongly CALIBFREE depends on the detector and whether the learned representation remains useful when the detector quality changes.

The results show that CALIBFREE is relatively robust to the detector used during pretraining, while detector quality at inference time has a larger effect on final tracking performance. This is

Feature	Camera ID Acc. (%) ↑	Intra-Cam IDF1 (%) ↑	Cross-Cam IDF1 (%) ↑
f_a (view-agnostic)	33.6	71.8	64.2
f_s (view-specific)	78.4	74.9	52.6
Merged ($f_a + f_s$)	61.3	73.8	58.6
$f_a \oplus f_s$	–	74.2	59.7

Table 8.5: **Camera-sensitivity probe.** The view-specific feature f_s is substantially more camera-dependent and performs better for intra-camera association, while the view-agnostic feature f_a suppresses camera cues and performs better for cross-camera matching. *Merged* uses a single shared embedding without feature separation, while $f_a \oplus f_s$ routes f_a and f_s to cross-camera and intra-camera association, respectively; camera-ID probing is therefore not applicable in that setting.

expected because poor detections at inference directly degrade both within-camera and cross-camera association, regardless of how good the representation is. At the same time, the relatively small variation across pretraining detectors suggests that the learned representation is not tightly overfit to one specific detector. This is an important practical property, since calibration-free tracking systems are likely to be deployed with imperfect or changing detectors in real-world camera networks.

Table 8.5 provides direct evidence for feature specialization through a camera probing experiment. In this experiment, we train a lightweight camera-ID classifier on top of each learned feature branch and measure how accurately the feature can predict from which camera a detection was captured. The purpose of this test is not to improve tracking directly, but to diagnose what kind of information each branch encodes. If a feature strongly preserves camera-dependent information, it should achieve high camera-ID accuracy. Conversely, if a feature suppresses camera-specific bias and becomes more view-invariant, its camera-ID accuracy should be lower.

The probing results show that the view-specific feature f_s achieves much higher camera-ID prediction accuracy, while the view-agnostic feature f_a is substantially less camera-sensitive. This is exactly the behavior intended by the design of CALIBFREE. The result indicates that f_s retains view-dependent appearance cues that are useful for within-camera temporal tracking, whereas f_a sup-

presses camera-specific information and is therefore better suited for cross-camera identity matching. In other words, the two branches do not simply duplicate one another; they learn measurably different functional roles. This experiment is especially important because it provides direct evidence that the proposed feature separation is not only conceptually motivated, but also empirically realized in the learned representation.

From the perspective of this dissertation, these ablation results are significant because they clarify what transferable identity consistency requires in dynamic camera networks. The results suggest that calibration-free MCMOT cannot be reduced to generic representation learning alone. Instead, successful transfer across cameras requires three ingredients: stable within-view learning, explicit cross-view alignment, and representation specialization that separates camera-dependent appearance cues from camera-robust identity information.

8.5.5 Qualitative Analysis

Figure 8.3 visualizes cross-view reconstruction in CALIBFREE.

Even when large image regions are masked, the model can recover appearance and identity-relevant structure using complementary observations from other views. This provides an intuitive illustration of why synchronized co-occurrence is a powerful supervisory signal for calibration-free MCMOT. Rather than relying on camera geometry, the model learns to use distributed visual evidence across views to preserve the information needed for identity association.

These visualizations are consistent with the quantitative results. They suggest that CALIBFREE is learning more than a benchmark-specific similarity function: it is learning how identity-relevant structure persists across views and can be reconstructed from complementary observations. This qual-

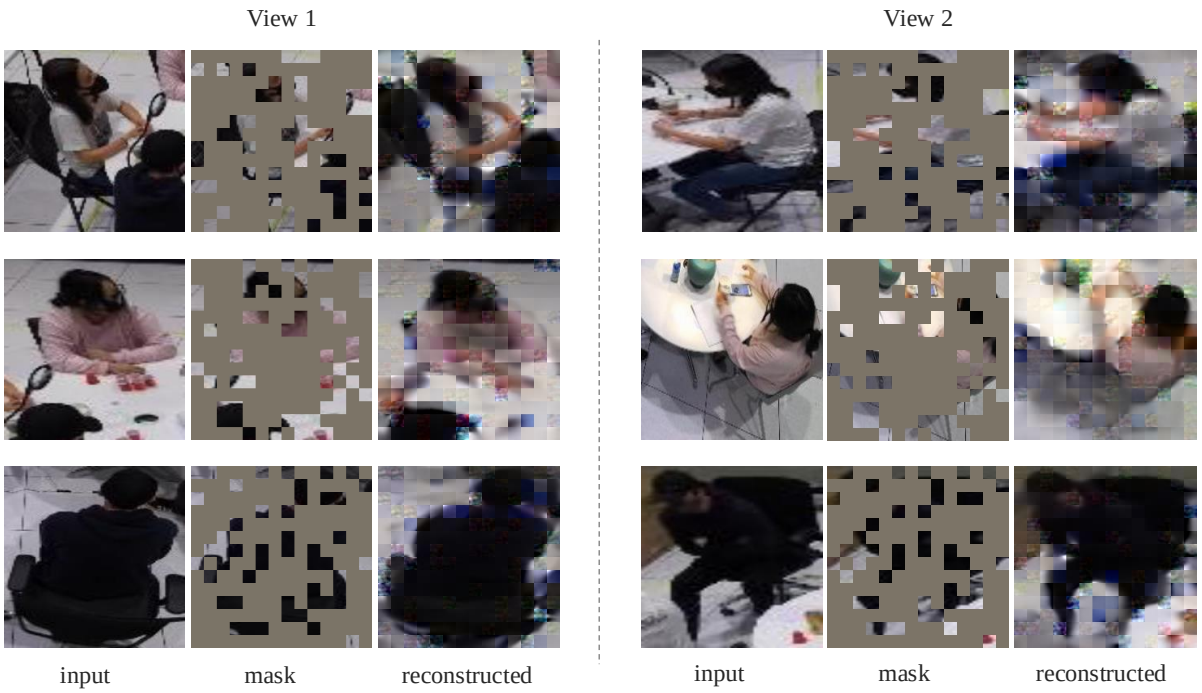


Figure 8.3: **Cross-view reconstruction results.** The input images show the original crops, while the masked images indicate regions removed for reconstruction. CALIBFREE reconstructs the masked regions by leveraging complementary observations across views, recovering consistent appearance and identity-relevant cues even when large portions are obscured.

itative behavior reinforces the main claim of the chapter that cross-camera identity consistency can be learned directly from synchronized RGB observations, without explicit calibration or annotation.

8.6 Discussion

CALIBFREE can be understood as a framework for learning transferable identity consistency across camera networks. Its main significance is therefore not only that it performs multi-camera tracking without calibration, but that it shows how task-relevant identity information can be preserved under changing viewpoints without relying on explicit geometry. In this setting, the challenge is not merely to build a stronger tracker, but to learn a representation that remains useful when the observation regime itself changes from one camera to another.

This perspective helps explain why the proposed design differs from conventional MCMOT formulations. Once calibration is removed, cross-camera association can no longer depend on camera parameters, epipolar constraints, or hand-designed transition structure. Instead, the model must learn correspondence directly from synchronized visual evidence. At the same time, within-camera temporal tracking and cross-camera identity matching do not place identical demands on the representation. The former benefits from retaining local appearance detail, while the latter requires suppressing camera-specific bias in favor of more stable identity-preserving cues. The central contribution of CALIBFREE is to address this tension explicitly through representation specialization.

The experimental results support this interpretation. The gains are not explained by a single stronger embedding or by a single auxiliary loss. Rather, they arise from the interaction between three ingredients: feature stabilization within a view, alignment across views, and separation between view-specific and view-agnostic information. In this sense, CALIBFREE suggests that calibration-free MCMOT is best understood not as a pure tracking problem, but as a representation transfer problem across camera networks.

From the perspective of Part III, this chapter complements Chapter 7 in an important way. AGL-Net showed that aerial observations can serve as transferable global references across sensing modalities for localization. CALIBFREE shows that identity representations can also remain transferable across camera views for tracking. Together, these chapters broaden the role of perception from understanding within a single sensing regime to preserving task-relevant consistency across changing observation conditions. This is the larger thesis-level significance of CALIBFREE.

More broadly, the results suggest that transferable perception in camera networks depends on both invariance and specialization. A representation that is too view-specific may fail to generalize across cameras, while one that is too aggressively view-invariant may discard information needed for

local temporal association. CALIBFREE indicates that these competing demands are better handled by explicitly allocating different representational roles rather than forcing a single embedding to solve both problems at once. This principle may be useful beyond MCMOT, including in multi-view activity understanding, distributed sensor fusion, and adaptive camera-network reasoning.

8.7 Limitations

Although CALIBFREE improves calibration-free MCMOT, it still has several limitations.

First, the current training formulation assumes at least some degree of synchronized cross-view overlap. Cross-view reconstruction relies on co-occurring observations from different cameras, and this signal weakens when overlap becomes extremely sparse or when synchronization quality deteriorates. As a result, the method may be less effective in camera networks where useful cross-view co-occurrence is rare.

Second, the framework still depends on the quality of the person detections used as input. While the detector ablation shows that CALIBFREE is reasonably robust to detector variation, poor detections at inference time can still degrade both within-camera and cross-camera association. In practice, this means that calibration-free identity learning alone cannot fully compensate for severe upstream detection errors.

Third, although the feature probing results support the intended specialization of f_a and f_s, the method does not guarantee strict disentanglement in a strong theoretical sense. The two branches are encouraged to play different functional roles, but they may still share overlapping information. For this reason, the claim of the chapter is best understood as one of *functional specialization* rather than exact disentanglement.

Fourth, the current formulation focuses on detection-level association in synchronized camera networks and does not explicitly model richer temporal context beyond the training objectives themselves. In more complex settings, longer-term memory, adaptive temporal reasoning, or explicit uncertainty modeling may further improve robustness, especially under prolonged occlusion or delayed cross-camera reappearance.

Finally, the empirical evaluation, although strong, remains limited to two multi-camera benchmarks. Broader validation across more diverse camera networks, more severe domain shifts, and different overlap regimes would be helpful for fully characterizing how well the learned identity representations generalize in evolving real-world deployments.

These limitations suggest that CALIBFREE should be viewed as a strong step toward transferable calibration-free MCMOT rather than a complete solution. Nevertheless, the results of this chapter indicate that self-supervised representation specialization is a promising direction for preserving identity consistency across camera networks without geometric assumptions.

8.8 Conclusion

This chapter presented CALIBFREE [115], a fully self-supervised calibration-free framework for multi-camera multi-object tracking. By combining single-view distillation, cross-view reconstruction, and feature-separation regularization, CALIBFREE learns complementary view-agnostic and view-specific embeddings that support both cross-camera and within-camera association without requiring camera calibration or manual identity annotations. Experiments on MMP-MvMHAT and MvMHAT demonstrated strong and robust tracking performance, and further analysis showed that the learned feature branches indeed specialize for different association roles.

The results support the main claim of this chapter: calibration-free MCMOT requires more than a single generic tracking representation. Once identity must remain consistent across camera networks, the model must simultaneously preserve temporal stability within a view and transfer identity cues across views. CALIBFREE addresses this challenge by learning specialized representations whose roles are aligned with these two distinct association problems.

From the perspective of this dissertation, CALIBFREE establishes the second part of the broader argument developed in Part III. AGL-Net showed that transferable correspondence can be learned across sensing modalities for localization. CALIBFREE shows that transferable identity consistency can also be learned across camera views for tracking. Together, these chapters demonstrate that effective perception depends not only on strong within-domain representations, but also on transferable representations that preserve task-relevant consistency when the observation condition changes.

More broadly, this chapter reinforces the central thesis theme that perception becomes substantially more powerful when it is designed to transfer across views, modalities, and sensing configurations rather than remaining confined to a single observation regime. In this sense, CALIBFREE extends the scope of transferable perception from spatial correspondence across modalities to identity correspondence across camera networks, and thereby completes the argument of Part III.

Chapter 9: Conclusion

9.1 Thesis Overview

This dissertation studied aerial perception as a spatial-temporal representation learning problem. The central premise was that effective aerial perception requires representations that are refined for aerial-specific visual understanding, practical under real-world deployment constraints, and transferable across heterogeneous sensing conditions. This perspective was motivated by the distinctive characteristics of aerial, overhead, and elevated-view observations, including weak foreground visibility, large background dominance, viewpoint and scale variation, camera motion, limited labeled data, and substantial shifts across sensing setups. Taken together, these factors make it difficult to directly apply standard perception pipelines developed for conventional ground-view settings.

To address these challenges, this dissertation developed three complementary themes. The first theme focused on refined spatial-temporal representation learning for aerial video understanding, showing that aerial action recognition benefits from stronger temporal alignment, more informative temporal evidence selection, and better semantic guidance. The second theme focused on practical and scalable aerial perception under real-world constraints, showing that aerial perception depends not only on representation quality, but also on efficient inference and supervision-efficient learning. The third theme focused on transferable aerial perception across views, modalities, and camera networks, showing that aerial perception becomes substantially more powerful when learned representations can

preserve task-relevant consistency across changing sensing conditions.

Across these themes, the dissertation advanced a unified view of aerial perception beyond isolated task-specific solutions. Rather than treating action recognition, self-supervised pretraining, localization, and multi-camera tracking as unrelated problems, this dissertation argued that they are connected by three recurring requirements: refinement, practicality, and transferability. Refinement is necessary because aerial observations are often weak, noisy, and dominated by irrelevant background. Practicality is necessary because aerial perception systems must operate under constraints on computation, resources, and annotation. Transferability is necessary because aerial perception increasingly involves heterogeneous sensing platforms, changing viewpoints, and distributed camera environments. In this sense, the dissertation contributes not only individual methods, but also a broader representation-centric perspective on what robust aerial perception should require.

9.2 Limitations of the Present Dissertation

Although this dissertation develops a unified perspective on aerial perception, several limitations remain. First, the proposed methods are still largely organized around specific task settings, including aerial action recognition, aerial-ground localization, and multi-camera tracking. While these tasks reveal common representation learning principles, the dissertation does not yet provide a single unified model that jointly addresses all of them within one aerial foundation framework.

Second, many of the proposed methods are evaluated in benchmark-driven settings with relatively well-defined task formulations. Although these benchmarks capture important aerial-specific challenges, they do not fully reflect the open-world, long-horizon, and continuously changing conditions faced by real deployed aerial systems. In particular, issues such as evolving semantic categories,

long-term environmental change, and large-scale multi-platform coordination remain only partially addressed.

Third, the dissertation primarily studies perception as an inference problem over given observations. While this is an important step, many real aerial systems require tighter coupling between perception and action, including viewpoint selection, motion planning, active sensing, and coordination across agents. These interactive aspects fall outside the main scope of the present dissertation.

Finally, although the dissertation studies transfer across views, modalities, and camera networks, transferability is still examined in relatively structured settings. A broader understanding of how aerial representations generalize across larger domain gaps, more diverse sensing configurations, and less controlled deployment environments remains an important open problem.

9.3 Broader Perspective and Future Directions

Beyond the individual tasks studied in this dissertation, a broader perspective emerges: aerial perception is fundamentally shaped by the tension between what is observable, what is practically learnable, and what remains transferable across changing sensing conditions. Unlike many conventional perception settings, aerial observations are rarely dense, stable, and semantically clean. Instead, they are often characterized by weak foreground cues, dynamic viewpoints, large scene coverage, sparse supervision, and substantial mismatch across platforms and sensing modalities. As a result, progress in aerial perception depends not only on improving benchmark performance for individual tasks, but also on developing representation learning principles that remain effective under these broader sources of variation.

One important future direction is the development of more unified aerial foundation models.

Much of the work in this dissertation focused on task-specific solutions for action recognition, localization, and tracking. While this was necessary to address the unique challenges of each setting, an important next step is to investigate whether these ideas can be integrated into a more general aerial representation learning framework. In particular, future models may need to jointly capture temporal dynamics, geometric structure, semantic salience, and cross-view consistency within a shared pretraining paradigm, rather than treating these components separately.

A second promising direction is open-world aerial perception. Many current benchmarks assume closed label spaces, fixed class definitions, and relatively constrained deployment settings. In practice, however, aerial perception systems must often reason under open vocabulary, evolving environments, and previously unseen events. Extending aerial perception from closed-set recognition toward open-world understanding will require models that can generalize beyond predefined training categories, adapt to new semantic concepts, and remain robust when the environment changes over time.

A third direction is the move from passive observation toward interactive and embodied aerial intelligence. Most problems studied in this dissertation assume that perception is applied to already captured observations. In real autonomous systems, however, perception and action are tightly coupled: where to look, how to move, which viewpoint to acquire, and when to coordinate with other sensing platforms all influence what can ultimately be understood. This suggests that future aerial perception research should increasingly study the interplay between representation learning, decision making, and active sensing, especially in long-horizon or multi-agent settings.

Finally, future work should further investigate multi-platform and cooperative perception. Several chapters in this dissertation already moved in this direction by studying aerial-ground correspondence and multi-camera identity consistency. A natural extension is to consider larger systems in which aerial platforms, ground robots, static cameras, and other sensing agents cooperate within a

shared perceptual framework. In such settings, the key challenge will be not only to transfer across modalities and viewpoints, but also to learn representations that support communication, coordination, and task-level consistency across distributed sensing systems.

In summary, this dissertation argued that aerial perception should be approached as a representation learning problem defined by refinement, practicality, and transferability. Looking forward, the field is likely to benefit most from methods that move beyond isolated task-specific solutions and instead support unified, adaptive, and collaborative perception across diverse aerial sensing environments. It is in this broader sense that aerial perception can progress from specialized recognition systems toward more general and robust visual intelligence.

Bibliography

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[2] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020.

[3] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1513–1522, 2021.

[4] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[5] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16266–16275, 2021.

[6] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1717–1726, 2020.

[7] Hazim Shakhatreh, Ahmad Sawalmeh, Ala Al-Fuqaha, Zuochao Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles: A survey on civil applications and key research challenges. *IEEE Access*, 7:48572–48634, 2019.

[8] Christoforos Kanellakis and George Nikolakopoulos. Survey on computer vision for uavs: Current developments and trends. *Journal of Intelligent & Robotic Systems*, 87(1):141–168, 2017.

[9] M. H. Rahman et al. A comprehensive survey of unmanned aerial vehicles (uavs). *Remote Sensing*, 16(5):879, 2024.

[10] Sami Hayat, Evsen Yanmaz, and Reza Muzaffar. Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint. *IEEE Communications Surveys & Tutorials*, 18(4):2624–2661, 2016.

[11] Yuncheng Lu, Zhi Xue, Gui-Song Xia, and Liangpei Zhang. A survey on vision-based uav navigation. *Geo-spatial Information Science*, 21(1):21–32, 2018.

[12] Zhen Zhang et al. A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications. *Drones*, 7(6):398, 2023.

[13] Abdulla Al-Kaff, David Martín, Fernando García, Arturo de la Escalera, and José María Armingol. Survey of computer vision algorithms and applications for unmanned aerial vehicles. *Expert Systems with Applications*, 92:447–463, 2018.

[14] A. V. R. Katkuri et al. Autonomous uav navigation using deep learning-based computer vision: A review. *Engineering Applications of Artificial Intelligence*, 132:107963, 2024.

[15] Jiaping Xiao, Rangya Zhang, Yuhang Zhang, and Mir Feroskhan. Vision-based learning for drones: A survey. *Proceedings of the IEEE*, 2025.

[16] S. Kapoor and A. Sharma. Diving deep into human action recognition in aerial videos. *Journal of Visual Communication and Image Representation*, 104:104298, 2024.

[17] Mohammadamin Barekatain, 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.

[18] 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.

[19] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[20] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[21] Elham Ravanbakhsh, Yongqing Liang, J. Ramanujam, and Xin Li. Deep video representation learning: A survey. *arXiv preprint arXiv:2405.06574*, 2024.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[23] 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.

[24] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021.

[25] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.

[26] 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.

[27] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[28] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[29] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[31] M. Y. Arafat and S. Moh. Vision-based navigation techniques for unmanned aerial vehicles: Review and challenges. *Drones*, 7(2):89, 2023.

[32] Y. Chang et al. A review of uav autonomous navigation in gps-denied environments. *Robotics and Autonomous Systems*, 168:104484, 2023.

[33] Asanka G Perera, Yee Wei Law, Titilayo T Ogunwa, and Javaan Chahl. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems*, 50(5):405–413, 2020.

[34] G. Tang et al. A survey of object detection for uavs based on deep learning. *Remote Sensing*, 16(1):149, 2023.

[35] D. Cazzato et al. A survey of computer vision methods for 2d object detection from unmanned aerial vehicle images. *Journal of Imaging*, 6(8):78, 2020.

[36] J. Leng et al. Recent advances for aerial object detection: A survey. *ACM Computing Surveys*, 2024.

[37] Xijun Wang, Ruiqi Xian, Tianrui Guan, Celso M. de Melo, Stephen M. Nogar, Aniket Bera, and Dinesh Manocha. Aztr: Aerial video action recognition with auto zoom and temporal reasoning. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1312–1318, 2023. URL `https://api.semanticscholar.org/CorpusID:257353660`.

[38] Ruiqi Xian, Xijun Wang, and Dinesh Manocha. Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6625–6634, 2024.

[39] Ruiqi Xian, Xijun Wang, Divya Kothandaraman, and Dinesh Manocha. Pmi sampler: Patch similarity guided frame selection for aerial action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6982–6991, January 2024.

[40] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. Adaframe: Adaptive frame selection for fast video recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1278–1287, 2019.

[41] Hanxiao Li et al. Vision-based learning for drones: A survey. *arXiv preprint arXiv:2312.05019*, 2024.

[42] H. Liu et al. A survey of sensors based autonomous unmanned aerial vehicle localization. *Complex & Intelligent Systems*, 2025.

[43] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021.

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[45] 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.

[46] 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.

[47] Hazar Mliki, Fatma Bouhlel, and Mohamed Hammami. Human activity recognition from uav-captured video sequences. *Pattern Recognition (PR)*, 100:107140, 2020.

[48] Divya Kothandaraman, Tianrui Guan, Xijun Wang, Sean Hu, Ming-Shun Lin, and Dinesh Manocha. Far: Fourier aerial video recognition. In *European Conference on Computer Vision*, 2022.

[49] 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, 2023. doi: 10.1109/ICRA48891.2023.10160271.

[50] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.

[51] Paul A. Viola and William M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:137–154, 1995.

[52] Frederik Maes, André M. F. Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16:187–198, 1997.

[53] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22:986–1004, 2003.

[54] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3192–3199, 2013.

[55] 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.

[56] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[58] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: Low-resolution video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7387–7394. IEEE, 2021.

[59] Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 126–133, 2020.

[60] 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 (ICCV)*, pages 6233–6243, 2019.

[61] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. Frameexit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15608–15618, 2021.

[62] Jintao Lin, Haodong Duan, Kai Chen, Dahua Lin, and Limin Wang. Ocsampler: Compressing videos to one clip with single-step sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13888–13897, 2022.

[63] Shreyank N. Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *AAAI*, 2021.

[64] Shweta Bhardwaj, Mukundhan Srinivasan, and Mitesh M Khapra. Efficient video classification using fewer frames. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 354–363, 2019.

[65] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

[66] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018.

[67] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[68] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[69] 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.

[70] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[71] Xijun Wang, Ruiqi Xian, Tianrui Guan, Fuxiao Liu, and Dinesh Manocha. Scp: Soft conditional prompt learning for aerial video action recognition. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10967–10974, 2024. doi: 10.1109/IROS58592.2024.10802074.

[72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021.

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[75] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[76] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[77] Xijun Wang, Ruiqi Xian, Tianrui Guan, and Dinesh Manocha. Prompt learning for action recognition. *arXiv preprint arXiv:2305.12437*, 2023.

[78] Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. *arXiv preprint arXiv:2303.15167*, 2023.

[79] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[80] Xingyi Zhou, Anurag Arnab, Chen Sun, and Cordelia Schmid. How can objects help action recognition? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2362, 2023.

[81] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. Motion saliency based multi-stream multiplier resnets for action recognition. *Image and Vision Computing*, 107: 104108, 2021.

[82] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[83] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[84] Seong Hyeon Park, Jihoon Tack, Byeongho Heo, Jung-Woo Ha, and Jinwoo Shin. K-centered patch sampling for efficient video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 160–176. Springer, 2022.

[85] Fan Yang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. A framework for knowing who is doing what in aerial surveillance videos. *IEEE Access*, 7:93315–93325, 2019.

[86] Abdullah M Algamdi, Victor Sanchez, and Chang-Tsun Li. Dronecaps: Recognition of human actions in drone videos using capsule networks with binary volume comparisons. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3174–3178. IEEE, 2020.

[87] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.

[88] Ping ZHang, Ping Wei, and SHuhuan Han. Capsnets algorithm. In *Journal of Physics: Conference Series*, volume 1544, page 012030. IOP Publishing, 2020.

[89] Santosh Kumar Yadav, Achleshwar Luthra, Esha Pahwa, Kamlesh Tiwari, Heena Rathore, Hari Mohan Pandey, and Peter Corcoran. Droneattention: Sparse weighted temporal attention for drone-camera based activity recognition. *Neural Networks*, 159:57–69, 2023.

[90] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020.

[91] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[92] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[93] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[94] Chengfei Lv, Chaoyue Niu, Renjie Gu, Xiaotang Jiang, Zhaode Wang, Bin Liu, Ziqi Wu, Qiulin Yao, Congyu Huang, Panos Huang, Tao Huang, Hui Shu, Jinde Song, Bin Zou, Peng Lan, Guohuan Xu, Fei Wu, Shaojie Tang, Fan Wu, and Guihai Chen. Walle: An End-to-End, General-Purpose, and Large-Scale production system for Device-Cloud collaborative machine learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 249–265, Carlsbad, CA, July 2022. USENIX Association. ISBN 978-1-939133-28-1. URL `https://www.usenix.org/conference/osdi22/presentation/lv`.

[95] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1502–1512, 2021.

[96] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pages 312–329. Springer, 2020.

[97] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023.

[98] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[99] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14733–14743, 2022.

[100] Rui Wang, Dongdong Chen, Yuheng Chen, Zhen Xiong, Zuxuan Wu, Alan Yuille, and Luowei Zhou. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[101] Gensheng Pei, Tao Chen, Xiruo Jiang, Yuan Li, Jilan Wang, and Limin Wang. Videomac: Video masked autoencoders meet convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[102] Ruiqi Xian, Xiyang Wu, Tianrui Guan, Xijun Wang, Boqing Gong, and Dinesh Manocha. Falcon: Future-aware learning with contextual object-centric pretraining for uav action recognition. *arXiv preprint arXiv:2409.18300*, 2024.

[103] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022.

[104] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3313–3322, 2022.

[105] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations (ICLR)*, 2022.

[106] Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[107] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[108] 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 (ICCV)*, 2019.

[109] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[110] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[111] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023.

[112] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems*, 36:40676–40693, 2023.

[113] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021.

[114] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021.

[115] Deep Anil Patel Sanjoy Kundu Martin Renqiang Min Ruiqi Xian, Iain Melvin and Dinesh Manocha. Calibfree: Self-supervised view feature separation for calibration-free multi-camera multi-object tracking. *arXiv preprint arXiv*, 2026.

[116] Tsung-Yu Lin, Yin Cui, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[117] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[118] Nam N. Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision (ECCV)*, 2016.

[119] Abhilash Durgam, Sidike Paheding, Vikas Dhiman, and Vijay Devabhaktuni. Cross-view geo-localization: A survey. *arXiv preprint arXiv:2406.09722*, 2024.

[120] Tim Y. Tang, Daniele De Martini, and Paul Newman. Get to the point: Learning lidar place recognition and metric localisation using overhead imagery. In *Robotics: Science and Systems (RSS)*, 2021.

[121] Huan Yin, Xuecheng Xu, Sha Lu, Xieyuanli Chen, Rong Xiong, Shaojie Shen, Cyrill Stachniss, and Yue Wang. A survey on global lidar localization: Challenges, advances and open problems. *International Journal of Computer Vision*, 2024.

[122] Ryan W. Wolcott and Ryan M. Eustice. Visual localization within lidar maps for automated urban driving. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[123] Weixin Lu, Yao Zhou, Guowei Wan, Shenhua Hou, and Shiyu Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[124] Sixing Hu, Mengdan Feng, Rang M. H. 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 (CVPR)*, 2018.

[125] Tianrui Guan, Aswath Muthuselvam, Montana Hoover, Xijun Wang, Jing Liang, Adarsh Jagan Sathyamoorthy, Damon Conover, and Dinesh Manocha. Crossloc3d: Aerial-ground cross-source 3d place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[126] Xipeng Wang, Steve Vozar, and Edwin Olson. Flag: Feature-based localization between air and ground. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[127] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[128] Tianrui Guan, Ruiqi Xian, Xijun Wang, Xiyang Wu, Mohamed Elnoor, Daeun Song, and Dinesh Manocha. Agl-net: Aerial-ground cross-modal global localization with varying scales. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8161–8161. IEEE, 2024.

[129] Tim Y. Tang, Daniele De Martini, and Paul Newman. Point-based metric and topological localisation between lidar and overhead imagery. *Autonomous Robots*, 2023.

[130] Ivan Moskalenko, Anastasiia Kornilova, and Gonzalo Ferrer. Visual place recognition for aerial imagery: A survey. *arXiv preprint arXiv:2406.00885*, 2024.

[131] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[132] Dong Yuan, Frederic Maire, and Feras Dayoub. Cross-attention between satellite and ground views for enhanced fine-grained robot geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[133] Royston Rodrigues and Masahiro Tani. Global assists local: Effective aerial representations for field of view constrained image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[134] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[135] Fabian Deuser, Christopher Hummel, Timo Luther, and Rainer Stiefelhagen. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[136] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[137] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[138] Zimin Xia and Alexandre Alahi. Fg2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[139] Younghun Cho, Giseop Kim, Sangmin Lee, and Jee-Hwan Ryu. Openstreetmap-based lidar global localization in urban environment without a prior lidar map. *IEEE Robotics and Automation Letters*, 7(2):4999–5006, 2022.

[140] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[141] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[142] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019.

[143] 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 (CVPR)*, pages 770–778, 2016.

[144] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[145] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, pages 1–16, 2017.

[146] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[147] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3559–3570, 2023.

[148] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6036–6046, 2018.

[149] Taiwo Ibrahim Amosa, Adebayo Abayomi-Alli, Sanjay Misra, Robertas Damasevicius, and Rytis Maskeliunas. Multi-camera multi-object tracking: A review of current trends and future advances. *Neurocomputing*, 565:126972, 2023.

[150] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951, 2019.

[151] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision (ECCV)*, pages 474–490, 2020.

[152] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12352–12361, 2021.

[153] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8844–8854, 2022.

[154] Longyin Wen, Weiyao Li, Junjie Yan, Zhen Lei, Dong Yi, and Stan Z. Li. Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision*, 122(2):313–333, 2017.

[155] Yu Gan, Ruize Han, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pages 3852–3860, 2021.

[156] Li Fei, Yajie Liu, Yexin Li, Weigang Tian, Yong Xu, and Min Wang. Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review. *Sensors*, 23(8):3852, 2023.

[157] Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13847–13863, 2023.

[158] Wei Feng, Ruize Han, Yu Gan, and Song Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. *arXiv preprint arXiv:2401.17617*, 2024.

[159] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

[160] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017.

[161] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3038–3046, 2017.

[162] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jiaqi Zheng, Wenqiang Xiong, Qi Tian, and Xinggang Wang. Transtrack: Multiple-object tracking with transformer. In *arXiv preprint arXiv:2012.15460*, 2020.

[163] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fuchen Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–21, 2022.

[164] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 17–35, 2016.

[165] Francesco Solera, Simone Calderara, Ergys Ristani, Carlo Tomasi, and Rita Cucchiara. Tracking social groups within and across cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2309–2321, 2017.

[166] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Dukemtmc: A benchmark for tracking multiple people in multi-camera video. In *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016.

[167] Mengran Gou, Zhe Wu, Angel Tong, Yuhang Liu, Dong Yi, Yadong Ma, Zhen Li, Qing Liao, Zhen Wang, Bing Dai, et al. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10–17, 2017.

[168] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5039, 2018.

[169] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8797–8806, 2019.

[170] Yue He, Judy Hoffmann, Trevor Darrell, and Kate Saenko. Multi-camera multi-target tracking with space-time-appearance graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[171] Ahmet Işçen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Multi-camera multi-object tracking with global graph modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[172] Chia-Chi Cheng, Yi-Hsuan Wu, Winston Hsu, and Yuheng Tseng. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21449–21459, 2023.

[173] Phuoc Nguyen, Trung Le, Thanh-Toan Nguyen, Minh Tran, Trung Bui, and Chang D. Yoo. Multi-camera multi-object tracking on the move via single-stage global association approaches. *Pattern Recognition*, 154:110594, 2024.

[174] Xiaoqian Zhou, Wanli Ouyang, Ping Luo, and Xiaogang Liu. Cross-camera trajectory prediction and association for multi-object tracking in complex camera networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[175] Yen-Liang Lin, Chih-Chung Hsu, and Hwann-Tzong Wang. Cross-camera tracklet association by constrained metric learning and graph optimization. *IEEE Transactions on Image Processing*, 28(12):5831–5844, 2019.

[176] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 737–753, 2018.

[177] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1770–1782, 2021.

[178] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[179] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa Narasimhan. Self-supervised multi-view person association and its applications. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[180] Junting Dong, Qi Fang, Wen Biao Jiang, Yurou Yang, Qi-Xing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6981–6992, 2021. URL `https://api.semanticscholar.org/CorpusID:236159249`.