Beyond the Horizon: Decoupling Multi-View UAV Action Recognition via Partial Order Transfer

Wenxuan Liu[1,2], Zhuo Zhou[3,#], Xuemei Jia[3], Siyuan Yang[4], Wenxin Huang[5], Xian Zhong[2,*], and Chia-Wen Lin[6]

[1] State Key Laboratory for Multimedia Information Processing, Peking University
[2] Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology
[3] National Engineering Research Center for Multimedia Software, Wuhan University
[4] College of Computing and Data Science, Nanyang Technological University
[5] Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University
[6] Department of Electrical Engineering, National Tsing Hua University
liuwx66@pku.edu.cn, {2023102110024, jiaxuemeiL}@whu.edu.cn, siyuan.yang@ntu.edu.sg, wenxinhuang_wh@163.com, zhongx@whut.edu.cn, cwlin@ee.nthu.edu.tw

Abstract

Action recognition using uncrewed aerial vehicles (UAVs) faces unique challenges due to substantial view variations along the vertical spatial axis. Unlike ground-based scenarios, UAVs capture actions from diverse altitudes, resulting in pronounced appearance discrepancies and reduced recognition robustness. To address this, we introduce a multi-view formulation tailored for UAV altitudes and empirically uncover a distinctive *partial order* among views, where recognition accuracy consistently declines as altitude increases. This key observation motivates the proposed **Aero Partial Order Guided Network (Aerorder)**, which explicitly models and exploits the hierarchical structure of UAV views to enhance cross-altitude action recognition. Aerorder comprises three main components: (1) a *View Partition (VP)* module that groups views by altitude using the head-to-body ratio; (2) an *Order-aware Feature Decoupling (OFD)* module that disentangles action-relevant and view-specific representations under partial order guidance; and (3) an *Action Partial Order Guide (APOG)* that progressively transfers knowledge from easier (low-altitude) to harder (high-altitude) views. Extensive experiments on DRONE-ACTION, MOD20, and UAV validate the superiority of Aerorder, achieving consistent improvements over state-of-the-art methods, up to 4.7% and 1.3% gains on DRONE-ACTION and MOD20, respectively.

Code — https://github.com/lwxfight/Aerorder

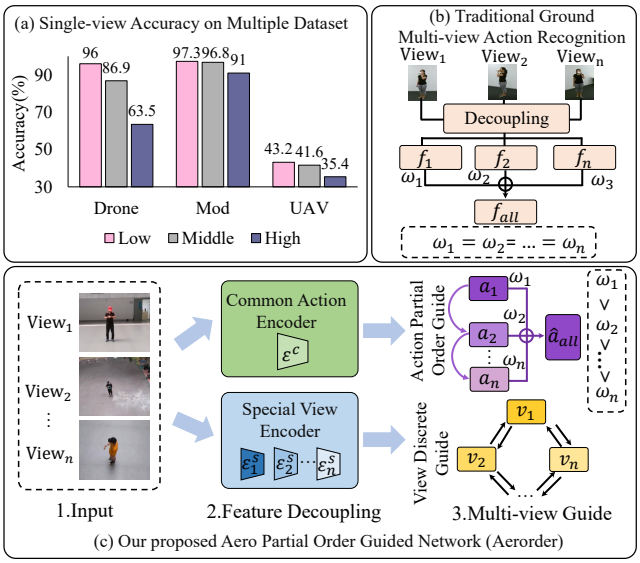

Figure 1: (a) Recognition accuracy per view on DRONE-ACTION, MOD20, and UAV, showing clear discrepancies across views. (b) Conventional ground-based multi-view recognition pipeline. (c) Our OFD module with feature decoupling and dual guidance (action and view).

Introduction

Human action recognition has been extensively studied (Feichtenhofer et al. 2019; Feichtenhofer 2020; Zhou et al. 2023; Yang et al. 2024), primarily focusing on videos captured by ground cameras (Liu et al. 2025a). With the rapid advancement of uncrewed aerial vehicles (UAVs), aerial action recognition has gained increasing attention, enabling diverse applications in surveillance, inspection, disaster response, intelligent transportation, and network security (Hong et al. 2025). Early studies (Perera, Law, and Chahl 2018, 2019; Li et al. 2021a) explored pose estimation (Guan and Zhao 2025) and attention mechanisms (Cheng et al. 2020) to handle view changes, yet they overlooked a crucial factor: UAVs operate in open aerial spaces with high mobility. Their fast motion leads to drastic shifts in viewing angle and distance, causing large appearance variations. Ignoring these variations often *biases models toward low-altitude views*, severely degrading recognition performance at higher altitudes.

Building upon this observation, we revisit view variation in UAV action recognition and uncover a systematic pattern along the vertical axis, in contrast to the random view changes

[#]Co-first author.
[*]Corresponding author.

typical of ground-based setups (Liu et al. 2023). As altitude increases, actions become more ambiguous and difficult to recognize due to severe visual distortions. To quantify altitude, we employ the head-to-body ratio as an interpretable metric. As illustrated in Fig. 1(a), recognition accuracy consistently decreases with higher ratios, revealing a distinctive *partial order* among UAV views.

Motivated by this finding, we propose the **Aero Partial Order Guided Network (Aerorder)**, a framework that explicitly models and exploits this partial order to enhance UAV action recognition. Our work addresses two key questions:

Q1: How can we isolate action-relevant features from view-induced variations under the partial order? Conventional multi-view action recognition approaches (Ullah et al. 2021; Xu et al. 2021; Zhong et al. 2022; Liu et al. 2023) generally assume all views contribute equally and rely on uniform feature fusion, as depicted in Fig. 1(b). This neglects the inherent hierarchy among views. In contrast, our *Order-aware Feature Decoupling (OFD)* module explicitly separates action-relevant and view-specific components. Through a shared encoder and multiple view branches, OFD disentangles these representations, while a generative constraint enforces clear separation and robustness, inspired by recent progress in direction-aware attention and mutual representation learning (Jiang et al. 2023, 2025). The observed partial order further guides feature decoupling and integration, ensuring structured learning across altitude levels.

Q2: How can we exploit the partial order to guide adaptive knowledge transfer from easier to harder views? Based on the disentangled features from OFD, we design an *Action Partial Order Guide (APOG)* to enable progressive knowledge transfer aligned with the partial order. APOG constructs a graph over view features to discretize the continuous view space and facilitates hierarchical adaptation from low- to high-altitude views. This process enhances feature alignment, reduces residual coupling between action and view, and ensures transfer follows the intrinsic UAV view hierarchy. By integrating OFD and APOG, Aerorder effectively leverages altitude-aware structure, leading to superior robustness under diverse viewing conditions.

In summary, our contributions are threefold:

- We introduce a multi-view formulation for UAV action recognition, employing the head-to-body ratio for altitude-based view partitioning and establishing a new recognition perspective.
- We reveal a *partial order* among UAV views and propose the *Order-aware Feature Decoupling (OFD)* module to disentangle view-specific and action-relevant features.
- We develop the *Aero Partial Order Guided Network (Aerorder)*, which leverages the partial order to guide adaptive learning and integration across multi-view UAV settings, achieving substantial gains over state-of-the-art methods.

Related Work

Multi-View Action Recognition

UAV View. Early studies (Perera, Law, and Chahl 2018, 2019; Li et al. 2021a) employed pose estimation (Cheng et al. 2020) to extract local action features from UAV videos. However, conventional feature extractors often struggle with small targets and wide viewing ranges, yielding noisy or incomplete representations. Subsequent works (Jin et al. 2022; Kothandaraman et al. 2022) emphasized temporal modeling to better capture motion dynamics, while recent research (Xian, Wang, and Manocha 2024; Xian et al. 2024) explored feature-based sampling strategies. More advanced designs (Peng et al. 2025) introduced transformer-based token compression to improve both efficiency and representational quality, whereas recent advances in spatio-temporal context memory networks further strengthened long-term temporal reasoning for UAV video understanding (Dang et al. 2024). Despite these developments, most approaches treat UAV views independently and overlook the structured relationships among varying altitudes.

Ground Multi-View. Cross-view action recognition aims to learn view-invariant representations (Ullah et al. 2021; Xu et al. 2021; Liu et al. 2023; Yang et al. 2023; You et al. 2024). Virtual camera simulation (Rahmani, Mian, and Shah 2018; Xiao et al. 2019) enables multi-view modeling but is annotation-intensive and computationally expensive. Alternative strategies, such as feature clustering (Shao, Li, and Zhang 2021; Ullah et al. 2021) or feature distillation (Vyas, Rawat, and Shah 2020; Zhong et al. 2022; Liu et al. 2023; Siddiqui, Tirupattur, and Shah 2024), enhance robustness but still neglect inter-view dependencies. Approaches like CVAM, VCD, and DRDN improve feature disentanglement yet assume uniform view contribution. In contrast, our method explicitly leverages the **partial order** among UAV views to guide discriminative feature decoupling and hierarchical learning, forming the central principle behind *Aerorder*.

Sample Classification

Instance differentiation has been explored through confidence-based learning (Han et al. 2018; Wan et al. 2022) and meta-learning (Li et al. 2021b; Wang et al. 2023; Zhong et al. 2023). Some methods use prediction variance or decision-boundary proximity to assess instance difficulty, while others encode ordinal relations via label recoding for classification and ranking (Niu et al. 2016; Wang et al. 2023). Distinct from these paradigms, we propose a data-driven partitioning strategy that utilizes the intrinsic structure of UAV multi-view data. By grouping samples based on the head-to-body ratio as an altitude proxy, we enhance the exploitation of view-specific cues for robust action recognition.

Transfer Learning

Our objective focuses on transferring knowledge across altitude levels rather than between networks. Conventional transfer learning techniques, such as knowledge distillation (Yu et al. 2020), pre-trained initialization (Bolya, Mittapalli, and Hoffman 2021), and rehearsal-based adaptation (Ji et al. 2023; Sun et al. 2023; Tian et al. 2023), target general classification tasks and disregard structural view variations. In contrast, we exploit the *partial order* among UAV views and partition instances by head-to-body ratio to enable structured,

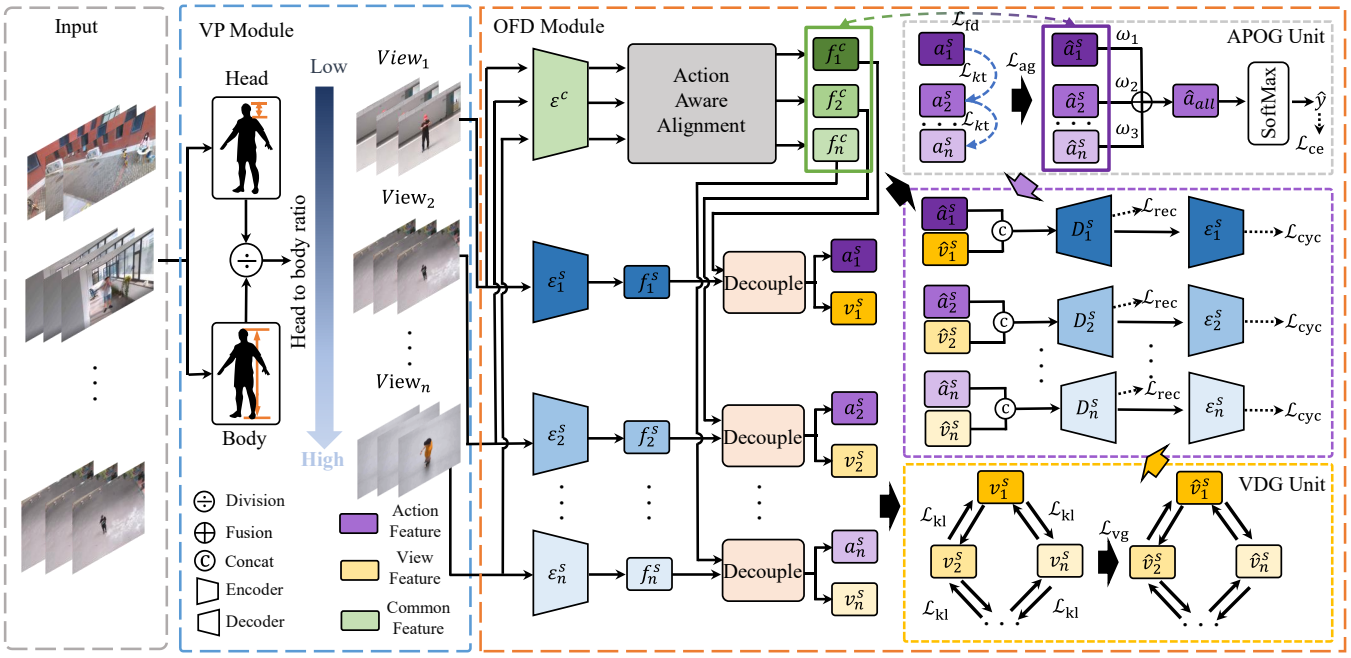

Figure 2: Framework of the Proposed Aerorder. The VP module divides UAV samples into altitude-based views using the head-to-body ratio. The OFD module disentangles and transfers action features from low- to high-altitude views via the partial order relation and view-specific weights. The APOG unit learns the partial order to guide feature transfer, and the VDG unit imposes discretization constraints on view features to decouple multi-view representations.

altitude-aware knowledge transfer. Moreover, inspired by external memory mechanisms for retrieval-augmented long-term video understanding (Dang et al. 2025), our framework equips *Aerorder* with enhanced adaptability and consistent performance across diverse viewing conditions.

Proposed Method

We propose the *Aero Partial Order Guided Network (Aerorder)*, illustrated in Fig. 2. Aerorder consists of two primary components: the *View Partition (VP)* module and the *Order-aware Feature Decoupling (OFD)* module. The VP module partitions UAV samples into low-, mid-, and high-altitude views, while the OFD module decouples multi-view features and learns discriminative action representations under the guidance of the *partial order* among views.

Notation. We adopt a classification backbone (*e.g.*, X3D (Feichtenhofer 2020)) as the feature extractor. Let ϵ_i^c denote the common extractor mapping the i-th view input X_i to its common feature:

$$f_i^c = \epsilon_i^c\left(X_i\right). \tag{1}$$

View-specific extractors ϵ_i^s further obtain:

$$f_i^s = \epsilon_i^s\left(X_i\right). \tag{2}$$

The VP module computes the head-to-body ratio H for each sample to guide view partitioning. The OFD module then decouples each f_i^s into an action feature a_i^s and a view feature v_i^s, refines a_i^s according to the partial order to obtain $\hat{a}_i^s$, and enhances v_i^s to produce $\hat{v}_i^s$. The refined action features are fused into a global feature $\hat{a}_{\text{all}}$ for classification.

View Partition Module

To model multi-view relations, we introduce the *View Partition (VP)* module, which groups UAV samples into low-, mid-, and high-altitude views based on the head-to-body ratio. Specifically, we employ YOLOv8[1] to detect the actor's head and body bounding boxes and compute:

$$H = \frac{B_{\text{head}}}{B_{\text{body}}}, \tag{3}$$

where B_{head} and B_{body} represent the detected bounding-box heights of the head and body, respectively. A smaller H implies a lower altitude (wider field of view) and thus finer action detail. All samples are sorted by H in ascending order and divided into n equal groups, each assigned a view index v_i:

$$v_i = \begin{cases} 0, & 0 < \frac{i}{m} \le \frac{1}{n}, \\ 1, & \frac{1}{n} < \frac{i}{m} \le \frac{2}{n}, \\ \vdots & \\ n-1, & \frac{n-1}{n} < \frac{i}{m} \le 1, \end{cases} \tag{4}$$

where v_i denotes the view index of the i-th sample, m is the total number of samples, and n is the number of altitude-based partitions. This partitioning enables altitude-aware analysis and establishes a structured *partial order* for downstream modules, facilitating a clearer understanding of UAV view hierarchy in vertical spatial contexts.

[1]https://github.com/ultralytics/ultralytics

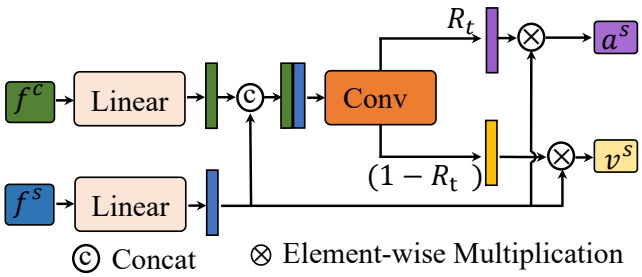

Figure 3: UAV Multi-View Feature Decoupling Unit. Given a common feature f^c and view-specific feature f^s, we compute the action correlation map R_t and decouple f^s into action feature a^s and view feature v^s.

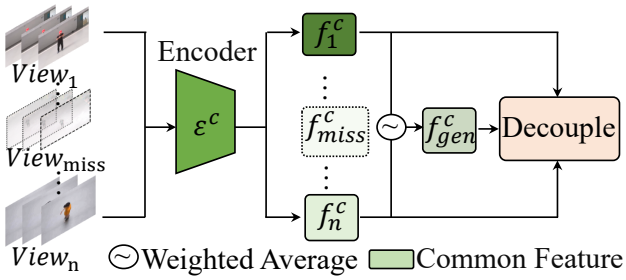

Figure 4: Action-Aware Alignment Unit. Missing common features f^c_{miss} are compensated by averaging common features from other views.

Order-Aware Feature Decoupling Module

UAV Multi-View Feature Decoupling. As shown in Fig. 3, given the common feature f_i^c and the view-specific feature f_i^s, we align their dimensions through a linear projection of f_i^c and concatenate:

$$R_t = \sigma\left(W_r\left[f_i^c, f_i^s\right]\right), \tag{5}$$

where $[\cdot,\cdot]$ denotes concatenation, W_r consists of two 1×1 convolutional layers with batch normalization (BN) and rectified linear unit (ReLU), and σ represents the sigmoid activation. The resulting correlation map R_t is used to decouple the view-specific feature:

$$a_i^s = f_i^s \otimes R_t, \tag{6}$$
$$v_i^s = f_i^s \otimes (1 - R_t), \tag{7}$$

where $\otimes$ denotes element-wise multiplication, yielding the action feature a_i^s and view feature v_i^s for each view i.

Action-Aware Alignment. Due to unrestricted UAV perspectives, some views may miss certain action representations (see Fig. 4). To preserve visual consistency, when the j-th view X_j lacks a valid action feature, we synthesize its common feature by averaging available ones:

$$f_j^c = \frac{1}{n-1}\sum_{i\neq j} f_i^c, \tag{8}$$

where f_j^c is the generated feature f^c_{gen} for the missing view.

Action Partial Order Guide. After view partitioning, we compute a credibility score for each view to quantify its reliability within the *partial order*:

$$c_i = \frac{n}{D}, \tag{9}$$
$$D = \sum_{k=1}^{n}(d_k + 1) = \sum_{k=1}^{n}\alpha_k, \tag{10}$$

n is the number of views. For view k, d_k is the ReLU-activated prediction score that quantifies reliability. c_i is normalized by the total Dirichlet strength D and scaled by α_k, which is computed from d_k. These confidence weights $\{c_i\}_{i=1}^n$ indicate each view's reliability. 1 in the Dirichlet prior ensures all class probabilities remain positive, preventing 0 probabilities. Let $A = \{a_1, a_2, \ldots, a_n\}$ denote the action features, and select a source feature $a_s \in A$ from an easier (low-altitude) view. Knowledge is transferred to target features $\{a_i\}$ via:

$$\mathcal{L}_{\text{kt}} = \frac{1}{n}\sum_{i=1}^{n} c_i \|a_s - a_i\|_F^2. \tag{11}$$

This confidence-weighted transfer enables low-altitude views to guide learning for higher-altitude ones. The refined features $\hat{A} = \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n\}$ are then used to optimize the common feature extractor through:

$$\mathcal{L}_{\text{fd}} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{a}_i - f_i^c\|_F^2. \tag{12}$$

The overall guidance loss combines both objectives:

$$\mathcal{L}_{\text{ag}} = \mathcal{L}_{\text{kt}} + \mathcal{L}_{\text{fd}}. \tag{13}$$

View Discrete Guide. To adaptively discretize the view space, we construct a directed graph where each node represents a view v_i, and the edge weights $w(v_i, v_j)$ encode discretization strength:

$$w\left(v_i, v_j\right) = \|v_i - v_j\|_F^2. \tag{14}$$

These distances form a discretization matrix E, and a learnable weight matrix W is initialized from E and refined through iterative graph updates. Each row of W is normalized using a softmax to mitigate scale variation. The graph discretization loss is defined as:

$$\mathcal{L}_{\text{vg}} = \|W \otimes E\|, \tag{15}$$

which encourages the learned edge weights to preserve the original inter-view relations encoded in E. The resulting graph models adaptive view interactions, forming the basis for altitude-aware refinement in *Aerorder*. Finally, each view feature v_i^s is updated into its discretized version $\hat{v}_i^s$ for subsequent fusion.

Reconstruction and Cycle-Consistency. To ensure distinct action and view representations, we reconstruct the input from refined features:

$$\mathcal{L}_{\text{rec}} = \|X_i - D_i\left([\hat{a}_i^s, \hat{v}_i^s]\right)\|_F^2, \tag{16}$$

where D_i is a lightweight deconvolution, and enforce consistency between the reconstructed and original features:

$$\mathcal{L}_{\text{cyc}} = \|f_i^s - \epsilon_i^s\left(D_i\left([\hat{a}_i^s, \hat{v}_i^s]\right)\right)\|_F^2. \tag{17}$$

It enforces faithful recovery of both input and view-specific features, alleviating overdominance of low-altitude features and maintaining the integrity of the *partial order* structure.

Training and Inference

After all modules, we aggregate the refined action features into a global representation:

$$\hat{a}_{\text{all}} = \sum_{i=1}^{n} c_i \hat{a}_i, \tag{18}$$

which is used to predict the action category via the cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\sum \hat{y} \log\left(\hat{a}_{\text{all}}\right), \tag{19}$$

where $\hat{y}$ denotes the one-hot ground truth label.

The complete training objective integrates three components: *(1) Classification:* $\mathcal{L}_{\text{ce}}$ for the common feature extractor. *(2) Decoupling:* $\mathcal{L}_{\text{dn}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}}$ to enforce the separation between action and view features. *(3) Guidance:* $\mathcal{L}_{\text{gn}} = \mathcal{L}_{\text{kt}} + \mathcal{L}_{\text{fd}} + \mathcal{L}_{\text{vg}}$ for partial-order-guided feature transfer and graph-based discretization.

The total objective is defined as:

$$\mathcal{L}_{\text{all}} = \gamma_{\text{ce}}\mathcal{L}_{\text{ce}} + \gamma_{\text{dn}}\mathcal{L}_{\text{dn}} + \gamma_{\text{gn}}\mathcal{L}_{\text{gn}}, \tag{20}$$

where γ_{ce}, γ_{dn}, and γ_{gn} control the relative importance of each term.

During inference, we compute $\hat{a}_{\text{all}}$ as above and determine the final action label via:

$$\hat{y}_{\text{pred}} = \arg\max\left(\hat{a}_{\text{all}}\right). \tag{21}$$

Experimental Results

Datasets and Implementation Details

High-Altitude UAV View. DRONE-ACTION (Perera, Law, and Chahl 2019) contains 240 high-altitude UAV videos of 13 outdoor actions, recorded at 1920×1080 resolution and 25 FPS. MOD20 (Perera et al. 2020) offers multi-view recordings of outdoor actions across 2,324 videos at 720×720 resolution and 29.97 FPS, combining aerial and ground footage. UAV (Li et al. 2021a) includes 67,428 sequences covering 155 actions from diverse aerial and vertical views.

Low-Altitude Ground View. NTU-RGB+D (Shahroudy et al. 2016) contains 56,880 samples across 60 action classes captured from ground-level cameras and serves as a standard benchmark for multi-view ground action recognition.

Settings and Metrics. We evaluate under two settings: *(1) UAV View*, using only UAV datasets to assess multi-altitude representations; and *(2) Cooperative UAV-Ground Views*, augmenting UAV data with NTU-RGB+D for cross-view guidance. We follow standard evaluation protocols and report Top-1 accuracy.

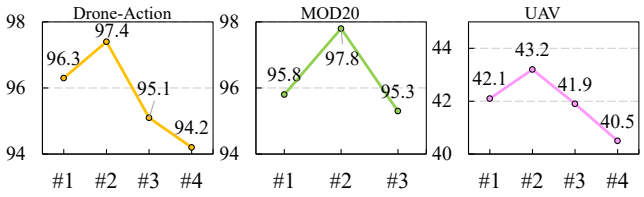

Figure 5: Cross-View Performance Evaluation on DRONE-ACTION, MOD20, and UAV. Models are trained on selected views (indexed by i) and tested on the remaining unseen ones. "#" stands for the view. MOD20 officially defines four views; hence, experiments with a fifth view were not conducted.

Implementation Details. We adopt X3D (Feichtenhofer 2020) as the backbone for its efficiency in both multi-view (Liu et al. 2023) and UAV-based (Kothandaraman et al. 2022) tasks. A 2D temporal convolution (kernel size 3) follows each 3D spatial convolution, with BN and ReLU activation. We apply a dropout rate of 0.9 before the final FC layer. Training uses Adam (Kingma and Ba 2015) with an initial learning rate of 1e-3, decayed by 1e-5, for 300 epochs on four NVIDIA Tesla V100 GPUs (16 GB each).

Comparison with State-of-the-Art Methods

Table 1 compares *Aerorder* with several baselines (Fayyaz et al. 2021; Cheng et al. 2022; Lin et al. 2022) on DRONE-ACTION, MOD20, and UAV. Aerorder consistently surpasses UniFormerV2 (Li et al. 2023), achieving accuracy gains of 16.5%, 1.6%, and 10.8% on DRONE-ACTION, MOD20, and UAV, respectively, demonstrating its adaptability to multi-view and altitude variations. Compared with ground-based multi-view methods, Aerorder outperforms DRDN by 4.7%, 1.5%, and 3.0%, and DVANet by 5.3%, 1.7%, and 3.7% across the three datasets, showing its strength in modeling structured view relations for more effective feature fusion. Moreover, Aerorder achieves this with only 15.2M parameters, significantly fewer than DRDN (35.8 M) and ASAT (61.8 M), while maintaining top-tier accuracy, highlighting its efficiency for real-world UAV deployment.

View Partition Analysis. We analyze the effect of varying the number of view partitions (2-5) across different datasets, as illustrated in Fig. 5. Using three views yields the best performance: employing only two views restricts cross-view information exchange and hampers the model's ability to bridge altitude gaps, whereas incorporating four or five views introduces redundant partitions that dilute feature transfer and reduce representation distinctiveness. Lower-numbered views (*e.g.*, view 1 and view 2) are typically closer to the ground, capturing richer and more discriminative cues for action understanding. In contrast, higher-numbered views (*e.g.*, view 4 and view 5) are more visually ambiguous due to increased background clutter and occlusion, thereby weakening their capacity to represent target actions.

Partial Order Relation Analysis. Fig. 6 presents recognition accuracy by view on DRONE-ACTION and UAV, without using view labels. With two views, accuracies are 96.2%

Type	Method	Backbone	Multi-view	DRONE-ACTION	MOD20	UAV	Params (M)
Vanilla	SlowFast (Feichtenhofer et al. 2019) †	ResNet-50	○	82.6	93.1	30.1	33.7
	X3D (Feichtenhofer 2020) †	ResNet-50	○	83.4	95.7	32.3	**3.6**
	3DResNet + ATFR (Fayyaz et al. 2021) †	X3D	○	79.5	94.2	28.2	21.1
	TSM (Lin et al. 2022) †	ResNet-50	○	75.4	<u>96.5</u>	24.3	24.3
	SBP (Cheng et al. 2022) †	Video Swin-T	○	81.2	96.3	29.1	<u>8.6</u>
	UniFormerV2 (Li et al. 2023) †	CLIP-ViT	○	81.9	93.5	32.4	354.0
	AIM (Yang et al. 2023) †	ViT-B	○	84.7	96.2	24.2	100.0
Multi-View	DRDN (Liu et al. 2023) †	SlowFast	●	<u>94.4</u>	96.3	38.5	35.8
	DVANet (Siddiqui et al. 2024) †	3D-CNN	●	93.1	96.1	37.9	45.3
UAVs	FAR (Kothandaraman et al. 2022)	X3D	○	92.7	-	38.6	14.4
	ASAT (Shi et al. 2023)	ResNet-50	○	-	- / <u>98.2</u>∗	39.7	61.8
	DAIL (Liu et al. 2024)	SlowFast	●	81.2	-	**45.1**	31.7
	StaRNet (Liu et al. 2025b)	SlowFast	○	85.4	-	40.8	33.7
	Aerorder (Ours)	X3D	●	**97.4**	**97.8 / 98.6**∗	<u>43.2</u>	15.2

Table 1: Comparison of Top-1 Accuracy (%) and Parameter Count (M) with State-of-the-Art Methods on DRONE-ACTION, MOD20, and UAV. Best and second-best results are in bold and underlined, respectively. † indicates reproduced results; ∗ denotes results on MOD20 subset.

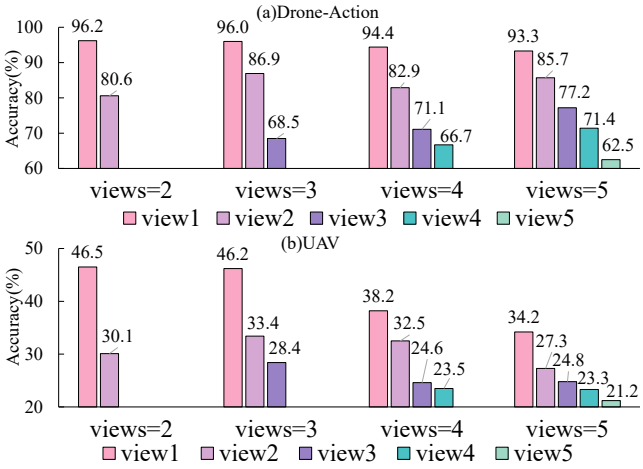

Figure 6: Per-View Performance across Different Partition Settings on DRONE-ACTION and UAV. "Views" denotes the training number of partitions.

Views	Guide Strategy	Acc
2	None	90.2
	$v_1 \rightarrow v_2$	94.2 (↑ 4.0)
	$v_2 \rightarrow v_1$	87.5 (↓ 2.7)
	NTU-RGB+D → DRONE-ACTION	95.8 (↑ 5.6)
	DRONE-ACTION → NTU-RGB+D	88.9 (↓ 1.3)
3	None	91.7
	$v_1 \rightarrow v_2$	95.8 (↑ 4.1)
	$v_1 \rightarrow v_3$	93.1 (↑ 1.4)
	$v_2 \rightarrow v_3$	94.5 (↑ 2.8)
	$v_1 \rightarrow v_2$ & $v_2 \rightarrow v_3$	**97.4** (↑ 5.7)
	$v_2 \rightarrow v_1$	90.2 (↓ 1.5)
	$v_3 \rightarrow v_1$	87.5 (↓ 4.2)
	$v_3 \rightarrow v_2$	90.2 (↓ 1.5)
	$v_3 \rightarrow v_2$ & $v_2 \rightarrow v_1$	88.9 (↓ 2.8)

Table 2: Top-1 Accuracy (%) under Different Guide Strategies in Aerorder on DRONE-ACTION and NTU-RGB+D. v_i denotes the i-th view.

(view 1) and 80.6% (view 2). Increasing to five views reduces performance to 93.3% (view 1) and 62.5% (view 5). This widening gap, *e.g.*, 68.5% for view 3 under three partitions, reveals a clear *partial order* along altitude, where lower views dominate and higher ones degrade. These findings underscore the importance of Aerorder's view-aware weighting for effectively balancing unequal contributions across views.

Guide Strategy Analysis. Table 2 summarizes multi-view transfer experiments for two- and three-view configurations. For two views, transferring $v_1 \rightarrow v_2$ yields a 4.0% improvement, while $v_2 \rightarrow v_1$ degrades accuracy by 2.7%, suggesting that top-down transfer introduces noise. Similarly, cross-dataset transfer from NTU-RGB+D (ground) to UAV improves by 5.6%, whereas the reverse direction decreases by 1.3%, confirming that low-altitude knowledge is more transferable. With three views, sequential transfer $v_1 \rightarrow v_2 \rightarrow v_3$ achieves the best accuracy (97.4%), aligning with the *partial order* hierarchy. All reversed transfers reduce performance (*e.g.*, $v_2 \rightarrow v_1$ by 1.5%), validating Aerorder's bottom-up guidance strategy.

Ablation Studies

Table 3 summarizes the module-wise ablation results. Both the APOG and VDG units outperform the X3D baseline (Feichtenhofer 2020), with APOG achieving the most significant improvement by facilitating efficient feature transfer from low- to high-altitude views. When combined, the two modules yield the highest overall accuracy, confirming their complementary effects and validating the design of the *Aerorder*.

Baseline	VDG	APOG	DRONE	MOD20	UAV
●	○	○	83.4	85.2	32.3
●	●	○	91.7	90.1	38.4
●	○	●	93.1	92.5	39.6
●	●	●	**97.4**	**97.8**	**43.2**

Table 3: Top-1 Accuracy (%) of Aerorder Component Variants on DRONE-ACTION, MOD20, and UAV.

Loss Weight		DRONE	MOD20	UAV
$\underline{\gamma_{ce}}, \gamma_{dn}, \gamma_{gn}$ ($\gamma_{dn}=1, \gamma_{gn}=1$)	$\gamma_{ce}=0.01$	90.2	91.2	34.5
	$\gamma_{ce}=0.1$	91.7	93.2	36.3
	$\gamma_{ce}=1$	95.8	97.2	42.5
	$\gamma_{ce}=10$	94.4	94.7	39.1
$\gamma_{ce}, \underline{\gamma_{dn}}, \gamma_{gn}$ ($\gamma_{ce}=1, \gamma_{gn}=1$)	$\gamma_{dn}=0.01$	91.7	91.4	35.1
	$\gamma_{dn}=0.1$	93.1	94.0	38.9
	$\gamma_{dn}=1$	95.8	97.2	42.5
	$\gamma_{dn}=10$	93.1	95.3	39.2
$\gamma_{ce}, \gamma_{dn}, \underline{\gamma_{gn}}$ ($\gamma_{ce}=1, \gamma_{dn}=1$)	$\gamma_{gn}=0.01$	94.4	95.1	41.2
	$\gamma_{gn}=0.1$	**97.4**	**97.8**	**43.2**
	$\gamma_{gn}=1$	95.8	97.2	42.5
	$\gamma_{gn}=10$	91.7	92.6	36.9
$\underline{\gamma_{ce}}, \gamma_{dn}, \gamma_{gn}$ ($\gamma_{dn}=1, \gamma_{gn}=0.1$)	$\gamma_{ce}=0.01$	91.7	92.3	36.5
	$\gamma_{ce}=0.1$	93.1	94.6	39.1
	$\gamma_{ce}=1$	**97.4**	**97.8**	**43.2**
	$\gamma_{ce}=10$	94.4	95.7	39.8
$\gamma_{ce}, \underline{\gamma_{dn}}, \gamma_{gn}$ ($\gamma_{ce}=1, \gamma_{gn}=0.1$)	$\gamma_{dn}=0.01$	93.1	92.5	37.2
	$\gamma_{dn}=0.1$	94.4	95.1	39.5
	$\gamma_{dn}=1$	**97.4**	**97.8**	**43.2**
	$\gamma_{dn}=10$	93.1	95.9	40.1

Table 4: Top-1 Accuracy (%) on DRONE-ACTION, MOD20, and UAV for Different Loss Weights.

Loss Parameter Analysis. Table 4 provides a detailed loss-weight analysis for *Aerorder* on DRONE-ACTION, MOD20, and UAV. The total objective $\mathcal{L}_{all}$ combines classification ($\mathcal{L}_{ce}$), decoupling ($\mathcal{L}_{dn}$), and guidance ($\mathcal{L}_{gn}$) losses. Following a Gibbs-sampling-inspired iterative scheme (Harrison et al. 2020), we vary each weight independently while fixing the others to determine optimal configurations. The best performance is obtained with $\gamma_{ce}=1$, $\gamma_{dn}=1$, and $\gamma_{gn}=0.1$, achieving 95.8%, 97.2%, and 42.5% on DRONE-ACTION, MOD20, and UAV, respectively. Decreasing γ_{dn} or γ_{gn} causes noticeable degradation, whereas moderate increases improve robustness, especially on UAV. These findings emphasize that carefully balancing the three loss terms is crucial for stable optimization and achieving robust multi-view recognition under the *partial-order* paradigm.

Visualization. Fig. 7 visualizes attention maps on UAV for the action "Look at the watch". In the low-altitude view (view 1), attention focuses sharply on the hand, the critical region for this action, capturing fine-grained motion cues due to proximity. Contrastly, the high-altitude view (view 3) exhibits a broader, more diffuse attention map caused by a wider field of view and lower spatial resolution, obscuring

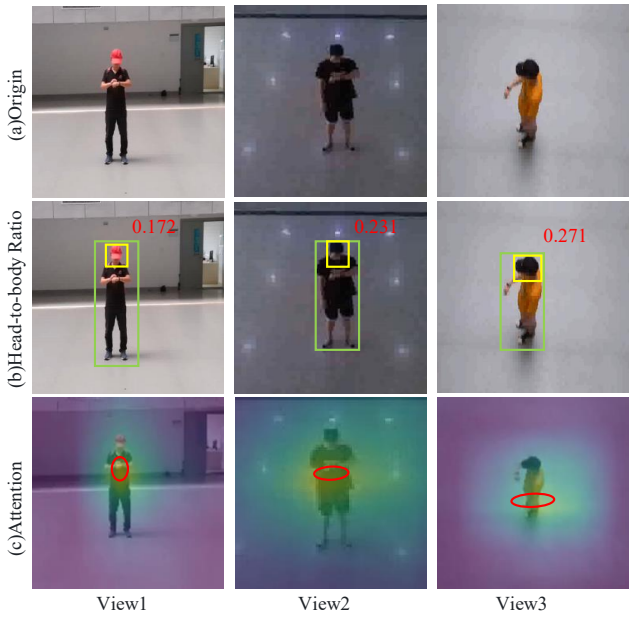

Figure 7: Visualization of Head-to-Body Ratio and Attention on UAV. (b) Green and yellow boxes denote body and head detections, respectively; red numbers represent head-to-body ratios (smaller values indicate lower-altitude views). (c) Red circles mark action attention centers.

detailed motion features. This comparison highlights the importance of low-altitude views in capturing precise action semantics and underscores how Aerorder effectively leverages altitude-dependent information through *partial-order-guided* learning.

Conclusion

In this paper, we uncover a previously overlooked *partial-order* structure among UAV views, revealing that action recognition accuracy systematically declines with increasing altitude due to amplified visual ambiguity and diminished motion cues. To address this issue, we propose the *Aero Partial Order Guided Network (Aerorder)*, which explicitly models the hierarchical relations across UAV altitudes to enhance cross-view action recognition. *Aerorder* integrates three key modules: a *View Partition (VP)* module that segments views based on the head-to-body ratio, an *Order-aware Feature Decoupling (OFD)* module that separates action-relevant and view-specific features under partial-order constraints, and an *Action Partial Order Guide (APOG)* unit that enables progressive, bottom-up knowledge transfer from easier (low-altitude) to more challenging (high-altitude) views. Extensive experiments on multiple UAV and ground multi-view benchmarks demonstrate that *Aerorder* significantly outperforms both single-view and multi-view baselines, validating the effectiveness of incorporating partial-order guidance for robust cross-view UAV action recognition.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants No. 62506011, 62271361, and 62301213), the Hubei Provincial Key Research and Development Program (Grant No. 2024BAB039), and the Postdoctoral Fellowship Program and the China Postdoctoral Science Foundation (Grant No. GZB20250388).

References

Bolya, D.; Mittapalli, R.; and Hoffman, J. 2021. Scalable Diverse Model Selection for Accessible Transfer Learning. In *Adv. Neural inform. Process. Syst.*, 19301–19312.

Cheng, F.; Xu, M.; Xiong, Y.; Chen, H.; Li, X.; Li, W.; and Xia, W. 2022. Stochastic Backpropagation: A Memory Efficient Strategy for Training Video Models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 8291–8300.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 180–189.

Dang, J.; Zheng, H.; Wu, X.; Jiao, J.; Wang, B.; Yang, J.; Hu, B.; Lai, J.; and Chua, T. S. 2025. External Memory Matters: Generalizable Object-Action Memory for Retrieval-Augmented Long-Term Video Understanding. In *Proc. Int. Joint Conf. Artif. Intell.*, 864–872.

Dang, J.; Zheng, H.; Xu, X.; Wang, L.; and Guo, Y. 2024. Beyond Appearance: Multi-Frame Spatio-Temporal Context Memory Networks for Efficient and Robust Video Object Segmentation. *IEEE Trans. Image Process.*, 33: 4853–4866.

Fayyaz, M.; Rad, E. B.; Diba, A.; Noroozi, M.; Adeli, E.; Van Gool, L.; and Gall, J. 2021. 3D CNNs with Adaptive Temporal Feature Resolutions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 4731–4740.

Feichtenhofer, C. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 200–210.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 6201–6210.

Guan, B.; and Zhao, J. 2025. Affine Correspondences between Multi-Camera Systems for Relative Pose Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–18.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *Adv. Neural inform. Process. Syst.*, 8536–8546.

Harrison, P. M. C.; Marjieh, R.; Adolfi, F.; van Rijn, P.; Anglada-Tort, M.; Tchernichovski, O.; Larrouy-Maestri, P.; and Jacoby, N. 2020. Gibbs Sampling with People. In *Adv. Neural inform. Process. Syst.*

Hong, S.; Yue, T.; You, Y.; Lv, Z.; Tang, X.; Hu, J.; and Yin, H. 2025. A Resilience Recovery Method for Complex Traffic Network Security Based on Trend Forecasting. *Int. J. Intell. Syst.*, 40(2): 1–13.

Ji, Z.; Hou, Z.; Liu, X.; Pang, Y.; and Li, X. 2023. Memorizing Complementation Network for Few-Shot Class-Incremental Learning. *IEEE Trans. Image Process.*, 32: 937–948.

Jiang, K.; Jiang, J.; Wang, Z.; Geng, Z.; and Liu, X. 2025. DAWN: Wavelet-Based Image Deraining Meets Direction-Aware Attention and Mutual Representation. *IEEE Trans. Neural Netw. Learn. Syst.*, 36(10): 18244–18258.

Jiang, K.; Liu, W.; Wang, Z.; Zhong, X.; Jiang, J.; and Lin, C.-W. 2023. DAWN: Direction-aware Attention Wavelet Network for Image Deraining. In *Proc. ACM Multimedia*, 7065–7074.

Jin, P.; Mou, L.; Hua, Y.; Xia, G.; and Zhu, X. X. 2022. FuTH-Net: Fusing Temporal Relations and Holistic Features for Aerial Video Classification. *IEEE Trans. Geosci. Remote Sens.*, 60.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proc. Int. Conf. Learn. Represent.*

Kothandaraman, D.; Guan, T.; Wang, X.; Hu, S.; Lin, M. C.; and Manocha, D. 2022. FAR: Fourier Aerial Video Recognition. In *Proc. Eur. Conf. Comput. Vis.*, 657–676.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*

Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; and Li, Z. 2021a. UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 16266–16275.

Li, W.; Huang, X.; Lu, J.; Feng, J.; and Zhou, J. 2021b. Learning Probabilistic Ordinal Embeddings for Uncertainty-Aware Regression. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 13896–13905.

Lin, J.; Gan, C.; Wang, K.; and Han, S. 2022. TSM: Temporal Shift Module for Efficient and Scalable Video Understanding on Edge Devices. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5): 2760–2774.

Liu, W.; Deng, Y.; Chen, K.; Zhong, X.; Yu, Z.; and Huang, T. 2025a. SOTA: Spike-Navigated Optimal TrAnsport Saliency Region Detection in Composite-Bias Videos. In *Proc. Int. Joint Conf. Artif. Intell.*

Liu, W.; Zhong, X.; Dai, Y.; Jia, X.; Wang, Z.; and Satoh, S. 2025b. Motion-Consistent Representation Learning for UAV-Based Action Recognition. *IEEE Trans. Intell. Transp. Syst.*

Liu, W.; Zhong, X.; Xu, X.; Zhou, Z.; Jiang, K.; Wang, Z.; and Bai, X. 2024. Discriminative Information Incremental Learning for Air-Ground Multi-View Action Recognition. *J. Image Graph.*

Liu, W.; Zhong, X.; Zhou, Z.; Jiang, K.; Wang, Z.; and Lin, C. 2023. Dual-Recommendation Disentanglement Network for View Fuzz in Action Recognition. *IEEE Trans. Image Process.*, 32: 2719–2733.

Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; and Hua, G. 2016. Ordinal Regression with Multiple Output CNN for Age Estimation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 4920–4928.

Peng, L.; Shu, X.; Yao, Y.; and Xie, G. 2025. 3D-Aware Select, Expand, and Squeeze Token for Aerial Action Recognition. In *Proc. AAAI Conf. Artif. Intell.*, 6479–6487.

Perera, A. G.; Law, Y. W.; and Chahl, J. 2019. Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition. *Drones*, 3(4): 82.

Perera, A. G.; Law, Y. W.; and Chahl, J. S. 2018. UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition. In *Proc. Eur. Conf. Comput. Vis. Worksh.*, 117–128.

Perera, A. G.; Law, Y. W.; Ogunwa, T. T.; and Chahl, J. S. 2020. A Multiviewpoint Outdoor Dataset for Human Action Recognition. *IEEE Trans. Hum. Mach. Syst.*, 50(5): 405–413.

Rahmani, H.; Mian, A. S.; and Shah, M. 2018. Learning a Deep Model for Human Action Recognition from Novel Viewpoints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3): 667–681.

Shahroudy, A.; Liu, J.; Ng, T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 1010–1019.

Shao, Z.; Li, Y.; and Zhang, H. 2021. Learning Representations From Skeletal Self-Similarities for Cross-View Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 31(1): 160–174.

Shi, G.; Fu, X.; Cao, C.; and Zha, Z. 2023. Alleviating Spatial Misalignment and Motion Interference for UAV-Based Video Recognition. In *Proc. ACM Multimedia*.

Siddiqui, N.; Tirupattur, P.; and Shah, M. 2024. DVANet: Disentangling View and Action Features for Multi-View Action Recognition. In *Proc. AAAI Conf. Artif. Intell.*, 4873–4881.

Sun, W.; Li, Q.; Zhang, J.; Wang, W.; and Geng, Y. 2023. Decoupling Learning and Remembering: A Bilevel Memory Framework with Knowledge Projection for Task-Incremental Learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 20186–20195.

Tian, C.; Zhang, X.; Liang, X.; Li, B.; Sun, Y.; and Zhang, S. 2023. Knowledge Distillation with Fast CNN for License Plate Detection. *IEEE Trans. Intell. Veh.*

Ullah, A.; Muhammad, K.; Hussain, T.; and Baik, S. W. 2021. Conflux LSTMs Network: A Novel Approach for Multi-View Action Recognition. *Neurocomputing*, 435: 321–329.

Vyas, S.; Rawat, Y. S.; and Shah, M. 2020. Multi-View Action Recognition Using Cross-View Video Prediction. In *Proc. Eur. Conf. Comput. Vis.*, 427–444.

Wan, Z.; Xu, X.; Wang, Z.; Yamasaki, T.; Zhang, X.; and Hu, R. 2022. Efficient Virtual Data Search for Annotation-Free Vehicle Reidentification. *Int. J. Intell. Syst.*, 37(5): 2988–3005.

Wang, C.; Jiang, Z.; Yin, Y.; Cheng, Z.; Ge, S.; and Gu, Q. 2023. Controlling Class Layout for Deep Ordinal Classification via Constrained Proxies Learning. In *Proc. AAAI Conf. Artif. Intell.*, 2483–2491.

Xian, R.; Wang, X.; Kothandaraman, D.; and Manocha, D. 2024. PMI Sampler: Patch Similarity Guided Frame Selection for Aerial Action Recognition. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 6967–6976.

Xian, R.; Wang, X.; and Manocha, D. 2024. MITFAS: Mutual Information Based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 6611–6620.

Xiao, Y.; Chen, J.; Wang, Y.; Cao, Z.; Zhou, J. T.; and Bai, X. 2019. Action Recognition for Depth Video Using Multi-View Dynamic Images. *inform. Sci.*, 480: 287–304.

Xu, C.; Wu, X.; Li, Y.; Jin, Y.; Wang, M.; and Liu, Y. 2021. Cross-modality online distillation for multi-view action recognition. *Neurocomputing*, 456: 384–393.

Yang, S.; Liu, J.; Lu, S.; Er, M. H.; Hu, Y.; and Kot, A. C. 2024. Self-Supervised 3D Action Representation Learning with Skeleton Cloud Colorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1): 509–524.

Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; and Li, M. 2023. AIM: Adapting Image Models for Efficient Video Action Recognition. In *Proc. Int. Conf. Learn. Represent.*

You, H.; Zhong, X.; Liu, W.; Wei, Q.; Huang, W.; Yu, Z.; and Huang, T. 2024. Converting Artificial Neural Networks to Ultra-Low-Latency Spiking Neural Networks for Action Recognition. *IEEE Trans. Cogn. Dev. Syst.*

Yu, L.; Twardowski, B.; Liu, X.; Herranz, L.; Wang, K.; Cheng, Y.; Jui, S.; and van de Weijer, J. 2020. Semantic Drift Compensation for Class-Incremental Learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 6980–6989.

Zhong, X.; Gu, C.; Ye, M.; Huang, W.; and Lin, C. 2023. Graph Complemented Latent Representation for Few-Shot Image Classification. *IEEE Trans. Multimedia*, 25: 1979–1990.

Zhong, X.; Zhou, Z.; Liu, W.; Jiang, K.; Jia, X.; Huang, W.; and Wang, Z. 2022. VCD: View-Constraint Disentanglement for Action Recognition. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2170–2174.

Zhou, Z.; Liu, W.; Xu, D.; Wang, Z.; and Zhao, J. 2023. Uncovering the Unseen: Discover Hidden Intentions by Micro-Behavior Graph Reasoning. In *Proc. ACM Multimedia*, 6623–6633.