Alleviating Spatial Misalignment and Motion Interference for UAV-based Video Recognition

Gege Shi
University of Science and Technology of China
Hefei, Anhui, China
shigg@mail.ustc.edu.cn

Xueyang Fu*
University of Science and Technology of China
Hefei, Anhui, China
xyfu@ustc.edu.cn

Chengzhi Cao
University of Science and Technology of China
Hefei, Anhui, China
chengzhicao@mail.ustc.edu.cn

Zheng-Jun Zha
University of Science and Technology of China
Hefei, Anhui, China
zhazj@ustc.edu.cn

ABSTRACT

Recognizing activities with Unmanned Aerial Vehicles (UAVs) is essential for many applications, while existing video recognition methods are mainly designed for ground cameras and do not account for UAV changing attitudes and fast motion. This creates spatial misalignment of small objects between frames, leading to inaccurate visual movement in drone videos. Additionally, camera motion relative to objects in the video causes relative movements that visually affect object motion and can result in misunderstandings of video content. To address these issues, we present a novel framework named Attentional Spatial and Adaptive Temporal Relations Modeling. First, to mitigate the spatial misalignment of small objects between frames, we design an Attentional Patch-level Spatial Enrichment (APSE) module that models dependencies among patches and enhances patch-level features. Then, we propose a Multi-scale Temporal and Spatial Mixer (MTSM) module that is capable of adapting to disturbances caused by the UAV flight and modeling various temporal clues. By integrating APSE and MTSM into a single model, our network can effectively and accurately capture spatiotemporal relations for UAV videos. Extensive experiments on several benchmarks demonstrate the superiority of our method over state-of-the-art approaches. For instance, our network achieves a classification accuracy of 68.1% with an absolute gain of 1.3% compared to FuTH-Net [20] on the ERA dataset.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**.

KEYWORDS

Unmanned aerial vehicles (UAVs), video recognition, action recognition and understanding, deep neural network, attention mechanism

ACM Reference Format:
Gege Shi, Xueyang Fu, Chengzhi Cao, and Zheng-Jun Zha. 2023. Alleviating Spatial Misalignment and Motion Interference for UAV-based Video Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611799

1 INTRODUCTION

Unmanned aerial vehicles (UAVs) have gained popularly due to their flexibility and wide range of applications, including moving object tracking [4, 33, 37, 62], search and rescue missions [11, 23, 24], action/event recognition and scene understanding [2, 7, 40], etc. The increasing use of UAVs has resulted in a growing number of aerial videos, making UAV-based video recognition crucial. However, as shown in Figure 1, due to the fast motion and constantly changing attitudes and heights, scenarios captured by UAVs contains complex appearances. This results in misalignment of small objects in spatial position between video sequences, leading to incorrect visual movement of static objects (Figure 1 (a)). Additionally, UAV videos typically present more diverse viewpoints with a larger receptive field, causing noteworthy objects to be much smaller than their corresponding background areas (Figure 1 (b)). Moreover, the motion of objects is hindered due to relative movement to flying UAVs, and different video sequences require adaptively capturing various temporal relations (Figure 1 (c)). These factors pose significant challenges in UAV-based video understanding, requiring the design of recognition methods specifically considering these characteristics.

Convolutional neural networks (CNN) have been successful in normal video action classification [16, 48]. Some methods [27, 63] leverage 2D convolutions to extract feature maps of frames and adopt RNNs or temporal convolutions to model temporal relations. However, these methods only consider temporal information at the last layer by aggregating feature maps of consecutive frames, ignoring the misalignment in spatial position of small objects between frames caused by changing attitudes of UAVs, which limits their effectiveness. Other studies [5, 53] use 3D convolutions to extract spatiotemporal features along both spatial and temporal dimensions, typically modeling short-term temporal relations, and are insufficient for complicated motion interference caused by UAV.

*Xueyang Fu is the corresponding author (xyfu@ustc.edu.cn). This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants 62225207, U19B2038, 62121002 and 62276243.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611799

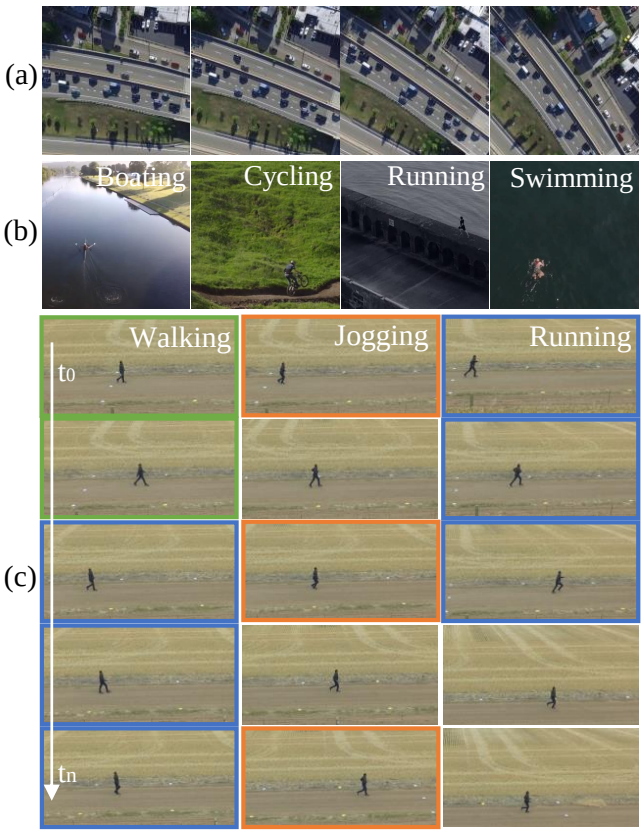

Figure 1: (a) demonstrates misalignment in spatial position of small objects between frames which leads to incorrect visual movement of static objects. (b) shows the objects in UAV videos lay in small regions compared to corresponding background. (c) represents motion of objects interfered on account of relative movement to flying UAVs. Various term temporal clues (boxes in different colors) are desired.

Recently, transformer-based methods [1, 3, 35] are developed for action classification using self-attention along temporal dimension. Benefitted from the huge amount of parameters, these methods achieve stat-of-the-art on action classification. Nevertheless, due to the lack usage of inductive bias which is crucial for UAV videos understanding, these approaches encounter an unexpected accuracy drop in drone datasets accompanied by diversified viewpoints, continuously changing attitudes and fast motion of UAVs.

Building upon our discoveries, we develop a customized framework for UAV-based video recognition called Attentional Spatial and Adaptive Temporal relations modeling (ASAT), which comprehensively considers the changing attitudes and fast motion of UAVs. Specifically, due to the continuously changing attitudes of UAVs and small object sizes in drone videos, we first design an Attentional Patch-level Spatial Enrichment (APSE) module to model dependencies of patches across multiple frames and enriches patch-level features to capture the appearance-based characteristics of videos. The enriched feature enables the network to focus on objects and their positions throughout the video, thereby alleviating spatial misalignment of small objects across consecutive frames. Furthermore, we propose a Multi-scale Temporal and Spatial Mixer (MTSM) module that captures various temporal relations through several parallel temporal convolution pathways with different kernel sizes. MTSM fuses these pathways by reweighting them according to the global information generated from all paths, thus adaptively varying the scale of temporal modeling depending on input video properties. As APSE fully mines spatial clues by modeling dependencies of patches across frames and MTSM models various temporal clues adaptively, our ASAT exhibits adequate spatial-temporal relations modeling capability for UAV-based video recognition.

The main contributions of our work are summarized as follows:

- We propose a novel Attentional Spatial and Adaptive Temporal Relations Modeling (ASAT) framework for UAV-based Video Recognition comprehensively considering the changing attitudes and fast motion of UAVs.
- We propose an Attentional Patch-level Spatial Enrichment module to alleviate misalignment in spatial position of small objects across frames caused by changing attitudes of UAVs by modeling dependencies of patches across consecutive frames and enrich patch-level features.
- We propose a novel Multi-scale Temporal and Spatial Mixer module to model complicated motion interfered by UAV's flight through various temporal kernel size along with enriched spatial features adaptively.
- Extensive experiments on four representative drone video datasets achieve superior performance over state-of-the-art methods, showing the effectiveness of our method. For example, our approach achieves an accuracy of 68.1% with an absolute gain of 1.3% over FuTH-Net [20] on the ERA dataset.

2 RELATED WORKS

2.1 Action Recognition

Due to the rapid development of convolutional neural networks (CNNs) [16, 17, 32, 38, 48, 52], they have now been applied to various fields of deep learning[26, 41, 45, 46, 56]. action recognition and localization has also been developed tremendously over the past decades benefitted from CNNs[15, 18, 21, 25, 27, 60]. According to the way how temporal clues are captured, the deep learning networks are roughly grouped into four categories.

Firstly, some studies capture temporal clues by encoding the motion of RGB frames in optical flows. Representatively, the two stream frameworks [50, 58] utilize two 2D CNNs separately on visual frames and stacked optical flows, which are trained separately and then averaged for final classification.

Secondly, to transfer CNNs superb capability from images to videos, several efforts attempted to extract frame features by using 2D convolutions and model temporal relations at the end. Hei Ng et al. [63] exploit an independent LSTM network before prediction. Jiang et al.[19] construct several modules to encode spatio-temporal and motion features. Lin et al.[31] shift the features in the temporal dimension to optimize information exchange among neighboring frames with no extra parameters. Li et al.[30] design separate blocks to capture both short- and long-range temporal evolution.

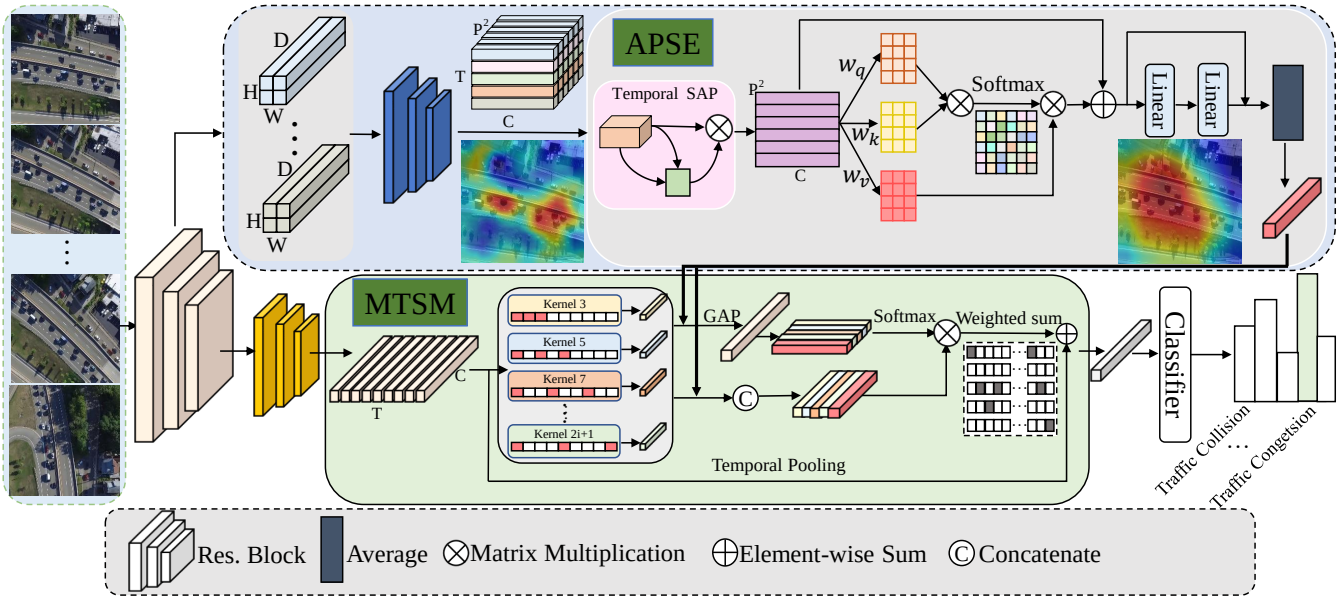

Figure 2: Our proposed framework consists of two specially designed modules: Attentional Patch-level Spatial Enrichment module (APSE) and Multi-scale Temporal and Spatial Mixer (MTSM). Given a video, Res-Blocks are firstly utilized to extract features of sampled frames. Then two different branches are designed for attentional spatial enrichment and multi-scale temporal relation modeling. APSE is utilized to alleviate spatial misalignment between frames through modeling dependencies of patches across multiple frames and enrich patch features to capture the appearance-based characteristics of videos. MTSM models the various-term temporal relations along with spatial features through several parallel pathways by an adaptive mode.

On the other hand, 3D CNNs are launched to extract unified spatio-temporal features. Tran et al.[53] firstly develop spatio-temporal feature learning using deep 3D convolutional networks. To make use of pretrained weights of 2D CNNs, Carreira et al.[5] inflate 2D convolutions to 3D ones. Some other efforts [54, 59] attempt to decompose 3D convolutions into 2D spatial convolutions and 1D temporal convolutions for decreasing heavy computational consumption in 3D CNNs. Feichtenhofe et al.[14] propose a SlowFast network, with the slow one to extract static spatial semantics and the fast one to capture dynamic motions at fine temporal resolution.

Most recently, inspired by the impressive performance of applying transformer architecture in computer vision [10, 34, 65], transformer-based mathods are developed [1, 3, 35, 36, 61]. TimeSformer [3] performs the self-attention along the temporal dimension for temporal modeling. ViViT [1] proposes factorised transformer encoder over spatial and temporal dimensions. Video Swin [35] extends 2D shifted windows to 3D and achieve superior performance through this transformed structure. Benefitted from the attention mechanism and huge amount of parameters, these methods achieve stat-of-the-art for action classification. However, these approaches still suffer from limited usage of inductive bias of aerial videos which is crucial when facing diversified viewpoints, continuously changing attitudes and fast motion of UAVs.

2.2 UAV-based Video Recognition

UAV-based video Recognition is drawing attention depending on the widespread use of unmanned aerial vehicles. UAV video databases [2, 7, 28, 40, 42, 43] have been developed to explore solutions for this task [9, 51]. Early researches [42, 44] propose UAV-recorded datasets and use Pose-based method [6] to extract local action from global receptive field to understand actions in aerial videos. Recently, Jin et al.[20] develop a two-path framework for holistic representation and multiple temporal scale relations learning. However, these solutions are directly based on techniques designed for ground-camera and doesn't give much consideration to the unique characteristics of UAVs. Recently, Kothandaraman et al.[22] uses Fourier theory to disentangle the small human agent from the background in the frequency domain, which offers a new solution for aerial video recognition but still struggles with misalignment in spatial position of small objects.

3 PROPOSED METHOD

In this section, we introduce our Attentional Spatial and Adaptive Temporal Relations Modeling (ASAT) framework for UAV-based video recognition. We first give an overview of our network and then elaborate the key modules.

3.1 Network Structure

Figure 2 presents the workflow of our proposed network. Given a sequence of video frames $F=\{f_1, f_2, ..., f_L\}$, the network firstly extracts features of each frame utilizing ResNet-50 architecture following [31] with shifted channels along the temporal dimension to make use of temporal information at an early stage. The structure

of our proposed network is a two-stream network to assemble observations from spatial and temporal views, which helps to extract more comprehensive representations in UAV videos. We deploy the parameter-shared shallow residual blocks of *layer1-layer3* in ResNet-50 for reducing network parameters, then two non-shared *layer4* are deployed after basic CNNs as the embedding network to output spatial and temporal video features respectively. After that, two customized modules are designed for further generating the holistic video features considering continuously changing attitudes and fast motion of UAVs. Specifically, the Attentional Patch-level Spatial Enrichment module is designed to firstly alleviate the misalignment in spatial position of small objects between frames through modeling dependencies of patches across consecutive frames and enrich patch-level features by attending to the spatial context to capture the appearance-based characteristics of the video. Next, the Multi-scale Temporal and Spatial Mixer is utilized to model the various term temporal relations along with enriched spatial feature by an adaptive mode and generates the holistic features of the video for final classification.

3.2 Attentional Patch-level Spatial Enrichment

We attach great importance to spatial information considering that the diversified viewpoints and changing attitudes in UAV videos. To alleviate misalignment in spatial position of objects between frames and enrich spatial features capturing the appearance-based characteristics of videos, it's necessary to model relations of patches in a single frame and dependencies across multi-frames. Inspired by attention mechanism [55], we propose a novel attentional patch-level spatial enrichment module to achieve this. Specifically, the extracted feature maps of the shared backbone $P_i \in R^{H\times W\times D}$ for video frames f_i ($i \in [1, L]$) are sent to an independent residual block for generating further spatial features $I_i \in R^{P\times P\times C}$, which are then spatially flattened to obtain $x_i \in R^{P^2\times C}$.

We firstly model dependencies of patches between frames through temporal self-attention pooling as shown in Figure 3. Given feature $x \in R^{T\times P^2\times C}$, a linear projection is applied to each spatial local feature $y_i \in R^{T\times C}, i \in [1, P\times P]$ and generates $F_i \in R^{T\times C}$, then we apply a matrix multiplication between F_i and its transposition to generate the self-attention matrix $M_i \in R^{T\times T}$ by

$$F_i = Wy_i, \quad M_i = F_i F_i^T, \tag{1}$$

where W is the parameter of the linear projection, $(\cdot)^T$ indicates the transpose operation. After that, we sum the self-attention matrix along one dimension and perform softmax operation along other dimension to infer the temporal attention vector $a_i \in R^T$.

$$a_i = Softmax(\sum_{j=1}^{T} M_i^j), j \in [1, T]\,. \tag{2}$$

Then a dot production is applied between each local spatial feature y_i and its temporal attention vector a_i. By this way, we obtain the attentive spatial feature $\hat{y_i}$ and generate the local patch spatial features $z_i \in R^C$.

$$\hat{y_i} = y_i \otimes a_i, \quad z_i = \sum_{t=1}^{T} \hat{y}_{i,t}, \tag{3}$$

where $\otimes$ represents the matrix dot multiplication.

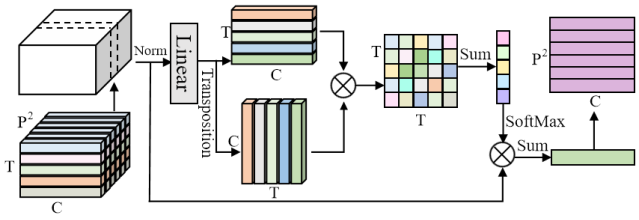

Figure 3: Temporal self-attention pooling for model relations of patches between frames.

Finally, we get the spatial feature sets $F_s = \{z_1, \cdots, z_i, \cdots, z_{p^2}\}$, ($i \in [1, P\times P]$) as the output of temporal self-attention pooling. To let the patch features attend to themselves by aggregating the congruent patch contexts, we then make patch-level enrichment employing self-attention [55], as shown in Figure 2. Weights $W_q, W_k, W_v \in R^{C\times C}$ project latent features F_s to obtain query-key-value triplets by

$$F_q = F_s W_q, \quad F_k = F_s W_k, \quad F_v = F_s W_v, \tag{4}$$

while the value embedding persists the current status of a patch $P \in \left[1, P^2\right]$, the query and key vectors score the pairwise similarity between P^2 patches. These value embeddings are reweighted by corresponding normalized scores to obtain attended features F_m, given by

$$F_m = Softmax(\frac{F_q F_k^T}{\sqrt{C}})F_v + F_s. \tag{5}$$

The position-wise feed-forward network is applied after the self-attention layer, which consists of two linear transformation layers and a ReLU activation function within them, denoted as the following function:

$$F = W_2\sigma\,(W_1 F_m) + F_m, \tag{6}$$

where W_1 and W_2 are the parameters of two linear transformation layers and σ is the ReLU activation function, leading to an improved aggregation of the appearance-based action context across patches between and within frames.

3.3 Multi-scale Temporal and Spatial Mixer

Considering the motion of objects in drone videos is typically interfered by UAVs flight, various term temporal clues (Figure 1 (c)) are desired to model complicated motion patterns. With the efficient APSE to fully enrich spatial features, we build Multi-scale Temporal and Spatial Mixer to model various temporal relations in UAV videos. Since the temporal relations with different scales have varying importance for different sequences, MTSM combines the multi-scale temporal relations in an adaptive way.

Specifically, MTSM takes a sequence of consecutive-frame feature maps $M = \{M_t\}_{t=1}^{T}$ and average of patch-level spatial feature $M_s = Average(F) \in R^C$ as input, where M_t is the feature map of the t^{th} frame. As shown in Figure 2, we conduct K parallel paths $\{\psi_i : M \rightarrow Y_i \in R^{T\times C\times h\times w}\}_{i=1}^{K}$, where ψ_i is 1D temporal convolution with kernel size $2i+1$. For efficiency, we replace the temporal convolution kernel of $(2i+1)\times 1\times 1$ with dilated convolution with a $3\times 1\times 1$ kernel and dilation size i. The basic idea of multi-scale is to use global information from all paths to determine assigned

Table 1: Comparision with the state-of-the-art approaches on ERA, Drone-Action and MOD20 datasets. † means the results are reported from [20], other details are explained in the supplementary meterials. "Kinetics-400", "ImageNet" and "Moments in Time" indicate the datasets named ImageNet [8][47], Kinetics400 [5] and Moments in Time [39] respectively.

Model	Backbone	Pretrain	Param.	Frames	ERA	Drone-Action	MOD20
I3D† [5]	Inception-v1	Kinetics+ImageNet	25.0M	16	58.5	85.5	92.9
TimeSformer [3]	ViT	Kinetics-400	121.4M	8	61.3	76.4	92.5
TRN† [64]	Inception-v3	Moments in Time	47.1M	16	64.3	85.0	93.0
TEA [30]	ResNet-50	ImageNet	24.2M	16	64.5	83.5	93.2
Video Swin [35]	ViT	Kinetics-400	88.1M	16	64.7	90.7	96.2
SlowFast† [14]	ResNet-50	ImageNet	34.6M	64	64.9	86.7	93.1
TDN [57]	ResNet-50	Kinetics-400	24.1M	16	65.0	91.7	95.8
TSM [31]	ResNet-50	Kinetics-400	23.5M	16	65.7	90.3	96.5
FuTH-Net† [20]	Inception-v1	Kinetics+ImageNet	77.1M	16	66.8	88.4	95.7
FAR [22]	X3D	Kinetics	3.8M	8	66.9	92.7	96.9
ASAT(ours)	ResNet-50	Kinetics-400	61.8M	16	**68.1**	**93.1**	**98.2**

weights to each path. We first fuse all paths by elemente-wise summation, then perform global average pooling operation to generate channel-wise statistics of global feature $G \in R^C$, formulated as:

$$G = GAP((\sum_{i=1}^{K} Y_i) + M_s), \tag{7}$$

where GAP denotes global average pooling. The we can obtain the channel selection weights$\{g_i \in R^C\}_{i=1}^K$ by:

$$g_i = \frac{exp(W_{i2}W_{i1}G)}{\sum_{j=1}^{K+1} exp(W_{j2}W_{j1}G)}, \quad i \in 1, \cdots, K, K+1, \tag{8}$$

where $W_{i1} \in R^{C \times c}$ and $W_{i2} \in R^{c \times C}$ are parameters to generate g_i. The aggreated feature Z is the obtained through the assigned weights on various temporal kernels and spatial enrichment path,

$$Z = \sum_{i=1}^{K} g_i \odot Y_i + g_{K+1} \odot M_s. \tag{9}$$

After that, an excite operation modulates the input feature map M by conditioning on Z with a residual scheme to fully capture channel-wise dependencies. Considering that most previous work benefit from taking the average along time, we add an independent pathway parallel to multiple timescales, executing average along the temporal dimension. The final global video feature $G_f = Z + T_{AP}(M)$, where T_{AP} represents temporal pooling.

It's worth pointing out that, different from using scale-wise weight to generate coarse fusion, we propose to use channel-wise weights as Eq. (8), which produces more fine-grained fusion. Moreover, the weights are dynamically computed depending on input sequences. This is crucial for UAV-based video recognition where different sequences may have different dominate temporal scales even belongs to the same category.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. Our approach is evaluated on four drone video recognition benchmarks: ERA [40], MOD20 [43], Drone-Action [42] and UAV-Human[29]. The ERA is an aerial event recognition dataset and consists of 1473 videos for training and 1391 videos for testing from 25 categories. MOD20 dataset is a multi-viewpoint outdoor dataset contains 2324 video clips from 20 action categories and provides 3 training/testing splits (split2 is used in our experiment settings). Drone-Action is a dataset for human action classification in aerial videos, in which 240 aerial videos are collected and defined by 13 different actions. Similar to MOD20, this dataset also offers 3 training/testing splits, each of which consists of 168 training clips and 72 testing clips. By following existing works, we report the accuracy on average over the three splits. Besides, we also evaluate the effectiveness of our proposed method on the the largest UAV-based human behavior understanding dataset UAV-Human[29]. To compare with existing works, we use split 1 which contains 15172 and 5556 videos for training and testing respectively.

During preprocessing, we transform video clips of the Drone-Action and MOD20 datasets into the same data structure as the ERA dataset. Since the resolution of videos is 720 × 720 in MOD20 dataset and 1920 × 1080 in Drone-Action dataset, each frame is cropped and resized to a size of 640 × 640. Afterward, durations of videos in the Drone-Action dataset range from 5 to 21 s and the number of frames ranges from 70 to 140 in the MOD20 dataset, we cut them to 64 frames and sample 16 frames from each video clip with a fixed sampling rate on all datasets as input.

Implementation Details. Our network is implemented using Pytorch on two NVIDIA RTX 3060Ti GPUs. In the training process, we uniformly sample 16 frames to generate the input sequence from each video and apply multi-scale crop and horizontal flip for data augmentation with the size of 256 × 256. We use a batch size of 6 and SGD optimizer using a momentum of 0.9 and a weight decay of $5e^{-4}$. The maximum training epoch is set to 80, with the initial learning rate 10^{-4}, decreased by a factor of 10 for every 20 epochs. The final softmax predictions of our method is constrained using multi-class cross entropy loss.

4.2 Comparison with State-of-the-art Methods

We compare our proposed ASAT with several start-of-the-art action recognition methods [3, 5, 14, 30, 31, 35, 57, 64] and UAV-based video recognition methods [20, 22]. Table 1 reports the quantitative

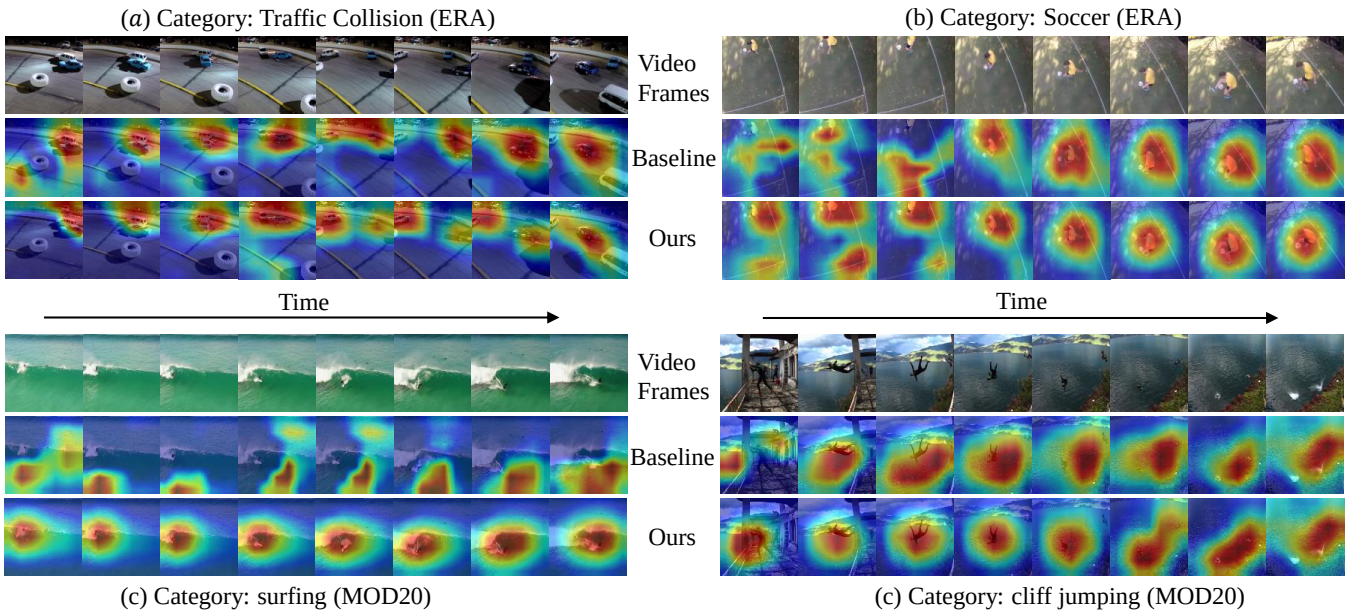

Figure 4: Example attention map visualizations obtained from the baseline (TSM[31]) and our proposed approach on four examples from the ERA and MOD20 dataset. The attention maps measure the activation magnitude of latent features. The baseline struggles in the case of spatial and temporal context variations that are commonly encountered in the drone scene, e.g., 2nd and 3rd frame from the left in (b), where the regions not corresponding to actions are emphasized improperly. Similarly, while background region is also emphasized in (c), the action in the 1st and 6th frame from the left in (d) is not accurately captured. Our proposed approach explicitly enhances class-specific feature discriminability through spatio-temporal context aggregation by alleviating spatial misalignment and motion interference caused by UAVs.

Table 2: Comparison with state-of-the-art methods on the ERA dataset. We report the per-class precision and overall accuracy (OA) on the test set. The best precision/overall accuracy is shown in bold.

Model	*post-earthquake*	*flood*	*fire*	*landslide*	*mudslide*	*traffic collision*	*trafficcongestion*	*harvesting*	*ploughing*	*constructing*	*police chase*	*conflict*	*baseball*	*basketball*	*boating*	*cycling*	*running*	*soccer*	*swimming*	*car racing*	*party*	*concert*	*parade/protest*	*religious activity*	*non-event*	OA
I3D†	60.0	67.1	65.7	29.0	60.4	51.5	52.2	67.1	66.7	54.2	64.8	57.9	85.0	61.9	**86.4**	**75.0**	44.4	77.6	64.1	65.2	53.7	50.0	47.8	65.1	43.0	58.5
TimeSformer	57.4	62.0	88.7	63.6	47.7	46.7	59.3	75.0	84.6	74.5	65.0	27.6	**88.1**	64.5	69.0	62.5	47.3	73.9	60.7	41.4	53.3	58.3	56.9	58.7	47.4	61.3
TRN†	69.2	87.8	88.9	**65.8**	60.0	44.1	58.3	78.1	**90.7**	70.8	73.3	28.6	83.3	72.7	73.7	60.0	66.7	73.6	70.6	63.6	65.1	47.7	42.7	65.1	47.9	64.3
SlowFast†	70.1	**88.0**	83.3	57.2	67.3	51.4	56.2	68.4	87.6	**82.0**	75.1	**75.5**	40.8	70.3	71.8	61.4	54.7	78.2	72.9	74.3	50.4	70.3	50.6	65.7	60.7	64.9
TSM	59.3	79.5	87.7	44.4	53.0	**54.3**	60.9	**81.7**	84.6	76.9	74.0	35.3	79.7	78.6	81.3	62.5	73.8	91.8	62.3	46.2	78.0	61.7	50.0	64.6	47.7	65.7
FuTH-Net†	**72.7**	75.5	87.5	57.1	**74.5**	34.0	56.0	76.6	71.2	81.4	76.5	36.0	78.0	**85.4**	80.4	73.6	16.3	64.5	**80.4**	**84.2**	56.0	**89.8**	**65.3**	63.0	**63.9**	66.8
ours	62.3	85.7	**91.4**	56.5	62.7	47.7	**66.0**	68.8	**90.8**	75.0	**80.5**	40.0	84.0	78.9	85.7	64.6	**78.8**	**94.0**	61.4	61.9	**87.1**	56.0	47.9	65.3	55.2	**68.1**

results in ERA, MOD20 and Drone-Action datasets. It is obvious that our proposed framework achieves outstanding improvements compared with corresponding competitors.

We first compare ASAT with several action recognition methods. These methods have demonstrated the effectiveness on normal video classification datasets (action recognition), but failed to produce competitive results on UAV-based video datasets. For example, the accuracy of [35] (the start-of-the-art on Kinetics400 dataset [5]) is 64.7% and 96.2% when deployed on ERA and MOD20 dataset. This is much lower than expected because these methods cannot suit for UAV-based video understanding very well due to the changing attitudes and fast motion as unique characteristics of UAVs. Next, we compare with UAV-based video recognition methods FuTH-Net[20], which fuse holistic features and temporal relations for aerial video classification benefitted from combining [64] and [5]. The results is superior to normal action recognition methods, obtaining 66.8% on

Table 3: Comparison with state-of-the-art methods on the MOD20 dataset. We report the per-class precision and OA on the test set of split2. The best precision/accuracy is shown in bold.

Model	*backpacking*	*chainsawing trees*	*cliff jumping*	*cutting wood*	*cycling*	*dancing*	*fighting*	*figure skating*	*fire fighting*	*jekskiing*	*kayaking*	*motorbiking*	*nfl catches*	*rock climbing*	*running*	*skateboarding*	*skiing*	*standup paddling*	*surfing*	*windsurfing*	OA
I3D	83.8	94.3	97.8	97.1	70.2	93.7	96.9	100	100	100	100	82.7	93.8	100	87.1	83.3	89.4	100	**100**	94.5	92.9
TimeSformer	85.3	78.9	97.7	87.5	**90.0**	100	97.1	100	100	100	97.1	78.0	97.8	100	80.0	82.9	85.0	100	**100**	97.2	92.5
TRN	90.6	87.5	100	78.9	85.7	96.6	94.4	97.1	100	100	97.1	87.8	100	96.7	76.9	89.6	91.9	100	94.7	92.1	93.0
SlowFast	83.3	96.0	97.8	71.7	81.4	100	100	100	100	100	94.3	84.8	100	100	83.3	96.8	89.5	100	97.3	97.2	93.1
TSM	90.0	100	100	100	72.5	100	100	100	100	100	100	85.7	100	100	96.8	100	**100**	97.4	97.2	100	96.5
FuTH-Net	90.2	100	100	100	82.5	96.2	100	81.0	100	100	100	78.4	100	100	**100**	100	**100**	100	92.3	100	95.7
ours	**97.3**	**100**	**100**	**100**	84.4	**100**	**100**	**100**	**100**	**100**	**100**	**93.8**	**100**	**100**	97.0	**100**	94.4	**100**	97.3	**100**	**98.2**

Table 4: Quantitative Comparisions on the UAV-Human dataset. The results with † are the reported from [29]

Model	Input Size	GFlops	OA
I3D-M †[5]	540 ×960	346.55	21.06
MVIT †[12]	-	70.8	24.3
TimeSformer †[3]	224 ×224	2380	33.9
X3D-M †[13]	540 ×540	14.39	36.6
FAR †[22]	540 ×540	14.41	38.6
Ours	224 ×224	72.9	**39.7**

Table 5: Quantitative ablation study on the two proposed modules. We evaluate the effect of two modules on overall accuracy values

Attentional Patch-level Spatial Enrichment	Multi-scale Temporal and Spatial Mixer	ERA	MOD20
✗	✗	65.7	96.5
✓	✗	67.4	97.3
✗	✓	66.6	97.9
✓	✓	**68.1**	**98.2**

ERA dataset in accuracy, however 1.3% lower than ours. Compared to these approaches, our proposed network clearly outperforms them on all datasets. Specifically, our method achieves accuracy of 68.1%, 93.1%, 98.2% in ERA, Drone-Action and MOD20 dataset respectively.

Tables 2 and 3 present pre-class precision of different models in ERA and MOD20 datasets. In particular, our model achieves the highest per-class precisions for some challenging categories, such as party (87.1%), soccer (94.0%), and running (78.8%) in Table 2 and backpacking (97.3%), motorbicking (93.8%) in Table 3. This is attribute to our ASAT can fully utilize spatial consistency between consecutive frames and capture complex dynamic temporal relations adaptively, which is crucial to distinguish events with significant interclass variances. We also visualize the attention maps of some examples in Figure 4. Our proposed method can focus more on noteworthy fields and suppress irrelevant areas.

4.3 Comparison on UAV-Human

we also evaluate the effectiveness of our proposed method on the the largest UAV-based human behavior understanding dataset UAV-Human[29]. We report state-of-the-art comparisons in Table 4. Our proposed method achieves 39.7% accuracy, which is 18.6% higher than I3D-M [5], while the GFlops is only one-fourth of it. When compared with a similar model in GFlops MVIT [12], our method can achieve 15.4% higher accuracy, demonstrating the effectiveness.

4.4 Ablation Study

To evaluate the effectiveness of the proposed components of ASAT, we extensively conduct the ablation study.

Effect of Proposed Modules. We validate the importance of Attentional Patch-level Spatial Enrichment module by training ASAT without temporal relations modeling over ERA and MOD20 datasets, and there is a great performance gap of quantitative results over two datasets in the first two rows of Table 5. It shows that modeling dependencies of patches across consecutive frames and then enrich patch-level features can efficiently mine the spatial clues in UAV videos and thus improve the recognition accuracy for UAV videos.

We also validate the effectiveness of Multi-scale Temporal and Spatial Mixer without enriched spatial feature. That's to say, only dynamic modeling various-term temporal relations by an adaptive mode and results are shown in the first and third row in Table 5. Obviously, dynamic temporal relation modeling achieves accuracy gains up to 0.9% and 1.4% on ERA and MOD20 dataset, respectively.

Most importantly, when both Attentional Patch-level Spatial Enrichment module and Multi-scale Temporal and Spatial Mixer are embedded in ASAT, we achieve even higher accuracy performance 68.1% on ERA dataset and 98.2% on MOD20 dataset.

The Number of Dynamic Temporal Kernels. We also investigate the effect of combining different scale temporal relations in ERA dataset. We consider five different scale kernels, called "K3" (standard $3 \times 1 \times 1$ 3D convolutional kernel), "K5", "K7", "K9", "K11" ($3 \times 1 \times 1$ convolution with dilation 2,3,4,5 to approximate $5 \times 1 \times 1$,

Table 6: Comparision results of ASAT with different combinations of multiple temporal kernels in MTSM. GAP means adding an independent pathway parallel to multiple timescales, executing average along the temporal dimension.

Kernel Size					ERA	
K3	K5	K7	K9	K11	OA	GAP
✓					67.2	66.7
	✓				66.7	67.4
		✓			66.9	66.8
			✓		66.3	66.1
				✓	65.9	65.7
✓	✓	✓			**67.3**	**68.1**
	✓	✓	✓		66.9	67.1
		✓	✓	✓	66.2	66.0
✓	✓	✓	✓		65.8	66.5
	✓	✓	✓	✓	65.9	66.6
✓	✓	✓	✓	✓	66.0	66.5

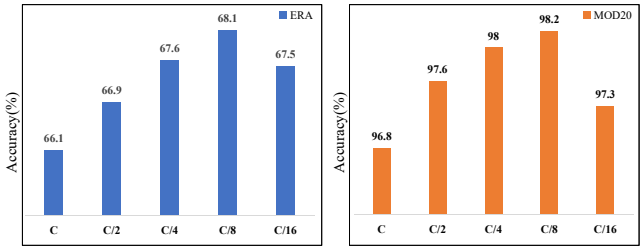

Figure 5: Accuracy on ERA and MOD20 datastes with different channel settings.

$7 \times 1 \times 1$, $9 \times 1 \times 1$, $11 \times 1 \times 1$ kernel size respectively). The results are reported in Table 6. GAP means adding an independent pathway parallel to multiple timescales, executing average along the temporal dimension. It can be observed that: (1) The overall accuracy in the first block of the Table 6 is generally getting lower with the kernel size increasing except 'K5'. That's to say, using large kernel size singly perhaps can't model motion patterns effectively. (2) Using multi-temporal kernels with different sizes is beneficial to modeling temporal relations. (3) Using more temporal kernels (K = 4 or 5) does not bring performance gain, showing three temporal kernels are enough to capture the temporal clues of video. The best performance is produced when K3, K5, K7 are combined with GAP operation, achieving 68.1% in ERA dataset.

Effect of the Middle Channels Note that when generating the channel selection weights $\{g_i \in R^C\}_{i=1}^K, (C = 2048)$, the number of channels generated through W_{i1} attaches great importance to the performance of ASAT. So we also evaluate the influence of the number of channels. Specifically, we set channel number $c \in (C, C/2, C/4, C/8, C/16)$. Figure 5 shows the overall accuracy performance on ERA and MOD20 datasets in different settings. The best results are produced when $c = C/8$.

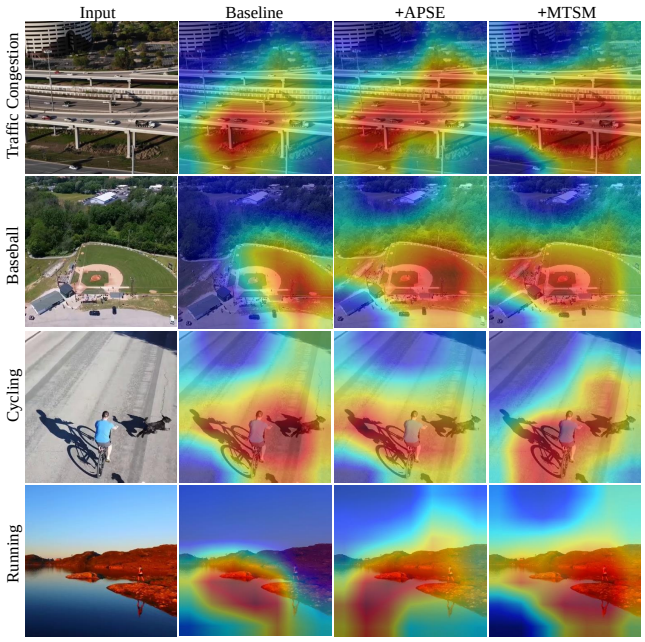

Figure 6: Visualization of feature maps by Grad-CAM on ERA dataset. For more visualization results, please refer to the Supplementary Material.

4.5 Visualization

For visulaization, we additionally visualize the activation maps by Grad-CAM [49] of Baseline, Baseline+APSE and Baseline + APSE + MTSM (ASAT) in Figure 6 on ERA dataset, and we can see that APSE and MTSM help to focus more on noteworthy fields as they suppress irrelevant areas even the foreground regions that taking more visual field, which is crucial for UAV-based video recognition task. Besides, We also visualize the attention maps of video sequences in Figure 4. Our proposed method can focus more on correct spatial position and capture corresponding actions even the regions encounter with massive deformation.

5 CONCLUSION

For UAV-based video recognition task, we propose a novel Attentional Spatial and Adaptive Temporal Relations Modeling (ASAT) framework considering the changing attitudes and fast motion of UAVs. On the one hand, ASAT employs the Attentional Patch-level Spatial Enrichment (APSE) module to alleviate spatial misalignment of small objects between frames and enrich patch-level features to capture the appearance-based characteristics of videos. On the other hand, Multi-scale Temporal and Spatial Mixer (MTSM) is designed to adaptively model complicated visual motion interfered by UAV's flight through capturing various term temporal relation along with enriched spatial feature. Benefiting from the two modules, our method achieves 68.1%, 93.1%, 98.2%, 39.7%, accuracy in ERA, Drone-Action, MOD20 and UAV-Human dataset respectively, validating the effectiveness of the proposed ASAT network.

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[2] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. 2017. 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*. 28–35.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *ICML*, Vol. 2. 4.

[4] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. 2022. TCTrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14798–14808.

[5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[6] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. 2015. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*. 3218–3226.

[7] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. 2020. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1717–1726.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[9] Meng Ding, Ning Li, Ziang Song, Ruixing Zhang, Xiaxia Zhang, and Huiyu Zhou. 2020. A Lightweight Action Recognition Method for Unmanned-Aerial-Vehicle Video. In *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*. IEEE, 181–185.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Milan Erdelj, Enrico Natalizio, Kaushik R Chowdhury, and Ian F Akyildiz. 2017. Help from the sky: Leveraging UAVs for disaster management. *IEEE Pervasive Computing* 16, 1 (2017), 24–32.

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6824–6835.

[13] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 203–213.

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[15] Guoqiang Gong, Liangfeng Zheng, Wenhao Jiang, and Yadong Mu. 2021. Self-Supervised Video Action Localization with Adversarial Temporal Transforms.. In *IJCAI*. 693–699.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[18] Yuko Iinuma and Shin'ichi Satoh. 2021. Video Action Retrieval Using Action Recognition Model. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 603–606.

[19] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2000–2009.

[20] Pu Jin, Lichao Mou, Yuansheng Hua, Gui-Song Xia, and Xiao Xiang Zhu. 2022. FuTH-Net: Fusing Temporal Relations and Holistic Features for Aerial Video Classification. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13.

[21] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[22] Divya Kothandaraman, Tianrui Guan, Xijun Wang, Shuowen Hu, Ming Lin, and Dinesh Manocha. 2022. FAR: Fourier Aerial Video Recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*. Springer, 657–676.

[23] Christos Kyrkou and Theocharis Theocharides. 2019. Deep-Learning-Based Aerial Image Classification for Emergency Response Applications Using Unmanned Aerial Vehicles.. In *CVPR workshops*. 517–525.

[24] Christos Kyrkou and Theocharis Theocharides. 2020. EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 1687–1699.

[25] Bing Li, Jiaxin Chen, Dongming Zhang, Xiuguo Bao, and Di Huang. 2022. Representation Learning for Compressed Video Action Recognition via Attentive Cross-modal Interaction with Motion Enhancement. *arXiv preprint arXiv:2205.03569* (2022).

[26] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. 2023. Edge-Aware Regional Message Passing Controller for Image Forgery Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8222–8232.

[27] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. 2016. Action recognition by learning deep multi-granular spatio-temporal video representation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 159–166.

[28] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16266–16275.

[29] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16266–16275.

[30] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. 2020. TEA: Temporal Excitation and Aggregation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[32] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. 2021. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4370–4379.

[33] Shuai Liu, Xin Li, Huchuan Lu, and You He. 2022. Multi-object tracking meets moving UAV. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8876–8885.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.

[36] Wenyang Luo, Yufan Liu, Bing Li, Weiming Hu, Yanan Miao, and Yangxi Li. [n. d.]. Long-Short Term Cross-Transformer in Compressed Domain for Few-Shot Video Classification. ([n. d.]).

[37] Murari Mandal, Lav Kush Kumar, and Santosh Kumar Vipparthi. 2020. Mor-uav: A benchmark dataset and baselines for moving object recognition in uav videos. In *Proceedings of the 28th ACM international conference on multimedia*. 2626–2635.

[38] Shaobo Min, Hantao Yao, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. 2020. Multi-objective matrix normalization for fine-grained visual recognition. *IEEE Transactions on Image Processing* 29 (2020), 4996–5009.

[39] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. 2020. Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2020), 502–508. https://doi.org/10.1109/TPAMI.2019.2901464

[40] Lichao Mou, Yuansheng Hua, Pu Jin, and Xiao Xiang Zhu. 2020. Era: A data set and deep learning benchmark for event recognition in aerial videos [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine* 8, 4 (2020), 125–133.

[41] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. 2020. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 191–207.

[42] Asanka G Perera, Yee Wei Law, and Javaan Chahl. 2019. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones* 3, 4 (2019), 82.

[43] Asanka G Perera, Yee Wei Law, Titilayo T Ogunwa, and Javaan Chahl. 2020. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems* 50, 5 (2020), 405–413.

[44] Asanka G Perera, Yee Wei Law, and Javaan Chahl. 2018. UAV-GESTURE: A dataset for UAV control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[45] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. *CoRR* abs/1506.02640 (2015). arXiv:1506.02640 http://arxiv.org/abs/1506.02640

[46] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. 2019. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing* 28, 9 (2019), 4364–4375.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[50] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).

[51] Waqas Sultani and Mubarak Shah. 2021. Human action recognition in drone videos using a few aerial training examples. *Computer Vision and Image Understanding* 206 (2021), 103186.

[52] Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. 2020. Learning to Discretely Compose Reasoning Module Networks for Video Captioning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

[53] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[54] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[56] Kunyu Wang, Xueyang Fu, Yukun Huang, Chengzhi Cao, Gege Shi, and Zheng-Jun Zha. 2023. Generalized UAV Object Detection via Frequency Domain Disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1064–1073.

[57] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1895–1904.

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[59] H. Wu, J. Liu, Z. J. Zha, Z. Chen, and X. Sun. 2019. Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*.

[60] Haoze Wu, Jiawei Liu, Xierong Zhu, Meng Wang, and Zheng-Jun Zha. 2021. Multi-scale spatial-temporal integration convolutional tube for human action recognition. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 753–759.

[61] Shuo Yang and Xinxiao Wu. [n. d.]. Entity-aware and Motion-aware Transformers for Language-driven Action Localization. ([n. d.]).

[62] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. 2022. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8896–8905.

[63] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.

[64] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*. 803–818.

[65] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. 2022. Two-Stream Transformer for Multi-Label Image Classification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3598–3607.