Saliency-aware Spatio-temporal Modeling for Action Recognition on Unmanned Aerial Vehicles

Xiaoxiao Sheng, Zhiqiang Shen, and Gang Xiao

***Abstract*—Action recognition on unmanned aerial vehicles (UAVs) must cope with complex backgrounds and focus on small targets. Existing methods usually use additional detectors to extract objects in each frame, and use the object sequence within boxes as the network input. However, for training, they rely on additional detection annotations, and for inference, the multi-stage paradigm increases the burden of deployment on UAV terminals. Therefore, we propose a saliency-aware spatio-temporal network (SaStNet) for UAV-based action recognition in an end-to-end manner. Specifically, the short-term and long-term motion information are captured progressively. For short-term modeling, a saliency-guided enhancement module is designed to learn attention scores for weighting the original features aggregated within neighboring frames. For long-term modeling, informative regions are first adaptively concentrated using a saliency-guided aggregation module. Then, a spatio-temporal decoupling attention mechanism is designed to focus on spatially salient regions and capture temporal relationships within all frames. Integrating these modules into classical backbones encourages the network to focus on moving targets, reducing interference from background noises. Extensive experiments and ablation studies are conducted on UAV-Human, Drone action, and something-something datasets. Compared to state-of-the-art methods, SaStNet achieves a 5.7% accuracy improvement on the UAV-Human dataset using 8-frame inputs.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9102

***Index Terms*—Deep learning, action recognition, attention mechanism, unmanned aerial vehicles.**

I. Introduction

Unmanned aerial vehicles (UAVs) equipped with cameras have demonstrated great potential in various tasks, leveraging the advantages of low-altitude flights. Recently, UAV-based action recognition has been applied in security, rescue, agriculture, and traffic management to enhance decision-making capabilities. Unlike fixed indoor cameras, varying altitudes and perspectives of UAVs result in videos with complex and dynamic backgrounds. Moreover, the objects in UAV videos are small in proportion and may even be imperceptible, as shown in Fig. 1.

The objective of action recognition is to classify the input video into its respective action category, and it is essential to model the appearance and motion information of the entire video. The methods for universal action recognition usually apply dual-stream architectures and 3D convolutions. The dual-stream networks [1]–[3] utilize inputs with different frame rates to model appearance and motion information respectively. 3D convolution networks [4]–[6] learn spatio-temporal representations jointly using one backbone. Besides, lightweight backbones [7]–[9] are designed to understand videos in a cascading or spatio-temporal decoupling manner. However, these typical methods often overlook the saliency of moving objects in complex environments and are susceptible to background noise, which are unignorable challenges for action recognition on UAVs.

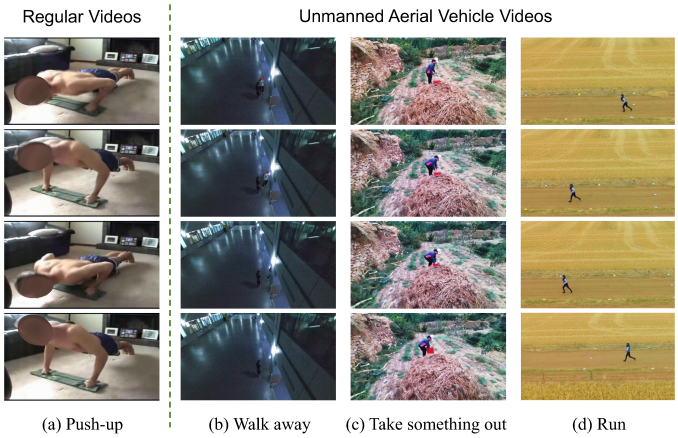


Fig. 1. Compared to regular action recognition samples, the foreground objects in UAV videos occupy a small proportion, and the background is complex and dynamic. These are the challenges of UAV action recognition.

To alleviate this issue, additional trained detectors are utilized to crop target objects in UAV videos. For example, MITFAS [10] first detects the object box and uses mutual information criterion to crop foreground sequences for action recognition. Drone-HAT [11] designs a drone system to detect, track, and classify the actions of object sequences separated from complex backgrounds. In addition, saliency detection is also introduced to action recognition to highlight the foreground areas in each frame. For example, Assefa *et al*. [12] proposed a pre-trained method to obtain salient regions by background patching; then, the pre-trained model is fine-tuned on action recognition. These methods enable networks effectively focus on the regions of interest, thereby achieving higher accuracy. However, additional preprocessing is required. Besides, Kong *et al*. [13] integrated the saliency detection task into the recognition network, creating a multi-task framework. By using saliency ground truth for supervi-

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo José da Luz (*Corresponding author: Xiaoxiao Sheng*).

X. Sheng, Z. Shen, and G. Xiao are with Shanghai Jiao Tong University, Shanghai, China (e-mails: shengxiaoxiao@sjtu.edu.cn, shenzhiqiang@sjtu.edu.cn, and xiaogang@sjtu.edu.cn).

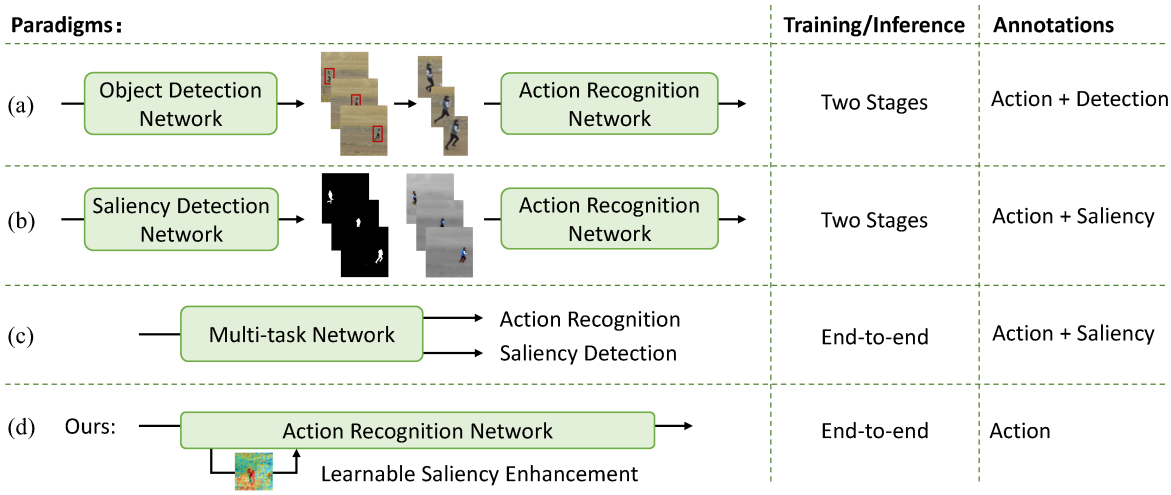


Fig. 2. To mitigate the impact of complex backgrounds, existing methods usually introduce additional trained objects or saliency detectors (a-c). This increases the cost of training or inference. Unlike these methods, we design a saliency-aware spatio-temporal modeling network for action recognition, which is end-to-end and does not require additional annotations (d).

sion, useful foreground information can be highlighted, further improving action recognition during training. However, additional annotations is required. Therefore, we design a saliency-aware paradigm to adaptively improve spatio-temporal modeling with end-to-end training. In Fig. 2, the differences between the paradigms for UAV action recognition are shown.

In this paper, we propose a Saliency-aware Spatio-temporal Network, termed SaStNet. Saliency-guided modules are developed to enhance short-term modeling and facilitate long-term spatio-temporal attention. Specifically, uniform sampling is first applied to the video, and the neighboring frames around each sampled frame form a segment. Then, short-term motion information is extracted by analyzing temporal differences within each segment. The saliency-guided enhancement module is designed to activate informative motions and capture prominent appearance features of the sampled center frames. Next, long-term spatio-temporal information is extracted from all sampled frames. An adaptive aggregation module is designed to condense key information from multi-level feature maps. Decoupled spatial and temporal attention mechanisms are then applied to the condensed features to focus on salient motion areas and model global temporal relationships. By embedding these modules into classical backbones, comprehensive spatio-temporal representations can be effectively learned for action recognition. Extensive experiments and ablation studies are conducted on UAV-Human [6] and Drone action [2] dataset for UAV action recognition. In addition, we also utilize universal action recognition benchmarks to demonstrate the effectiveness of our SaStNet. The main contributions of this paper are as follows:

- We propose a saliency-aware spatio-temporal modeling network for UAV action recognition. Our designed modules can seamlessly integrate with classic backbones in a plug-and-play manner.
- A saliency-guided short-term module is proposed to activate key motion information and capture prominent appearance features.
- A long-term module is developed to learn global information through spatio-temporal decoupled attention with adaptively aggregated multi-level feature maps.
- Our method achieves state-of-the-art results compared to other advanced works on UAV and universal action recognition datasets. We also perform ablation studies to show the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II introduces the framework of our proposed method. In Section III, we present the performance comparison, ablation studies, and visualizations. Finally, Section IV concludes the paper.

II. Method

A. Overview

This section provides an overview of our proposed SaStNet, and its framework is shown in Fig. 3. First, uniform sampling is utilized to obtain a video with T frames. Using these sampled frames as centers, their neighboring frames are extracted to form T short segments. Within each segment, short-term modeling with saliency-guided enhancement modules is conducted on appearance and motion features to improve the spatio-temporal representation. For long-term modeling, a saliency-guided aggregation module is designed to adaptively aggregate key information from the feature maps. The spatial feature is further enhanced by applying attention mechanisms between the condensed keys and the original feature queries. Then, temporal attention is used to learn long-term relationships within the entire video.

Our designed modules can seamlessly integrate with existing networks. In this paper, we adopt the classic ResNet [14] as the backbone to universally evaluate our method. The Conv1 and Stage 2 of ResNet are used for short-term modeling, and the Stages 3-5 are utilized for long-term modeling.

B. Saliency-guided Short-term Enhancement Module

To capture short-term motion information, existing methods usually utilize inputs or low-level features for elaborate

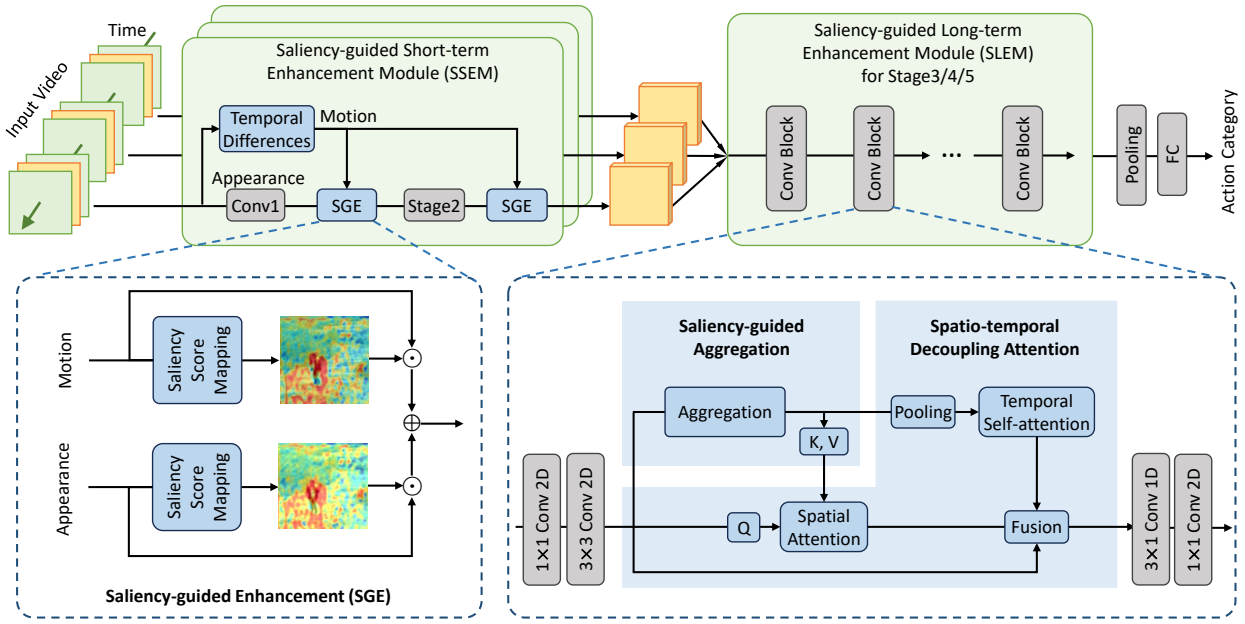


Fig. 3. The framework of SaStNet. Uniform sampling is first used to obtain multiple frames, and their neighboring frames are extracted to form short segments. The saliency-guided short-term enhancement module (SSEM) is designed to learn the spatio-temporal information within the segment. Then, the saliency-guided long-term enhancement module (SLEM) is proposed to learn global information across all sampled frames. Finally, the learned comprehensive representations are utilized for classification. The SSEM is used for Conv1 and Stage 2, and the SLEM is inserted into Stages 3-5 of the backbone.

modular design. Although these methods achieve good performance, they do not fully leverage learnable saliency guidance. Besides, while existing works use temporal differences to capture motion information, they also introduce dynamic background interference caused by flying UAVs. To alleviate this, we design a saliency-guided short-term enhancement module (SSEM) to focus on salient foreground regions and motions. The network of our SSEM is shown in Fig. 4.

Specifically, the uniformly sampled frames are regarded as appearance branch inputs and denoted as $\boldsymbol{X} \in \mathbb{R}^{N\times T\times 3\times H\times W}$, where N is the batchsize, T is the number of frames, H is height, and W is the width of each frame. In addition, the inputs of the motion branch are short segments, denoted as $\boldsymbol{S} \in \mathbb{R}^{N\times 5T\times 3\times H\times W}$, where each segment contains 5 frames surrounding the sampled frame. Temporal differences within each segment are calculated to capture short-term motions of adjacent frames as follows:

$$\Delta T_i = \boldsymbol{S}_{T_{i+1}} - \boldsymbol{S}_{T_i}, \quad T_i = 0, 1, \ldots, 4, \tag{1}$$

where $\Delta T_i \in \mathbb{R}^{N\times 3\times H\times W}$ is the temporal difference between T_{i+1}-th and T_i-th frames. Then, the temporal differences are stacked and processed by a motion branch as follows:

$$\Delta \boldsymbol{T}_{avg} = Conv1(Avgpool(Concat([\Delta T_0, \ldots, \Delta T_4]))), \tag{2}$$

where the average pooling is performed to filter the background noise. Following ResNet [14], the $Conv1$ contains the Convolution, BatchNorm, and ReLu. Meanwhile, the spatial appearance features of each sampled center frame are extracted as follows:

$$\boldsymbol{F} = Conv1(\boldsymbol{X}). \tag{3}$$

As the appearance and motion branch focus on different information, the saliency scores are respectively extracted for $\boldsymbol{F}$ and $\Delta \boldsymbol{T}_{avg}$ as follows:

$$\hat{\boldsymbol{F}} = Mapping(\boldsymbol{F}), \tag{4}$$

$$\Delta \hat{\boldsymbol{T}}_{avg} = Mapping(\Delta \boldsymbol{T}_{avg}), \tag{5}$$

where the $Mapping$ is a lightweight multilayer perceptron (MLP) followed by a Sigmoid function to obtain the saliency scores, and its framework is shown in Fig. 4.

The comprehensive spatio-temporal representation $\tilde{\boldsymbol{F}} \in \mathbb{R}^{N\times T\times C\times H\times W}$ is obtained through saliency weighted fusion between $\hat{\boldsymbol{F}}$ and $\Delta \hat{\boldsymbol{T}}_{avg}$ as follows:

$$\tilde{\boldsymbol{F}} = \hat{\boldsymbol{F}} \times \boldsymbol{F} + \Delta \hat{\boldsymbol{T}}_{avg} \times \Delta \boldsymbol{T}_{avg}. \tag{6}$$

The same enhancement is conducted in Stage 2. The motion branch is merged into the spatial appearance branch during short-term modeling.

C. Saliency-guided Long-term Enhancement Module

The saliency-guided aggregation module first extracts key tokens from spatio-temporal feature maps. Then, decoupled spatial and temporal attention mechanisms are employed to emphasize salient features and capture temporal relationships. The network of SLEM is illustrated in Fig. 5.

1) Saliency-guided Aggregation: Besides focusing on salient foreground objects during the short-term stage, it is also crucial to attend to informative features during long-term stages. To achieve this with acceptable computational complexity, the saliency-guided aggregation module is used to extract sparse primary tokens. First, the sine and cosine positional encoding [15] is added to the feature map $\tilde{\boldsymbol{F}}$ to

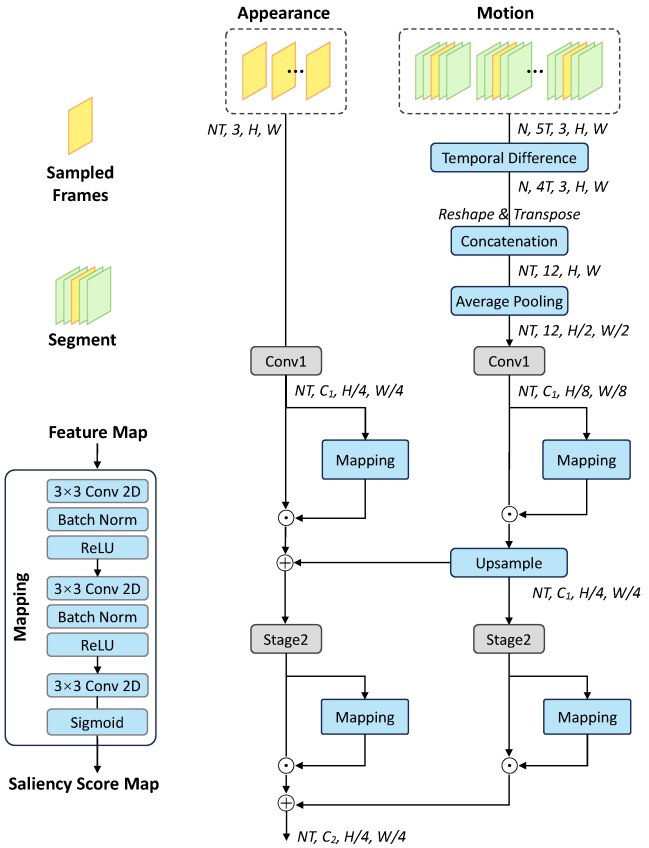


Fig. 4. The network of Saliency-guided Short-term Enhancement Module (SSEM). The mapping network is used to obtain saliency score maps. The learned appearance and motion saliency are applied to the original spatial features and the temporal differences, respectively.

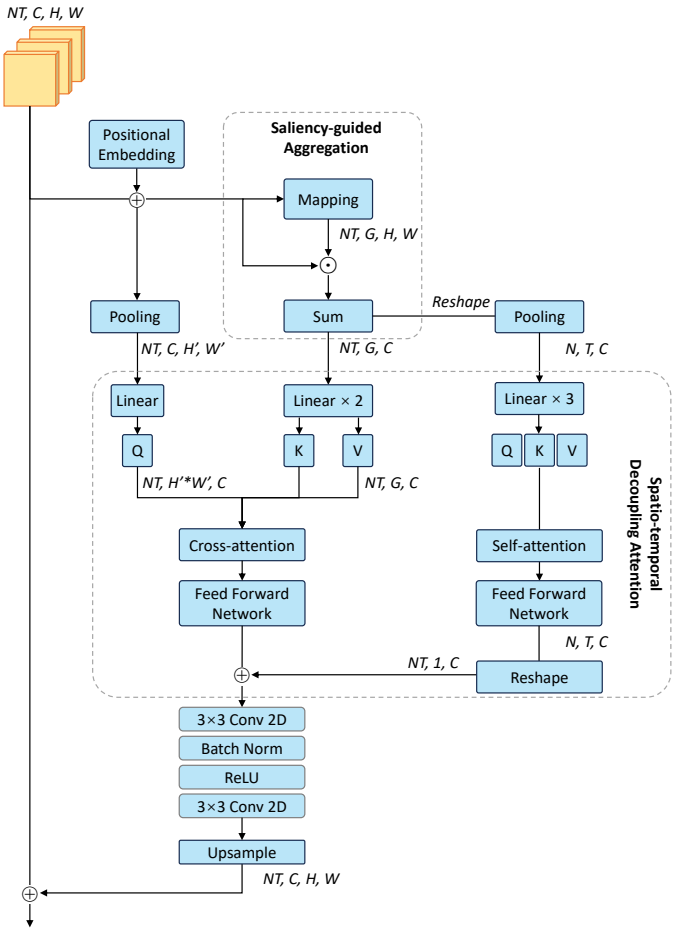


Fig. 5. The network of Saliency-guided Long-term Enhancement Module (SLEM). The saliency scores learned from the mapping network is used to aggregate prominent spatio-temporal tokens. Then, spatial attention is employed to focus on these informative features, and temporal attention is used to model global temporal evolution. For simplicity, we omit the Norm and Residual operation in attention mechanism.

provide positional cues. Then, the $Mapping$ module is applied to $\tilde{\boldsymbol{F}}$ to obtain saliency weights $\boldsymbol{W} \in \mathbb{R}^{N\times T\times G\times H\times W}$, where G is the number of aggregated tokens. Next, these weights are operated on spatio-temporal feature maps as follows:

$$\boldsymbol{F}_s = \tilde{\boldsymbol{F}}_r \times \boldsymbol{W}, \tag{7}$$

where $\tilde{\boldsymbol{F}}_r \in \mathbb{R}^{N\times T\times 1\times H\times W\times C}$ is the reshaped map, $\boldsymbol{F}_s \in \mathbb{R}^{N\times T\times G\times C}$ is the tokens aggregated after summation along H and W dimensions, and G is much smaller than $H \times W$.

2) Spatio-temporal Decoupling Attention: The decoupled spatio-temporal attention mechanism is implemented to effectively focus on significant spatial regions at the feature level and model long-term temporal information across entire videos. For the spatial attention, the pooled feature maps are used as queries, while the aggregated tokens $\boldsymbol{F}_s$ serve as keys and values. This not only concentrates on salient features but also reduces computational complexity due to our saliency-guided aggregation module.

Specifically, three linear layers are employed to generate the query, key, and value for multi-head attention [15]. Then, the attention scores computed from the query and key are used to weight the value, calculated as follows:

$$\boldsymbol{Q} = Linear(\tilde{\boldsymbol{F}}),\ \boldsymbol{K} = Linear(\boldsymbol{F}_s),\ \boldsymbol{V} = Linear(\boldsymbol{F}_s), \tag{8}$$

$$\boldsymbol{O} = Softmax(\frac{\boldsymbol{Q}^T\boldsymbol{K}}{\sqrt{d}})\boldsymbol{V}, \tag{9}$$

$$\tilde{\boldsymbol{F}}' = FFN(\boldsymbol{O}), \tag{10}$$

where $\tilde{\boldsymbol{F}}'$ is the updated feature map, d is the feature dimension of one head, and FFN is the feed forward network [15]. For simplicity, we omit the Norm and Residual operation.

After applying spatial attention, temporal attention is then employed to model long-term dynamics. To emphasize important features while avoiding less informative ones, we use the aggregated $\boldsymbol{F}_s$ instead of the original feature maps. Specifically, $\boldsymbol{F}_s$ are pooled to obtain the representation of each frame, denoted as $\boldsymbol{F}_t \in \mathbb{R}^{N\times T\times C}$. The temporal attention is calculated as follows:

$$\boldsymbol{Q}_t = Linear(\boldsymbol{F}_t),\ \boldsymbol{K}_t = Linear(\boldsymbol{F}_t),\ \boldsymbol{V}_t = Linear(\boldsymbol{F}_t), \tag{11}$$

$$\boldsymbol{O}_t = Softmax(\frac{\boldsymbol{Q}_t^T\boldsymbol{K}_t}{\sqrt{d}})\boldsymbol{V}_t, \tag{12}$$

$$\boldsymbol{F}_t' = FFN(\boldsymbol{O}_t), \tag{13}$$

TABLE I
ACTION RECOGNITION ACCURACY (%) ON UAV-HUMAN DATASET USING 8 FRAMES

Methods	Backbone	Frames	Inputs	Accuracy (%)
I3D [18]	ResNet101	8	540×960	21.1
FNet [19]	I3D	8	540×960	24.3
FAR [20]	I3D	8	540×960	29.2
FAR [20]	X3D-M	8	620×620	39.1
DiffFAR [21]	X3D-M	8	540×540	41.9
MITFAS [10]	X3D-M	8	620×620	46.6
SaStNet (Ours)	ResNet50	8	224×224	**52.3**

TABLE II
ACTION RECOGNITION ACCURACY (%) ON UAV-HUMAN DATASET USING 16 FRAMES

Methods	Backbone	Frames	Inputs	Accuracy (%)
X3D-M [22]	-	16	224×224	30.6
MViT [23]	-	16	224×224	24.3
FAR [20]	X3D-M	16	224×224	31.9
AZTR [24]	X3D-M	16	224×224	47.4
MITFAS [10]	X3D-M	16	224×224	50.8
MG Sampler [25]	-	16	224×224	53.8
PMI Sampler [26]	-	16	224×224	55.0
SaStNet (Ours)	ResNet50	16	224×224	**57.1**

TABLE III
ACTION RECOGNITION ACCURACY (%) ON DRONE ACTION DATASET USING 16 FRAMES

Methods	Frames	Inputs	Accuracy (%)
X3D-M [22]	16	224×224	83.4
FAR [20]	16	224×224	92.7
AP-TransNet [27]	32	224×224	97.2
SaStNet (Ours)	16	224×224	**97.5**

where $\boldsymbol{F}_t'$ represents the features learned from temporal attentions. Finally, the decoupled spatio-temporal attention features are added and aligned with the original maps $\tilde{\boldsymbol{F}}$ using a lightweight MLP as follows:

$$\tilde{\boldsymbol{F}}_o = Upsample(MLP(\tilde{\boldsymbol{F}}' + \boldsymbol{F}_t')) + \tilde{\boldsymbol{F}}, \tag{14}$$

where $\tilde{\boldsymbol{F}}_o$ is the comprehensive spatio-temporal representations, and MLP is consisted of Convolution, BatchNorm, and ReLu. By integrating the SLEM into the backbone, our model can adaptively attend to regions of interest across the video.

The cross-entropy loss $\mathcal{L}_{ce}$ is used to optimize the classification of actions as follows:

$$\mathcal{L}_{ce}(\hat{p}, p) = -\sum_{i=1}^{M} p_i \log \hat{p}_i, \tag{15}$$

where p is the ground truth, $\hat{p}$ is the prediction for a video, and M is the total number of classes.

III. EXPERIMENT

A. Datasets

UAV-Human [6] is currently the largest benchmark dataset for UAV action recognition. It includes 67,428 videos performed by 119 subjects and covers various urban and rural scenes. There are 155 categories in the UAV-Human dataset, which includes daily and interaction activities, such as walking together closely. We use the same training and test splits as specified in [6].

The Drone Action dataset [2] is collected using low-altitude flight UAVs and includes 13 categories, such as walking, jogging, and running. It contains 240 videos and 66,919 frames. The challenges of this dataset stem from fine-grained actions and dynamic background. The same training and test splits are used as specified in [2].

Something-Something [16] is a large-scale action recognition dataset comprising 174 fine-grained classes, such as opening something or closing something. Following [17], our method is trained on 86K videos and evaluated on 11K videos.

B. Training Details

All experiments are conducted on 8 NVIDIA 2080Ti GPUs. The Pytorch and Python versions are 1.7.1 and 3.8.5, respectively. Following the previous works [17], [28], [29], the pretrained ResNet50 [14] on ImageNet is used for initialization, and the same data augmentation strategies are utilized. We train the model using a batch size of 64 and the optimizer of SGD. The initial learning rate is 0.01. The model is trained for 60 epochs, with the learning rate decreasing at epochs 30, 45, and 55. The resolution of the videos is 224×224 pixels. The number of sampled frames is 8 or 16. Following [30], the number of aggregated key tokens G is set to 8. The Drone Action dataset needs about 1 hour for training. The training time for the UAV-Human dataset and Something-Something dataset are about 0.5 days and 2 days, respectively.

C. Evaluation Metrics

We use accuracy as the evaluation metric for action recognition. Accuracy measures the proportion of correct predictions out of all predictions and is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{16}$$

where TP is the number of correctly identified positive samples, TN represents the number of correctly identified negative samples, FP indicates the number of negative samples incorrectly classified as positives, and FN is the number of positive samples incorrectly classified as negatives.

D. Results and Analysis

1) UAV-Human: In Table I, we compare our SaStNet with other state-of-the-art methods using 8 frames as the input. FNet [19] and FAR [20] employ 3D convolution-based backbones, and their performances are suboptimal. This may be attributed to the difficulty in optimizing 3D networks during training. Additionally, X3D-M [22] is a backbone optimized through the architecture search of ResNet. MITFAS [10] utilizes X3D-M as the backbone and achieves an accuracy of 46.6%. Our SaStNet uses the classic ResNet50 as the backbone and obtains an accuracy of 52.3%, marking a 5.7% improvement over MITFAS [10]. This demonstrates the effectiveness of our saliency-guided spatio-temporal modeling network.

In Table II, the results of the UAV-Human dataset using 16 frames are presented. The recent method MITFAS [10]

TABLE IV
ACTION RECOGNITION ACCURACY (%) ON SOMETHING-SOMETHING DATASET USING 8 FRAMES

Methods	Backbone	Frames	Inputs	Accuracy (%)
TSM [29]	ResNet50	8	224×224	45.6
TEINet [31]	ResNet50	8	224×224	47.4
TEA [32]	Res2Net50	8	224×224	48.9
SmallBig [33]	ResNet50	8+16	224×224	50.4
CT-Net [34]	ResNet50	8	224×224	50.1
MDAF [35]	ResNet50	8	224×224	49.1
STBiM [36]	ResNet50	8	224×224	50.4
SaStNet (Ours)	ResNet50	8	224×224	**52.4**

TABLE V
ACTION RECOGNITION ACCURACY (%) ON SOMETHING-SOMETHING DATASET USING 16 FRAMES

Methods	Backbone	Frames	Inputs	Accuracy (%)
TSM [29]	ResNet50	16	224×224	47.2
TEINet [31]	ResNet50	16	224×224	49.9
TEA [32]	Res2Net50	16	224×224	51.9
CT-Net [34]	ResNet50	16	224×224	52.5
SaStNet (Ours)	ResNet50	16	224×224	**53.5**

initially employs a detector to locate foreground objects in each frame and then utilizes the optimized backbone to process the cropped sequence. In contrast, we propose multiple saliency-guided feature enhancement modules for end-to-end spatio-temporal modeling. Our SaStNet outperforms MITFAS [10] by 6.3% in the recognition accuracy. The recent MG Sampler [25] and PMI Sampler [26] achieve an accuracy of 53.8% and 55.0%, respectively. Our proposed SaStNet using 16 frames as inputs achieves an accuracy of 57.1%. Moreover, these two methods utilize an additional large dataset for pre-training [26], while our proposed SaStNet trains the model from scratch. This demonstrates that our learnable saliency modules effectively model the dynamics of foreground objects.

2) Drone Action: In Table III, we present the performance comparison with other advanced methods on the Drone Action dataset using 16 frames. The recent methods X3D-M [22] and FAR [20] achieve accuracies of 83.4% and 92.7%, respectively. With the same number of frames and input resolution, our proposed SaStNet achieves an accuracy of 97.5% Additionally, AP-TransNet [27] employs 32 frames as inputs, and our method also achieves competitive results. These demonstrate the effectiveness of our saliency-guided sptio-temporal modeling network.

3) Something-Something: Table IV compares the performance of our SaStNet with other methods on Something-Something dataset. The classic method TSM [29] introduces temporal shift convolution within the blocks of later ResNet stages, achieving an accuracy of 45.6%. TEA [31] and CT-Net [34] design multiple pluggable modules for spatio-temporal modeling, achieving an accuracy of 48.9% and 50.1%, respectively. The MDAF [35] method captures effective information with spatial, temporal, and channel attention, achieving an accuracy of 49.1%. The STBiM [36] method proposes the adaptive and bidirectional motion difference aggregation module for local and global modeling, obtaining an accuracy of 50.4%. The accuracy of our proposed SaStNet is 52.4%, which integrates saliency-guided modules and effectively captures the spatio-temporal evolution of objects.

Table V shows the results on the Something-Something dataset using 16 frames. Under the same setting, our proposed SaStNet achieves competitive results. Compared to the advanced methods TEA [31] and CT-Net [34], our SaStNet achieves improvements in accuracy of 1.6% and 1%, respectively. This demonstrates the effectiveness of the saliency-guided mechanism in action recognition.

TABLE VI
STUDY ON THE SALIENCY ENHANCEMENT IN SSEM

	Saliency Module	Accuracy (%)
A0	Without saliency	50.7
A1	Spatial saliency	51.4
A2	Temporal saliency	51.1
A3	Spatial and temporal saliency	**52.3**

TABLE VII
STUDY ON THE ATTENTION MECHANISM IN SLEM

	Attention Mechanism	Accuracy (%)
B1	Spatial attention without aggregation	51.2
B2	Spatial attention with aggregation	51.7
B3	Temporal attention	51.6
B4	Spatial and temporal attention	**52.3**

E. Ablation Studies

The ablation studies are conducted on the UAV-Human dataset using 8 frames.

1) The saliency enhancement in SSEM: In short-term modeling, the spatial and temporal saliency enhancement are evaluated independently to demonstrate their effectiveness. As shown in Table VI, without the saliency enhancement module, the accuracy of action recognition is 50.7%. Introducing spatial and temporal saliency increases performance by 0.7% and 0.4%, respectively. Combining all the saliency enhancement modules results in the highest accuracy, emphasizing the importance of capturing both appearance and motion information in short-term modeling.

2) The attention mechanism in SLEM: In Table VII, the effectiveness of decoupled spatio-temporal attention in long-term modeling is evaluated. First, introducing the saliency-guided aggregation improves the accuracy by 0.5% (B1→B2). This indicates that the selection and aggregation of spatial features help the model focus on significant foreground information. Additionally, using temporal attention alone brings an accuracy of 51.6%. Finally, experimental results demonstrate that combining spatial and temporal attention achieves the highest accuracy. This indicates the importance of exploring spatial saliency information and modeling global temporal relationships.

3) The effectiveness of SSEM and SLEM: In Table VIII, we further study SSEM and SLEM on action recognition. Following [29], C1 employs only 1D temporal convolution in the backbone of ResNet50, achieving an accuracy of 46.4%. Introducing the SSEM increases recognition accuracy by 5.1%. Combining SSEM and SLEM achieves the highest

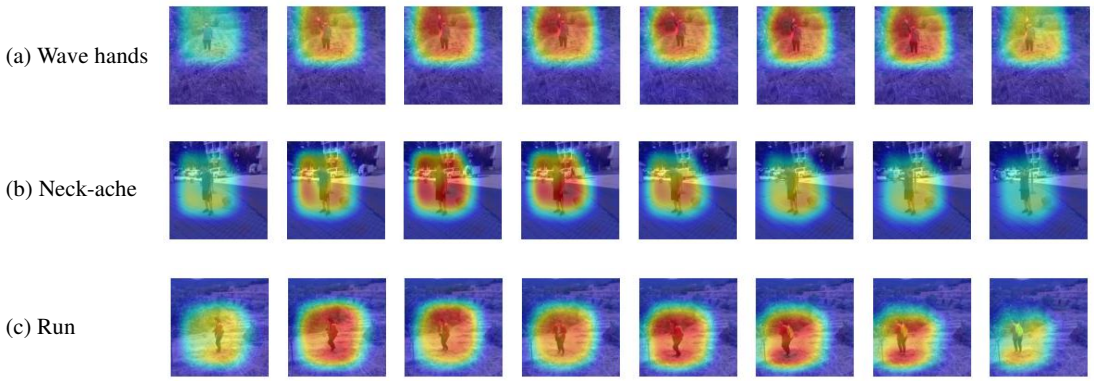


Fig. 6. The visualization of class activation mappings for three examples of the UAV-Human dataset. Our method explicitly focuses on foreground objects.

TABLE VIII
STUDY ON SSEM AND SLEM ON RESNET50 BACKBONE

	SSEM	SLEM	GFLOPs	Accuracy (%)
C1	×	×	33	46.4
C2	✓	×	34	51.5
C3	×	✓	35	49.7
C4	✓	✓	37	**52.3**

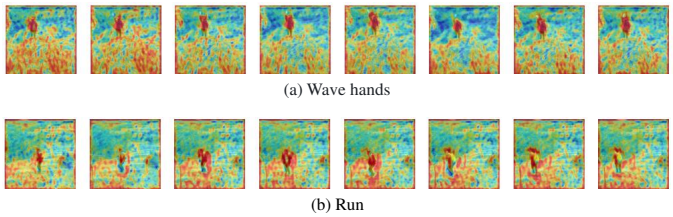


Fig. 7. The visualization of learned spatial saliency in the SSEM. The foreground objects in small proportion can be effectively captured.

accuracy with acceptable complexity. This indicates that short-term modeling within segments and long-term modeling across the entire video are both essential for action recognition.

F. Visualization

Fig. 6 illustrates the class activation mappings for uniformly sampled 8 frames. The three examples are sourced from the UAV-Human dataset, where foreground objects occupy a small proportion in the complex backgrounds. The visualization demonstrates our method effectively focuses on these foreground targets.

Furthermore, in Fig. 7, we visualize the learned spatial saliency in the short-term module. We can see that small targets can be effectively captured in a learnable manner, facilitating action recognition. This also demonstrates the effectiveness of our proposed saliency-guided spatio-temporal modeling network. Importantly, our method achieves this without additional bounding box and saliency annotations, effectively learning object representations within complex backgrounds.

IV. CONCLUSIONS, LIMITATIONS AND FUTURE WORKS

In this paper, we propose a saliency-guided spatio-temporal modeling network for UAV action recognition, termed SaSt-Net. Multiple learnable saliency-guided modules are designed to capture prominent information during short-term and long-term modeling. These modules are seamlessly integrated with classic backbones in a plug-and-play manner. With acceptable computational cost, our method achieves performance improvement on three benchmarks. Extensive ablation studies demonstrate the effectiveness of our method.

Although we achieve state-of-the-art results on the UAV-Human dataset, the accuracy is still limited due to numerous challenges, such as occlusion, distortion, and viewpoint variation. In the future, we will continue investigating these issues and explore more efficient architectures for UAV action recognition.

REFERENCES

[1] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-Action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 28–35, doi:10.1109/CVPRW.2017.267.

[2] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-Action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, 2019, doi:10.3390/drones3040082.

[3] A. G. Perera, Y. W. Law, T. T. Ogunwa, and J. Chahl, "A multiviewpoint outdoor dataset for human action recognition," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 405–413, 2020, doi:10.1109/thms.2020.2971958.

[4] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1717–1726, doi:10.1109/WACV45572.2020.9093511.

[5] U. Demir, Y. S. Rawat, and M. Shah, "TinyVIRAT: Low-resolution video action recognition," in *International Conference on Pattern Recognition*, 2021, pp. 7387–7394, doi:10.1109/icpr48806.2021.9412541.

[6] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021, pp. 16 266–16 275, doi:10.1109/cvpr46437.2021.01600.

[7] O. L. Barbed, P. Azagra, L. Teixeira, M. Chli, J. Civera, and A. C. Murillo, "Fine-grained pointing recognition for natural drone guidance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1040–1041, doi:10.1109/cvprw50498.2020.00528.

[8] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Predicting the Future: A jointly learnt model for action anticipation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5562–5571, doi:10.1109/iccv.2019.00566.

[9] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from uav-captured video sequences," *Pattern Recognition*, vol. 100, p. 107140, 2020, doi:10.1016/j.patcog.2019.107140 .

[10] R. Xian, X. Wang, and D. Manocha, "MITFAS: Mutual information based temporal feature alignment and sampling for aerial video action recognition," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2024, pp. 6625–6634, doi:10.1109/wacv57701.2024.00649.

[11] M. Khan, J. Ahmad, A. El Saddik, W. Gueaieb, G. De Masi, and F. Karray, "Drone-HAT: Hybrid attention transformer for complex action recognition in drone surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4713–4722.

[12] M. Assefa, W. Jiang, K. Gedamu, G. Yilma, B. Kumeda, and M. Ayalew, "Self-supervised scene-debiasing for video representation learning via background patching," *IEEE Transactions on Multimedia*, pp. 5500–5515, 2022, doi:10.1109/TMM.2022.3193559.

[13] Y. Kong, Y. Wang, and A. Li, "Spatiotemporal saliency representation learning for video action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1515–1528, 2021, doi:10.1109/tmm.2021.3066775.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778, doi:10.1109/cvpr.2016.90.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008, doi:10.48550/arXiv.1706.03762.

[16] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850, doi:10.1109/iccv.2017.622.

[17] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021, pp. 1895–1904, doi:cvpr46437.2021.00193.

[18] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, doi:10.1109/cvpr.2017.502.

[19] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with fourier transforms," *arXiv preprint arXiv:2105.03824*, 2021, doi:10.48550/arXiv.2105.03824.

[20] D. Kothandaraman, T. Guan, X. Wang, S. Hu, M. Lin, and D. Manocha, "FAR: Fourier aerial video recognition," in *European Conference on Computer Vision*, 2022, pp. 657–676, doi:10.1007/978-3-031-19836-6_37.

[21] D. Kothandaraman, M. Lin, and D. Manocha, "Differentiable frequency-based disentanglement for aerial video action recognition," *arXiv preprint arXiv:2209.09194*, 2022, doi:10.48550/arXiv.2209.09194.

[22] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 203–213, doi:cvpr42600.2020.00028.

[23] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE international conference on computer vision*, 2021, pp. 6824–6835, doi:10.1109/iccv48922.2021.00675.

[24] X. Wang, R. Xian, T. Guan, C. M. de Melo, S. M. Nogar, A. Bera, and D. Manocha, "AZTR: Aerial video action recognition with auto zoom and temporal reasoning," in *2023 IEEE International Conference on Robotics and Automation*, 2023, pp. 1312–1318, doi:icra48891.2023.10160564.

[25] Y. Zhi, Z. Tong, L. Wang, and G. Wu, "Mgsampler: An explainable sampling strategy for video action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1513–1522, doi:10.1109/ICCV48922.2021.00154.

[26] R. Xian, X. Wang, D. Kothandaraman, and D. Manocha, "Pmi sampler: Patch similarity guided frame selection for aerial action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6982–6991, doi:10.1109/WACV57701.2024.00683.

[27] C. Dhiman, A. Varshney, and V. Vyapak, "AP-TransNet: A polarized transformer based aerial human action recognition framework," *Machine Vision and Applications*, vol. 35, no. 3, p. 52, 2024, doi:10.1007/s00138-024-01535-1.

[28] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal Segment Networks: Towards good practices for deep action recognition," in *European conference on computer vision*, 2016, pp. 20–36, doi:10.1007/978-3-319-46484-8_2 .

[29] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 7083–7093, doi:10.1109/iccv.2019.00718 .

[30] M. S. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "TokenLearner: What can 8 learned tokens do for images and videos?" *arXiv preprint arXiv:2106.11297*, 2021, doi:arXiv.2106.11297.

[31] Z. Liu, D. Luo, Y. Wang, L. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and T. Lu, "TEINet: Towards an efficient architecture for video recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 669–11 676, doi:10.1609/aaai.v34i07.6836.

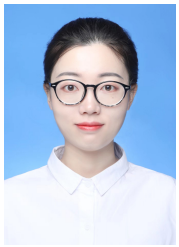
[32] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 909–918, doi:10.1109/cvpr42600.2020.00099.

[33] X. Li, Y. Wang, Z. Zhou, and Y. Qiao, "SmallBigNet: Integrating core and contextual views for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 1092–1101, doi:10.1109/cvpr42600.2020.00117.

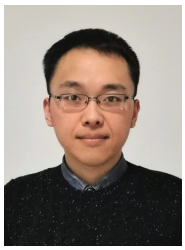
[34] K. Li, X. Li, Y. Wang, J. Wang, and Y. Qiao, "CT-Net: Channel tensorization network for video classification," *arXiv preprint arXiv:2106.01603*, 2021, doi:10.48550/arXiv.2106.01603.

[35] B. Wang, F. Chang, C. Liu, W. Wang, and R. Ma, "An efficient motion visual learning method for video action recognition," *Expert Systems with Applications*, vol. 255, p. 124596, 2024, doi:10.1016/j.eswa.2024.124596.

[36] L. Li, M. Tang, Z. Yang, J. Hu, and M. Zhao, "Spatio-temporal adaptive convolution and bidirectional motion difference fusion for video action recognition," *Expert Systems with Applications*, p. 124917, 2024, doi:10.1016/j.eswa.2024.124917.



Xiaoxiao Sheng is currently pursuing the Ph.D. degree in Shanghai Jiao Tong University, China. She received the master's degree with the School of Control Science and Engineering, Shandong University, China, in 2020. Her research interests include action recognition and video understanding.



Zhiqiang Shen is currently pursuing the Ph.D. degree in Shanghai Jiao Tong University, China. He received the master's degree with the School of Control Science and Engineering, Shandong University, China, in 2018. His current research interests include self-supervised representation learning and point cloud understanding.

Gang Xiao received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a full professor with the school of aeronautics and astronautics, Shanghai Jiao Tong University, director of Advanced Avionics and Intelligent Information Laboratory. His current research interests include image fusion and target tracking, avionics integration and simulation. From 2008 to 2016, he had published 40 papers and 2 books. He received the title of Shanghai Pujiang talent in 2016. He is a member of China aviation society information fusion branch. He was a Visiting Scholar with Cranfield University, UK (2006), University of California, San Diego, USA (2010), Southern Illinois University Edwardsville, USA (2014-2015), respectively.