SFMVIT : SLOWFAST MEET VIT IN CHAOTIC WORLD

Jiaying Lin[*12], *Jiajun Wen*[*12], *Mengyuan Liu*[†1], *Jinfu Liu*[2], *Baiqiao Yin*[2], *Yue Li*[2]

[1]National Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School
[2]School of Intelligent Systems Engineering, Sun Yat-sen University
[*] means co-first authors with equal contributions.
The corresponding author is Mengyuan Liu (e-mail: liumengyuan@pku.edu.cn).

ABSTRACT

The task of spatiotemporal action localization in chaotic scenes is a challenging task toward advanced video understanding. Paving the way with high-quality video feature extraction and enhancing the precision of detector-predicted anchors can effectively improve model performance. To this end, we propose a high-performance dual-stream spatiotemporal feature extraction network SFMViT with an anchor pruning strategy. The backbone of our SFMViT is composed of ViT and SlowFast with prior knowledge of spatiotemporal action localization, which fully utilizes ViT's excellent global feature extraction capabilities and SlowFast's spatiotemporal sequence modeling capabilities. Secondly, we introduce the confidence maximum heap to prune the anchors detected in each frame of the picture to filter out the effective anchors. These designs enable our SFMViT to achieve a mAP of 26.62% in the Chaotic World dataset, far exceeding existing models. Code is available at https://github.com/jfightyr/SlowFast-Meet-ViT.

***Index Terms*—** spatiotemporal action localization, chaotic scenes, dual-stream network, ViT, anchor pruning

1. INTRODUCTION

Spatiotemporal action localization, which involves localizing persons and recognizing their actions from videos, is an important task that has gained attention in recent years [7, 12, 6, 2, 13, 10]. Spatiotemporal feature modeling refers to learning the temporal and spatial characteristics present in videos as the most fundamental and critical part of action recognition tasks[21]. However, the subpar quality of feature extraction by current methods hampers the advancement of this field, leaving room for improvement in model performance.

The Chaotic World [11], a newly proposed dataset [8, 14, 7] in this task, presents more complex and chaotic scenes compared to previous datasets, thus imposing greater demands on comprehending intricate scenes and interactions among individuals. It's not uncommon to encounter difficulties when applying traditional video feature extraction modules such as I3D [1] or SlowFast [6] directly to datasets like Chaotic World. The complexity and chaotic nature of scenes in the Chaotic World dataset may necessitate more advanced or customized approaches for feature extraction and modeling to attain satisfactory results.

On the one hand, VideoMAEv2 [18] has achieved state-of-the-art performance across multiple downstream tasks by employing a straightforward yet effective video self-supervised pretraining paradigm. It is reasonable to expect that robust video feature extraction networks pre-trained on large-scale video datasets will perform effectively. SlowFast [6], a standard backbone network in the Video Recognition field, excels at capturing action features with strong temporal correlations, but it may be sensitive to disruptions in complex environments, resulting in reduced robustness across different scenes. Conversely, Vision Transformer [4], trained on extensive video data, demonstrates robustness in spatiotemporal feature modeling, capable of capturing diverse spatiotemporal features in complex scenes.

Considering the strengths of both models, we introduce the dual-stream **SFMViT** video spatiotemporal modeling network. This network utilizes the advantages of SlowFast and ViT to enhance the robustness and effectiveness of spatiotemporal feature modeling in video recognition tasks. In addition, the action detection of the Chaotic World dataset needs to focus on the action recognition of specific people, so effective actor detection is also very critical. To this end, we designed a **Confidence Pruning Strategy** to avoid redundant actor extracted by Yolo, thereby ensuring that ACAR [12] works better and improving the generalization and accuracy of the SFMViT model.

We conduct extensive experiments on the challenging Chaotic World dataset for spatiotemporal action localization. Our proposed SFMViT leads to significant improvements in localizing spacial-temporal actions. Our contributions are summarized as three-fold:

- We have introduced the dual-stream spatiotemporal feature modeling network SFMViT, which integrates SlowFast's ability to capture temporal features with Vision Transformer's capability in complex scene spatiotemporal modeling. This fusion enhances overall spatiotemporal model-

ing capabilities in complex scenes.

- We introduce a Confidence Pruning Strategy to find the most suitable upper anchor counts of the instances, which can be used to prune and filter out the optimal anchors while taking into account the efficiency and performance, increasing the mAP by nearly 2%.
- We achieve SOTA performances with significant margins on the Chaotic World datasets. At the time of submission, our method ranks first on the 2024 MMVRAC leaderboard.

2. RELATED WORK

2.1. Spatiotemporal Action Localization

The spatiotemporal action localization problem is typically composed of a detector for participant localization on keyframes and a 3D backbone for video feature extraction. Currently, it can be mainly classified into two types based on whether the detector is embedded within the backbone model. One approach implements localization and feature classification as two independent backbones in a two-stage pipeline, while the other employs a unified backbone for spatiotemporal localization and feature classification. Unified backbone models often face challenges such as high complexity, difficulty in optimization, and high training costs [3, 23, 20, 2]. Therefore, the most advanced methods currently [5, 12, 16, 18] tend to adopt a two-stage pipeline model. In our model, considering that unified backbone models incur higher training costs compared to two-stage pipeline models for achieving the same performance, we select the two-stage pipeline model as the baseline.

2.2. Video Foundation Models

Video Foundation Models (ViFMs) have emerged as a pivotal area of research in artificial intelligence [9, 19, 18], with a focus on constructing models capable of comprehending and generating video content through extensive learning from video data. These models have exhibited remarkable performance across various downstream tasks, including action recognition, video-text retrieval, video question answering, and video generation.

InternVideo2 [19] represents a recent advancement achieved through progressive training paradigms, demonstrating optimal performance in action recognition, video-text tasks, and video-centric dialogues. Its integration of diverse self-supervised learning frameworks highlights the potential for enhancing the performance of video foundation models. VideoMAEv2 [18] has trained a more generalized video foundation model capable of adapting to various downstream tasks by scaling and diversifying pre-training video data. These models and methods play a crucial role in advancing video foundation model research and applications, greatly enhancing the model's performance in capturing video features.

3. METHOD

3.1. Overall Architecture

We first introduce our overall framework for action localization, where the proposed dual-stream spatiotemporal feature modeling network SVMViT is the key module as the backbone part. The framework is designed to detect all individuals in an input video clip and estimate their action labels. It is constructed using a pre-existing person detector (e.g. YOLOv9 [17]) and a video backbone network (e.g. I3D [1]). The features of individuals and their surroundings are subsequently analyzed by the established ACAR module [12] , which incorporates a long-term Actor-Context Feature Bank for the final action prediction.

In detail, the person detector operates on the key frame of the input clip and obtains N detected actors. The detection boxes obtained from the person detector will be further used as annotations for our backbone network, guiding our final action classification of individuals. Simultaneously, our proposed visual feature extraction model SFMViT will extract global spatiotemporal features from input video clips. These feature maps will undergo average pooling along the temporal dimension to reduce computational costs, resulting in feature maps like $X \in \mathbb{R}^{C\times H\times W}$, and C, H, W are channel, height and width respectively. Guided by annotations, these feature maps will further produce a series of N actor features, each containing spatiotemporal context features. Subsequently, these actor features will be fed into the ACAR module for higher-order relationship reasoning, culminating in the classification of the inferred features.

3.2. SFMViT Module

SFMViT. Leveraging the inherent advantages of the aforementioned architectures, we have adopted the ViT model as our video feature extractor. Additionally, we have incorporated SlowFast as an auxiliary network with domain knowledge in Video Recognition. This integration has led to the development of our proposed SFMViT model architecture. More specifically, we obtain the fused features of the SlowFast and ViT models as:

$$f_{x,y}^{\text{SFMViT}} = [f_{x,y}^{V}, f_{x,y}^{S}, f_{x,y}^{F}] \tag{1}$$

x,y represent the spatial location (x, y). $[\cdot,\cdot,\cdot]$ denotes concatenation along the channel dimension. f^S and f^F refer to feature maps obtained from Slow pathway and Fast pathway respectively and both of them have already undergone average pooling along the temporal dimension. Especially, f^V obtained from the ViT model needs to undergo 3D convolution first to achieve the same spatial dimensions as f^S and then undergo temporal average pooling. Finally, we can get a fused feature map like $f_{x,y}^{SFMViT} \in \mathbb{R}^{C\times H\times W}$, where C equals the sum of the channel numbers of the three original feature maps.

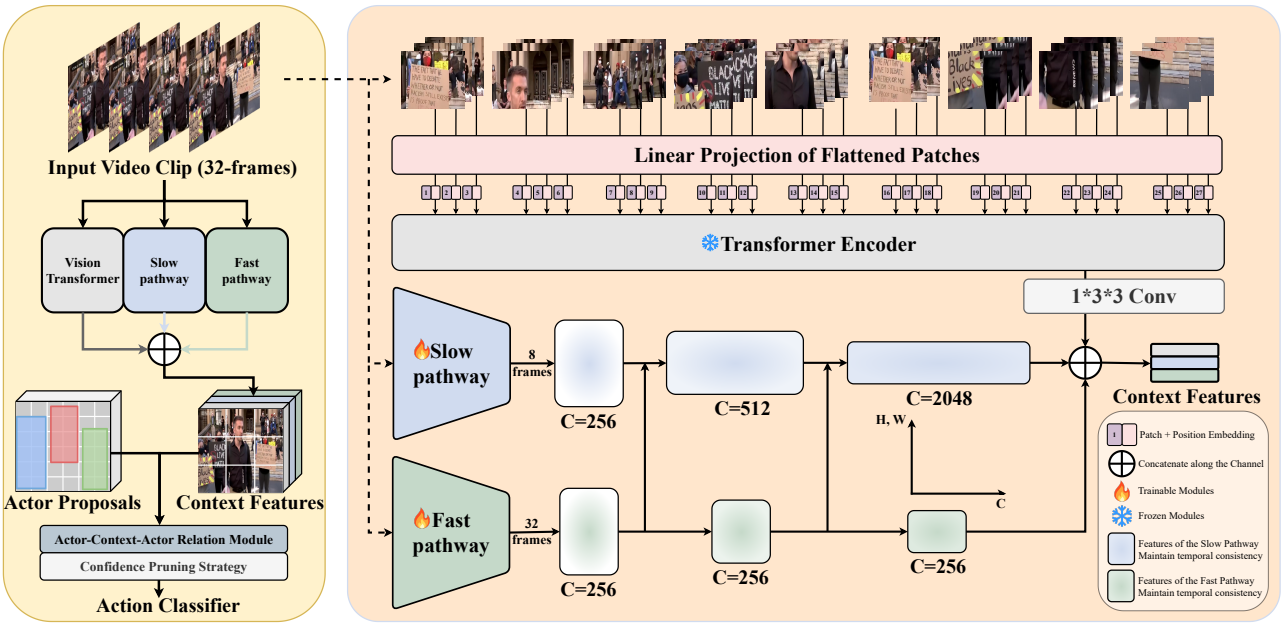

Fig. 1. **Framework of our proposed SFMViT**. The left yellow background box diagram shows the architecture flow of our entire spatiotemporal action localization task, and the right orange background box diagram shows the details of the SFMViT module we propose. The 32-frame input video clip goes through the ViT and SlowFast dual-stream networks to obtain spatiotemporal context features. The coordinate axes illustrate the changes in the spatial size and channel dimensions of the feature maps in the SlowFast branch. Anchors are obtained by YOLO series object detectors based on the detection results at keyframes to get individual features from context features. These features of the target individuals go through high-order relation reasoning by the ACAR module [12] and our proposed Confidence Pruning Strategy before the final action category classification.

Vision Transformer. ViT employs a global attention mechanism by transforming videos into patches across spatial and temporal dimensions. Combined with the Transformer architecture, it can effectively captures the overall information available in videos. We select the pretrained ViT models on extensive datasets such as Kinetics700 [8] and AVA [7] to form a robust foundation for transfer learning across diverse downstream tasks.

SlowFast Pathways. The SlowFast feature extraction branch can capture strongly correlated temporal action signals and assist in extracting domain-specific features in our model. We follow the basic instantiation of SlowFast 8×8, R50 as stage res_5 defined in [6]. For the Fast pathway, it has a lower channel capacity, such that it can better capture motion while trading off model capacity. Spatial downsampling is performed with stride 2^2 convolution in the center filter of the first residual block in each stage of both the Slow and Fast pathways.

3.3. Confidence Pruning Strategy

It needs to be clarified that mAP is the mean average precision of each action category:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP(a_i) \tag{2}$$

N represents the total number of action categories ; a_i represents the i-th action class; and $AP()$ denotes the average precision for a particular class:

$$AP = \frac{1}{M}\sum_{k=1}^{k} Precision(k) \tag{3}$$

M denotes the number of positive samples in the test set, and $Precision(s)$ denotes the precision for the top k test samples.

Because the task of spatiotemporal action localization focuses solely on the actions of key individuals, the detector, if left unprocessed, tends to generate a significant number of unnecessary anchors, leading to resource wastage. Previous methods typically involve extensive training of the detector on the target dataset, which requires substantial resources, takes a long time, and exhibits poor transferability. To address this problem, SFMViT adopts a strategy to store the predicted anchors of each $\{img\_id, time\}$ pair in the maximum heap based on the confidence score, and set the maximum number of nodes as the capacity, so that keep the top capacity anchors with the highest confidence.

It is worth noting that setting the capacity too high or too low can lead to performance degradation and resource wastage, as we will demonstrate in Section 4.4. Furthermore, the optimal capacity value exhibits randomness across different datasets and models, making it inappropriate to fix a

single capacity value, as done in prior approaches [12]. Our SFMViT employs a multi-threaded iterative approach (Algorithm 1) , which empirically sets the range for capacity enumeration (abbreviated as capacity_range) and quickly identifies the optimal capacity within this range. This approach prunes the predicted results to retain favorable anchors for the model, thereby capturing the optimal mAP value. Our strategy significantly improves model performance in a short time, aligning with the requirement for fast results in competitive scenarios.

Algorithm 1: Confidence Pruning Strategy

Input : Dataset, capacity_range
Output: mAP for each capacity value

```
for cap ∈ capacity_range do
    Initialize a tree structure with capacity = cap;
    foreach data point dp in Dataset do
        if number of entries in the tree < cap then
            Insert (data point into the tree);
        end
        else if score of data point > score of least confident
          entry in the tree then
            Replace (least confident entry with current data
              point);
        end
    end
    CalculateMAP();
end
```

4. EXPERIMENTS

4.1. Dataset

We evaluate our model on the Chaotic World dataset [11], a chaotic scene dataset containing multi-task annotations, with 299,923 annotated instances for spatiotemporal action localization, with 50 action categories, and with time ranges from 5 to 200 seconds per action (10 seconds on average). Officially, the data labels are divided into training sets and test sets according to AVA [7] format, and we experimented according to officially provided training and test sets.

4.2. Experimental Setup

Person Detector. We used Yolov8 as our base object detector, which is pre-trained on AVA [7] and WiderPerson [22], then we fine-tuned it on our Chaotic World dataset. Due to different training strategies and varying confidence threshold settings, there was a significant impact on the results. Therefore, besides testing on the detector, we also conducted ablation experiments using Ground Truth anchors as the result of the first-stage object detection.

Backbone Network. We will use SFMViT proposed above as our video feature extractor. Specifically, we freeze the ViT model branch and set the SlowFast model branch to a trainable state.

Training and inference. All experiments were conducted on 4 * GeForce RTX 3090 GPUs. For the Chaotic World dataset, all given segments were cropped to the central 32 frames. Our model sets the capacity enumeration range to [50, 2200] and utilizes the stochastic gradient descent (SGD) optimizer for training, with Nesterov momentum of 0.9 and weight decay of 10^{-7}. We employed cross-entropy as the loss function. The training epochs and learning rate were set to 90 and 0.01, respectively, with a batch size of 3 per GPU for each training epoch.

Model	Backbone	Pre-training Dataset	mAP
ACAR [12]	SF-R50-NL	K400(ft:AVA v2.2)	9.88
Hiera [13]	Hiera-Large	K400	10.94
ACRN [15]	I3D [1]	/	11.91
ACRN [15]	SlowFast [6]	/	12.90
SlowFast [6]	SF-R101-NL	K600	13.24
ACAR [12]	SF-R101-NL	K600	13.99
IntelliCare [11]	I3D [1]	/	16.25
SFMViT‡	ViT-Giant+SF-R50	K400(ft:AVA v2.2)	**26.62**

Table 1. Comparison of the performance and efficiency improvements of EVAD with higher resolutions. "ft" denotes fine-tuning on this dataset. "/ " represents unclear about its pre-training dataset.

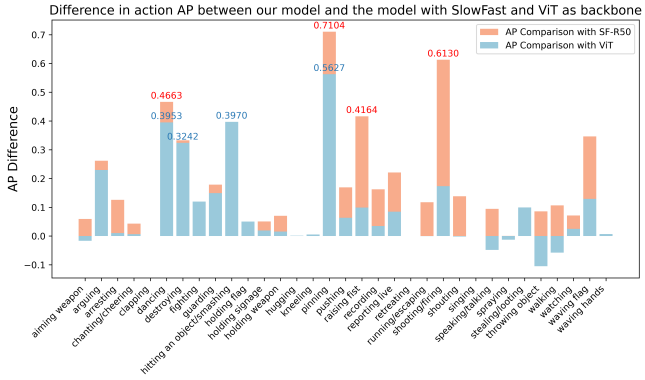

Fig. 2. Difference in action AP between our model and the model with SlowFast and ViT as backbone. The top 5 action categories in terms of absolute value of AP change in the table are labeled with spec change values.

4.3. Comparison with the state-of-the-art methods

We compared our SFMViT with state-of-the-art methods on the Chaotic World dataset, and SFMViT outperformed all previous models, as shown in Table 1. Currently, the best-performing model on the Chaotic World dataset is IntelliCare proposed in its paper, which is a multi-task model that leverages information exchange between tasks to aid performance. Our model achieves a performance improvement of 10.37% over IntelliCare in the task of spatiotemporal action localization, even without additional information.

4.4. Ablation Studies

We performed an in-depth ablation study to investigate the effectiveness of our design in SFMViT. All results are reported on the Chaotic World dataset. Table 2a shows the

Ablation object	Method	frame-mAP
Backbone (capacity = 50)	w/o SF	18.25
	w/o ViT	9.88
	SFMViT	**24.67**
Capacity (backbone: SFMViT)	cap=10	22.46
	cap=50	24.67
	cap=500	25.87
	cap=1599(OURS)	**26.62**
	cap=3000	22.18

(a) **Comparison of frame-mAP under different ablation settings.**

Backbone Type	Model	Backbone	Pre-training Dataset	mAP
Single-stream	ACAR [12]	SF-R50-NL	K400(ft:AVA v2.2)	31.25
	Hiera [13]	Hiera-Large	K400	32.88
	ViT [4]	Hiera-Large	K400	37.57
Multi-stream	SlowFast+ViT +Hiera‡	SF-R50-NL+ViT-Giant +Hiera-Large	K400(ft:AVA v2.2) +K400	29.54
	Hiera+ViT‡	Hiera-Large+ViT-Giant	K400,AVA v2.2	33.89
	SFMViT‡	ViT-Giant+SF-R50-NL	K400(ft:AVA v2.2)	**42.28**

(b) **Comparison of Chaotic World datasets using ground truth instead of detector results.**

Table 2. Comparison with state-of-the-art models on the Chaotic World dataset. We have bolded the best mAP."ft" denotes fine-tuning on this dataset.‡ denotes the model contains a multi-stream backbone. "cap" is the abbreviation of capacity. Specifically, in Table (b), the SlowFast+ViT+Hiera model, the Hiera-Large backbone is pre-trained on K400, while SF-R50-NL and ViT-Giant are pre-trained on K400 and fine-tuned on AVA v2.2.

frame-mAP comparisons under different ablation settings. Here, "cap" stands for capacity, and the ablation study regarding capacity is conducted with SFMViT as the baseline.

Backbone Comparison. As shown in Table 2a, under the condition of capacity=50, SFMViT's mAP is significantly higher than that of single-stream backbone models. This illustrates that our dual-stream backbone complements each other, and the absence of either stream would result in a significant decrease in mAP. In Fig.2, we show a comparison of the action APs of the dual-stream backbone and each stream backbone (SlowFast-ResNet50 and ViT-Giant) in the Chaotic World dataset. Hereafter, we refer to the model with SlowFast-ResNet50 as the backbone as SlowFast*, and the model with ViT-Giant as the backbone as ViT*.

SFMViT outperforms single-stream backbone models in most action categories and even surpasses all action APs of SlowFast*. This is because the dual-stream backbone combines the powerful global feature extraction capability of ViT with the excellent spatiotemporal modeling capability of SlowFast, a conclusion confirmed in [4, 6].

However, compared to ViT*, SFMViT exhibits a slight decrease in AP for five action categories such as "throwing object" and "walking". This is primarily because SlowFast performs poorly in these categories, leading to a decline in the performance of the fusion model. The reasons can be summarized as follows: I. These actions are not time-sensitive. Taking the "throwing object" action as an example, the key recognition features mainly focus on the movement of the arm from backward swing to forward throwing, where the speed or accuracy at specific time points is not critical, but rather the trajectory and posture of the arm. II. SlowFast exhibits a strong bias in action judgment, resulting in a high misclassification rate. SlowFast* has a high misclassification rate for "walking", often confusing it with categories such as "holding signage" that have more training data instances or high temporal similarity. Although the introduction of SlowFast may cause some action APs to decrease slightly, due to SlowFast's advantages in spatiotemporal modeling and prior knowledge

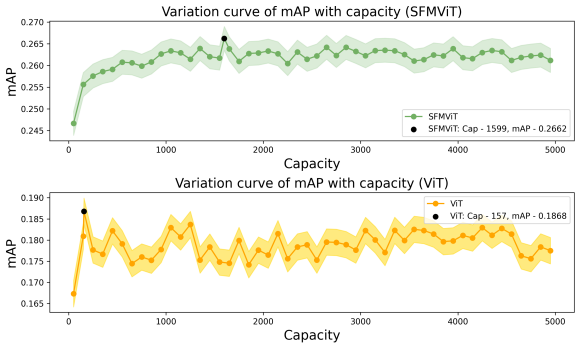

Fig. 3. Curve of mAP with capacity for SFMViT and ViT*. The envelope outside the curve is centered on the curve, with the curve value ± standard deviation at that point as the upper and lower limits.

of spatiotemporal action localization, it can assist ViT and improve model expression capabilities in a short time.

Comparison of backbone combination methods. Not all advanced video feature extraction models or multi-stream model combinations can achieve the effectiveness of SFMViT. We compared various advanced video backbones such as Hiera and ViT as single-stream backbones and further explored combinations of these backbones into dual-stream or triple-stream configurations. To minimize the bias introduced by detectors on different models and datasets, we replaced detector annotations with ground truth, and conducted comparisons between SFMViT and other methods on a unified baseline, as shown in Table 2b. The mAP of SFMViT exceeds that of all single-stream and multi-stream backbone models, indicating that the dual-stream configuration of SFMViT is the most appropriate. Additionally, some backbone models (Hiera+ViT, SlowFast+ViT+Hiera) exhibit lower performance after fusion, suggesting that incompatible backbones may interfere with each other, further validating the effectiveness of SFMViT's dual-stream configuration.

Confidence Pruning Strategy: Capacity Comparison.

Fig. 3 shows how the model's mAP changes with capacity, with the size of the envelope reflecting the standard deviation of the mAP change. Capacity is the upper limit of the number of anchors predicted for each $\{img\_id, time\}$ key-value pair. As the capacity increases, mAP first rises and then slowly decreases. SFMViT obtains a maximum mAP value of 26.62% when capacity=1599, and ViT* obtains a maximum mAP value of 18.68% when capacity=157. If the upper limit of the number of anchors is set too small, correctly predicted anchors may be missed, thus affecting performance; if it is set too large, it will affect efficiency and consume memory. In addition, for different data and models, the mAP value is random, so it is not suitable to fix a capacity value to find the best mAP. Therefore, we can find an optimal capacity in a short time through multi-thread iteration, thereby capturing the optimal mAP value.

5. CONCLUSIONS

The Chaotic World dataset contains unique fine-grained complex actions such as destructive behaviors and lifesaving behaviors. In response to this characteristic: I. SFMViT leverages expert knowledge, pre-trained on large-scale datasets, and fine-tuned on Chaotic World. II. Its dual-stream backbone combines ViT for global features and SlowFast for spatiotemporal modeling. III. SFMViT introduces a Confidence Pruning Strategy, which can prune and select the optimal anchors, unleashing the potential of the model.

However, due to time constraints: I. Our fusion method of the dual-stream backbone is simplistic. II. Our model's performance relies heavily on the detector. In the future, we will have more time to optimize these shortcomings to fully realize the potential of our model.

6. FOUNDATION

This work was supported by National Natural Science Foundation of China (No. 62203476), Natural Science Foundation of Shenzhen (No. JCYJ20230807120801002).

References

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

[2] Lei Chen et al. "Efficient video action detection with token dropout and context refinement". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 10388–10399.

[3] Shoufa Chen et al. "Watch only once: An end-to-end video action detection framework". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 8178–8187.

[4] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

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

[6] Christoph Feichtenhofer et al. "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6202–6211.

[7] Chunhui Gu et al. "Ava: A video dataset of spatio-temporally localized atomic visual actions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6047–6056.

[8] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[9] Kunchang Li et al. "Unmasked teacher: Towards training-efficient video foundation models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 19948–19960.

[10] Mengyuan Liu, Fanyang Meng, and Yongsheng Liang. "Generalized Pose Decoupled Network for Unsupervised 3d Skeleton Sequence-based Action Representation Learning". In: *Cyborg and Bionic Systems* 2022 (2022), p. 0002.

[11] Kian Eng Ong et al. "Chaotic World: A Large and Challenging Benchmark for Human Behavior Understanding in Chaotic Events". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 20213–20223.

[12] Junting Pan et al. "Actor-context-actor relation network for spatio-temporal action localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 464–474.

[13] Chaitanya Ryali et al. "Hiera: A hierarchical vision transformer without the bells-and-whistles". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29441–29454.

[14] Gunnar A Sigurdsson et al. "Hollywood in homes: Crowdsourcing data collection for activity understanding". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 510–526.

[15] Chen Sun et al. "Actor-centric relation network". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 318–334.

[16] Jiajun Tang et al. "Asynchronous interaction aggregation for action detection". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer. 2020, pp. 71–87.

[17] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information". In: *arXiv preprint arXiv:2402.13616* (2024).

[18] Limin Wang et al. "Videomae v2: Scaling video masked autoencoders with dual masking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14549–14560.

[19] Yi Wang et al. "InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding". In: *arXiv preprint arXiv:2403.15377* (2024).

[20] Tao Wu et al. "Stmixer: A one-stage sparse action detector". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14720–14729.

[21] Saining Xie et al. "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 305–321.

[22] Shifeng Zhang et al. "WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild". In: *IEEE Transactions on Multimedia* 22.2 (2020), pp. 380–393. DOI: 10.1109/TMM.2019.2929005.

[23] Jiaojiao Zhao et al. "Tuber: Tubelet transformer for video action detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13598–13607.