Advances in Social Science and Culture
ISSN 2640-9682 (Print) ISSN 2640-9674 (Online)
Vol. 7, No. 5, 2025
www.scholink.org/ojs/index.php/assc

Original Paper

A Review of Spatio-Temporal Action Detection Models in Computer Vision Based on Deep Learning

Mingzhu Zhang[1,a*]

[1] College of Information Engineering, Xi'an FanYi University, Xi'an 710105, Shaanxi, China

[a] Email: 33046404@qq.com

[*] Corresponding Author

Received: August 29, 2025 Accepted: October 08, 2025 Online Published: October 28, 2025

doi:10.22158/assc.v7n5p140 URL: http://dx.doi.org/10.22158/assc.v7n5p140

Abstract

Spatio-Temporal Action Detection (STAD) is a cutting-edge task in computer vision, aiming to recognize and localize action categories, temporal intervals, and spatial positions from videos. With the development of deep learning, methods based on 3D CNNs, two-stream networks, and Transformers have emerged continuously, greatly advancing this field. This paper systematically reviews the mainstream spatiotemporal action detection models in recent years, analyzes their key modules, advantages, disadvantages, and development trends, and summarizes commonly used datasets and evaluation metrics.

Keywords

Spatiotemporal Action Detection, Deep Learning, 3D CNN, Transformer, Feature Fusion, Interaction Modeling

1. Introduction

With the rapid development of computer vision technology and deep learning algorithms, video understanding has become one of the core research directions in the field of artificial intelligence. As a key task in video understanding, spatiotemporal action detection aims to simultaneously recognize action categories, localize spatial positions, and detect temporal intervals from continuous video sequences, integrating the technical capabilities of object detection, action recognition, and temporal analysis. In intelligent surveillance scenarios, spatiotemporal action detection can real-time detect abnormal behaviors such as fighting and falling and trigger early warnings; in sports analysis, it can accurately locate the spatiotemporal range of athletes' technical movements to assist in training optimization; in human-computer interaction, it can realize natural device control by recognizing users'

gestures and body movements; in medical and health fields, it can analyze the standardization of patients' movements in rehabilitation training to assist in rehabilitation evaluation.

Traditional spatiotemporal action detection methods rely on handcrafted features (such as HOG and optical flow). Limited by the expressive power of these features, they exhibit poor robustness in complex scenarios involving dynamic backgrounds, occlusions, and multi-object interactions. With the breakthroughs in deep learning, methods based on architectures like Convolutional Neural Networks (CNNs) and Transformers have enabled automatic learning of action features, significantly improving detection accuracy and generalization ability. However, three core challenges remain: first, the lack of global contextual information—most models focus on local features while ignoring interactive information between actors, scenes, and objects, leading to high misjudgment rates for actions with similar postures; second, insufficient modeling of multi-dimensional interaction relationships in complex scenarios, making it difficult to handle spatial interactions between “people and people” or “people and objects” as well as temporal interactions between “frames and frames,” which limits detection accuracy in complex environments; third, the dilemma of balancing accuracy and real-time performance—models such as 3D CNNs can effectively capture spatiotemporal features but have high computational complexity, while lightweight models tend to lose key information, making them difficult to deploy on edge devices.

Based on existing research results, this paper systematically sorts out the technical evolution of deep learning-based spatiotemporal action detection models in computer vision, categorically analyzes mainstream model architectures (CNN-based, Transformer-based, and multi-interaction relationship modeling-based), summarizes key technical innovations (feature enhancement, multi-scale fusion, and interaction relationship modeling), and verifies model performance through comparative analysis of experimental data.

2. Classification of Technical Architectures for Spatiotemporal Action Detection Models

Based on differences in core architectures and feature modeling methods, current deep learning-based spatiotemporal action detection approaches can be categorized into three types: Convolutional Neural Network (CNN)-based methods, Transformer-based methods, and multi-interaction relationship modeling-based methods. Each type exhibits significant differences in feature extraction capability, interaction modeling ability, and applicable scenarios.

2.1 Methods Based on Convolutional Neural Networks (CNNs)

Benefiting from the advantages of local feature extraction and parameter sharing, CNN has become the fundamental architecture for spatiotemporal action detection, which can be further subdivided into 2D CNN-based methods, 3D CNN-based methods, and two-stream network-based methods.

(1) methods Based on 2D CNNs

The 2D convolutional neural network takes static images as input and gradually extracts spatial features from low-level to high-level through multiple layers of convolution operations. The core idea is to

significantly reduce the number of network parameters through the local connection of convolution kernels and the sharing of weights, while retaining the hierarchical information of the spatial structure.

Kalogetion et al. proposed the ACT model, which uses a shared-weight 2D CNN backbone to extract features from consecutive K-frame image sequences, addressing the issue of temporal discontinuity in a single frame and reducing the ambiguity in action prediction. On the UCF101-24 dataset, the frame-mAP reached 67.1% and the FPS was 25 (Kalogeiton, Weinzaepfel, Ferrari, & Schmid, 2017). Li et al. proposed the MOC (No Anchor Box Action Detector) based on ACT, modeling action instances as "temporal motion trajectories of frame-level action centers", abandoning the traditional anchor box design. On the UCF101-24 dataset, the frame-mAP was improved to 78.0% and the FPS reached 53 (Li, Wang, Wang, & Wu, 2020).

(2) method based on 3D CNN

The 3D CNN directly models the "space-time" combined features through three-dimensional convolution kernels, naturally capturing the motion information between frames, and is suitable for complex action detection.In 2013, Ji et al. were the first to introduce 3D CNN into the task of temporal-spatial feature extraction. By extending the traditional 2D CNN operations, they enabled it to perform convolution directly on video sequences, capturing the temporal-spatial correlations of actions (Ji, Xu, Yang et al., 2012). Subsequently, Tran et al. proposed the C3D model, which employed 3×3×3 convolutional and pooling kernels. The effectiveness of 3D CNN was verified on datasets such as UCF101 and HMDB-51 (Carreira & Zisserman, 2017). Carreira et al. proposed the I3D model, which expands the 2D CNN convolution kernels into 3D. While retaining the spatial feature extraction capability of 2D CNN, it enhances the temporal feature capture. On the Kinetics dataset, the accuracy of action recognition reached 78.4% (Qiu, Yao, & Mei, 2017).

To address the high computational cost issue of 3D CNN, researchers have proposed an optimization solution: Qiu et al. proposed the P3D model, which decomposes the 3D convolution into 1×3×3 spatial convolution and 3×1×1 temporal convolution, reducing the parameter quantity by 40% (Qiu, Yao, & Mei, 2017). Tran et al. proposed the R (2+1) D model, which further decomposed the 3D convolution into 2D spatial convolution and 1D temporal convolution, while maintaining accuracy and reducing computational complexity (Tran, Wang, Torresani et al., 2018).

(3) Two-Stream Network-Based Approaches

The two-stream network integrates the temporal modeling capability of 3D CNNs and the spatial feature extraction advantages of 2D CNNs. Its two branches process "spatiotemporal features" and "spatial features" respectively, balancing computational cost and model performance. In 2019, Köpüklü et al. proposed the YOWO model (Köpüklü, Wei, & Rigoll, 2019), which for the first time combined 3D CNN (3D-ResNeXt-101) with a 2D object detection network (Darknet-19): the 3D CNN branch extracts spatiotemporal features from video sequences, while the 2D CNN branch extracts spatial features from the current frame. Feature integration is achieved through channel fusion and an attention mechanism (CFAM), pioneering the one-stage spatiotemporal action detection paradigm.

Jiang, Yang, Jiang et al. (2024) proposed YOWOv2 based on YOWO, constructing a feature pyramid network and a decoupled head to separate classification and localization features: the feature pyramid enhances small target detection capability through multi-scale feature fusion, while the decoupled head optimizes classification loss and localization loss respectively. Its lightweight version (with 3D-ShuffleNetv2 as the backbone) achieves only 2.9 GFLOPs and a frame-mAP of 15.6%. Zhu, Wang, Yang et al. (2024) proposed YOWOv3, which adopts two label assignment mechanisms to match predicted and ground-truth labels, further improving bounding box regression accuracy, with a frame-mAP of 81.44% on the UCF101-24 dataset. Such methods make up for the shortcomings of 2D CNNs and 3D CNNs through a dual-branch structure, striking a balance between accuracy and speed, and have become the mainstream solution for real-time spatiotemporal action detection.

2.2 Methods Based on Transformers

Benefiting from the global dependency modeling capability of the self-attention mechanism, Transformers break through the limitation of CNNs' local receptive fields and have become a research hotspot in spatiotemporal action detection in recent years, especially performing prominently in long-temporal action and multimodal interaction scenarios.

Gritsenko et al. proposed the STAR model with an all-Transformer architecture, which incorporates temporal inductive bias into the encoder and captures long-range temporal dependencies through the self-attention mechanism. It can accurately predict the spatiotemporal range of actions even in weakly supervised scenarios, achieving a frame-mAP of 24.93% on the AVA dataset (Gritsenko, Xiong, Djolonga et al., 2024). Faure et al. proposed the HIT model (Holistic Interaction Transformer Network), which fuses RGB stream and pose stream data and designs an Intra-Modal Aggregation (IMA) module and an Attentive Fusion Module (AFM): the IMA module selectively merges "human-object-hand" interaction features within a single modality, while the AFM module adaptively fuses multi-modal features. It achieves a frame-mAP of 83.8% and a video-mAP of 9.7% on the JHMDB-21 dataset.

Wang, Huang, Zhao, Tong et al. (2023) proposed the VideoMAE model, which performs pre-training through a "dual masking strategy": the encoder processes a subset of video tokens, while the decoder reconstructs another subset. This enables efficient spatiotemporal feature learning at the billion-parameter scale, with action detection accuracy improved by over 15% after fine-tuning. Lu et al. proposed the FTP framework, which combines Vision Transformer (ViT) with vision-language models (such as LLaVA). It retains the global modeling capability of ViT and enhances the semantic expression of visual encoding by aligning with VLM outputs, achieving a frame-mAP of 32.6% on the AVA dataset (Lu, Jian, Poppe & Salah, 2024).

In addition, Zhao et al. proposed the Tuber model, which combines 3D CNNs with Transformers. It predicts the spatiotemporal boundaries of action tubes through a Transformer decoder, addressing the limitation of 3D CNNs with fixed clip lengths, and achieves a video-mAP of 58.4% on the UCF101-24 dataset. The advantages of such methods lie in their strong global temporal modeling capability, making them suitable for long video action detection. However, they have high computational costs,

strong dependence on large-scale annotated data, and face significant challenges in deployment on edge devices.

2.3 Methods Based on Multi-Interaction Relationship Modeling

In complex scenarios, the semantic understanding of actions highly depends on spatial interactions between "people and people" or "people and objects", as well as temporal interactions between "frames and frames." Most existing methods focus on a single type of interaction, which limits detection accuracy. Methods based on multi-interaction relationship modeling explicitly capture multi-dimensional correlations, significantly enhancing adaptability to complex scenarios.

Tang, Xia, Mu, Pang, and Lu (2020) proposed the AIA Network (Asynchronous Interaction Aggregation Network), which serially stacks interaction modules for "person-person", "person-object", and "person-temporal memory" interactions, integrating interaction relationships into a hierarchical structure for collaborative work: the "person-person interaction" module models multi-target limb correlations; the "person-object interaction" module parses the spatial dependencies between actors and scene objects; the "person-temporal memory" module captures long-term action dynamics, resulting in an 8.7% improvement in video-mAP on the THUMOS'14 dataset. Pan, Chen, Shou, Liu, Shao, and Li (2021) proposed ACAR-Net, which not only models direct interactions between person-object pairs but also designs a High-Order Relation Reasoning Operator (HR2O) to capture indirect correlations among multiple objects. By storing full-video high-order relationship information through an Actor-Context Feature Bank (ACFB), it achieves a frame-mAP of 33.3% on the AVA dataset.

Wu et al. proposed the Long-Term Feature Bank (LFB), which provides the model with long-temporal context support by integrating RoI features over fixed time steps: the LFB module records action features of the entire video, and the feature bank operator calculates the interaction relationship between short-term RoI features and long-term features through an attention mechanism, achieving a frame-mAP of 61.6% on the THUMOS'14 dataset (Wu, Feichtenhofer, Fan, He, Krahenbuhl, & Girshick, 2019).

3. Key Technologies

Focusing on the three goals of "accuracy improvement", "efficiency optimization", and "complex scenario adaptation", researchers have proposed a number of technological innovations, forming practical optimization paths. The details are as follows.

3.1 Feature Enhancement Technology

Feature enhancement improves the discriminability and robustness of action information by optimizing the feature extraction process, such as the enhancement of key features based on attention mechanisms. The attention mechanism focuses on key action regions and suppresses background interference by dynamically adjusting feature weights. Hou, Zhou, and Feng (2021) proposed the Coordinate Attention (CA) module, which converts feature maps into coordinate representations along the spatial dimension, focusing on horizontal and vertical spatial dependencies respectively to enhance the model's sensitivity

to target positions.

Hu, Shen, and Sun (2018) proposed the Squeeze-and-Excitation (SE) module, which obtains channel-level statistical features through global average pooling and then learns channel weights via fully connected layers to enhance key channel features. Woo, Park, Lee et al. (2018) proposed the CBAM module, which combines channel attention and spatial attention: channel attention optimizes channel weights, while spatial attention locates key spatial regions. After introducing CBAM into the YOWO model, the frame-mAP on the UCF101-24 dataset increased by 1.2% (Wang, Ga, Zhang, and Sang, 2021).

3.2 Multi-Scale Feature Fusion Technology

Multi-scale feature fusion enhances the model's adaptability to multi-scale actions by integrating features from different levels and modalities, with common methods based on feature pyramid fusion. The Feature Pyramid Network (FPN) fuses multi-level features through top-down and bottom-up paths, improving the detection capability for small targets. Wang, Gao, Zhang and Sang (2020) proposed the Multi-Layer Temporal Pyramid Network (MLTPN), which designs a Temporal H-module (THM) to generate multi-scale temporal features. A pyramid structure composed of cascaded multiple THMs and Convs is constructed to enhance the model's ability to capture actions of different durations, achieving a video-mAP of 59.9% on the THUMOS'14 dataset.

Weng, Pan, Han, Chang, and Zhuang (2022) proposed the Spatiotemporal Pyramid Transforme(STPT). Its first two stages use Local Spatiotemporal Attention blocks (LSTA) to capture local patterns, while the last two stages use Global Spatiotemporal Attention blocks (GSTA) to handle long-range dependencies. Finally, a temporal feature pyramid network is used to reduce spatiotemporal dimensions and obtain multi-scale temporal features, achieving a frame-mAP of 30.1% on the AVA dataset. By complementing multi-scale features, such methods effectively improve action detection accuracy in complex scenarios.

3.3 Interaction Relationship Modeling Technology

Interaction relationship modeling enhances the ability to understand actions in complex scenarios by explicitly capturing associations between "people and people", "people and objects", and "frames and frames". For example, spatial interaction modeling focuses on the spatial correlations between actors, the environment, and objects. Sun C, Shrivastava, Vondrick, Murphy, Sukthankar and Schmid (2018) proposed the Actor-Centered Relation Network (ACRN), which uses 1×1 convolution to calculate the relationship feature maps between actors and surrounding objects, and then accumulates correlation information of adjacent positions through 3×3 convolution to strengthen the modeling of "person-object" interactions, achieving a frame-mAP of 82.4% on the UCF101-24 dataset.

3.4 Loss Function Optimization Technology

Loss function optimization improves the model's prediction accuracy by refining bounding box regression loss and classification loss. Zheng, Wang, Liu et al. (2020) proposed the CIoU loss, which comprehensively considers the IoU between the predicted box and the ground-truth box, the distance

between center points, and the aspect ratio, thereby addressing the problem of bounding box regression bias.

Gevorgyan et al. (2022) proposed the SIoU loss, which adds angle loss (direction difference between the ground-truth box and the predicted box), distance loss (distance between center points), and shape loss (aspect ratio difference) to further optimize bounding box regression. After the EAD model adopted the SIoU loss, the frame-mAP on the UCF101-24 dataset increased by 0.8%.

4. Experimental Verification and Performance Comparison

4.1 Core Datasets and Evaluation Metrics

(1) Core Datasets

- UCF101-24: It contains 24 action categories and 3207 untrimmed videos, with numerous dynamic backgrounds. It features frame-level annotations of spatiotemporal boundaries, making it suitable for evaluating the model's adaptability to complex scenarios.
- JHMDB-21: It contains 21 action categories and 31,838 annotated images, with significant differences in action durations, making it suitable for evaluating the model's temporal modeling capability.
- AVA2.2: It contains 80 categories of atomic actions and 430 videos, with dense annotations at 1 frame per second. It supports multi-label action detection and serves as an authoritative benchmark for spatiotemporal action detection.
- THUMOS'14: It contains 20 action categories and 2584 test instances. The test set provides only partial annotations, posing challenges to the model's open-world detection capability.

(2) Evaluation Metrics

- Frame-mAP: The area under the precision-recall curve of per-frame detection results is calculated to measure the accuracy of frame-level action classification and localization.
- Video-mAP: It is calculated based on the overall video prediction results. A prediction is considered correct if the average IoU between the action tube and the ground-truth annotation exceeds the threshold and the category is accurate, which is used to measure the temporal modeling capability.
- FPS: The number of video frames processed per second, used to evaluate the real-time performance of the model.
- GFLOPs/Params: It measures the computational complexity and parameter quantity of the model, and is used to evaluate the model's computational complexity and scale.

4.2 Performance Comparison of Representative Models

Table 1. Performance of CNN-based Models

Models	Datasets	Frame-mAP（%）	Video-mAP（%）	FPS	GFLOPs	Params（M）
C3D	UCF101-24	68.1	45.3	18	35.2B	87.6
I3D	UCF101-24	76.3	59.9	22	42.5B	94.3
SlowFast	AVA2.2	24.2	-	15	308B	304.0
YOWO	UCF101-24	80.4	48.8	34	43.7B	121.4
YOWOv2-T	UCF101-24	80.5	51.3	50	2.9B	10.9
EAD	UCF101-24	80.93	50.41	53	2.77B	10.92

As can be seen from the Table, the YOWO series models based on two-stream networks strike a balance between accuracy and speed: compared with YOWO, YOWOv2-T reduces the number of parameters by 90.9%, increases FPS by 47.1%, and simultaneously improves frame-mAP by 0.1%; after introducing LSDAM and MSFFU, the EAD model achieves a 0.43% improvement in frame-mAP and a 0.13B reduction in GFLOPs compared with YOWOv2-T.

Table 2. Performance of Transformer-based Models

Models	Datasets	Frame-mAP（%）	Video-mAP（%）	FPS	GFLOPs	Params（M）
STAR	AVA2.2	24.93	-	8	128B	156.0
Tuber	UCF101-24	83.2	58.4	12	132B	142.0

Models	**Datasets**	Frame-mAP（%）	Video-mAP（%）	FPS	GFLOPs	Params（M）
HIT	JHMDB-21	83.8	89.7	12.37	165B	187.0
Video MAE	AVA2.2	32.6	-	10	256B	210.0

Transformer-based models exhibit strong global modeling capabilities but incur high computational costs: the HIT model achieves a frame-mAP of 83.8% on the JHMDB-21 dataset, yet its FPS is only 12.37, making real-time deployment challenging; VideoMAE improves accuracy through pre-training, but its parameter count reaches 210M, resulting in high hardware requirements.

Table 3. Performance of Models Based on Multi-Interaction Relationship Modeling

Models	**Datasets**	Frame-mAP（%）	Video-mAP（%）	FPS	GFLOPs	Params（M）
AIA	THUMOS’14	78.8	54.4	20	85.6B	112.0
ACAR	AVA2.2	33.3	-	18	252B	208.0
LFB	THUMOS’14	61.6	64.8	25	92.4B	135.0
IREAD	AVA2.2	19.34	-	50	16.04B	23.44

Models based on multi-interaction relationship modeling perform excellently in complex scenarios: the IREAD model, equipped with the IRMM module, achieves a 3.42% higher frame-mAP than the EAD model on the AVA2.2 dataset while maintaining real-time performance of 50 FPS; although the ACAR model delivers the highest accuracy, its GFLOPs reach 252B, making edge deployment difficult.

5. Conclusion

This paper systematically reviews the technological evolution of deep learning-based spatiotemporal action detection models in computer vision. It categorically elaborates on three mainstream architectures: CNN-based, Transformer-based, and multi-interaction relationship modeling-based models. It summarizes key technological innovations such as feature enhancement, multi-scale fusion, and interaction relationship modeling, and verifies model performance with experimental data.

The research shows that two-stream network-based models (e.g., YOWOv2, EAD) strike a balance between accuracy and speed, making them suitable for real-time deployment; Transformer-based models (e.g., HIT, VideoMAE) possess strong global modeling capabilities, making them suitable for complex scenarios; and multi-interaction relationship modeling-based models (e.g., IREAD, ACAR) perform excellently in multi-object interaction scenarios.

Future efforts need to further break through technologies such as multi-modal fusion and lightweight inference to promote the application of spatiotemporal action detection in more practical scenarios.

References

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kineticsdataset. *Proceedings of the IEEE Conference on Computer Vision and PatternRecognition*, 6299-6308.

Faure, G. J., Chen, M. H., & Lai, S. H. (2023). Holistic interaction transformer network for actiondetection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3340-3350.

Gevorgyan, Z. (2022). SIoU loss: More powerful learning for bounding box regression. *arXivpreprint arXiv:2205.12740*.

Gritsenko, A. A., Xiong, X., Djolonga, J. et al. (2024). End-to-end spatio-temporal action localisationwith video transformers. *Proceedings of the IEEE/CVF Conference on ComputerVision and Pattern Recognitio*, 18373-18383.

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design[C].Proceedings of the IEEE. *CVF conference on computer vision and pattern recognition*, 13713-13722.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEEConference on Computer Vision and Pattern Recognition*, 7132-7141.

Ji, S., Xu, W., Yang, M. et al. (2021). 3D convolutional neural networks for human actionrecognition. *IEEE transactions on pattern analysis and machine intelligence*, *35*(1), 221-231.

Jiang, Z., Yang, J., Jiang, N. et al. (2024). Yowov2: A stronger yet efficient multi-level detectionframework for real-time spatio-temporal action detection. International Conferenceon Intelligent Robotics and Applications. *Singapore: Springer Nature Singapore*, 33-48.

Kalogeiton, V., Weinzaepfel, P., Ferrari, V., & Schmid C. (2017). *Action tubelet detector for*

spatiotemporal action localization. *Proceedings of the IEEE International Conference onComputer Vision*, 4405-4413.

Köpüklü, O., Wei, X., & Rigoll, G. (2019). You only watch once: A unified cnn architecture for real-timespatiotemporal action localization. *arXiv preprint arXiv:1911.06644*.

Li, Y., Wang, Z., Wang, L., & Wu, G. (2020). Actions as moving points. *Proceedings of the European Conference on Computer Vision*, 68-84.

Lu, H., Jian, H., Poppe, R., & Salah, A. A. (2024). Enhancing Video Transformers for ActionUnderstanding with VLM-aided Training. *arXiv preprint arXiv:2403.16128*,

Pan, J., Chen, S., Shou, M. Z., Liu, Y., Shao, J., & Li, H. (2021). Actor-context-actor relation network forspatio-temporal action localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 464-474.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residualnetworks. *Proceedings of the IEEE International Conference on Computer Vision*, 5533-5541.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residualnetworks. *Proceedings of the IEEE International Conference on Computer Vision*, 5533-5541.

Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., & Schmid, C. (2018). Actor-centricrelation network. *Proceedings of the European Conference on Computer Vision*, 318-334.

Tang, J., Xia, J., Mu, X., Pang, B., & Lu, C. (2020). Asynchronous interaction aggregation for actiondetection. *Proceedings of the European Conference on Computer Vision*, 71-87.

Tran, D., Bourdev, L., Fergus, R. et al. (2015). Learning spatiotemporal features with 3dconvolutional networks. *Proceedings of the IEEE international conference oncomputer vision*, 4489-4497.

Tran, D., Wang, H., Torresani, L. et al. (2018). A closer look at spatiotemporal convolutions foraction recognition. *Proceedings of the IEEE conference on Computer Vision andPattern Recognition*, 6450-6459.

Wang, L., Huang, B., Zhao, Z., Tong, Z. et al. (2023). Videomae v2: Scaling video masked autoencoderswith dual masking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14549-14560.

Wang, X., Gao, C., Zhang, S., & Sang, N. (2020). Multi-level temporal pyramid network for actiondetection. *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision*, 41-54.

Weng, Y., Pan, Z., Han, M., Chang, X., & Zhuang, B. (2022). An efficient spatio-temporal pyramidtransformer for action detection. *Proceedings of the European Conference on Computer Vision*, 358-375.

Woo, S., Park, J., Lee, J. Y. et al. (2018). Cbam: Convolutional block attention module. *Proceedings of*

the European conference on computer vision (ECCV), 3-19.

Wu, C. Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., & Girshick, R. (2019). Long-term feature banksfor detailed video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 284-293.

Zhao, J., Zhang, Y., Li, X., Chen, H., Shuai, B. et al. (2022). Tuber: Tubelet transformer for video actiondetection. *Proceedings of the IEEE/CVF Conference on Computer Vision and PatternRecognition*, 13598-13607.

Zheng, Z., Wang, P., Liu, W. et al. (2020). Distance-IoU loss: Faster and better learning forbounding box regression. *Proceedings of the AAAI conference on artificialintelligence*, *34*(07), 12993-13000.

Zhu, A., Wang, Y., Yang, J. et al. (2024). YOWOv3: A Lightweight Spatio-Temporal JointNetwork for Video Action Detection. *IEEE Transactions on Circuits and Systems forVideo Technology*, *34*(9), 8148-8160.

Funding:

This work was supported by the the Shaanxi Provincial Sports Bureau Project under Grant No. 20250153

Author Information

Mingzhu Zhang was born in Taian,Shandong,China, in 1981.She is a associate in Xi'an FanYi university. She received the master's degree from Xi'an Technological University. Her research interest include computer image processing and pattern recognition.