Dean&Francis

ISSN 2959-6157

# Research and Analysis of Action Recognition Based on Video

**Yuexin Dai**[1*,]

[1]David Game College, London, EC3N 2ET, the United Kingdom
*Corresponding author: daiyuexin11@126.com

**Abstract:**

In recent years, along with the swift development of computer vision and deep learning technologies, video-based action recognition has turned into one of the core research directions within the field of artificial intelligence. Its accomplishments are extensively applied in such real scenarios as intelligent monitoring, human-computer interaction, and autonomous driving. This paper first presents the research background and practical significance of video-based action recognition in recent years then analyzes the current challenges such as those of complex background motion blur and similarity among categories. Next it expounds upon the principle's structures and application effects of representative models such as 3D convolutional neural networks two - stream networks and models based on the Transformer and also analyzes the advantages and disadvantages of various models. Finally, it summarizes the application scenarios of video action recognition, explores the existing technical difficulties, and also looks forward to the development trends in the future like lightweight models and few - shot learning. This paper offers comprehensive references for researchers in relevant fields, making them able to grasp the research current situation and carry out in - depth research.

**Keywords:** Video action recognition; Deep learning; Convolutional neural network; Transformer; Computer vision.

## 1. Introduction

In the era of big data videos that are core data carriers are full of a large amount of information about human actions and interactions with the environment. The video-based action recognition technology which is the key part of intelligent video analysis can automatically identify the categories of human actions in video sequences. This not only enriches the theoretical framework of computer vision but also has urgent practical requirements in multiple fields and also has a broad application prospect.

From the perspective of application, this technology has indeed already become a core module of various intelligent systems. IN the field of intelligent monitoring, it can automatically identify abnormal

behaviors like fighting and falling and can give real-time warnings, thus solving the problems of low efficiency and easy miss detection in manual monitoring and improving the level of public safety management; in the field of human-computer interaction, it makes machines understand human body language, breaking through the limitations of traditional input devices and enhancing the immersion and interactivity of VR/AR systems; in the field of autonomous driving, it can assist in identifying the actions of pedestrians and cyclists (for example, judging whether to cross the road) and improve the accuracy and safety of driving decisions. Moreover, this technology also plays important roles in the fields of intelligent medical care, sports analysis and smart home, and is driving the changes in production and life.

Owing to the breakthroughs in deep learning techniques and the growth of large-scale video datasets the research on video action recognition has made quite some progress yet still faces numerous challenges. IN the real environment, factors like complex backgrounds, light variations, and motion blurs will interfere with the accuracy of feature extraction; the diversity (different styles of the same action) and ambiguity (similar actions are easy to confuse) of human actions increase the difficulty of recognition; at the same time, the high computational complexity of video data has high requirements for the real-time performance of advanced models, restricting the application of advanced models on edge devices. Thus, designing models with high precision, strong robustness, and low computational complexity has become the core issue of current research.

At present this field is in a stage of rapid development and a large number of new methods emerge. In 2014 Carreira and Zisserman put forward the two-stream convolutional neural network the spatial stream extracts appearance features the temporal stream extracts motion features and then they are fused for classification laying the foundation for deep learning action recognition and solving the problem that the traditional two-dimensional convolutional neural network is very difficult to capture time information In 2016 Feichtenhofer and others...A time - interval network is put forward, which samples multiple video clips and fuses features, effectively capturing long - term time information and solving the problem of sparse sampling in long videos [1].

The development of 3D convolutional neural networks has propelled a new phase in this field. In 2015, the C3D model was developed by Tran et al. which extends the 2D convolution kernels to 3D ones so that spatial-temporal features can be extracted yet it has the problems of a large number of parameters and high computational complexity. In the year of 2017, Feichtenhofer et al. An expanded 3D convolutional neural network (I3D) is put forward, which extends the parameters of the pre - trained 2D CNN to 3D ones, thus reducing the training difficulty, decreasing the computation cost, and also increasing the accuracy, and becoming one of the models that are widely used [1]. In recent years the Transformer model grounded in the self-attention mechanism has been cross - applied to the computer vision field. In 2021, Bertasius et al. put forward the video version of the Swin Transformer which introduced the hierarchical window self - attention mechanism to capture the spatio-temporal context information achieved the optimal results on large - scale datasets presented great potential but also faced the demand of high computational complexity needing optimization [2].

This paper means to systematically review the research progress of video action recognition in the recent years and then summarize the evolution process and the technical characteristics of the mainstream models. The structure of this paper is as follows: The second part elaborates in detail the mainstream technologies like the dual-stream network, the 3D convolutional neural network, and the model based on Transformer, analyzing their principles, structures, and advantages and disadvantages; the third part summarizes the application scenarios combined with case studies, analyzing the application effects and problems; the fourth part discusses the current challenges and looks forward to the future development trends; the fifth part summarizes the whole paper, pointing out the core directions of future research.

## 2. Overview of Main Technologies for Video Action Recognition

### 2.1 Double-stream Convolutional Neural Network and Its Improved Model

Before the emergence of the dual - stream convolutional neural network, action recognition research mainly depended on the traditional way of extracting features manually. For instance, features such as the histogram of oriented gradients (HOG) and the scale - invariant feature transform (SIFT) was extracted from video frames, and then classifiers like the support vector machine (SVM) were combined to be used for classification and so forth. These methods have the characteristics of low recognition accuracy and poor robustness, which makes it difficult to adapt to complex real - world environments. In 2014, Carreira and Zisserman developed the two - stream convolutional neural network, which started a new era of deep - learning application in the field of action recognition and became a milestone achievement in the field of video

ISSN 2959-6157

action recognition.

The core idea of the dual-stream convolutional neural network is to separate the spatial information and the temporal information in the video then extract them respectively and then fuse these two types of information so as to realize action recognition. This model incorporates two separate two-dimensional convolutional neural network branches namely the spatial stream and the temporal stream. The spatial stream takes RGB video frames as input and its core task is to extract the appearance features of actions like human forms clothing colors and the spatial positions of actions. At this point the temporal stream takes optical flow maps as input where optical flow maps reflect the motion information between consecutive frames and the temporal stream is used to extract the motion features of actions such as motion directions and speeds. After each data flow finishes feature extraction, the classification results of the two data flows are fused through methods like average voting, thereby generating the final action recognition result.

Another significant improvement direction is also to achieve the end-to-end training of the dual - stream network. The original dual - stream network has to pre - calculate the optical flow map which is not very beneficial to the overall integration of the model. To solve this problem, a motion representation method by means of flow-guided convolution was put forward according to the citation „Motion Representation via Flow-Guided Convolution". Within the flow of time this model substitutes traditional convolutional layers with flow - guided convolutional layers which can directly extract motion features from RGB video frames without calculating optical flow maps in advance enabling the whole two - stream network to be trained end - to - end resulting in a simplified training process and an enhanced generalization ability of the model.

## 2.2 3D Convolutional Neural Network

Although the double - flow network and its improved models have achieved good results, this way of extracting spatial and temporal features separately is not the most natural way of processing video data. Videos are a kind of spatio-temporal data which integrates space and time. So then, it is fairly reasonable to simultaneously extract spatio-temporal features as well. The 3D convolutional neural network (3D CNN) extends the 2D convolution kernel in the spatial domain to the 3D convolution kernel in the spatiotemporal domain and can also capture the spatial and temporal information in videos synchronously and effectively as well [3].

The C3D model that Tran et al. put forward. In the year 2015, the first 3D CNN model which was widely applied in the action recognition field came into being. Its structure is simple and clear, consisting of 8 3D convolution layers, 5 3D pooling layers and 2 fully connected layers. A 3×3×3 3D convolution kernel in the model is set which can balance the extraction of spatial features and temporal features to a certain extent. After having been trained on the large - scale video dataset Sports - 1M, the C3D model is fine - tuned on small - scale datasets like UCF101 and HMDB51, and this has shown a remarkable improvement in comparison with the dual - stream network at that time. The success of the C3D model demonstrates the unique advantage of 3D CNN in dealing with the temporal-spatial features of videos, which provides the foundation for the subsequent development of action recognition methods based on 3D CNN.

However, the C3D model also has obvious drawbacks: because of the use of 3D convolution kernels, the number of parameters of the model and the computational complexity are much higher than those of 2D CNN, which makes the training of the model quite difficult and cannot meet the real-time requirements of edge devices; additionally, the C3D model uses fixed - size convolution kernels, which are difficult to adapt to the differences in the time length and spatial scale of different actions. To address these issues a series of optimized 3D CNN models have been proposed by scholars among which the dilated 3D convolutional neural network (I3D) put forward by Fichtenholfer et al. in 2017 has a rather significant impact [1].

The core idea of the I3D model is to expand the parameters of the pre - trained 2D CNNs (such as Inception - V1, VGG and the like) into 3D parameters. Specifically speaking it is to add a time dimension and turn the two - dimensional convolution kernel into a three - dimensional convolution kernel then initialize the weights of the new time dimension with relatively small values. This method can make full use of the pre - trained parameters from the image data set to solve the problem of insufficient video training data and can effectively reduce the number of model parameters and computational complexity. Meanwhile the I3D model makes use of multi-scale convolution kernels to extract features of different spatial-temporal scales thus improving the model's adaptability to various actions. This model that has obtained the best results on multiple benchmark datasets has turned into a classic model in the field of action recognition based on 3D CNN.

In recent years, along with the development of lightweight neural networks, a number of scholars have begun to conduct research on lightweight 3D CNN models so as to meet the application requirements of edge devices. For example, in 2019, Zhang and his colleagues developed the MobileNet3D model, and also introduced the depth-wise

separable convolution from MobileNet into the 3D CNN as well [4]. The deep separable convolution that decomposes the 3D convolution into depthwise convolution and pointwise convolution will, while ensuring the recognition accuracy, greatly reduce the number of parameters and computational complexity. These lightweight models which have established a solid foundation for the practical application of video action recognition technology are very portable.

### 2.3 Action Recognition Model Based on Transformer

Owing to its characteristics of local receptive fields and shared weights, the convolutional neural network (CNN) that is quite remarkable in extracting spatial features within the image space. However, it has inherent limitations in capturing long-distance spatio-temporal dependencies. Convolutional neural networks (CNNs) need to stack numerous convolutional layers to achieve global feature perception, which not only increases the model complexity but also may lead to the problem of gradient disappearance. The self-attention mechanism of the Transformer can directly calculate the association between any two elements in the sequence without relying on the local sliding window, thus providing new ideas for solving the long-term spatio-temporal modeling problem in video action recognition [3].

Regarding the research on video action recognition based on transformers, it mainly focuses on optimizing the spatio-temporal attention mechanism and enhancing the model efficiency in a certain manner. In the early research, video frames were often input into the Transformer as a sequence, for instance, the ViViT (Video Vision Transformer) model, which divides video frames into a number of image patches and arranges them in chronological order, and makes use of the standard Transformer structure to extract spatio-temporal features [5]. This model was the first one to verify the feasibility of the pure Transformer architecture in the video action recognition task, but because it was not optimized for the spatio-temporal characteristics of video, its capacity to capture motion information was rather weak, and the accuracy in action recognition did not go beyond the mainstream CNN models.

To enhance the video spatio-temporal feature modeling ability of the Transformer, improvements that researchers put forward from the design of the attention mechanism are implemented. A hierarchical window self-attention mechanism is generated by Video Swin Transformer. At the bottom layer of the model, a relatively small spatio-temporal window is employed to compute attention, concentrating on local fine - grained features; at a higher level, the receptive field is expanded through merging windows to capture global spatio-temporal dependencies [2]. This hierarchical structure guarantees the efficiency of feature extraction and can also effectively model spatio - temporal information of different scales as well. Experiments show that this model, which achieves a much higher recognition accuracy than classic CNN models like I3D on large - scale datasets such as Kinetics - 400 and Something - Something V2, has become one of the current mainstream video action recognition frameworks.

Additionally, in order to deal with the issue of the high computational complexity of Transformers, scholars have come up with all sorts of optimization strategies. For example, Times former separates spatial attention and temporal attention. First, spatial attention is computed in a single frame to extract appearance features, and then attention is computed in the temporal dimension to capture motion information. In this way, the computational complexity is reduced from $O((T\times H\times W)^2)$ to $O(T\times(H\times W)^2 + (H\times W)\times T^2)$, greatly improving the inference speed of the model [6]; the X3D model uses a unified scaling strategy to balance the depth, width, and resolution dimensions of the model, making the model lighter while ensuring the recognition accuracy, which provides the possibility for the application of Transformer on edge devices [7].

Although the action recognition models based on Transformers have achieved remarkable progress there are still some issues firstly the model training depends on large-scale annotated data and has rather poor generalization ability in small-sample scenarios secondly the self-attention mechanism is quite sensitive to positional information while the spatial-temporal positional flexibility of actions in videos is fairly strong. How to design a more effective position encoding method is still a research puzzle; thirdly, the interpretability of the model is relatively weak, leading to it being hard to clearly distinguish which spatio-temporal features are crucial for action recognition. These problems need to be further dealt with in subsequent research.

## 3. The Scene and Analysis of Action Recognition Technology based on Video

### 3.1 The Implementation Status of the Main Application Scenarios

With the video-based action recognition technology which has the ability to precisely capture temporal-spatial information, large-scale applications have been carried out in

ISSN 2959-6157

several core domains. Different technological approaches each having distinct characteristics show obvious scene-adapted situations because of their own unique features.

In the domain of intelligent monitoring, the lightweight 3D CNN and the dual - stream attention fusion model that possess a light - weight characteristic have truly become the mainstream choices. Optimized by Chen Ming et al. in 2023, the MobileNet3D architecture has an abnormal action recognition accuracy rate exceeding 92% in scenarios such as urban road monitoring and mall security. It can quickly detect dangerous behaviors like fighting, falling, and climbing, with a response delay controlled within 500 milliseconds, satisfying the real-time early warning requirement [8]. This scenario indeed has an extremely high requirement for the robustness of the model. The problem of complex backgrounds, lighting changes and occlusions can be effectively alleviated through the multi-feature fusion technology, yet in the low - light nighttime environment, the recognition accuracy still approximately drops by 8 - 12%.

In the domain of human - computer interaction, the models that are based on Transformers are quite dominant, simply because they do have advantages in capturing global features. The video Swin Transformer and the time Swin model are widely applied in places such as VR/AR devices, smart cockpits and the like, supporting interactive methods like gesture control and limb instruction recognition and so forth. In the scenario of intelligent vehicles, this model can identify things such as the driver‘s fatigue like frequent nodding and closing of eyes, as well as dangerous operations like hands leaving the steering wheel. With an accuracy rate surpassing 95%, it offers auxiliary support for driving safety. However, this situation does have rather strict requirements on the real - time performance of the model. Through optimization methods such as separate spatio-temporal attention and model quantization, there can be an improvement in inference speed of around 30%, which basically meets the computing power requirements of vehicle - mounted equipment.

In the domains of intelligent medical care and sports analysis, the key consists in identifying the specific details of actions as well as creating what is standardized and so forth. Put forward by Zhang Siyuan and the like, the PoseConv3D model is the one that can in real time monitor how standard the limb rehabilitation movements of patients in rehabilitation medicine are [9]. By comparing with the preset action templates, the quantitative indicators like joint angle deviations and action completion degrees will be output, thereby assisting doctors in adjusting the rehabilitation plans. In the sports training, this model can analyze the biomechanical characteristics like those of athletes‘ running and shooting. The problem of improper force identification has an accuracy rate exceeding 93%. The characteristic of resisting posture estimation noise makes its practicability in clinical and training scenarios much stronger than that of traditional GCN models.

## 3.2 Core Technology Performance Comparison Analysis

Through conducting testing on mainstream benchmark datasets (UCF101, Kinetics - 400, NTU RGB+D), the performance of core technologies in the past 5 years is as presented in table 1.

**Table 1. The performance of the core technology**

| technological type | Representative type | Date set | Recognition accuracy rate | Parameter quantity(M) | Inference speed(FPS) | Advantageous scenario |
|---|---|---|---|---|---|---|
| 3D CNN | X3D | Kinetics-400 | 86.7% | 6.7 | 58 | Edge devices, real-time monitoring |
| | PoseConv3D | NTU RGB+D | 94.2% | 8.3 | 45 | Bone movement recognition, medical sports |
| Transformer | Video Swin Transformer | Kinetics-400 | 89.5% | 28.1 | 22 | Complex scenarios, precise interaction |
| | TimeSformer | UCF101 | 98.3% | 25.6 | 27 | Human-computer interaction, action detail recognition |
| Improved Two-stream Network | Two-stream Attention Fusion Model | UCF101 | 97.8% | 15.2 | 41 | Security monitoring, low-cost scenarios[10] |

Just as is shown in Table 1, in respect of the identification accuracy, those models that are based on Transformer are more splendidly manifested in the complex datasets. On the Kinetics - 400 dataset, Video Swin Transformer functions more efficiently, attaining a precision which is 2.8 percentage points higher than that of X3D, this showing the advantage of the global spatio-temporal modeling. In terms of parameters and inference speed the 3D CNN model does indeed possess more prominent advantages. The parameters of X3D are only one - fourth of the ones of the video Swin Transformer, and the inference speed has been doubled. Six operations were conducted to make it more appropriate for deployment on edge devices [7]. As for the adaptability of specific tasks, PoseConv3D performs exceptionally well on the NTU RGB+D dataset for skeletal action recognition, with an accuracy rate 5 to 8 percentage points higher than other models, thereby verifying the effectiveness of the skeletal representation based on 3D heatmaps [9].

# 4. Challenges and Prospects of Video-Based Action Recognition Technology and Its Applications

## 4.1 The Current Core Challenge

### 4.1.1 Technical challenges

The core bottleneck that still affects the recognition accuracy under complex scenarios is the absence of sufficient robustness because factors such as light changes in the actual environment, background interference, motion deformation, and occlusion often disrupt effective feature extraction and the existing models also have to laboriously adapt to multiple complex situations as well. In addition, achieving a balance between efficiency and accuracy has become another critical challenge in the application of technology, because highly precise models such as those based on the Transformer architecture usually have a relatively large quantity of computation, while lightweight models often lack sufficient accuracy when handling complex tasks. Moreover, few-shot and zero-shot learning also bring more obstacles, because labeling video data is both time-consuming and laborious, especially in the action data of professional fields where the number of annotated samples is extremely small. Although existing few-shot learning methods like Meta-Transformer have made certain breakthroughs, their performance in zero-shot situations still needs to be further enhanced [11]. Finally, the poor interpretability of the model also restricts the in - depth application of video action recognition technology. Most of the current deep learning models are „black boxes“ which are difficult to clearly find out the key features that drive the action recognition decision, such as specific joint movements and action time patterns, and this in turn restricts their deployment in high - reliability fields like healthcare and security.

### 4.1.2 Industrial implementation challenge

Adaptation across devices poses a notable challenge, as the computing power of hardware devices varies significantly across different application scenarios ranging from edge sensors to cloud servers, and customized models need to be developed for each specific device type, which inevitably increases the overall cost of technology implementation. Furthermore, data security and privacy protection have become an indispensable consideration for industrial applications, given that action data contains rich personal behavioral characteristics and is classified as sensitive information, and ensuring the security of such data throughout the entire lifecycle of collection, transmission, and storage while preventing unauthorized abuse and leakage has emerged as a critical issue that must be addressed. Additionally, the lack of unified industry standards has led to prominent technical incompatibility issues, since the training datasets, evaluation metrics, and interface specifications of models developed by different manufacturers remain fragmented, which not only hinders the interoperability of heterogeneous systems but also makes it difficult to form a standardized and large-scale application ecosystem

## 4.2 Future Development Trends

### 4.2.1 Technological innovation trends

The deepening of the multi-modal fusion technology will make use of the complementary advantages of the integrated multi-modal data including visual, audio, and inertial sensor information to enhance the robustness of recognition in complex scenarios. The intelligent monitoring scenario is a typical case, for example, integrating audio data such as distress calls and fighting sounds, which can assist in action recognition and reduce the error rate. Moreover, the optimization of the light - weight and efficient model needs to rely on continuous innovation in the model structure such as sparse attention mechanisms dynamic convolutions and model distillation techniques and so on. These can further reduce the parameter size of the Transformer - based model and the demand for computing power while maintaining high recognition accuracy thus enabling them to be efficiently deployed on edge devices with limited resources. In addition the progress in few - shot and unsupervised learning will utilize technologies

ISSN 2959-6157

like meta - learning contrastive learning and generative artificial intelligence to reduce the model's dependence on large - scale labeled data improve its generalization ability in scenarios with limited or no labeled samples and also expand the application range of video action recognition technology [11]. Moreover the enhancement of the model interpretability which combines visualization techniques attention weight analysis and causal reasoning methods to identify the core decision-making basis of the action recognition models will then lead to an increase in the trust of such technologies in high-reliability fields like healthcare and safety. Finally, the customized modeling for specific scenarios will lay emphasis on developing specialized models for such extreme environments as deep-sea exploration and high-altitude mountain climbing as well as such special groups as the elderly and children, and optimizing the extraction methods of action features so as to improve the adaptability of the technology to the target application scenarios.

#### 4.2.2 Industrial application trends

The architecture of „edge real-time inference plus cloud model update" will be constructed through edge computing and cloud collaboration. At this moment, edge devices are in charge of quickly processing simple action recognition tasks, while the cloud optimizes models via large-scale data training and then pushes the optimized versions to edge devices, thereby achieving a dynamic balance between real-time performance and the iterative improvement of accuracy. In addition, the promotion of industry standards will result in the formulation of standardized datasets, evaluation metrics, and safety standards for the developed action recognition technology, which then helps with the technical interoperability among different manufacturers and lays a solid basis for the formation of a large-scale industrial ecosystem. Moreover, the cross-domain comprehensive application will progress because the video action recognition technology is deeply integrated with emerging technologies like the Internet of Things, big data, and blockchain; for instance, in the smart city situation, the combination of action recognition and urban big data can make possible the intelligent upgrading of fields such as public safety and traffic management, while in the privacy protection field, the application of blockchain technology can make possible data encryption and sharing, effectively reducing the risk of privacy leakage.

## 5. Conclusion

This article carries out a systematic review of the research progress of video-based action recognition technologies in the past five years, with a strong emphasis on mainstream technical approaches like 3D convolutional neural networks, Transformers, and improved two-stream networks. It delves deeply into the principles, the structural advantages, and the applicable scenarios of various models, and also combines specific application cases and performance comparisons to demonstrate the implementation effects and the characteristic differences of this technology. It is shown by research that the 3D CNN model, with its light and highly real-time advantages, occupies a dominant position in edge devices and monitoring scenarios; the Transformer model, having powerful global spatial-temporal modeling capabilities, performs well in precise interaction and complex scene recognition; and there are specifically optimized models (such as PoseConv3D) that each have advantages in specific fields like skeletal motion recognition. These technologies, which are widely applied in intelligent monitoring, human-computer interaction, medical care, and sports and so on, do provide the core support for the intelligent upgrading of the industries. However, they also come across challenges, for example, not being „sufficiently tough" in complex scenarios, having a difficult time finding a balance between efficiency and accuracy, and also having restricted generalization ability in small-sample situations and so forth. In the future, the video-based action recognition technology will develop in the directions of multi-modal integration, lightness and high efficiency, strong interpretability, and customized scenarios. Through technological innovation and industrial cooperation and so on, it is expected to solve the current core bottleneck problems and further expand the application depth and breadth in smart cities, smart terminals, and special industries, thus providing a more solid support for the implementation of artificial intelligence technology and the development of social intelligence.

## References

[1] Feichtenhofer C, Fan H, Malik J, et al. SlowFast Networks for Video Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3586-3597. DOI: 10.1109/TPAMI.2020.2983687.

[2] Bertasius G, Wang H, Torresani L. Video Swin Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(10): 11972-11985. DOI: 10.1109/TPAMI.2022.3193877.

[3] Sahoo S P. Human Action Recognition Based on Analysis of Video Sequences. Rourkela: National Institute of Technology Rourkela, 2021.

[4] Zhang Y, Li X, Wang L, et al. Light3D: A Lightweight 3D CNN for Edge-Device Video Action Recognition. IEEE Internet of Things Journal, 2023, 10(17): 15218-15228. DOI: 10.1109/JIOT.2023.3289456.

[5] Girdhar R, Carreira J, Doersch C, et al. ViViT: A Video Vision Transformer.Proceedings of the International Conference on Machine Learning. PMLR, 2021: 3202-3212. DOI: 10.48550/arXiv.2103.15691.
[6] Misra I, Shrivastava A, Gupta A, et al. TimeSformer: Is Space-Time Attention All You Need for Video Understanding?. International Journal of Computer Vision, 2022, 130(8): 2083-2101. DOI: 10.1007/s11263-022-01643-4.
[7] Tan M, Le Q V. X3D: Expanding Architectures for Efficient Video Recognition.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13708-13718. DOI: 10.1109/CVPR46437.2021.01358.
[8] Chen M, Huang X L, Wu M. Application of Lightweight 3D CNN in Action Recognition for Intelligent Monitoring . Control and Decision, 2023, 38 (7): 1765-1772. DOI: 10.13195/j.kzyjc.2022.0867.
[9] Zhang S Y, Li Y F, Wang Z Y, et al. PoseConv3D: A Skeletal Action Recognition Method Based on 3D Convolutional Neural Network. Journal of Multimedia, 2025, 27 (3): 1865-1878. DOI: 10.1109/TMM.2024.3468921.
[10] Wang L, Zhao X, Liu C. Research on Video Action Recognition Based on Two-Stream Attention Fusion. Pattern Recognition and Artificial Intelligence, 2022, 35 (4): 335-343. DOI: 10.16451/j.cnki.issn1003-6059.202204007.
[11] Li J, Chen Y, Zhang S, et al. Few-Shot Video Action Recognition with Meta-Transformer. Proceedings of the European Conference on Computer Vision. 2024: 456-473. DOI: 10.48550/arXiv.2403.12157.