Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061376

ARTICLE

Lightweight Classroom Student Action Recognition Method Based on Spatiotemporal Multimodal Feature Fusion

Shaodong Zou[1], Di Wu[1], Jianhou Gan[1,2,*], Juxiang Zhou[1,2] and Jiatian Mei[1,2]

[1]Key Laboratory of Education Informatization for Nationalities, Ministry of Education, Yunnan Normal University, Kunming, 650500, China

[2]Yunnan Key Laboratory of Smart Education, Yunnan Normal University, Kunming, 650500, China

*Corresponding Author: Jianhou Gan. Email: ganjh@ynnu.edu.cn

Received: 23 November 2024; Accepted: 13 January 2025; Published: 26 March 2025

ABSTRACT: The task of student action recognition in the classroom is to precisely capture and analyze the actions of students in classroom videos, providing a foundation for realizing intelligent and accurate teaching. However, the complex nature of the classroom environment has added challenges and difficulties in the process of student action recognition. In this research article, with regard to the circumstances where students are prone to be occluded and classroom computing resources are restricted in real classroom scenarios, a lightweight multi-modal fusion action recognition approach is put forward. This proposed method is capable of enhancing the accuracy of student action recognition while concurrently diminishing the number of parameters of the model and the Computation Amount, thereby achieving a more efficient and accurate recognition performance. In the feature extraction stage, this method fuses the keypoint heatmap with the RGB (Red-Green-Blue color model) image. In order to fully utilize the unique information of different modalities for feature complementarity, a Feature Fusion Module (FFE) is introduced. The FFE encodes and fuses the unique features of the two modalities during the feature extraction process. This fusion strategy not only achieves fusion and complementarity between modalities, but also improves the overall model performance. Furthermore, to reduce the computational load and parameter scale of the model, we use keypoint information to crop RGB images. At the same time, the first three networks of the lightweight feature extraction network X3D are used to extract dual-branch features. These methods significantly reduce the computational load and parameter scale. The number of parameters of the model is 1.40 million, and the computation amount is 5.04 billion floating-point operations per second (GFLOPs), achieving an efficient lightweight design. In the Student Classroom Action Dataset (SCAD), the accuracy of the model is 88.36%. In NTU RGB+D 60 (Nanyang Technological University Red-Green-Blue-Depth dataset with 60 categories), the accuracies on X-Sub (The people in the training set are different from those in the test set) and X-View (The perspectives of the training set and the test set are different) are 95.76% and 98.82%, respectively. On the NTU RGB+D 120 dataset (Nanyang Technological University Red-Green-Blue-Depth dataset with 120 categories), the accuracies on X-Sub and X-Set (the perspectives of the training set and the test set are different) are 91.97% and 93.45%, respectively. The model has achieved a balance in terms of accuracy, computation amount, and the number of parameters.

KEYWORDS: Action recognition; student classroom action; multimodal fusion; lightweight model design

1 Introduction

With the rapid advancement of deep learning technology, the domain of student classroom action recognition has experienced swift development. Nowadays, this technology is gradually integrating into

practical application scenarios such as intelligent education, learning behavior analysis, and online teaching monitoring, showing its great potential in the education field. The core task of student classroom behavior recognition is to accurately identify and classify the classroom behaviors of students presented in the videos. This allows for an accurate evaluation of students' behaviors, offers data to enhance educational tactics, and advances the scientific development of educational research. Through this technology, teachers and educational administrators can have a deeper understanding of students' learning habits, interaction patterns, and learning outcomes, thereby providing more personalized and precise teaching guidance.

Traditionally, action recognition mostly adopts a single-modal strategy, that is, it relies on a single data source, such as RGB video frames or human keypoint data for action recognition. However, in the face of the intricate and variable classroom environment, these single-modal approaches invariably exhibit certain limitations. Although the methods based on RGB video frames can intuitively reflect the scene appearance and context information [1–4], their recognition performance is extremely vulnerable to external factors such as changes in lighting conditions and interference from occlusions; while the methods based on keypoint [5–8], although they show strong robustness in responding to environmental changes, in scenarios that require recognition relying on appearance features, their accuracy is often not satisfactory.

In order to overcome these limitations of single-modal methods, multimodal action recognition methods have gradually come into the spotlight of research [9–12] and have shown significant advantages. By fusing multi-source information such as RGB video frames and keypoint data, multimodal methods can not only effectively capture rich visual appearance features and scene context information, but also use human skeleton information to reduce the negative impact of occlusion and background changes on recognition performance, thereby significantly improving the accuracy and robustness of action recognition. However, while multimodal methods achieve high-precision recognition, they also have the disadvantages of high computational costs and high resource consumption. This poses a non-negligible challenge for practical applications, especially in the resource-constrained classroom environment.

In view of the background described above, this article proposes an innovative lightweight multimodal student action recognition method, aiming to fully utilize the respective advantages of the RGB image modality and the keypoint modality to improve recognition accuracy. At the same time, it focuses on addressing the computational efficiency problem faced by multimodal methods in practical applications. This method achieves the goal of significantly reducing model complexity and computational resource consumption while ensuring recognition accuracy by constructing a dual-branch model based on RGB images and keypoint data and optimizing the network architecture and algorithm design. In the model, the RGB image branch focuses on extracting visual features, while the keypoint branch uses the human skeleton heatmap for accurate action recognition. In addition, we have also designed an efficient multimodal feature encoding and fusion module to fully utilize the unique complementary information of the two modalities to generate a more comprehensive and rich action representation. Specifically, our main contributions can be summarized as follows: (1) We propose a lightweight multimodal action recognition method that takes RGB information and keypoint heatmaps generated based on the keypoint as input. This method includes a feature extraction network with a bidirectional fusion module, which can simultaneously process information from different modalities and effectively fuse them together. (2) We design a Feature Encoding Fusion Module (FFE). After multi-modal information is fused through multiple layers of convolution, the unique features of each modality may be lost. To fully utilize the unique information of these two modalities, we designed a feature encoding fusion module to fuse the early feature information of the modalities. (3) To explore the application of multimodal action recognition technology in the field of education, we have carefully constructed a Student Classroom Action Dataset (SCAD). This dataset covers six common action categories

for students in the classroom and contains more than 10,000 data samples, providing data support for subsequent research.

The structure of this article is as follows: Section 2 presents the related work, specifically, action recognition based on RGB images, action recognition based on keypoint, and action recognition based on multimodal fusion. Section 3 elaborates on the lightweight classroom student action recognition method proposed in this article. Section 4 details some experiments conducted in this article. Section 5 concludes the paper and provides some future research directions.

2 Related Work

2.1 Action Recognition Based on RGB Images

Currently, action recognition based on RGB images is a highly regarded field, aiming to utilize visual information, especially color image data, to recognize different human actions. In action recognition based on RGB images, it is usually necessary to use the color image sequence captured by the camera to extract the key information about human motion. Then, action recognition is performed by analyzing and understanding the spatiotemporal features contained in these dynamic images. Tran et al. [1] pioneered an efficient three-dimensional convolutional architecture specifically designed for accurately extracting spatiotemporal features in video data. Compared with the two-dimensional convolutional network, it shows significant advantages in handling video tasks. Through the use of a deep three-dimensional convolutional network, the spatiotemporal information in the video can be captured more accurately. Lin et al. [13] proposed an innovative temporal shift module, which realizes the efficient exchange of inter-frame information by shifting the channels in the temporal dimension. This design enables TSM to achieve a performance level comparable to that of the 3D convolutional network while maintaining the computation amount of the 2D convolutional network. Feichtenhofer [14] achieves excellent performance by multi-dimensional expansion of the tiny 2D image classification architecture, including space, time, width, and depth. And by adopting a method of gradually expanding the network, focusing on the expansion of only one dimension at a time, a good balance between accuracy and complexity is achieved, maintaining high accuracy while being extremely lightweight in terms of network width and parameters. Li et al. [15] integrated the advantages of 3D convolution and the spatiotemporal self-attention mechanism to achieve a balance between computational efficiency and accuracy. Through a unique relation aggregator, local and global token affinities are learned in the shallow and deep layers respectively, effectively handling the spatiotemporal redundancy and complex dependencies between video frames and achieving a balance between computational cost and accuracy.

In conclusion, the action recognition method based on RGB images uses technologies such as 3D convolution, which can efficiently process temporal and spatial information simultaneously and performs well in accurately capturing image details for action recognition. However, because the image itself is easily interfered by complex background factors such as lighting and occlusion, especially in specific environments such as classrooms, the recognition accuracy of this kind of method may be affected to some extent.

2.2 Action Recognition Based on Keypoint

Action recognition based on keypoint is a popular research direction in the field of computer vision. This method captures the keypoint and skeletal structure of the human body, and uses deep learning and other algorithms to process and analyze the keypoint data, thereby achieving the recognition and classification of human actions. Compared with the traditional action recognition methods based on RGB images or videos, action recognition based on keypoint is not affected by environmental factors such as lighting and background, and has higher accuracy and robustness. Yan et al. [16] proposed a deep learning

network specifically for processing spatiotemporal data. It combines the advantages of graph convolution and spatiotemporal convolution, and can capture the spatiotemporal correlation features in the input graph structure data. The node features are extracted through the graph convolution layer. and the temporal dynamics and spatial correlation of the keypoint data are captured through the spatiotemporal convolution layer. Shi et al. [5] designed a two-stream adaptive graph convolutional network, which solves the problem of the fixed and inflexible graph topology structure in the traditional GCN method. It uses a data-driven approach to learn the topology of the graph, which can not only be learned uniformly to adapt to the global characteristics, but also be learned separately for each sample to capture the subtle differences, thereby enhancing the flexibility and versatility of the model. In addition, a two-stream framework is designed to simultaneously process first-order and second-order information, effectively improving the recognition accuracy. Chen et al. [7] proposed an innovative graph convolutional network CTR-GCN, which realizes the dynamic and effective modeling of the channel topology by introducing the channel topology refinement graph convolution, thereby being able to accurately capture the complex spatiotemporal correlations in the keypoint actions. Different from the traditional GCN, CTR-GCN breaks the limitation that all channels share the same set of topological structures. By refining the training topological structure, the upper limit of the model's ability is further improved. Zhao et al. [17] designed an innovative part-based graph convolutional network. The main idea is to finely divide the skeleton graph into four subgraphs, and these subgraphs share keypoint. Compared with the model using the entire skeleton graph, this model has a significant improvement in recognition performance.

In conclusion, the keypoint-based method shows strong robustness in dealing with background factors such as lighting and occlusion. With the effective use of keypoint, it can more accurately capture the temporal information of actions. However, when this method recognizes some actions, its recognition accuracy decreases due to the neglect of the specific details in the image.

2.3 Action Recognition Based on Multimodal Fusion

Recently, action recognition techniques that fuse features of different modalities have received widespread attention. Traditional methods that rely only on single-modal features, such as RGB images or only keypoint data, often struggle to achieve the desired recognition accuracy in action recognition tasks due to their inherent limitations. In contrast, multimodal methods can effectively utilize the advantages of each modality by fusing features of different modalities, significantly improving recognition accuracy. Multimodal methods can simultaneously consider multiple modality information such as RGB images and keypoint data, and more comprehensively describe the characteristics of actions through complementarity and enhancement. Vaezi Joze et al. [9] proposed an MMTM module that can be conveniently embedded in the feature hierarchy to achieve modal fusion. It can utilize information from multiple modalities to accurately recalibrate channel features inside CNN and complete the fusion of feature modalities in the convolutional layers of different spatial dimensions. Guo et al. [18] designed an innovative bidirectional synchronous cross-spatial attention fusion model, and at the same time introduced a novel motion-oriented human pose representation method, Limb Flow Field (LFF). This method effectively alleviates the temporal ambiguity of human poses during the recognition process, thereby improving the accuracy and robustness of action recognition. Duan et al. [19] used 3D heat maps as the basic representation of human keypoint. Compared with the traditional GCN method, it shows a higher efficiency in spatiotemporal feature learning. It not only captures the dynamic changes of human actions more accurately, but also significantly enhances the resistance to keypoint estimation noise, improving the robustness of the model. At the same time, PoseC3D can fuse the spatiotemporal information of the keypoint heat map and the visual data in the video frame to achieve a more comprehensive and accurate recognition of human actions. Shah et al. [10]

proposed an action recognition method based on multiview videos, which adopts a supervised contrastive learning framework and uses multi-view data to learn view-robust feature embeddings. By improving the contrastive loss and increasing the synchronous view positive samples to improve the model performance, and innovatively using the classifier probability to guide the selection of hard negative samples to enhance the feature discrimination, this method shows stronger domain generalization ability compared to the standard supervised training of synthetic multiview data.

In conclusion, the multimodal-based recognition methods can significantly improve the recognition accuracy by virtue of their ability to integrate different modality information. However, in practical applications, more computational resources are required to process multiple modality information simultaneously, which to some extent increases the difficulty of practical applications and the burden on system resources. Therefore, the model constructed in this study has fully exploited the strength of RGB images to capture details and the advantage that keypoints are less affected by background interference. Meanwhile, the model has been designed to be lightweight, enabling it to connect with the actual classroom needs more effectively and provide strong support and guarantee for classroom teaching.

3 Method

In this section, we will elaborate on our research method. The model architecture we propose consists of two main branches: the RGB branch and the keypoint branch. Both branches are dedicated to performing action recognition by learning the temporal and spatial information of human actions. By fusing the outputs of these two branches, our network can more comprehensively understand and recognize various complex human actions.

3.1 Overall Model Structure

The model structure, as illustrated in Fig. 1, comprises two branches. During the feature extraction stage, the features of different branches are fused, and then the fused features are transmitted to two classification heads for action recognition. We use RGB image data and joint data as the inputs of two different modalities. First, in order to reduce the influence of irrelevant areas in the video image, we use the keypoint as the guiding information to crop the video image. By calculating the maximum and minimum coordinates of the keypoint in the horizontal and vertical directions in each frame, the actual coverage area of the character's actions is determined, thereby retaining the key action information and significantly reducing the unnecessary background area. This process not only improves the recognition accuracy but also reduces the computational cost. For each keypoint, a two-dimensional Gaussian Map is generated with its coordinate as the center. Its intensity is set by the confidence of the keypoint, and its spatial distribution is determined by the Gaussian function. After all the Gaussian Maps of the keypoint are superimposed, a complete pose heatmap is formed, which not only reflects the positions of the keypoint but also includes their confidence information. Then, the cropped video image and the keypoint heatmap are input into the feature extraction network together for in-depth feature extraction. During the feature extraction process, the fusion between modalities is realized through lateral connection to fully integrate the information of the two modalities. Then, the extracted features are sent to the first classifier for action recognition. At the same time, in order to make more comprehensive use of the unique information of the two modalities, we introduce the FFE module. This module fuses and encodes the modal information in the middle layer of the feature extraction network to capture rich spatiotemporal features. The fused features processed by the FFE module are input into the second classifier. Finally, the decision results of the two classifiers are combined through the fusion strategy to obtain a more accurate and comprehensive action recognition result.

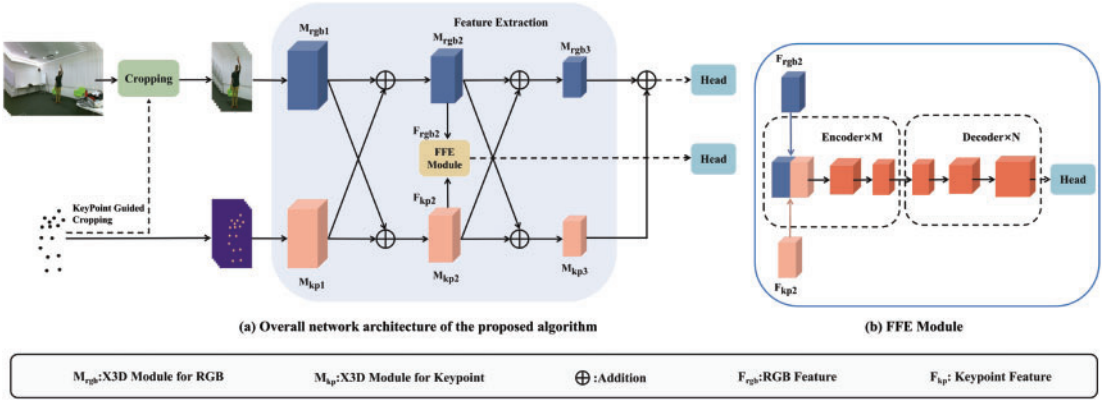

Figure 1: Lightweight multimodal action recognition model structure

3.2 Multimodal Feature Extraction

To improve the model's computational efficiency and reduce the number of parameters, the model first utilizes the first three layers of the X3D network to extract features from the RGB modality and the keypoint modality. Meanwhile, a feature-layer fusion strategy is introduced during the feature extraction process. This strategy is implemented before the feature extraction in the second and third layers, aiming to more effectively integrate information from different modalities. The schematic diagram of fusion visualization is presented in Fig. 2. The fusion operation is carried out through addition and is divided into the following two steps: Step 1: For the RGB feature map, a convolutional operation is used to resize it to the same size as the keypoint feature map. Then, it is added to the feature map of the keypoint modality. Simultaneously, a copy of the keypoint feature map that has not been fused with the RGB feature map is reserved. Step 2: A de-convolution operation is performed on the keypoint feature map that has not been fused with the RGB feature map, rendering its size identical to that of the RGB feature map. Subsequently, the keypoint feature map is incorporated into the RGB branch for fusion.

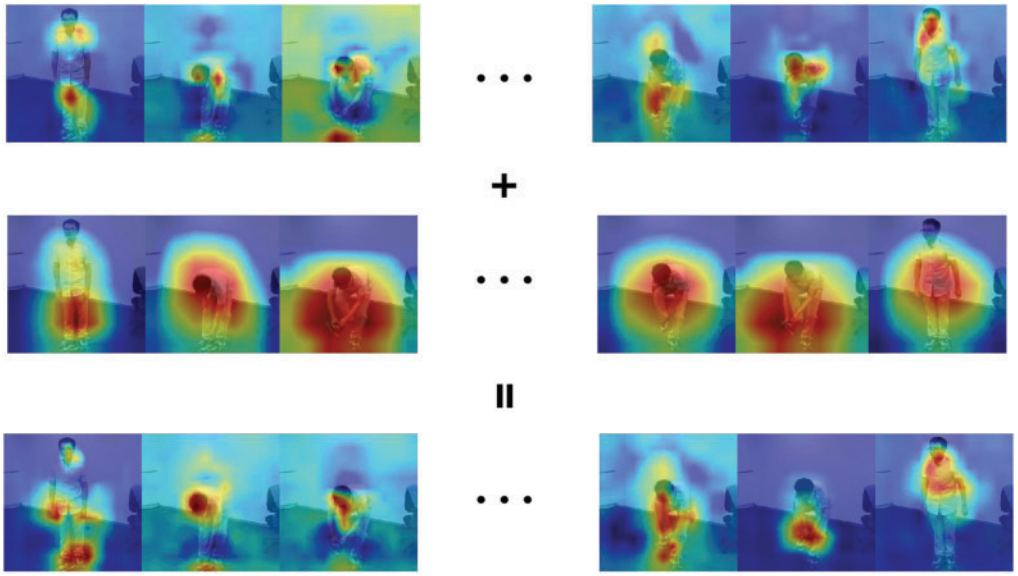

Figure 2: Modal fusion visualization diagram

3.3 Feature Fusion Encoding Module

Although additive fusion of the RGB modality and the keypoint modality can improve the recognition accuracy to a certain extent, with the progression of multilayer convolutional fusion, the unique features of each modality are highly likely to be gradually lost during this process. To make full use of the unique information of these two modalities, we have designed a Feature Fusion Encoding Module (FFE). The core idea of this module is to retain more unique information of each modality by capturing and fusing the shallow features of the RGB modality and the keypoint modality output from the second layer of the feature extraction network. Specifically, it can be divided into three steps: Step 1: Obtain the features of the RGB image modality and the keypoint modality output from the second layer of the feature extraction network, and concatenate them in the Channels dimension as the input of the FFE module. Step 2: Inside the module, an encoder is used to encode the input features. Inter-modality fusion is achieved by reducing the number of Channels. Step 3: A decoder is used to restore the encoded feature map to the original number of Channels for subsequent operations. In addition, to further enhance the fusion effect, we have introduced a residual connection in the decoder part to preserve more original information. The design of this Feature Fusion Encoding Module enables the fusion of the two modalities while retaining the unique information of the RGB and keypoint modalities.

4 Experiments

4.1 SCAD Dataset

Within the constructed Student Classroom Action Dataset (SCAD), we have meticulously collected and labeled six principal student action categories in the classroom, namely, Raising hands, Standing up, Turning back, Listening, Reading, and Taking notes. Each category corresponds to a typical student action, as shown in Fig. 3. Raising hands records the body language when students try to attract the teacher's attention or request to speak. The number of samples in this category in the dataset may reflects the students' active participation and the frequency of their questions. Standing up includes all instances of students getting up from their seats, which may be for answering questions, moving to the blackboard, or participating in classroom activities. The amount of data in this category can reveal the classroom activity level and the students' participation in the course content. Turning back records the action of students changing the direction of their sitting position in the classroom, which may indicate their communication with classmates. Listening represents the focused action that students show when the teacher is lecturing, and it corresponds to the main action mode of students in regular teaching activities. Reading captures the situation where students Reading textbooks, extracurricular books, or other text materials. The amount of data for the Reading action can provide clues about students' self-study habits. Taking notes reflects the action of students Taking notes when Listening or Reading. The data distribution of this category may show students' preferences for information recording and organization. Among these six categories, the data volume of each category is shown in Fig. 4, where there are 1404 data for Raising hands, 1829 data for Standing up, 1918 data for Turning back, 1798 data for Listening, 1869 data for Reading, and 1873 data for Taking notes.

Figure 3: Sample of SCAD dataset

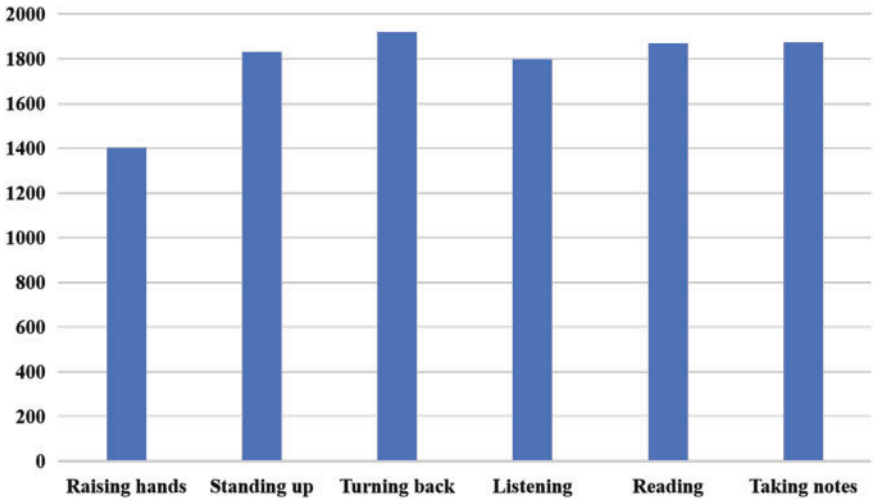

Figure 4: SCAD dataset category data volume

4.2 Experimental Analysis

To thoroughly and comprehensively evaluate the effectiveness of our proposed method, tests were conducted on the NTU RGB+D 60 dataset and the NTU RGB+D 120 dataset. The NTU RGB+D 60 dataset contains more than 50,000 video samples, covering 60 different action categories. Each sample contains data in four modalities: RGB video, depth map sequence, 3D skeleton data, and infrared (IR) video. The resolution of RGB videos is 1920 × 1080. Such high-resolution videos can capture more detailed information and help improve the accuracy of action recognition. The NTU RGB+D 120 dataset is expanded to 120 categories on the basis of the NTU RGB+D 60 dataset, and the number of samples is expanded to more than 100,000. In addition, to further verify the performance of our method in the classroom, we also conducted experiments on the student action dataset SCAD. To evaluate the performance of the algorithm, we evaluated the accuracy as well as the number of parameters and the computation amount of the model.

4.2.1 NTU RGB+D Comparative Experiment

To evaluate the performance of the proposed action recognition method in the public dataset, we performed detailed experiments on the NTU RGB+D 60 dataset and the NTU RGB+D 120 dataset and compared it with existing methods. The experimental results of the NTU RGB+D 60 dataset are shown in Table 1. In the X-Sub test set, our method achieved a Top-1 accuracy rate of 95.76%, while on the X-View test set, the accuracy rate was 98.82%. The experimental results of the NTU RGB+D 120 dataset are shown in Table 2. In the X-Sub test set, our method achieved a Top-1 accuracy rate of 91.97%, while on the X-Set test set, the accuracy rate was 93.45%, exceeding other compared methods, showing the accuracy and efficiency of our method.

Table 1: Comparison of Top-1 accuracy on NTU RGB+D 60 dataset

Methods	K	R	X-Sub	X-View
ST-LSTM [20]	✓	-	69.2%	77.7%
View-invariant [21]	✓	-	80.0%	87.2%
STGCN [16]	✓	-	81.5%	88.3%
2s-AGCN [5]	✓	-	88.5%	95.1%
DGNN [22]	✓	-	89.9%	96.11%
MS-G3D [23]	✓	-	91.5%	96.2%
CTR-GCN [7]	✓	-	92.4%	96.8%
C3D [1]	-	✓	63.5%	70.3%
HybridNet [24]	-	✓	86.5%	88.5%
Glimpse clouds [25]	-	✓	86.6%	93.2%
STAR-Transformer [26]	✓	✓	92.0%	96.5%
TSMF [27]	✓	✓	92.5%	97.4%
ViewCon [10]	✓	✓	93.7%	**98.9%**
PoseC3D [19]	✓	✓	94.1%	97.1%
Our	✓	✓	**95.76%**	98.82%

Table 2: Comparison of Top-1 accuracy on NTU RGB+D 120 dataset

Methods	K	R	X-Sub	X-Set
ST-LSTM [20]	✓	-	58.2%	60.9%
GCA-LSTM [28]	✓	-	58.3%	59.2%
STGCN [16]	✓	-	83.5%	85.2%
2s-AGCN [5]	✓	-	84.2%	86.0%
MS-G3D [23]	✓	-	86.9%	88.4%
CTR-GCN [7]	✓	-	88.9%	90.6%
ViewCon [10]	-	✓	85.6%	87.5%
DVANet [29]	-	✓	91.6%	90.4%
VT-BPAN [30]	✓	✓	86.3%	88.2%
TSMF [27]	✓	✓	87.0%	89.1%
VPN++ +3D Pose [31]	✓	✓	90.7%	92.5%
STAR-Transformer [26]	✓	✓	90.3%	92.7%
Our	✓	✓	**91.97%**	**93.45%**

4.2.2 Accuracy, Parameter Number and Computation Amount Evaluation on the SCAD Dataset

To evaluate the performance of the proposed action recognition method in the classroom environment, we performed experiments on the self-constructed dataset and compared it with several existing advanced methods. These methods include methods based on skeletal data, methods based on RGB images, and multimodal PoseC3D. Our experiment aims to evaluate the performance of each method on three key metrics: Action Recognition Accuracy (Accuracy), Number of Model Parameters (params), and Computation Amount (GFLOPs). We use Acc, P, and G to represent these metrics, respectively. The experimental results are shown in Table 3. Our method achieved a Top-1 accuracy of 88.36%, exceeding other comparison methods. In addition, the number of our model parameters is only 1.40 M, and the computation amount is 5.04 GFLOPs, showing high computational efficiency. Compared with the methods based on skeletal data, our method not only has a significant improvement in accuracy, but also reduces the number of model parameters while maintaining a low computation amount. For example, the accuracy of MS-G3D is relatively low, and the computation amount is 6.8 GFLOPs, which is higher than our method. When compared with the methods based on RGB data, our method also shows better performance. Especially when compared with TimesFormer, which uses a large number of parameters and has a high computation amount, our method significantly reduces the requirements for the number of parameters and computation amount while maintaining a similar or even higher accuracy.

Table 3: Performance comparison on the SCAD dataset

Methods	K	R	Acc	P(M)	G
2s-AGCN [5]	✓	-	80.58%	4.4	3.5
STGCN [16]	✓	-	82.40%	3.8	3.1
CTR-GCN [7]	✓	-	82.85%	1.436	**1.95**
MS-G3D [23]	✓	-	83.69%	2.95	6.8
C3D [1]	-	✓	85.62%	78.4	38.5
R2plusld [32]	-	✓	82.67%	63.8	53.1

(Continued)

Table 3 (continued)

Methods	K	R	Acc	P(M)	G
SlowFast [33]	-	✓	86.60%	34.47	66.1
TimesFormer [34]	-	✓	87.76%	86.11	141
PoseC3D [19]	✓	✓	87.75%	36.0	58.25
Our	✓	✓	**88.36%**	**1.40**	5.04

4.2.3 Ablation Experiment

In this study, we designed a series of ablation experiments to evaluate the impact of different input modalities and their combinations on the performance of the video action recognition model. The experimental results are shown in Table 4. The method using RGB images alone generally performs better on the three datasets than the method using only poses. This may be because RGB images provide richer environmental and texture information. However, the experiment shows that fusing pose and RGB data can further improve the accuracy of the model. It is worth noting that when the model uses only single-frame RGB and single-frame pose data, the performance drops significantly, which emphasizes the importance of multi-frame data in capturing the dynamic characteristics of actions. We also compared removing the FFE module and using the Cat to concatenate the feature maps to replace the FFE module. In comparison, our complete model, by comprehensively using multi-frame data and the FFE module, shows the best performance on all datasets. This further confirms the key role of the multimodal fusion method in improving the accuracy and robustness of video action recognition.

Table 4: Performance comparison on the SCAD dataset

Methods	SCAD	X-Sub	X-View
Only Pose	79.12%	90.75%	94.99%
Only RGB	87.39%	94.60%	98.39%
Our (w/o FFE)	87.59%	95.15%	98.42%
Our (w/o FFE+Cat)	87.92%	95.62%	98.43%
Our (1 frame)	85.72%	65.81%	61.71%
Our	**88.36%**	**95.76%**	**98.82%**

4.2.4 Visualization Analysis

To verify the performance of our multimodal method in capturing key actions, we used the Gradient-weighted Class Activation Mapping (Grad-CAM) method to conduct an in-depth visualization analysis of the model on the NTU RGB+D dataset and the SCAD dataset. As shown in Fig. 5, the method that only relies on keypoint information appears to have a relatively broad focus range and lacks a clear focus; while the method that uses RGB images alone has a more concentrated area of attention, but it may miss other information that is crucial for action recognition. In contrast, our multimodal method fuses the areas of concern of both, not only precisely focusing on the important areas of the key actions, but also taking into account other secondary areas that contribute to the recognition, thereby achieving a more comprehensive and accurate action recognition. In Fig. 6, we show some recognition cases. After analysis, we found that the keypoint-based method has better recognition accuracy for actions with large amplitudes, while the RGB

image-based method is more superior in recognizing subtle actions. Our method combines the advantages of both, so the recognition accuracy is higher. In addition, we also present the cases where the algorithm makes mistakes. We analyzed these errors and believe that the reason may be the interference caused by other secondary actions when the person in the video is performing the main action.

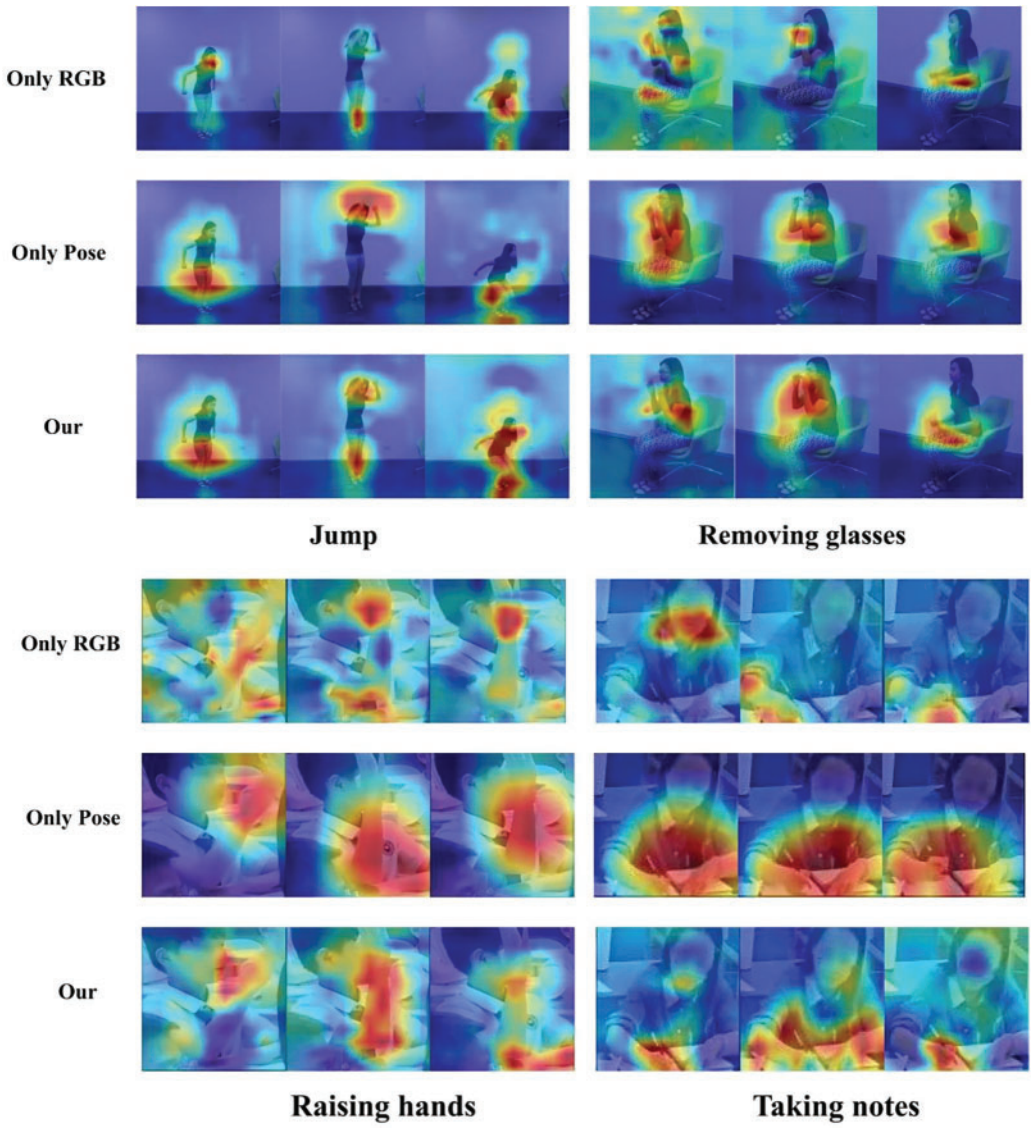

Figure 5: Data set attention visualization chart

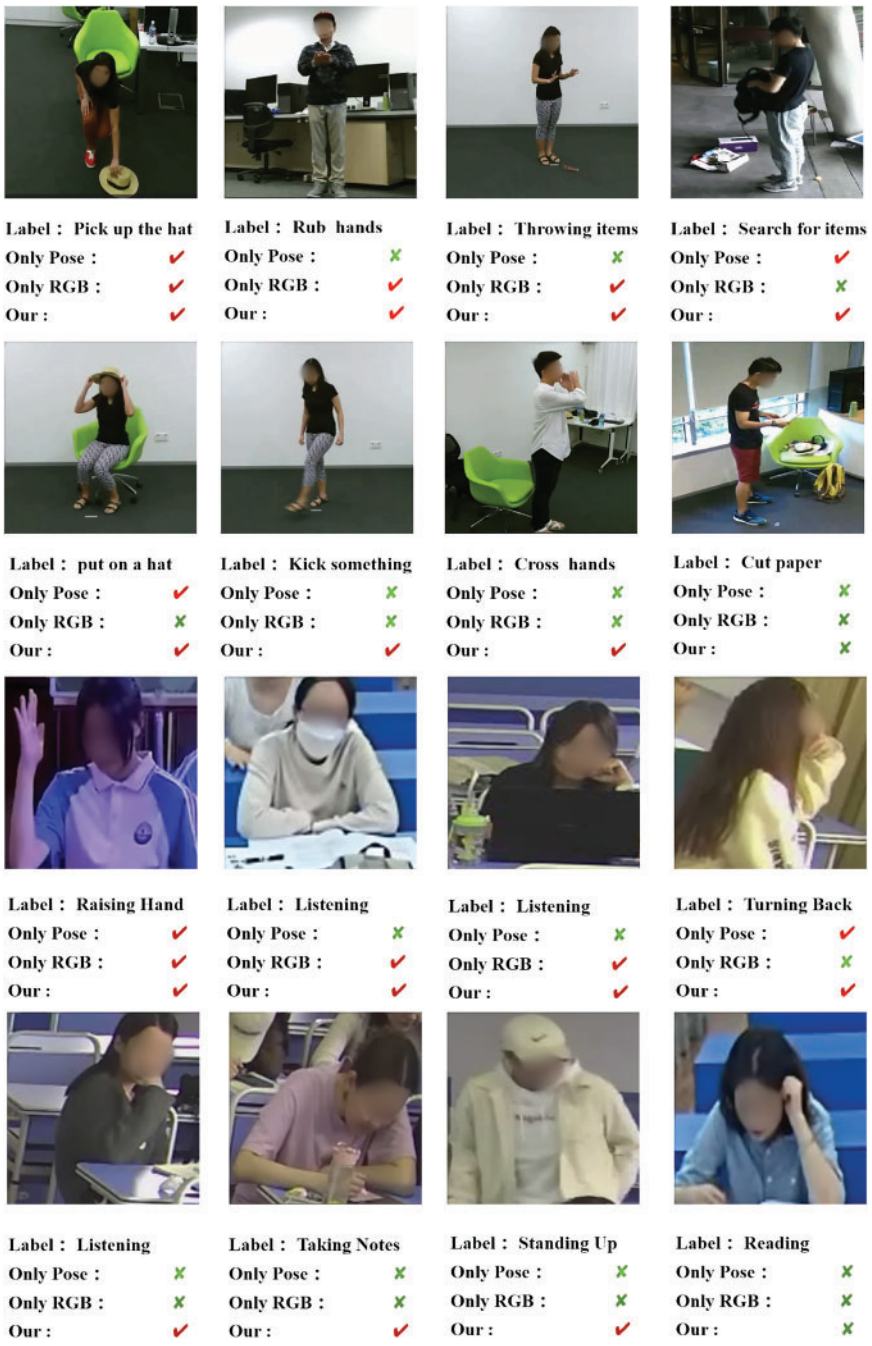

Figure 6: Recognition case demonstration

5 Conclusion

This article proposes a lightweight classroom student action recognition method based on multimodal feature fusion and complementarity. This method fuses RGB image information and keypoint data and effectively integrates the unique information of the two modalities through the Feature Fusion Encoding (FFE) module, significantly enhancing the accuracy of action recognition and the computational efficiency of the model. Experimental results demonstrate that on the self-constructed SCAD dataset, our method achieves an accuracy rate of 88.36%, exceeding other comparison methods. On the public datasets NTU

RGB+D 60 and NTU RGB+D 120, our method also outperforms other comparison methods. In addition, the number of parameters of our model is 1.40 M, and the computational cost is 5.04 GFLOPs. Compared with other methods with similar accuracy, it has a significant advantage in terms of model complexity and exhibits higher resource utilization efficiency. Moreover, we have also constructed a student classroom action dataset covering a variety of student classroom action categories, laying a solid foundation for further research. In future research, introducing large-scale models and implementing specific lightweight improvement strategies for these large-scale models can be considered. Meanwhile, the range of action categories in the classroom can be further expanded.

Acknowledgement: I would like to express my sincere gratitude to everyone who has made efforts for this article. It is with everyone's concerted efforts that this article has been accomplished.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grant 62107034, the Major Science and Technology Project of Yunnan Province (202402AD080002), and Yunnan International Joint R&D Center of China-Laos-Thailand Educational Digitalization (202203AP140006).

Author Contributions: Shaodong Zou and Di Wu carried out the experiments and composed the paper. Jianhou Gan, Juxiang Zhou and Jiatian Mei offered data support for this paper. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to personal privacy concerns, the datasets generated and/or analyzed during the current study are not publicly available at the moment.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE/CVF International Conference on Computer Vision (ICCV); 2015; Santiago, Chile. p. 4489–97. doi:10.1109/ICCV.2015.510.
2. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 7794–803. doi:10.1109/CVPR.2018.00813.
3. Ahmed M, Ramzan M, Ullah Khan H, Iqbal S, Attique Khan M, Choi JI, et al. Real-time violent action recognition using key frames extraction and deep learning. Comput Mater Contin. 2021;59(2):2217–30. doi:10.32604/cmc.2021.018103.
4. AlQaralleh EA, Aldhaban F, Nasseif H, Alksasbeh MZ, Alqaralleh BAY. Smart deep learning based human behaviour classification for video surveillance. Comput Mater Contin. 2022;72(3):5593–605. doi:10.32604/cmc.2022.026666.
5. Shi L, Zhang Y, Cheng J, Lu H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA. p. 12026–35. doi:10.1109/CVPR.2019.01230.
6. Wang Y, Xia Y, Liu S. BCCLR: a skeleton-based action recognition with graph convolutional network combining behavior dependence and context clues. Comput Mater Contin. 2024;79(3):4489–507. doi:10.32604/cmc.2024.048813.
7. Chen Y, Zhang Z, Yuan C, Li B, Deng Y, Hu W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021; Montreal, QC, Canada. p. 13359–68. doi:10.1109/ICCV48922.2021.01311.
8. Dong J, Liu W, Zheng Z, Xie W, Wang L, Mao L, et al. Intercity rail platform abnormal action recognition based on a skeleton tracking and recognition framework. Mach Vis Appl. 2024;35(6):131. doi:10.1007/s00138-024-01608-1.

9. Vaezi Joze HR, Shaban A, Iuzzolino ML, Koishida K. MMTM: multimodal transfer module for Cnn fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA. p. 13289–99. doi:10.1109/CVPR42600.2020.01330.
10. Shah K, Shah A, Lau CP, de Melo CM, Chellapp R. Multi-view action recognition using contrastive learning. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023; Waikoloa, HI, USA. p. 3381–91. doi:10.1109/WACV56688.2023.00338.
11. Wu D, Wang J, Zou W, Zou S, Zhou J, Gan J. Classroom teacher action recognition based on spatio-temporal dual-branch feature fusion. Comput Vis Image Understanding. 2024;247:104068. doi:10.1016/j.cviu.2024.104068.
12. Liang W, Xu X. HgaNets: fusion of visual data and skeletal heatmap for human gesture action recognition. Comput Mater Contin. 2024;79(1):1089–103. doi:10.32604/cmc.2024.047861.
13. Lin J, Gan C, Han S. TSM: temporal shift module for efficient video understanding. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019; Seoul, Republic of Korea. p. 7083–93. doi:10.1109/I-CCV.2019.00718.
14. Feichtenhofer C. X3D: expanding architectures for efficient video recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA. p. 203–13. doi:10.1109/CVPR42600.2020.00028.
15. Li K, Wang Y, Gao P, Song G, Liu Y, Li H, et al. Uniformer: unified transformer for efficient spatiotemporal representation learning. arXiv:2201.04676. 2022.
16. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proc AAAI Conf Artif Intell. 2018;32(1). doi:10.1609/aaai.v32i1.12328.
17. Zhao M, Dai S, Zhu Y, Tang H, Xie P, Li Y, et al. PB-GCN: progressive binary graph convolutional networks for skeleton-based action recognition. Neurocomputing. 2022;501:640–9. doi:10.1016/j.neucom.2022.06.070.
18. Guo F, Jin T, Zhu S, Xi X, Wang W, Meng Q, et al. B2C-AFM: bi-directional co-temporal and cross-spatial attention fusion model for human action recognition. IEEE Trans Image Process. 2023. doi:10.1109/TIP.2023.3308750.
19. Duan H, Zhao Y, Chen K, Lin D, Dai B. Revisiting skeleton-based action recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; New Orleans, LA, USA. p. 2969–78. doi:10.1109/CVPR52688.2022.00298.
20. Liu J, Shahroudy A, Xu D, Kot AC, Wang G. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. IEEE Trans Pattern Anal Mach Intell. 2017;40(12):3007–21. doi:10.1109/TPAMI.2017.2771306.
21. Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit. 2017;68:346–62. doi:10.1016/j.patcog.2017.02.030.
22. Shi L, Zhang Y, Cheng J, Lu H. Skeleton-based action recognition with directed graph neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA. p. 7912–21. doi:10.1109/CVPR.2019.00810.
23. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA. p. 143–52. doi:10.1109/CVPR42600.2020.00022.
24. Wang H, Song Z, Li W, Wang P. A hybrid network for large-scale action recognition from rgb and depth modalities. Sensors. 2020;20(11):3305. doi:10.3390/s20113305.
25. Baradel F, Wolf C, Mille J, Taylor GW. Glimpse clouds: human activity recognition from unstructured feature points. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 469–78. doi:10.1109/CVPR.2018.00056.
26. Ahn D, Kim S, Hong H, Ko BC. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023; Waikoloa, HI, USA. p. 3330–9. doi:10.1109/WACV56688.2023.00333.
27. Yu BXB, Liu Y, Chan KCC. Multimodal fusion via teacher-student network for indoor action recognition. Proc AAAI Conf Artif Intell. 2021;35(4):3199–207. doi:10.1007/s00138-024-01598-0.
28. Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC. Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans Image Process. 2017;27(4):1586–99. doi:10.1109/TIP.2017.2785279.

29. Siddiqui N, Tirupattur P, Shah M. DVANet: disentangling view and action features for multi-view action recognition. Proc AAAI Conf Artif Intell. 2024;38(5):4873–81. doi:10.1609/aaai.v38i5.28290.
30. Sun Y, Xu W, Yu X, Gao J. VT-BPAN: vision transformer-based bilinear pooling and attention network fusion of rgb and skeleton features for human action recognition. Multimed Tools Appl. 2023;83:1–15. doi:10.1007/s11042-023-17788-3.
31. Das S, Dai R, Yang D, Bremond F. VPN++: rethinking video-pose embeddings for understanding activities of daily living. IEEE Trans Pattern Anal Mach Intell. 2021;44(12):9703–17. doi:10.1109/TPAMI.2021.3127885.
32. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 6450–9. doi:10.1109/CVPR.2018.00675.
33. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019; Seoul, Republic of Korea. p. 6202–11. doi:10.1109/I-CCV.2019.00630.
34. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? arXiv:2102.05095. 2021.