IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 November 2023, accepted 5 December 2023, date of publication 8 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3341035

RESEARCH ARTICLE

Action Behavior Learning Based on a New Multi-Scale Interactive Perception Network

CHUNLIN ZHENG[1,2], JUN GU[3], AND SHENG XU[4], (Member, IEEE)
[1]Teacher's Training Center of Jiangsu Province, Jiangsu Second Normal University, Nanjing 210013, China
[2]School of Philosophy, Beijing Normal University, Beijing 100875, China
[3]Wuxi Education Informatization and Equipment Management Service Center, Wuxi 214000, China
[4]Department of Computer Science and Technology, Nanjing Forestry University, Nanjing 210037, China

Corresponding author: Sheng Xu (xusheng@njfu.edu.cn)

This work was supported by the Major National Social Science Project under Grant 19ZDA041.

ABSTRACT Action recognition is a fundamental research topic in the field of video understanding, but classical action recognition relies on a large amount of manually annotated video data, which limits its development. Small sample action recognition is a promising topic that can overcome the dependence on large-scale annotated data. However, the current small sample action recognition has a series of shortcomings such as temporal singularity and a lack of perception of global information. Therefore, this work proposes an interactive perception network by designing a multi-scale temporal feature extraction module to capture global temporal dependencies between all frames and local temporal information between frames. Then the algorithm proposes a cross scale matching strategy to achieve robust matching between videos with different motion speeds, maximizing the consistency between local and global features of the same type of action. Finally, the experimental results on the SSV2 dataset and the HMDB51 dataset show that the proposed method outperforms the current mainstream methods. Compared with the backbone network, which is the most advanced method of ResNet, this method achieved performance improvements of 0.8% and 0.4% on HMDB51 and SSV2, respectively.

INDEX TERMS Action recognition, small sample learning, video understanding, meta learning.

I. INTRODUCTION

With the development of the internet, there has been an increasing number of video creators and platforms, leading to the growing popularity of the live streaming industry and the generation of a large amount of video data. Video understanding tasks such as action recognition can greatly benefit users and the live streaming industry. Moreover, autonomous driving and robotics technology also rely on the support and assistance of action behavior recognition. However, traditional recognition methods heavily rely on a large volume of manually annotated video data. The sensitivity of tags and the time-consuming process of video collection have limited the progress in this field. For instance, there is often a scarcity of video types, making it particularly important to reduce dependence on a large amount of video data. The education industry can also benefit from advancements in video understanding tasks.

In the field of online learning, with the widespread use of the internet, there has been a surge in video creators and platforms providing educational content, resulting in a significant amount of video data. However, there has been a lack of interactivity in courses, and platforms have inadequate understanding of learners' behaviors, leading to the inability to provide timely and relevant teaching recommendations. Action recognition, which involves understanding human actions and movements, can greatly enhance the experience of online learning. Additionally, in offline learning, the identification of human behavior is particularly important in various educational scenarios, such as understanding the behaviors of teachers and students in classrooms and identifying risky behaviors in laboratories. However, traditional methods of action recognition heavily depend on a large volume of manually annotated video data. The process

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

of manual tagging and video collection is not only time-consuming but also sensitive to the accuracy of the assigned labels. Moreover, the availability of labeled video types is often limited, further constraining the development of action recognition in the education industry.

Therefore, reducing reliance on a large amount of video data is crucial. Innovative approaches using action recognition algorithms and machine learning techniques, such as transfer learning, semi-supervised learning, and active learning, can help overcome these limitations. By efficiently utilizing existing annotated data and minimizing the need for extensive manual annotation, the education industry can harness the potential of action recognition technology to provide an exceptional learning experience for students and educators.

Action recognition has made significant progress in recent years. To address the fact that the action recognition relies on a large amount of manually annotated video data, few-shot action recognition is a promising direction that can overcome the dependence on large-scale annotated data. Currently, there are two issues with current few-shot action recognition:

(1) due to different motion speeds in videos, information from a single temporal scale cannot fully represent the semantic behavior of actions;

(2) due to the lack of perception of global information in video actions, matching between local information is inaccurate.

Besides, in small sample action recognition, a high-quality classifier is established from limited labeled data. When faced with the same video data information, compared with the existing sample action recognition methods, there are fewer datasets available for small sample action recognition. Therefore, improving the recognition accuracy and generalization performance of small sample action recognition models has become the key work in this study.

To address the above-mentioned issues, this work focuses on an innovative and effective method called multi-scale interaction perception network. The proposed work has three advantages: (1) to obtain multiple temporal scale semantics for action behavior, a multi-scale temporal feature extraction module is designed to capture global temporal dependencies between all frames and local temporal information between frames. (2) compared with the commonly used alignment matching between single scales or granularities in previous methods, the cross-scale matching strategy can achieve robust matching between videos with different motion speeds. (3) a global information interaction matching module is designed to match global information and local frame-level features of videos, thus maximizing consistency between local features and global features within the same category of actions.

In summary, the main content of this work is to conduct research on small sample action recognition in the case of limited labeled video samples. The proposed network framework has completed algorithm validation on two action recognition datasets, demonstrating the effectiveness of the core module and its superiority over other advanced methods. The research contributions of this article can be summarized as follows:

(1) This study strengthens the matching between videos from two aspects: the fusion of long-term and short-term temporal dependencies, as well as the mutual perception of global and local information.

(2) This study designs a multi-scale Temporal Feature Extraction Module (MTFEM) and a Cross Scale Alignment Module (CSAM) to enhance the temporal understanding of actions from both characterization and measurement aspects, achieving matching between actions with different motion speeds.

In experiments, this work compares the test results of the multi-scale temporal interactive perception network model on the SSV2 dataset and HMDB51 dataset with the current state-of-the-art methods, demonstrating the superiority and effectiveness of the proposed method.

The main research content and structural arrangement are as follows: Section I is an introduction, which introduces the background and significance of action behavior recognition based on small sample learning. Section II is related works, which discusses the research on small sample image recognition and action recognition, providing a solid theoretical basis for the research and design of the methods. Section III is small sample action recognition based on multi-scale temporal perception. Section IV is the experiment and analysis, which introduces the dataset and evaluation, clarifies the experimental environment and training details, conducts detailed comparative experiments and analysis, proves the superiority of the proposed method. Section V is the conclustion and outlook, summarizing the work content, analyzing the certain shortcomings of the method, and the directions for future improvement and research in small sample action recognition.

II. RELATED WORKS

Small shot action recognition aims to classify action categories that have never appeared in the training set through a very small number of annotated video samples, thereby overcoming the dependence on a large number of annotated samples. The input for small sample action recognition is usually a video sequence of multiple samples, divided into a query set and a support set, and the output is the classification results of the query set video sequence.

The existing methods are relatively systematic, mostly based on metric meta learning paradigms, and then align or pool the video representation in time series into a vector to measure the similarity between videos. Figure 1 illustrates the basic process of small sample action recognition methods. At present, small sample action recognition based on meta learning paradigms, which can be roughly classified into four categories: classifier based classification methods, temporal alignment based methods, spatio temporal relationship modeling methods, and methods combined with other modalities.

In small sample image recognition tasks, deep learning models often adopt a meta learning based paradigm.

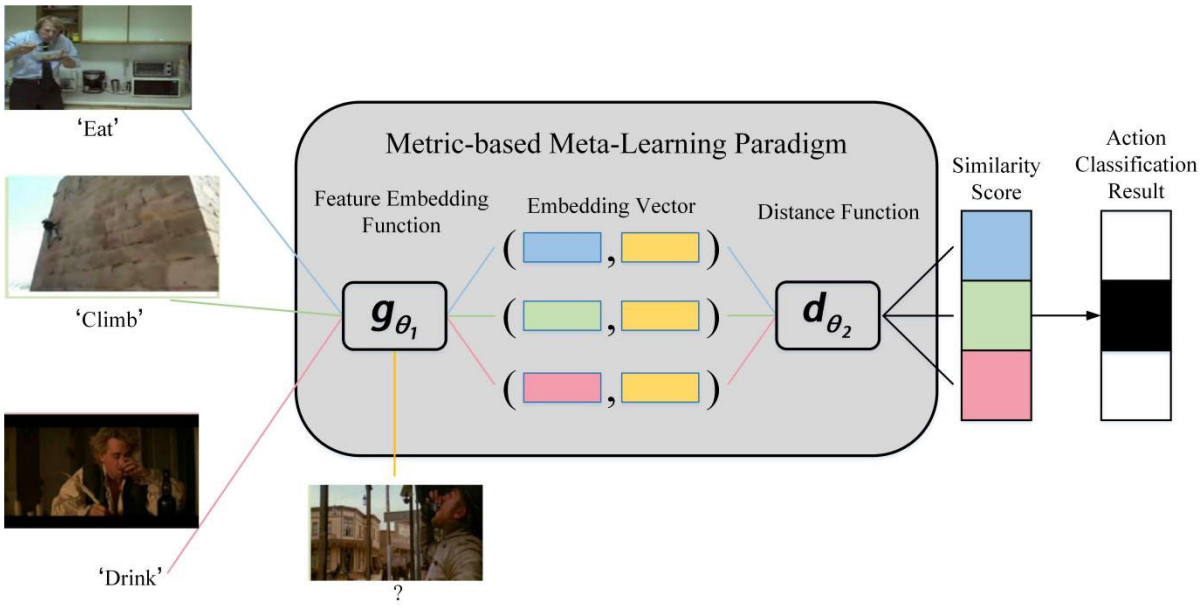


FIGURE 1. The framework of the general small sample action recognition algorithm, including feature embedding function, distance function, similarity score calculation and the classification.

It learns a feature representation with certain generalization performance, and then uses a classifier for classification. Inspired by the idea of small sample learning, The method CMN [1] proposed a composite memory network structure to store category features in the support set. Then, a multi-level saliency encoding algorithm was used to encode video sequences with unfixed length in the query set into fixed length video representations. Then, the nearest neighbor category was searched from the composite memory network to match with the videos in the query set. Although [1] has achieved certain results, it has not fully considered the importance of complex temporal information in video understanding.

In small sample action recognition, feature alignment based on time series is often complex and multi-level. The method TARN [2] began to utilize the temporal information of videos, and in the vectorization stage, TARN extracted video features using 3D CNN (C3D), and strengthened temporal information using Gated Recurrent Unit (GRU) module structure. In TARN, this temporal alignment is relatively rough and does not achieve accurate alignment. The true temporal alignment should be complex. The method OTAM [3] maintains the sequence of frames and uses an improved dynamic time warping algorithm to align two videos in time series. This method displays the alignment at the framelevel for modeling, but the computational complexity of displaying modelingtiming information is enormous. The method TRX [4] matches actions through a large number of video subsequence. This subsequence matching method can deal with the diversity of actions, such as different timing intervals and different action start times. However, TRX may experience performance degradation in action recognition with multiple objects or complex backgrounds, and the tuple representation in TRX is fixed and not flexible enough. The method MTFAN [5] proposed a segmented attention approach to achieve multi-level temporal alignment. Therefore, how to choose the appropriate granularity alignment method for semantic alignment is crucial for small sample action recognition.

In video, in addition to model temporal information, capturing spatial information within a single frame and temporal relationships between frames is of great help in encoding spatio temporal context information. Strengthening spatio temporal information at the same time is beneficial for obtaining discriminative features of specific categories, as well as focusing on the motion information of objects and objects highly related to categories in the video. The previous method OTAM [3] explicitly performed frame level alignment, which had a significant computational burden. In order to reduce this huge computational burden, The method ITANet [6] models spatio temporal information and achieves more robust video matching through an implicit frame level temporal alignment strategy. However, due to different action instances, achieving semantic alignment is relatively difficult. The problem comes from two aspects: (1) the start and duration of the action are different; (2) the evolution process of actions may be inconsistent. The previous semantic alignment work did not address the issue of significant action differences. Therefore, the method TA2N [7] proposed a two-stage action behavior alignment network to achieve more accurate semantic alignment. The method STRM [8] designed a module for enhancing spatio temporal information, utilizing two sub modules: local patch module and global frame level module to help understand

contextual information at both temporal and spatial levels. In the end, STRM obtained richer spatio temporal representations, enabling better learning and matching of the relationships between query targets and actions.

In the field of video, fusing RGB information with other modal information often yields richer information, which is of great help for various video comprehension tasks. In small sample action recognition tasks, there have been some works that combine multiple modalities. The method AMeFu Net [9] introduces depth information as a supplement to scene information and integrates it with RGB information. Huang et al. [10] introduced object information as multimodal information and used multiple relational encoders to encode object information, temporal information, and global information. Although the introduction of multimodal information can greatly improve performance, it often brings significant time and computational overhead.

Although the above-mentioned methods have achieved high performance by conducting temporal matching on videos at various granularities, a single granularity temporal matching ignores the complex temporal information of the video. Especially when facing videos with different motion speeds, a single temporal matching often leads to temporal misalignment, which affects matching performance.

III. A NEW MULTI-SCALE TEMPORAL INTERACTION PERCEPTION NETWORK

A. SMALL SAMPLE LEARNING NETWORKS

In the process of small sample learning, model training randomly samples a batch of data from the training set, and such a training task is called a meta task. Each meta task is divided into a support set and a query set. Finally, model testing evaluates the performance of the model by executing meta tasks on the test set. The following is an introduction to several small sample learning classical networks based on metric learning.

The prototype network [11] (as shown in Figure 2) uses a 4-layer CNN as the feature embedding function. The prototype of each category is defined as the average of the feature vectors of all supporting set images. Expressed as follows by Eq.(1):

$$V_c = \frac{1}{|S^c|} \sum_{(x_k, y_k) \in S^c} g(x_k) \tag{1}$$

Among them, S^c is the number of samples in a certain category and 'g()' is the feature embedding function. The classification result is obtained by calculating the Euclidean distance between the support set sample prototype and the query set sample prototype. In addition, in order to generate more discriminative embedded features, Zhang et al. [12] proposed using comparative loss to bring samples of the same class closer in the feature space and samples of different classes farther away.

The matching network [13] (as shown in Figures 3) uses cosine similarity measurement to learn the feature representations of each category. Given the support set and query image x, the matching network obtains the probability distribution of output label y by calculating attention scores. The attention score is generally obtained by calculating the cosine similarity between the query image and the support set image in the feature space, and then normalizing the attention score. The calculation formula is as follows:

$$a(x, x_k) = e^{\cos(f(x), g(x_k))} \Big/ \sum_{k=1}^{t} e^{\cos(f(x), g(x_k))} \tag{2}$$

Although progress has been made in the field of small sample image recognition, it is unreasonable to directly apply these methods to the field of small sample action recognition due to the complex structure and richer information of videos.

Various deep learning based network models have made tremendous progress in the field of image recognition, and at the same time, researchers have also attempted to transfer image recognition models to videos. In the field of video understanding, action recognition is a fundamental task and also one of the evaluation tasks for various video algorithms. Action recognition is also a very challenging task, videos contain complex temporal information. The current action recognition algorithms have not yet reached the level of human visual perception systems, but in recent years, due to the development of deep neural networks, there has also been some development in the field of action recognition.

Convolutional operation is the most fundamental component of deep neural networks for action recognition tasks, mainly divided into 2D convolution and 3D convolution. Image based 2D convolution is the fundamental operation of deep neural networks. The method based on 2D convolutional networks [14], [15] can also directly load pre-trained weights on large-scale image datasets. However, 2D convolutional networks cannot model temporal information in videos, so additional network design is needed for reasonable temporal modeling.

Due to the large number of frames in the video, 3D convolution can capture temporal information over a short period of time. But the method based on 3D convolutional neural network [16], [17] has more parameters than 2D convolutional network. Furthermore, these methods cannot load pre-trained weights for large-scale image datasets.

Time series modeling is important in video action recognition. Generally speaking, temporal modeling can be divided into three types. The most direct method is to directly use 3D convolution on adjacent frames. Therefore, the temporal dimension in 3D convolution can capture temporal motion information from adjacent frames, but its limitation lies in the inability to use pre-trained weights on large-scale image datasets.

Another type of method is to model temporal information in videos through a multi-stream approach. One of the streams is trained using optical flow frames to capture

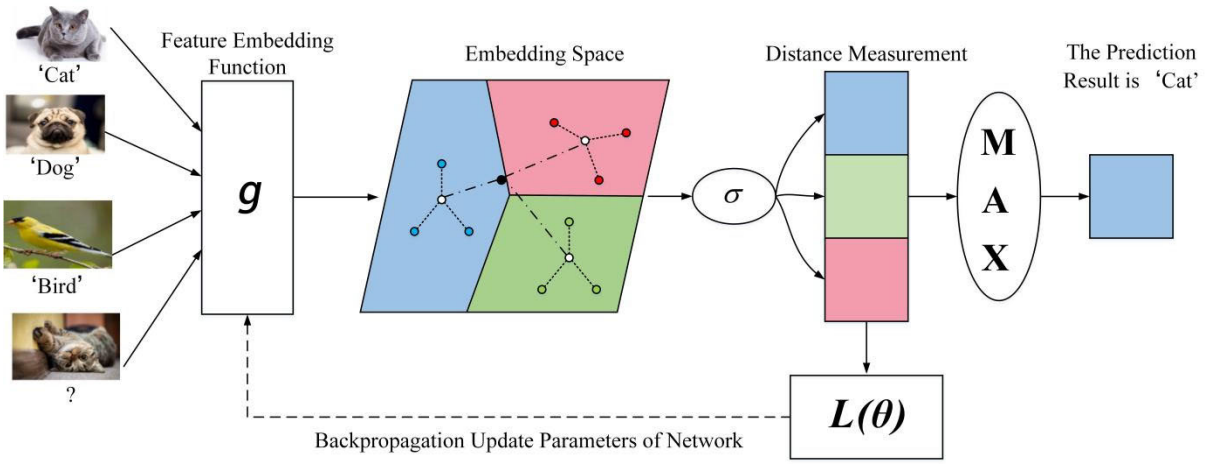


FIGURE 2. The structure of prototypical network, including feature embedding function, distance measurement and prediction results.

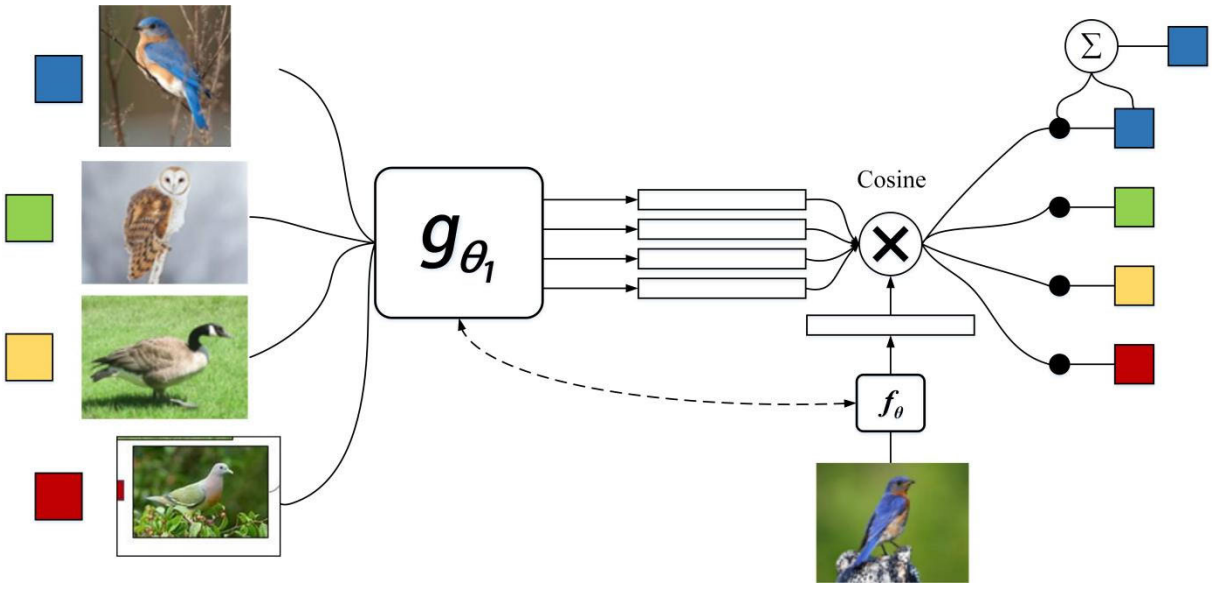


FIGURE 3. The structure of the matching network structure, including a key cosine similarity calculation stage.

motion information between adjacent frames. However, these methods cannot be well used to model long-term temporal information in long videos. Moreover, due to the use of 2D convolutional networks, the dual stream network can directly load pre-trained weights on large-scale image datasets. In response to the disadvantages of the previous two types of methods, some methods adopt the (2+1) D CNN approach [18], [19], which uses two-dimensional convolution to extract spatial information and one-dimensional convolution to extract temporal information.

Due to the need to train deep neural networks, the recently proposed video datasets are all very large. For example, the Youtube-8M dataset has over 8 million videos. In such a large-scale dataset, annotating such a large number of video datasets is time-consuming and almost impossible. Even if search engines assign certain labels to these video data for retrieving videos, there is still a high probability of errors. One solution is to perform action recognition in an unsupervised or weakly supervised manner [20], [21]. Thus, the model does not require complete annotation data, only partial annotation data is needed to complete training.

Faced with the problem that a fully supervised action recognition model requires a large number of labeled samples, and the performance of the model will significantly decrease with only a small number of samples, small sample learning can effectively solve this problem. The development of small sample action recognition is of great significance. Small sample action recognition aims to classify categories that have not appeared in the training set through a very small number of labeled samples. The input to a task is usually a sequence of multiple video samples. The following will present our method in details..

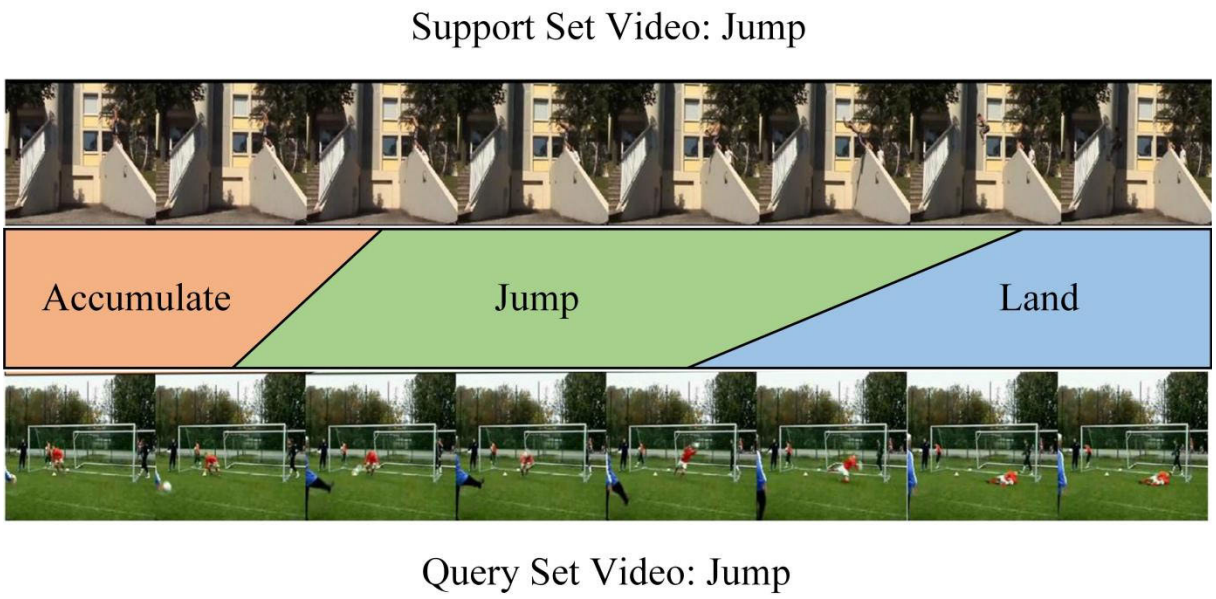


FIGURE 4. Example of the sub-action misalignment problem. Different instances exhibit distinct temporal differences.

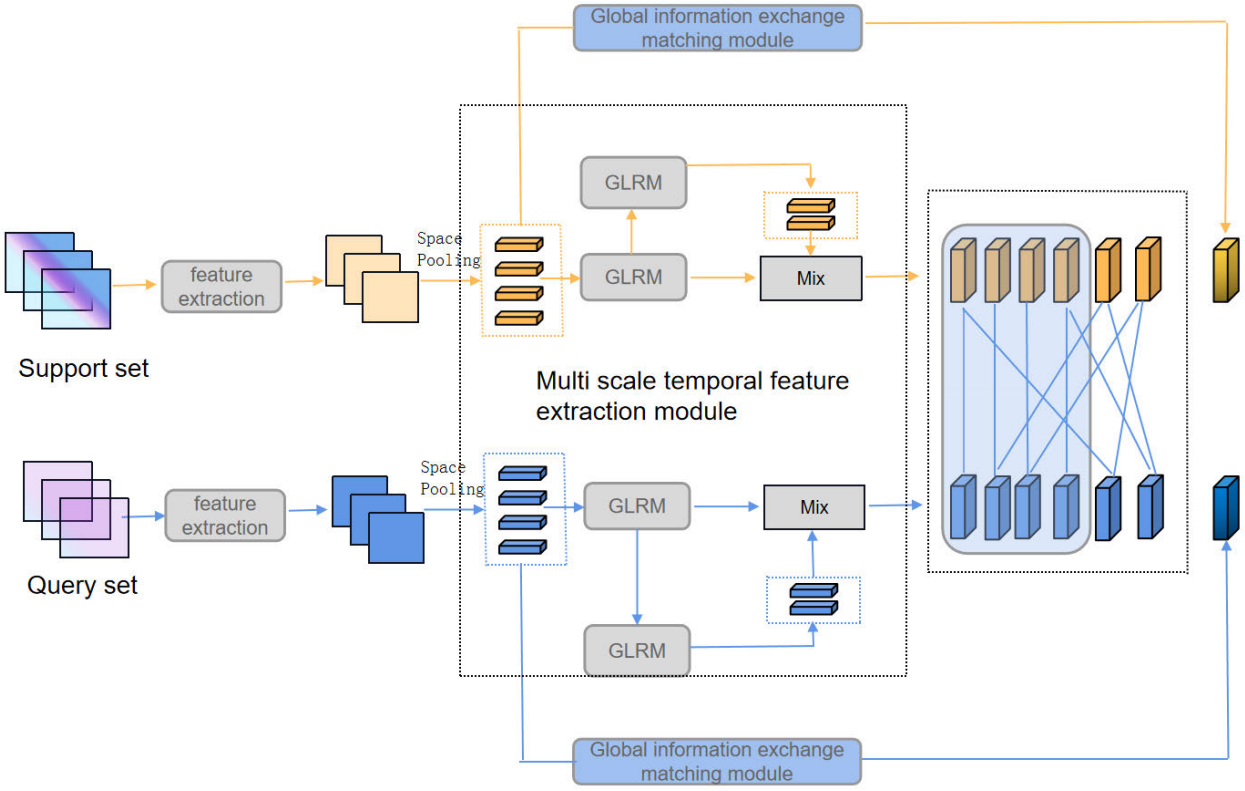


FIGURE 5. The structure of the proposed multi-scale temporal interactive perception network, including a support set and query set for the feature extraction and matching.

B. A NEW MULTI-SCALE TEMPORAL INTERACTIVE PERCEPTION NETWORK MODEL

In videos, different actions are often visually complex, and precise temporal matching between actions cannot be achieved solely through a single temporal scale feature. As shown in Figure 4, although they all belong to the action category of 'long jump', different instances exhibit distinct temporal differences. The temporal intervals occupied by

sub-actions of different instances vary greatly. Therefore, if only a single frame level matching is performed, it may cause misalignment matching between sub-actions and affect the effectiveness of small sample action recognition. We believe that obtaining multi-scale temporal features is beneficial for understanding actions from different temporal scales, which has never been explored in previous work.

In small sample action recognition, single temporal matching cannot accurately achieve temporal matching between actions. If only a single frame level matching is performed, it may cause misalignment matching between subactions and affect the effectiveness of small sample action recognition. If there are different temporal scales, then fine-grained temporal features of subactions can be learned in the early stage, and coarse-grained temporal relationships of combined actions can be learned in the later stage. Therefore, we design a Multi-scale Temporal Feature Extraction Module (MTFEM) and a Cross Scale Alignment Module (CSAM) to enhance the temporal understanding of actions from representation and measurement. More specifically, in the multi-scale temporal feature extraction module, the temporal features of each scale are enhanced by the Global Local Relationship Module (GLRM), which captures both the global temporal dependencies between all frames and the local temporal information between frames, thus facilitating the modeling of complex temporal information in videos. At the same time, a cross scale alignment module was designed to fuse multi-scale temporal feature information, and then perform cross scale alignment to achieve robust matching between videos with different motion speeds. Based on the multi-scale temporal feature extraction module and cross scale alignment module, this section proposes a novel small sample action recognition framework, namely the multi-scale temporal interactive perception network.

The goal of small sample action recognition is to learn a model with good generalization performance using only a small number of labeled samples, which can recognize new categories that have never been seen before. In order to make the training and testing stages highly similar, we adopt a meta task approach for training, similar to the previous small sample learning method. There are two sets in each meta task, namely the support set S and the query set Q. Support set S includes $N \times K$ samples, each from N action categories, and each category has K samples. This setting is called the N-way K-shot problem. The trained model needs to classify the videos in query set Q into one of N action categories.

Figure 5 shows the overall framework of a multi-scale temporal interactive perception network. For each input video sequence, the video is divided into T temporal segments and one frame is randomly selected from each temporal segment. Thus, in each meta task, the support set S can be represented as $\mathrm{S} = \{s_1, s_2, \ldots, s_{\mathrm{N}\times\mathrm{K}}\}$ where s_{i} is each video sequence, and s_{i} can be represented as $s_{\mathrm{i}} = \{s_{\mathrm{i}1}, s_{\mathrm{i}2}, \ldots, s_{\mathrm{i}\mathrm{T}}\}$. For the convenience of describing the model and method, the N-way 1-shot problem should be discussed in the network, where $K = 1$ and only includes one video q in the support set Q. Firstly, according to the previous method, a universal feature extraction network Resnet50 is used to extract visual features for each video sequence, in order to obtain support set features $\mathrm{F_s} = \{\mathrm{f_{s1}}, \mathrm{f_{s2}}, \cdots, \mathrm{f_{n*k}}\}$ and query video features. The extracted video features are input into the multi-scale temporal feature extraction module and the global information interaction matching module, respectively, to obtain the multi-scale temporal features and global features of the video. Then, the obtained multi-scale temporal features are input into the cross scale alignment module to obtain the matching score for cross scale alignment. At the same time, the obtained global features are matched with local frame level features of other videos to obtain a global matching score.

This module attempts to model temporal relationships at different temporal scales, which can capture different aspects of behavioral actions. The temporal scale in the early stage can capture the slow and comprehensive motion characteristics of actions, while the temporal scale in the later stage can capture the fast and general motion characteristics of actions. Given a series of video frame level features, using multi-scale temporal design to fully utilize the advantages of one-dimensional convolution and transformer architectures, the model combines these two structures to capture both global and local information in the video. As shown in Figure 6, the multi-scale temporal extraction module includes a global local temporal relationship module and a temporal downsampling module. In the early stages, there are more temporal tokens, which can represent richer action information; In the later stage, there are fewer timing tokens, but they can represent more general action information.

The global local temporal relationship module can be further decomposed into a global temporal relationship module and a local temporal relationship module. In the global temporal relationship module, in order to capture long-term temporal dependencies, this model uses standard multi-head self attention to model global contextual temporal relationships. In the local temporal relationship module, this model uses a temporal convolutional layer (kernel k) to enhance the feature representation of each temporal token by fusing the information of adjacent temporal tokens. Therefore, by capturing both short-term and long-term temporal dependencies simultaneously, more discriminative temporal features can be obtained.

In the global local temporal module, for each self attention head $i \in \{1, 2, \ldots, H\}$, the network converts the input feature X into the corresponding Q, K, V:

$$Q_i = W_i^Q X \tag{3}$$

$$K_i = W_i^K X \tag{4}$$

$$V_i = W_i^V X \tag{5}$$

Among them, $\mathrm{W_i^Q}$, $\mathrm{W_i^K}$ and $\mathrm{W_i^V} \in \mathbb{R}^{C\times D}$ represent the weight of the linear layer. Therefore, the self attention

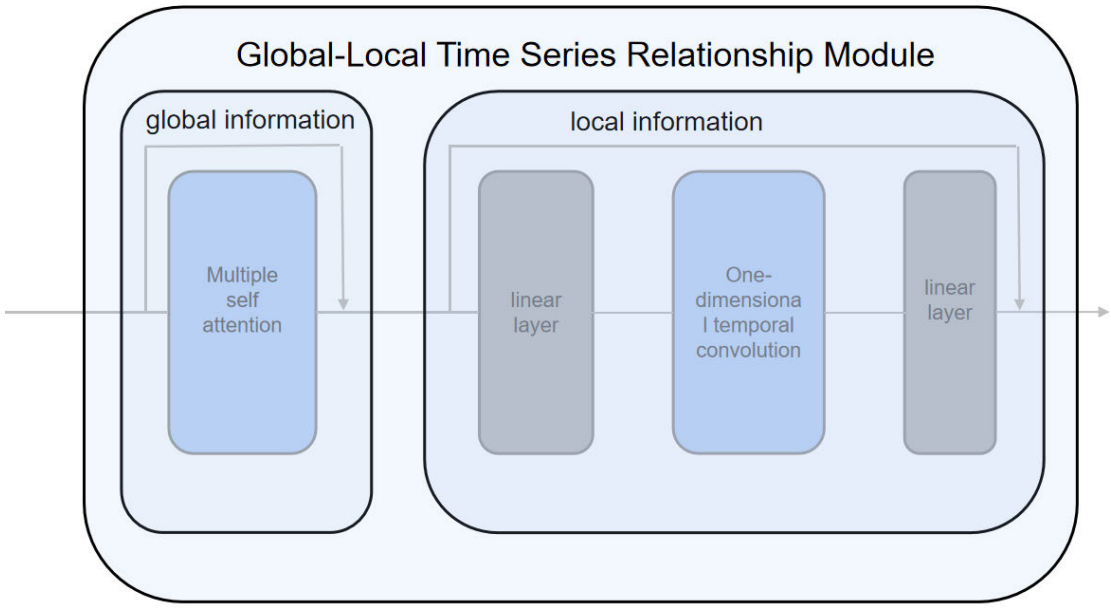


FIGURE 6. The structure of global local relationship module.

calculation process is as follows:

$$\mathrm{Att}_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{C}}\right) V_i \tag{6}$$

where *softmax*() represents the softmax Activation function, and C the feature dimension of each self attention head. Next, the outputs of different self attention heads are concatenated and passed through a linear layer, as shown in Eq.(7):

$$\mathrm{M} = \mathrm{W^o concat}\left(\mathrm{Att}_1, \ldots, \mathrm{Att}_\mathrm{H}\right) + \mathrm{X} \tag{7}$$

Among them, $\mathrm{W^o} \in \mathbb{R}^{D\times D}$ is the weight of the linear layer. Next, multiple temporal tokens obtained by the global temporal relationship module are extracted through the feature extraction of a local temporal relationship module. The local temporal relationship module includes two linear layers and 1 one-dimensional temporal convolutional layer. As shown in Figure 6, multiple temporal tokens pass through a linear layer to reduce the feature dimension from D to D1, and then use a one-dimensional temporal convolution to capture local information of adjacent temporal tokens to enhance the information of the current token. Multiple temporal tokens pass through a linear layer again to restore the feature dimension to D, thus obtaining the final temporal feature.

The temporal downsampling module achieves the connection between temporal features at different scales, which can reduce the temporal resolution of the video. It can be regarded as the average pooling of adjacent time series tokens. In the actual operation process, we use a temporal convolution with a convolution kernel of k and a step size of 2 to obtain the temporal tokens of the later stage.

C. CROSS SCALE ALIGNMENT MODULE

In order to achieve matching between videos with different motion speeds, this section designs a cross scale alignment module. We have obtained two features of different time scales in the multi-scale temporal feature extraction module, namely early temporal features $\mathrm{H}_1 = \{f_1, f_2, \ldots, f_\mathrm{T}\}$ and later time series characteristics $\mathrm{H}_2 = \{s_1, s_2, \ldots, s_\mathrm{J}\}$. In order to achieve cross scale alignment, this section concatenates temporal features of different scales to obtain mixed temporal features $\mathrm{H} = \{\mathrm{H}_1, \mathrm{H}_2\}$. In order to complete local feature matching, the Hausdorff distance [22] improved by Eq.(8) as:

$$\mathrm{D}\left(\mathrm{H_S}, \mathrm{H_Q}\right) = \mathrm{M}\left(\left[\mathrm{H_{S1}}, \mathrm{H_{S2}}\right], \left[\mathrm{H_{Q1}}, \mathrm{H_{Q2}}\right]\right) \tag{8}$$

Eq.(8) represents the distance between the mixed scale features of the support set video and the mixed scale features of the query set video. At the same time, some movements have strong timing, such as the long jump, which inevitably involves running up and taking off. The sequence of movements before and after is very important. Therefore, on the basis of cross scale alignment, we introduced frame level alignment as in Eq.(9)

$$\mathrm{D}\left(\mathrm{H_{S1}}, \mathrm{H_{Q1}}\right) = \mathrm{M}\left(\left[\mathrm{H_{S1}}\right], \left[\mathrm{H_{Q1}}\right]\right) \tag{9}$$

Based on the above statement, combining cross scale alignment and frame level alignment can better achieve metrics between support set videos and query set videos [23]. The measurement formula is as follows:

$$\mathrm{D} = \mathrm{D}\left(\mathrm{H_S}, \mathrm{H_Q}\right) + \alpha \mathrm{D}\left(\mathrm{H_{S1}}, \mathrm{H_{Q1}}\right) \tag{10}$$

The α is the balance parameters for two alignment methods. The calculated distance between videos can be used to further obtain the output probability for action classification.

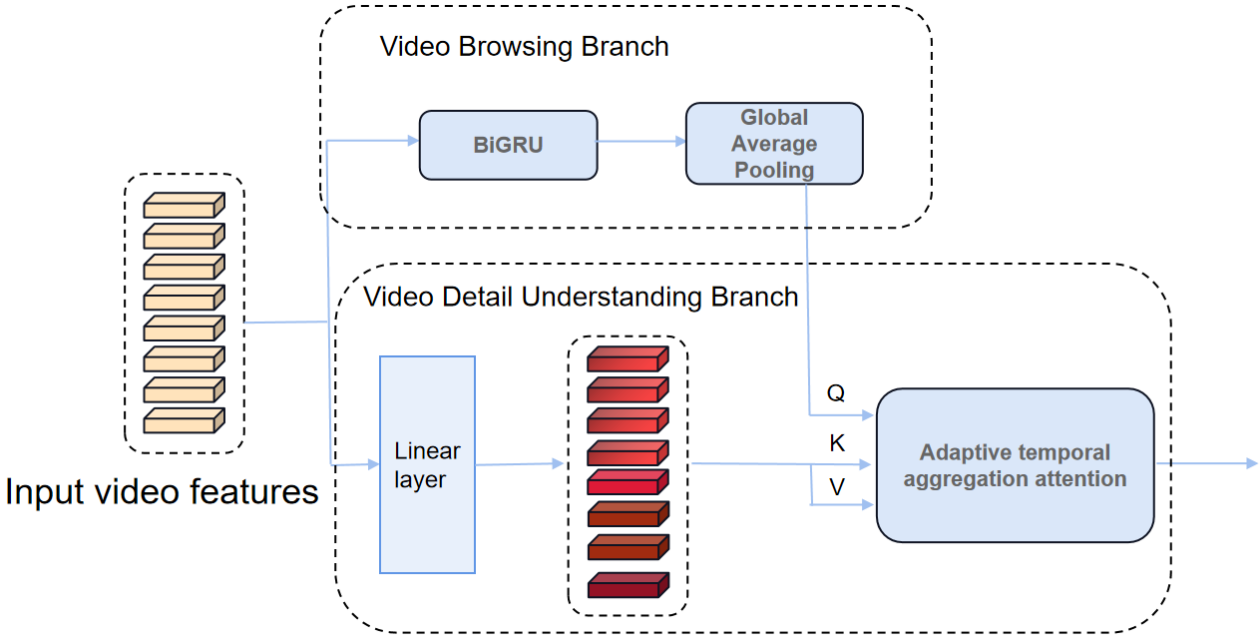


FIGURE 7. Global information interaction matching module for matching betweeen different frames and actions.

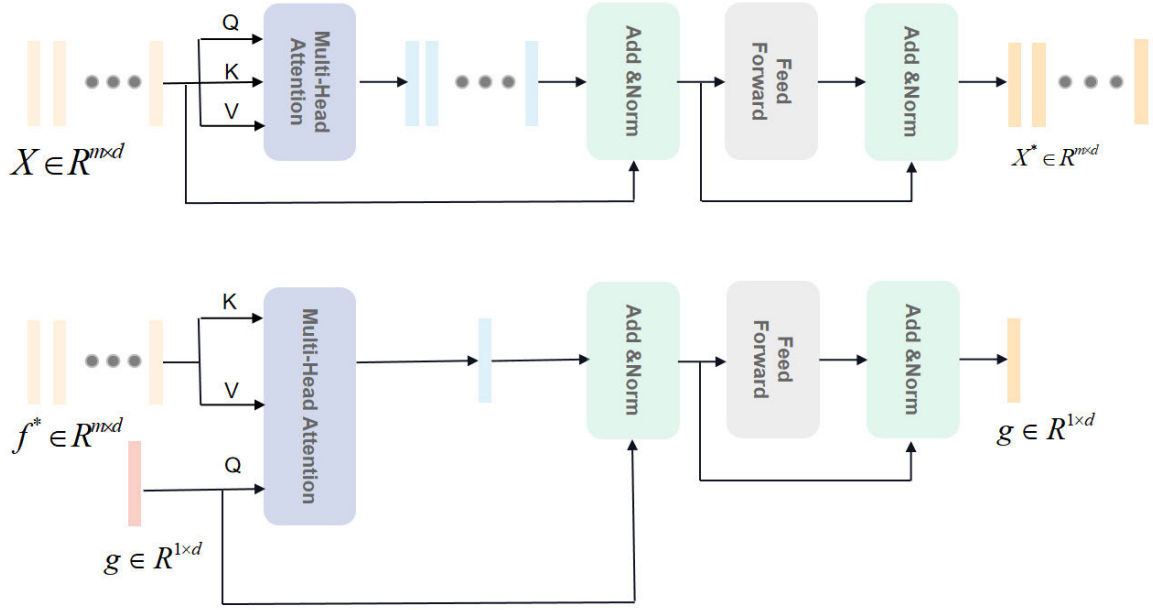


FIGURE 8. Multihead self-attention and adaptive temporal convergent attention for global feature representation.

D. GLOBAL INFORMATION EXCHANGE MATCHING MODULE

In the real world, people often have a reading habit of first obtaining an overall understanding of the text information by browsing it, and then carefully reading and understanding the text based on the obtained overall information. This often deepens the understanding of the text information. If the model obtains the overall features of the video, it will help deepen the understanding of the video action content. Therefore, the global information interaction matching module proposed includes two branches, namely the video browsing branch and the video detail understanding branch (as shown in Figure 7).

The video browsing branch is composed of a lightweight video encoder to obtain a summary of the content in the video. Using a Bidirectional Gated Recurrent Unit (BiGRU) encoder to extract the overall information of the video, the input video sequence V passes through the bidirectional GPU encoder

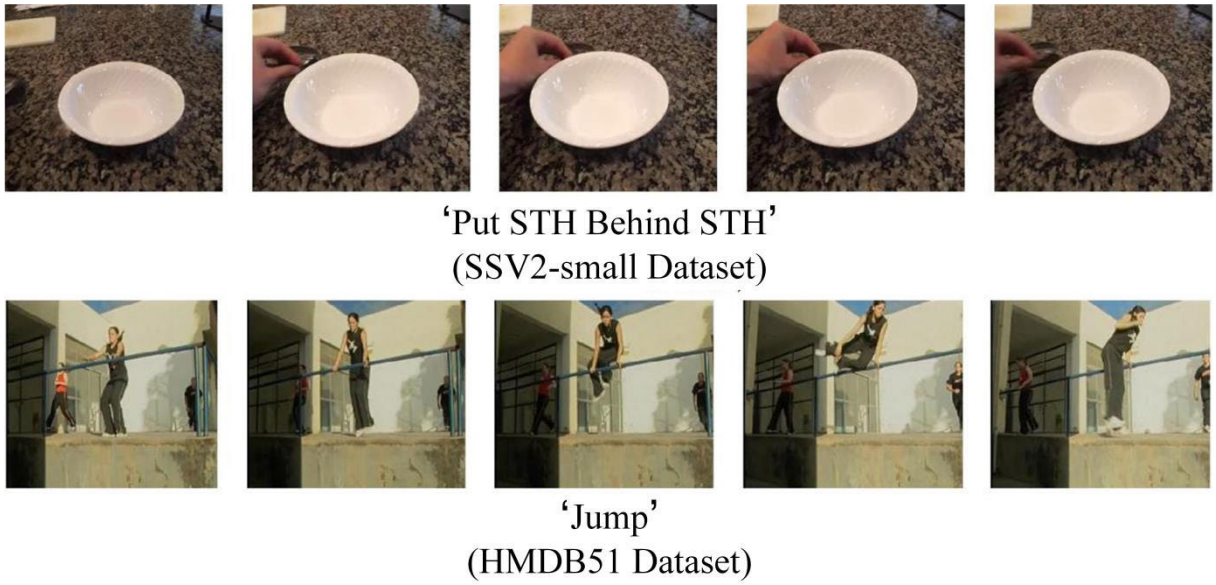


FIGURE 9. Samples for training and testing process from SSV2-small dataset and HMDB51 dataset.

and becomes a feature sequence $H = \{h_1, h_2, \ldots, h_T\}$. Next, the overall feature representation of the video is obtained by averaging pooling along the temporal dimension on the feature sequence F, expressed as follows:

$$g = \frac{1}{T}\sum_{i=1}^{T} h_i \tag{11}$$

where $h_i \in \mathbb{R}^{1\times D}$, $g \in \mathbb{R}^{1\times D}$, D is the feature dimension of the overall video information. On the contrary, the video detail understanding branch consists of a video encoder with a relatively large number of parameters, thereby obtaining deeper levels of video information. Usually, videos contain a lot of target information and complex scenes, and it is not possible to fully obtain the overall information of the actions in the video solely through video browsing branches. Therefore, the branch of video detail understanding can be further introduced to obtain deeper information. Given the frame level features $F = \{f_1, f_2, \ldots, f_T\}$, in order to obtain semantic information highly related to the overall video features from the frame level feature sequence, we design an adaptive temporal aggregation attention. Specifically, in order to design specific adaptive temporal aggregation attention, the idea of self attention mechanism in the transformer architecture is borrowed. Unlike the self attention mechanism, we use the output features g of the video browsing branch as a ''query'' and the frame level feature F as a ''key'' and ''value'' (as shown in Figure 8). Thus, the video detail understanding branch can further fuse fine-grained video action information with global feature representations, making the obtained global information more discriminative.

Previous work only considered frame level matching, neglecting to enhance the perception ability of frame level features to the overall information of actions during the matching process. If two videos belong to the same category, then the matching distance between the global information of videos is close. Moreover, especially in cases where similar frames exist, the introduction of global information is helpful for local semantic matching. Therefore, we use a comparison metric distance method [24] to match global information and local frame level information by

$$D\left(f_i, f_q^g\right) = M\left(\left[f_i^1, \ldots, f_i^T\right], f_q^g\right) \tag{12}$$

Among them, f_i represents the local frame level features of the video, while f_q^g represents the global features of the video, $M()$ representing the distance measurement method based on comparative learning. It can explicitly make local frame level features perceive global contextual information, thereby promoting more robust small sample action recognition.

IV. EXPERIMENT AND ANALYSIS

A. DATASETS AND EVALUATIONS

Experiments used two action behavior video datasets, namely the SSV2 small dataset and the HMDB51 dataset. This section provides examples of corresponding samples in Figure 9, and details each dataset below. The SSV2 small dataset contains a total of 193690 action videos of human object interactions with complex temporal information, involving a total of 174 behavioral categories. According to the dataset partitioning strategy of the previous small sample action recognition method, 64 action categories were randomly selected from the dataset as the training set and 24 categories as the test set. At the same time, select 100 samples for each category. The HMDB51 dataset contains a total of 51 types of actions, with a total of 6849 videos. Each type of action consists of at least 50 videos from websites such as YouTube and Google. The HMDB51

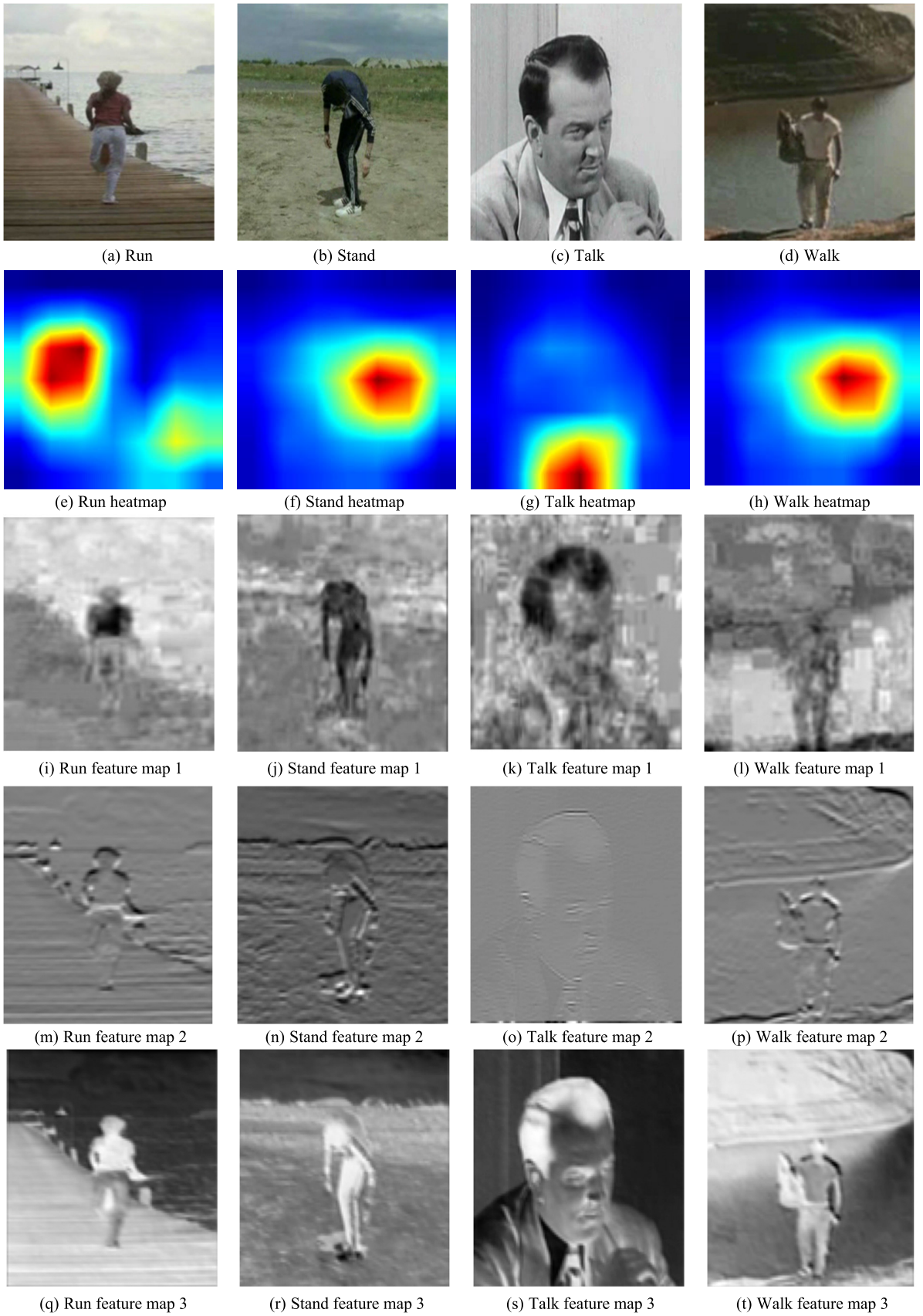


FIGURE 10. Examples of heatmaps and feature maps in the shallow layers.

TABLE 1. Quantitative comparison with SOTA methods on the SSV2 dataset.

Method	Reference	Backbone	SSV2-small
OTAM[3]	CVPR2020	Resnet-50	36.4
ITANet[6]	IJCAI2021	Resnet-50	39.8
TRX[4]	CVPR2021	Resnet-50	36.0
STRM[8]	CVPR2022	Resnet-50	37.1
HCL[25]	ECCV2022	Resnet-50	38.7
MIAN (ours)	-	Resnet-50	40.2

TABLE 2. Quantitative comparison with SOTA methods on the HMDB51 dataset.

Method	Reference	Backbone	HMDB51
ProtoGAN[26]	CVPR2019	C3D	34.7
TARN[2]	ECCV2020	C3D	44.6
TRX[4]	CVPR2021	Resnet-50	53.1
MTFAN[5]	CVPR2022	Resnet-50	59.0
STRM[8]	CVPR2022	Resnet-50	52.3
HCL[25]	ECCV2022	Resnet-50	59.1
Nguyen[27]	ECCV2022	Resnet-50	59.6
MIAN (ours)	-	Resnet-50	60.4

dataset has extremely strong appearance properties, that is, there are generally frames with high similarity in a video. We followed the dataset partitioning strategy of the previous small sample action recognition method, using 31 categories as the training set and 10 categories as the test set.

In order to quantitatively evaluate the effectiveness of the design model, the evaluation index commonly used in computer vision recognition tasks, namely multi-class accuracy, is adopted. Multi-class accuracy is defined as the proportion of correctly classified samples to the number of samples, and the formula is as follows:

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{total}}} \tag{13}$$

Among them, n_{correct} is the number of correctly classified samples and n_{total} is the total number of samples.

In experiments, we use ResNet as the backbone network to extract each frame feature of the video and load pretrained weights on ImageNet. Like the previous method, each video is sparsely sampled by 8 frames. During the training phase, basic data augmentation methods such as horizontal flipping are used, with each frame cropped to 224 × 224. With the Stochastic Gradient Descent (SGD) optimizer, the initial learning rate is set to 0.001. Due to the significant memory overhead required for each meta task, one meta task is run at a time,

TABLE 3. Ablation study under 5-way 1-shot.

Method	SSV2-small	HMDB51
Baseline	39.2	58.1
Baseline w/ MTFEM	39.8	58.7
Baseline w/ MTFEM+CSAM	40.0	60.3
MIAN（ours）	40.2	60.4

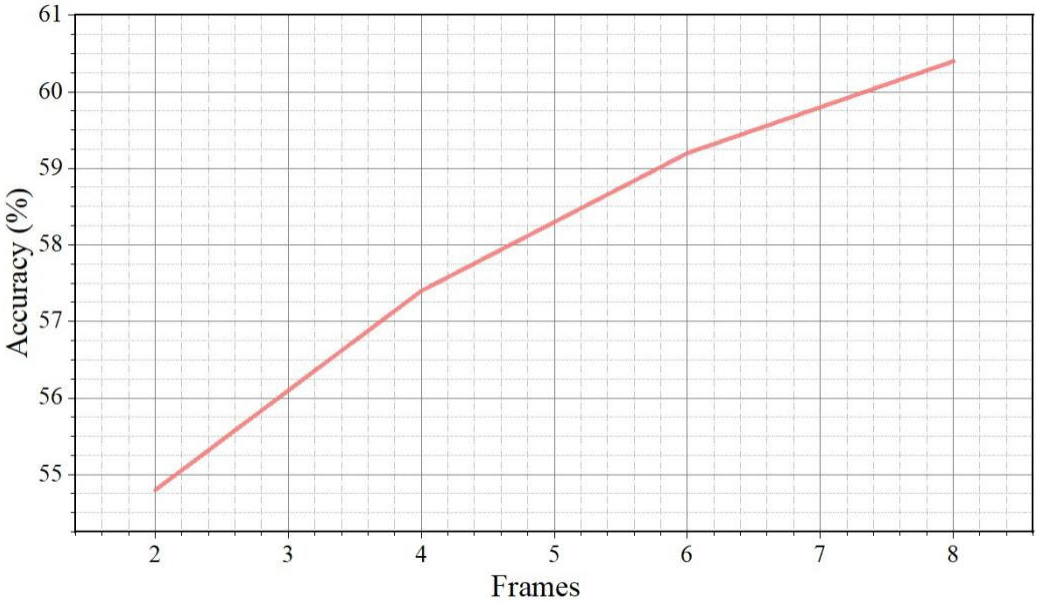


FIGURE 11. Ablation study on the effect of changing the number of input video frames under the 5-way 1-shot.

with gradient averaging and backpropagation performed for every 16 meta tasks. During testing, randomly perform 10000 dimensional tasks on the test set for evaluation, and then report the average accuracy of the 10000 dimensional tasks.

This experiment is carried out on the Windows 10 operating system, with a memory size of 64G. We utilize a GPU powered by an NVIDIA GeForce 3060. For learning frameworks, Pytorch-GPU 1.8.1, Cuda11.1, and Cudnn 8.0.5 are employed.

Figure 10 demonstrates the feature maps in the learning stage. We show the salient features in the second row of Figure 10, which shows that the network focuses on the salient object in the classification rather than the whole scene. From the third to last row of Figure 10 shows feature maps in the shallow layers. It is worth noting that in the deep layers, the network collects the global features which seems not intuitive to human vision.

B. COMPARATIVE ANALYSIS AND DISCUSSION

This section compares the proposed model Multi-Scale Interactive Adaptive Network (MIAN) with various latest methods on different datasets. As shown in Table 1, the model in this work surpasses all the current latest methods on both datasets under the 5-way 1-shot setting. The experimental results demonstrate the effectiveness of the method. Based on the experimental results in Tables 1 and 2, the following findings are made.

(1) Compared with the most advanced method of C3D using backbone networks, such as TARN, our method achieved a significant improvement of 15.8% in HMDB51. Because 3D CNN networks introduce a large number of model parameters, it is easy to cause overfitting of the model under the setting of small sample learning. In the cross scale alignment module, we did not introduce additional parameters, but only spliced the temporal features of different scales, and used the improved Hausdorff distance to measure,

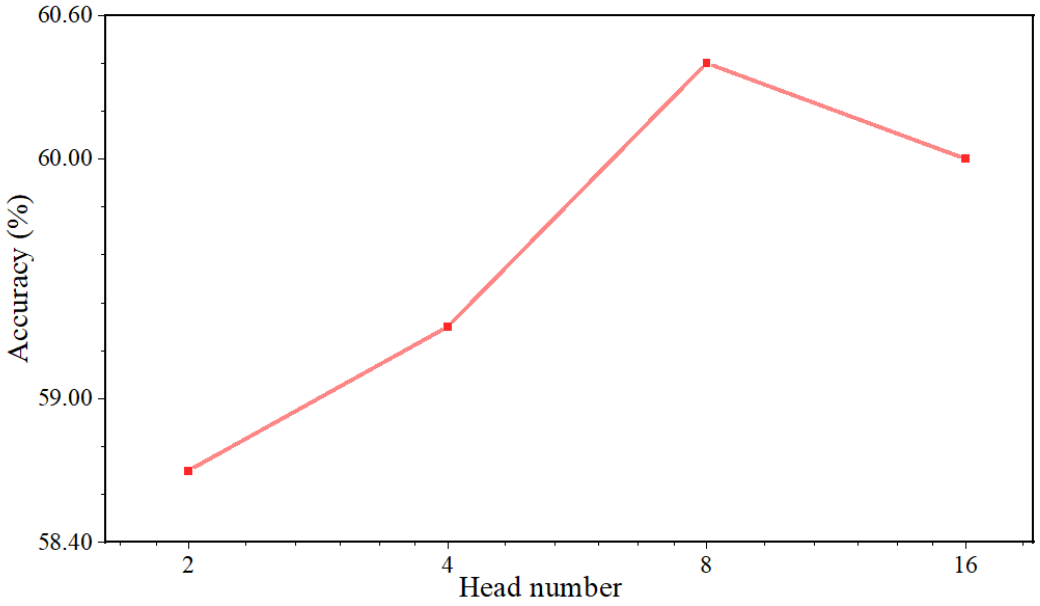


FIGURE 12. Ablation study on the effect of changing the number of heads under the 5-way 1-shot.

which is more conducive to the generalization of the model.

(2) Compared with the backbone network being ResNet's most advanced method, our method achieved performance improvements of 0.8% and 0.4% on HMDB51 and SSV2-small, respectively. The experimental results demonstrate that our model can learn rich and effective features even when the sample size is very small, thus exhibiting high generalization performance. Our model can achieve interactive perception of local frame level information and global video information, as well as interactive perception between different temporal scales, further enhancing the model's generalization ability.

Compared with the most advanced methods of single scale temporal alignment, our method achieved the best results, thus proving the effectiveness of the multi-scale mixed alignment strategy in small sample action recognition.

C. ABLATION EXPERIMENT

To verify the effectiveness of each module proposed in the proposed algorithm, the following variant methods were designed. Baseline: the baseline method only includes single scale temporal feature extraction and single scale temporal feature matching.

Baseline w/MTFEM: this variant method uses the multi-scale feature extraction module MTFEM to obtain multi-scale temporal features, which are then measured and matched separately.

MIAN: this is the complete version of the model in this work, which explores small sample action recognition through cross scale alignment matching between multi-scale temporal information and matching between video global information and local frame level information.

Baseline w/MTFEM+CSAM: based on the variant method of Baseline w/MTFEM, the multi-scale temporal feature information obtained is input into the cross scale alignment module CSAM for temporal matching.

To demonstrate the effectiveness of each module in the proposed method, the performance results of each variant method on two datasets are reported in Table 3. Based on the experimental results, there are several discussions.

After introducing multi-scale temporal features, Baseline w/MTFEM achieved some performance improvement on both datasets, indicating that in small sample action recognition tasks, the acquisition of multi-scale temporal features helps to deepen a comprehensive understanding of action behavior.

Compared with the previous variant method, Baseline w/MTFEM+CSAM achieved improvements of 0.3% and 1.6% on two datasets, respectively, indicating that cross scale alignment helps achieve robust matching between videos with different motion speeds.

The performance improvement of the final method in this work indicates that the fusion of information at different temporal scales, as well as the fusion of global and local information, can coordinate with each other and jointly promote performance improvement. This is consistent with the work [28] that features are required to be considered jointly.

D. ANALYSIS OF THE IMPACT OF DIFFERENT VIDEO FRAME NUMBERS

For fair comparison, the model MIAN in this work is compared under the input frame number $T = 8$. In order to analyze the impact of inputting different video frames under

(a) board writing-01 (b) board writing-05 (c) Teaching-01

(d) Teaching-12 (e) reading-01 (f) Reading-23

(g) taking note-01 (h) taking note-08 (i) answering-01

(j) answering-16 (k) listening-01 (l) listening-18

FIGURE 13. Qualitative results on the collected classroom teaching videos.

the small sample learning setting on the experimental results, 2, 4, 6, 7 and 8 frames were sampled for the experiment. As shown in Figure 11, as the number of input frames increases, the model performance increases from fast to slow, and gradually saturates. The experimental results show that the more input frames, the more significant the performance improvement. At the same time, the model in this work can achieve performance comparable to many advanced methods when the input frame number is 4. Previous studies have shown that multi-head self attention can focus on different patterns within features and thus capture different features [29], [30]. In Figure 12, the effect of changing the number of heads in multi-head self-attention on performance is studied. The experiment shows that the impact of the number of multiple heads is significant, with a performance impact far exceeding one point.

Compared with other methods, we introduce MTFEM and CSAM in the multi-scale interactive adaptive network. A rise in the complexity seems inevitable. In terms of the MTFEM as shown in Figure 5, it includes two GLRM and the feature

TABLE 4. Quantitative results on the collected classroom teaching videos.

Action	board writing	teaching	reading	taking note	answering	listening	Average
Accuracy	0.84	0.65	0.52	0.92	0.86	0.68	0.75

maps mix stage. The key module in GLRM is the multi-self attention, which depends on the length of the input sequence n and dimension d, i.e. $O(n^2 \cdot d)$. In terms of the CSAM as shown in Figure 7, the key module is the BiGRU to propagate information between neurons in the forward and backward stages. The complexity of BiGRU depends on its layers which can be reduced to be 3-4 layers, which can meet the requirements of Action Behavior classification in video sequence. The learning stages take us about 81.6 hours, and the classification stage relies on the length of the sequence.

In order to show our scalability, we conduct the proposed on our collected video sequences. We have 138 videos, containing 6 actions related to the classroom teaching, namely board writing, teaching, reading, taking note, answering and listening. Each action has no less than 20 videos, the video resolution is 460 × 380. Our qualitative results are shown in Figure 13. Experiments show that the proposed algorithm obtains the correct labels in each video when it has no ambiguity in actions. The quantitative results are calculated based on the ratio of accurately classified as shown in Table 4, which shows the average accuracy is up to 0.75.

V. CONCLUSION

This work conducts research on small sample action recognition tasks. Small sample action recognition tasks can overcome the dependence on large-scale annotated data and use an extremely small number of annotated samples to classify categories that have not appeared in the training set. However, there are two issues in the current small sample action recognition: (1) a single temporal scale of information cannot fully display the semantics of action behavior; (2) the matching between local information is not effective.

This work proposes a multi-scale interactive perception network to fully utilize the multi-scale and global information of videos to achieve robust matching between different action behaviors. We enhance robust matching between videos through the fusion of long-term and short-term temporal dependencies, as well as the mutual perception of global and local information. The multi-scale temporal feature extraction module (MTFEM) and cross scale alignment module (CSAM) enhance the temporal understanding of actions from two aspects: representation and measurement, to achieve matching between actions with different motion speeds. At the same time, the global information interaction matching module promotes the matching of global information and local frame level features of videos, thereby maximizing the consistency between local and global features of the same type of action. The multi-scale temporal interactive perception network proposed in this work is compared with the current state-of-the-art methods on multiple datasets, demonstrating the superiority and effectiveness of the proposed method.

Our multi-scale temporal interactive perception network model provides a new solution for small sample action recognition and achieves advanced results. However, the algorithm still has certain shortcomings. Future work improvements mainly focus on two aspects: (1) our multi-scale temporal extraction introduces a certain amount of parameters, and in the future, we try to reduce the module's parameter quantity without reducing performance. (2) the current small sample action recognition method is only applicable to a single dataset and lacks domain generalization ability. In the future, the cross domain recognition ability of small sample action recognition models can be further enhanced.

ACKNOWLEDGMENT

The authors would like to thank Qu, Master of Computer Science, for his excellent support and assistance with the experiments. *(Chunlin Zheng and Jun Gu are co-first authors.)*

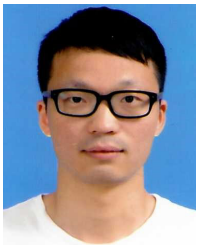
REFERENCES

[1] L. Zhu and Y. Yang, "Compound memory networks for few-shot video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 751–766.
[2] M. Bishay, G. Zoumpourlis, and I. Patras, "TARN: Temporal attentive relation network for few-shot and zero-shot action recognition," 2019, *arXiv:1907.09021*.
[3] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10615–10624.
[4] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational CrossTransformers for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 475–484.
[5] J. Wu, T. Zhang, Z. Zhang, F. Wu, and Y. Zhang, "Motion-modulated temporal fragment alignment network for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9141–9150.
[6] S. Zhang, J. Zhou, and X. He, "Learning implicit temporal alignment for few-shot video classification," 2021, *arXiv:2105.04823*.
[7] S. Li, H. Liu, R. Qian, Y. Li, J. See, M. Fei, X. Yu, and W. Lin, "TA2N: Two-stage action alignment network for few-shot action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1404–1411.
[8] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19926–19935.
[9] Y. Fu, L. Zhang, J. Wang, Y. Fu, and Y.-G. Jiang, "Depth guided adaptive meta-fusion network for few-shot video recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1142–1151.
[10] Y. Huang, L. Yang, and Y. Sato, "Compound prototype matching for few-shot action recognition," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 351–368.
[11] J. Wang and Y. Zhai, "Prototypical Siamese networks for few-shot learning," in *Proc. IEEE 10th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2020, pp. 178–181.

[12] X. Zhang, J. Nie, L. Zong, H. Yu, and W. Liang, "One shot learning with margin," in *Proc. 23rd Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, Macau, China. Cham, Switzerland: Springer, Apr. 2019, pp. 305–317.
[13] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3637–3645.
[14] Y. Zhang, Y. Bai, H. Wang, Y. Xu, and Y. Fu, "Look more but care less in video recognition," 2022, *arXiv:2211.09992*.
[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 568–576.
[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
[18] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 906–915.
[19] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1895–1904.
[20] W. Lin, X. Liu, Y. Zhuang, X. Ding, X. Tu, Y. Huang, and H. Zeng, "Unsupervised video-based action recognition with imagining motion and perceiving appearance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2245–2258, May 2023.
[21] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue, "Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance," 2022, *arXiv:2204.10312*.
[22] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[23] D. Wu, X. Dong, L. Shao, and J. Shen, "Multi-level representation learning with semantic alignment for referring video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4986–4995.
[24] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.
[25] S. Zheng, S. Chen, and Q. Jin, "Few-shot action recognition with hierarchical matching and contrastive learning," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 297–313.
[26] S. K. Dwivedi, V. Gupta, R. Mitra, S. Ahmed, and A. Jain, "ProtoGAN: Towards few shot learning for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1308–1316.
[27] K. D. Nguyen, Q.-H. Tran, K. Nguyen, B.-S. Hua, and R. Nguyen, "Inductive and transductive few-shot video classification via appearance and temporal alignments," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 471–487.
[28] C. P. O. Reyer, N. Brouwers, A. Rammig, B. W. Brook, J. Epila, R. F. Grant, M. Holmgren, F. Langerwisch, S. Leuzinger, W. Lucht, B. Medlyn, M. Pfeifer, J. Steinkamp, M. C. Vanderwel, H. Verbeeck, and D. M. Villela, "Forest resilience and tipping points at different spatio-temporal scales: Approaches and challenges," *J. Ecol.*, vol. 103, no. 1, pp. 5–15, Jan. 2015.
[29] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 181–185.
[30] Y. Zhang, Y. Gong, H. Zhu, X. Bai, and W. Tang, "Multi-head enhanced self-attention network for novelty detection," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107486.

CHUNLIN ZHENG is currently pursuing the Ph.D. degree with the School of Philosophy, Beijing Normal University, with a focus on artificial intelligence security strategies. He started working at Jiangsu Second Normal University, in 2006, where he is an Associate Researcher. His research interest includes empowering educational informatization with artificial intelligence.

JUN GU graduated from the Jiangsu University of Education. Currently, he is the Director of the Wuxi Education Informatization and Equipment Management Service Center and the Education Committee of the Wuxi Big Data Association. He explores combining artificial intelligence and information technology with education and teaching to improve the students' learning experience. His research interests include the development and application of intelligent education platforms and campus intelligence security.



SHENG XU (Member, IEEE) received the B.Eng. degree in computer science and technology from Nanjing Forestry University, Nanjing, China, in 2010, and the Ph.D. degree in digital image systems from the University of Calgary, Calgary, AB, Canada, in 2018. In 2018, he joined the College of Information Science and Technology, Nanjing Forestry University, where he is currently an Associate Professor. His research interests include mobile mapping, vegetation mapping, and computer vision.