Computer Science and Information Systems 21(4):1823–1843 https://doi.org/10.2298/CSIS240418060W

Spatio-Temporal-based Multi-level Aggregation Network for Physical Action Recognition

Yuhang Wang

School of Physical Education, Harbin University
Harbin, Heilongjiang 150086, China
wangyuhang1978@hrbu.edu.cn

Abstract. This paper introduces spatio-temporal-based multi-level aggregation network (ST-MANet) for action recognition. It utilizes the correlations between different spatial positions and the correlations between different temporal positions on the feature map to explore long-range spatial and temporal dependencies, respectively, generating the spatial and temporal attention map that assigns different weights to features at different spatial and temporal locations. Additionally, a multi-scale approach is introduced, proposing a multi-scale behavior recognition framework that models various visual rhythms while capturing multi-scale spatiotemporal information. A spatial diversity constraint is then proposed, encouraging spatial attention maps at different scales to focus on distinct areas. This ensures a greater emphasis on spatial information unique to each scale, thereby incorporating more diverse spatial information into multi-scale features. Finally, ST-MANet is compared with existing approaches, demonstrating high accuracy on the three datasets.

Keywords: Action recognition, spatial and temporal attention, multi-level aggregation network.

1. Introduction

With the development of information technology, videos have become an indispensable medium in our lives due to their powerful content delivery and expressive capabilities. The widespread deployment of image capture devices and the growth of computer networks have led to an explosive increase and dissemination of videos. These massive video datasets hold significant value, but traditional manual analysis is hindered by subjective factors and inefficiency. Therefore, utilizing computer technology for the analysis, understanding, and information mining of video content is a crucial topic in the field of computer vision and has been a forefront direction of research in recent years. Due to various factors such as different video data collection methods, diverse scenes, and varied content, designing manual features for algorithms is time-consuming and challenging to apply widely across multiple tasks. Deep learning technologies have undergone rapid development on computer vision, face recognition, and semantic segmentation, demonstrating superiority compared to traditional algorithms [2,6,15].

In human action recognition, behavior instances exhibit a complex temporal structure, especially with different behavior categories naturally having distinct visual rhythms, such as clapping and walking. This poses challenges for behavior recognition [13,11,10]. In certain situations, different behavior categories may look very similar in appearance, such

as walking, jogging, and running. The visual information of these behaviors is closely related, making accurate classification challenging based solely on appearance. In such cases, visual rhythm becomes a crucial factor in distinguishing them. Moreover, even for the same behavior, individual differences in factors like age, energy, and mood mean that each performer may execute these actions according to their own rhythm. For instance, older individuals typically move more slowly than younger ones, and individuals with greater weight generally exhibit slower movements than those with lesser weight. Therefore, precisely modeling the intra-class differences in the visual rhythm of action behaviors significantly enhance the performance of behavior recognition. In recent years, researchers have begun to explore the aspect of visual rhythm. The SlowFast Network is one such example, consisting of two pathways known as the Fast pathway and the Slow pathway [5]. It incorporates hard-coded visual rhythm at the input level, enabling the two pathways to sample frames at different rates. At each stage of the network, there is a lateral connection between the two pathways, facilitating the fusion of features from the Fast pathway into the Slow pathway. However, this hard-coded approach requires more input frames, and the temporal scale is limited to manually set and tuned values. While employing more pathways may enhance performance, it also comes with a significant increase in computational complexity, potentially making the network overly intricate. Considering a lot of utilized methods, i.e., C3D [22] and I3D [1], usually involve many operations of temporal convolution. That is, as the network depth increases, the temporal receptive field also expands. Therefore, multi-depth features in a method can be captured information covering various visual rhythms and possess different temporal receptive fields, indicating diverse temporal scales. Additionally, these features at different depths also encompass multi-scale spatial information. Meanwhile, the success of multi-scale features has been demonstrated in tasks like object detection [8] and pedestrian re-identification [9]. Feature extraction in deep CNNs is a process that transforms features from low levels into high levels. While deeper network layers can extract global information with highly discriminative features, they inevitably lose detailed information. Shallow features, on the other hand, maintain high resolution and weak semantics, serving as a complement to deep features.

Inspired by these considerations, we introduce a spatio-temporal-based multi-level aggregation network (ST-MANet) for action recognition. Specifically, ST-MANet, based on the self-attention mechanism, designs a temporal aggregation module to leverage the temporal correlations within videos. It aims to extract more temporal-related information and aggregate temporal information effectively. Subsequently, a temporal aggregation network is constructed based on the temporal aggregation module, consisting of two parts: the main branch and the temporal aggregation branch. The interaction between these branches further explores temporal information in the video, enhancing the network's temporal modeling capabilities and improving the accuracy of behavior recognition. Following this, a spatial attention module based on the self-attention mechanism is introduced. It evaluates the correlations between different spatial positions in the feature map, assessing the importance of each position and assigning different weights to different spatial locations. Furthermore, a multi-scale approach is incorporated, proposing a multi-scale behavior recognition framework that models various visual rhythms while capturing multi-scale spatiotemporal information. On this basis, a spatial diversity constraint is introduced to focus spatial attention at different scales on distinct spatial locations. This

ensures that different scales of spatial attention concentrate on spatial information unique to each scale, allowing for a more diverse representation of spatial information in multi-scale features. Finally, experimental results demonstrate the effectiveness of ST-MANet compared to existing approaches, showcasing high accuracy across three datasets.

The contributions of ST-MANet are threefold:

(1) A temporal aggregation module, designed through the self-attention mechanism, effectively extracts and aggregates temporal information from videos, significantly enhancing the network's ability to capture action sequence changes and thereby improving the accuracy of behavior recognition.

(2) By introducing a spatial attention module and a multi-scale approach, ST-MANet is capable of capturing unique spatial information at different scales, and through the spatial diversity constraint, ensures the comprehensiveness and diversity of this information, further strengthening the model's ability to recognize actions in complex scenes.

(3) Experimental results demonstrate the effectiveness of ST-MANet compared to existing approaches, showcasing high accuracy across three datasets in action recognition tasks.

The remaining sections of ST-MANet are organized as follows: Section II offers an extensive review of current methods in human action recognition. Section III detail the key components and the methodology behind ST-MANet. Section IV presents a detailed account of the experimental results and evaluations conducted to validate the efficacy of ST-MANet. Finally, Section VI concludes the paper by summarizing the key findings and contributions of ST-MANet.

2. Related Works

2.1. Dual-stream based methods

Dual-stream based methods incorporate both spatial and temporal information, which enhances behavior recognition in video analysis. Karpathy et al. propose the dual-stream network architecture to extract action features at different resolutions in two streams, extending convolutional neural networks into the realm of videos [14]. Simonyan et al. introduce the use of optical flow information to extract temporal features and provide a new direction for action recognition, which divides the model into two parts, with one part processing individual video frames for spatial information extraction, and the other part taking dense optical flow from consecutive frames for temporal feature extraction [19]. The classification results from both streams are combined with the Softmax to capture final action recognition results. Although the dual-stream network utilizing complex optical flow achieved superior results on public datasets at the time, it still had room for improvement as the network only utilized a small portion of information from the entire video. Building upon this, Wang et al. design a network based on the temporal segment (TSN) [26]. In the data processing stage, TSN divides video data into multiple segments. Each segment undergoes feature recognition based on the dual-stream network. The spatial and temporal stream scores from each segment are then aggregated to obtain the overall spatial and temporal scores for the entire video. Finally, the two scores are fused to yield the ultimate result. TSN effectively utilizes video information by uniformly extracting action features across different slices of the entire video, enhancing recognition accuracy.

However, TSN's fusion of information from different segments is only evident in the model's final stage, simply consolidating recognition results without capturing more nuanced temporal information from the feature extraction results. Sun et al. propose a novel LSTM structure and multimodal training approach to train networks [20]. They employed two streams to jointly train the input and optical flow gates in the network, departing from traditional methods that treated the two streams as independent training networks. Tran et al. introduce 3D CNNs in video recognition, asserting that 3D convolution kernels are more suitable for temporal features in video data [22]. They conducted tests on the optimal size for the third dimension (depth) based on Simonyan et al.'s recommended 3×3 2D convolution kernel and provided the most suitable three-dimensional convolution kernel and network architecture for video convolution. Carreira et al. present a dual-stream Inflated 3D ConvNet (I3D) based on dilated 2D convolutions [1]. They applied 3D convolutions in both streams, expanding the design and parameters of deep image classification convolutional networks seamlessly for video spatiotemporal feature extraction.

Dual-stream based methods have been a cornerstone in the field of action recognition due to their ability to capture both spatial and temporal information from videos. The spatial stream typically processes individual frames to extract features from the visual content, while the temporal stream exploits optical flow to analyze motion patterns over time. This dual approach has proven to be highly effective in distinguishing complex actions and interactions. However, the process of extracting optical flow is computationally intensive. It involves estimating the motion of objects between consecutive frames, which can be particularly challenging in videos with large motion variations, occlusions, or changes in illumination. Despite the significant progress in developing efficient optical flow algorithms, the computational demands remain a bottleneck for real-time applications and deployment on resource-constrained devices.

2.2. Behaviour rhythm based methods

Behaviour rhythm based methods aim to leverage the temporal structure and variations of actions within video sequences to enhance the performance of action recognition in videos. Wang et al. point out that the spatial stream of the dual-stream network is influenced by samples with similar backgrounds, and the temporal stream struggles to differentiate actions with similar trajectories in a short period [27]. Moreover, the two streams fail to deeply integrate spatial and temporal features. To address these issues, they proposed the Spatiotemporal Pyramid Network. In the temporal module, fixed-interval sampling is applied to the optical flow of the entire video, with the sampling intervals randomly determined to obtain features under different action rhythms. Zhang et al. introduce the Dynamic Temporal Pyramid Network (DTPN), based on multiscale modeling [30]. They directly employed inputs with five different random frame rates to extract features, designed a dual-branch multiscale temporal feature structure to handle intrinsic time scale differences in actions, and created a fusion module for different scale feature maps. This demonstrated the significance of both local and global features for video recognition. Feichtenhofer et al. propose the SlowFast network, training two different structures with data at two frame rates, sampling input frames with a larger interval to extract spatiotemporal features [5]. The latter, at a higher frame rate, samples frames with a smaller time gap on the same video segment. Similar to dual-stream networks, a fusion mechanism

is employed to integrate information from both pathways to obtain the final classification result. Yang et al. introduce the concept of visual rhythm and statistically analyzed the visual rhythms of different actions [28]. They proposed the Temporal Pyramid Network (TPN), which samples image features at different ratios to extract hierarchical action rhythms. A feature aggregation module is utilized to fuse the rhythm features. Inspired by machine learning feature selection methods, Facebook AI Research introduced the X3D extension structure for efficient video recognition [4]. This structure expands the model in three dimensions based on a 2D convolutional network, providing a meaningful reference example for further extension of efficient video recognition models.

The behavior rhythm-based methods for video action recognition represent a significant advancement in the field by focusing on the temporal dynamics of actions within a video sequence. These methods analyze the variations in human behavior over time to capture the rhythm and patterns that define different actions. By sampling videos at multiple rates and at different time intervals, these approaches can extract a rich set of features that encapsulate the nuances of human motion, leading to more accurate recognition of actions. However, the computational expense of such an approach is a significant drawback. The need for multiple video samplings and the subsequent processing to extract features at various rates results in a substantial increase in the data volume that needs to be handled. This not only slows down the training process but also requires substantial computational resources, which can be a limiting factor for large-scale video analysis.

2.3. Differential feature based Methods

This type of method directly acquires action features by utilizing the differential information between adjacent data points and features, significantly reducing the computational load involved in data processing and model training. Differential calculations are divided into RGB differential and feature differential. RGB differential, resembling optical flow in representing motion information, can capture certain action details, and its extraction method is often more efficient. It is frequently used to complement spatial features extracted based on RGB frames. For example, in the Temporal Segment Network (TSN) [26], differential feature extraction is added to the original image features in the spatial stream, and experimental results demonstrate the effectiveness of their combined features for video action recognition. Lin et al. propose the Temporal Shift Module (TSM), suitable for spatiotemporal feature extraction, where features of certain channels are temporally shifted to fill gaps in the time dimension, achieving an effect similar to extracting differential features [17]. Liu et al. introduced TEINet, which includes an action enhancement module and a time-increasing interaction module [18]. The action enhancement module represents motion information by computing the differential feature maps of adjacent frames, and the time module further performs convolution operations on features channel-wise, capturing temporal information. Jiang et al. propose a channel-wise action module for action feature extraction [12]. It calculates the differential features between different dimensional channel features, connects the differential results of each channel, and applies 2D convolution to obtain the final action feature information. Li et al. presented a temporal excitation and aggregation network, similar to the aforementioned methods, with the distinction that the aggregation module divides features into multiple slices [16]. Each slice's features undergo convolutional operations before subtraction. The action extraction module also conducts differential calculations on the features in the initial network layer,

enhancing the model's video recognition performance via multiple uses of differential features. In the temporal difference network proposed by Wang et al., appearance features are extracted in the lower-level network, introducing the short-term temporal difference module [25]. Meanwhile, temporal features are extracted in the higher-level network, introducing the long-term temporal difference module. Both modules employ differential techniques for RGB frames and features separately.

Despite the notable achievements in video action recognition through the use of differential information for feature extraction, these methods encounter several challenges. The RGB differential, which excels at capturing motion details, may fall short when it comes to complex actions that require a deeper understanding of semantic information. Additionally, while differential methods effectively capture local changes, they may struggle with long-term dependencies in videos, such as those addressed by the Temporal Segment Network (TSN). The challenge of effectively fusing features from multiple modules within network architectures like TEINet and others is significant, as is the design of these fusion strategies, which are critical to the final recognition performance but prone to information loss or redundancy. Computational resource optimization remains an issue, especially when dealing with large-scale video data, despite the computational benefits offered by differential calculations. The modular design of networks, although successful for specific tasks, may have limited generalization capabilities when applied to other action recognition tasks. Furthermore, the short-term and long-term temporal difference modules introduced in temporal difference networks, while handling features at different time scales, may not adapt well to rapid or nonlinear motion dynamics in videos. Lastly, the performance of these methods is heavily dependent on the quality and diversity of the training datasets, with biases or insufficiencies potentially leading to reduced generalization capabilities. In summary, while differential information provides valuable insights for action feature extraction, a range of challenges must be addressed to ensure robust and generalizable models in practical applications.

3. Spatio-Temporal-based Multi-level Aggregation Network for Action Recognition

A spatio-temporal-based multi-level aggregation network (ST-MANet) is proposed for action recognition, which contains the baseline network, the temporal-based deep aggregation network, and the spatio-based deep aggregation network. The main mathematical notations used in the paper are listed in Table 1.

3.1. Baseline Network

ST-MANet utilizes 3D ResNet-50 as the baseline network for human action recognition, which is composed of five stages, as shown in Fig. 1.

Specifically, stage 1 of 3D ResNet-50 consists of a convolutional layer, a time step of 1, and a spatial step of 2, followed by BN and ReLU layers in sequence, where the convolutional layer extracts features and reduces spatio-temporal dimensions. The Batch Normalization layer prevents the occurrence of gradient vanishing and accelerates network convergence. The ReLU layer enhances the expressive power of the extracted features.

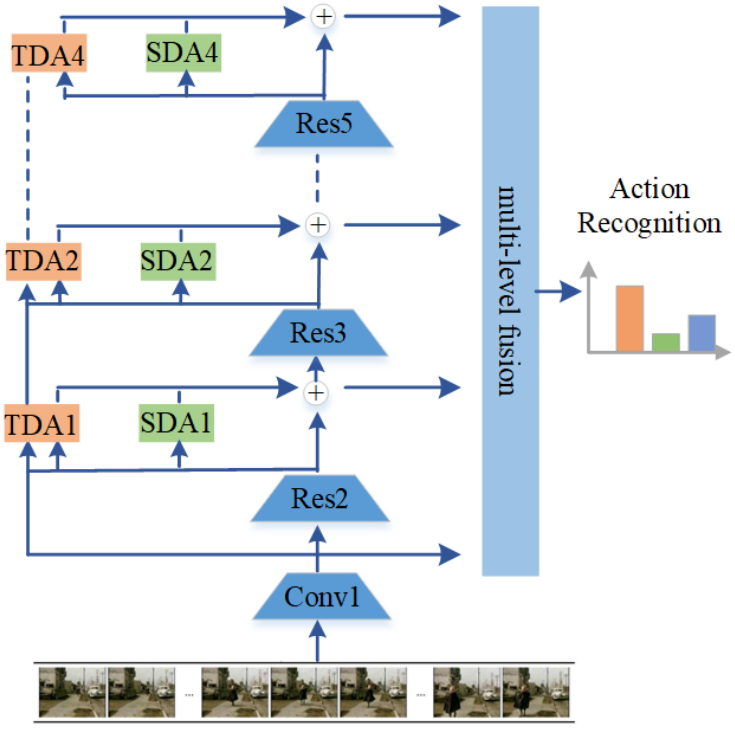


Fig. 1. The illustration of the proposed ST-MANet

Table 1. Frequently used notations

Notations	Description
Z^i	the feature map from the i-th stage of 3D ResNet-50
Z_{TDA}^{i-1}	the feature map from the i-th stage of temporal-based deep aggregation network
Z_c^i	the concatenation feature of Z^i and Z_{TDA}^{i-1}
A_t^i	the temporal attention matrix
A_s^i	the spatio attention matrix
$\bar{Z}_t^i$	the temporally aggregated feature map
$\bar{Z}^i$	the final output of the i-th stage of 3D ResNet-50
$f(\cdot)$	the cosine similarity function
$S^i(\cdot)$	the downsampling function
$Con(\cdot)$	the fusion function
L_{ce}	the cross-entropy loss
L_s	the spatial diversity constraint
L_t	the temporal diversity constraint
λ_1,λ_2	the balance coefficients

The subsequent four stages contain 3, 4, 6, and 3 Bottleneck Residual Blocks, respectively, where each block is a concatenation of three convolutional blocks. The first convolutional block consists of a convolutional layer, BN layer, and a ReLU layer. The output channel number is the same as the input channel number, aiming to promote interaction between different channels. The second convolutional block consists of a convolutional layer, a BN layer, and ReLU layer. The channel number is again the same as the input channel number, intended to expand the spatio-temporal receptive field of the network. The third convolutional block consists of a convolutional layer, a BN layer, and a ReLU layer. The output channel number is four times the input channel number. Each bottleneck residual block is constructed with a residual structure, stacked together to form the different stages of 3D ResNet-50. The feature space and temporal resolution of the output from each stage are reduced by a factor of two, while the channel dimension is doubled, giving the network powerful abstract feature extraction capabilities.

3.2. Temporal-based Deep Aggregation Network

Considering the varying contributions of video frames at different time points to action recognition, ST-MANet employs a visual self-attention mechanism to construct a temporal-based deep aggregation network (TDA), aiming to enhance the capturing of temporal information in action videos.

Specifically, TDA employs a dual-branch design, with inputs originating from the ResNet-50 branch and the temporal-based deep aggregation network branch, which is shown in Fig. 2. The ResNet-50 branch retains the relative temporal relationships of the original input and transfers information to the temporal-based deep aggregation network branch. TDA makes full use of a self-attention mechanism to aggregate and enhance the mining of temporal information about actions, which not only is temporal information reinforced, but also contributes to the extraction of holistic features.

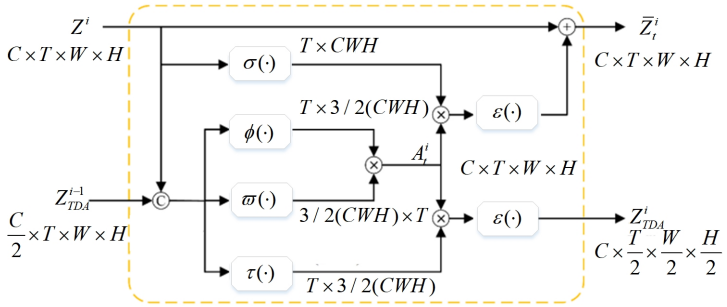


Fig. 2. Temporal-based deep aggregation network

More specifically, given the feature map from the i-th stage of 3D ResNet-50, $Z^i \in \mathbb{R}^{C\times T\times W\times H}$, and the feature map from the previous temporal-based deep aggregation network, $Z_{TDA}^{i-1} \in \mathbb{R}^{C/2\times T\times W\times H}$, as the input of the i-th temporal-based deep aggregation network, where C is channel number. T, W, and H are the time steps, the width, and

the height, respectively. TDA concatenates Z^i and Z^{i-1}_{TDA} along the channel dimension to obtain the concatenation feature $Z^i_c \in \mathbb{R}^{3C/2\times T\times W\times H}$, and then computes the temporal attention matrix A^i_t via:

$$A^i_t = softmax(\phi(Z^i_c) \otimes (\varpi(Z^i_c)) \tag{1}$$

where $\phi(\cdot)$ and $\varpi(\cdot)$ denote reshaping operations to obtain the feature map with the size of $(T \times 3/2(CWH))$ and $(3/2(CWH) \times T)$, respectively. The elements in A^i represent the degree of correlation between features at each moment in time. Moments with strong correlations to other moments in the feature map have higher time attention scores, and vice versa. It is worth noting that the convolutional kernel sizes used here are $3 \times 1 \times 1$, which, compared to the commonly used $1 \times 1 \times 1$ kernels in visual self-attention mechanisms, have a larger size along the time dimension. This results in a larger temporal receptive field, providing stronger temporal modeling capabilities.

Next, the feature map Z^{i-1}_{TDA}, after the $\tau(\cdot)$ operation, is multiplied element-wise with the temporal attention matrix A^i_t, normalized using the softmax function, and further processed by the $\epsilon(\cdot)$ operation to obtain the temporally aggregated feature map Z^i_{TDA}:

$$Z^i_{TDA} = \epsilon(softmax(A^i_t \otimes \tau(Z^{i-1}_{TDA}))) \tag{2}$$

where $\tau(\cdot)$ denote the reshaping operation to obtain the feature map with the size of $(T\times 3/2(CWH))$. $\epsilon(\cdot)$ first reshapes the feature map back to a size of $3C/2\times T\times W\times H$, then passes it through a convolutional layer with a kernel size of $3 \times 1 \times 1$, followed by a pooling layer to output the temporally aggregated feature X^i_{TDA}. It is worth noting that for the final TDA, $\epsilon(\cdot)$ does not include a pooling layer to ensure that the ResNet-50 branch and the temporal-based deep aggregation network branch outputs have the same dimensions for subsequent feature fusion.

Similarly, after the $\sigma(\cdot)$ operation, the feature map Z^i is multiplied element-wise with the attention map A^i_t. Subsequently, it undergoes the softmax activation function and the $\epsilon(\cdot)$ operation, and then is element-wise added with the X^i to obtain the temporally aggregated feature map $\bar{Z}^i_t$:

$$\bar{Z}^i_t = Z^i + \epsilon((softmax(A^i_t \otimes \sigma(Z^i)))) \tag{3}$$

where $\sigma(\cdot)$ denotes the reshape operation to obtain the feature map with the size of $(T \times CWH)$.

TDA uncovers additional temporal correlation information within videos, exploring long-range dependencies over time and consolidating temporal information. Notably, at each stage, the output from the upper part of the temporal aggregation module is integrated into the main branch, reinforcing temporal information in the primary pathway. The introduction of the temporal aggregation module enhances the temporal modeling capabilities of 3D ResNet-50. Additionally, the interaction between the two branches delves into more information within the video.

3.3. Spatio-based Deep Aggregation Network

To focus on regions containing more discriminative information, ST-MANet designs the spatio-based deep aggregation network(SDA) to calculate the correlation between different spatial positions in the feature map, as illustrated in Fig. 3.

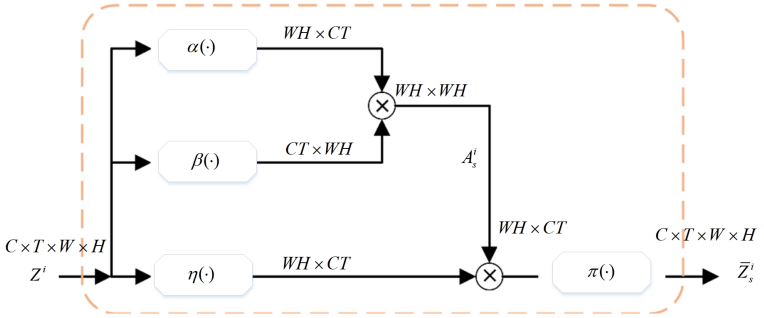


Fig. 3. Spatio-based deep aggregation network

Specifically, given the feature map from the i-th stage of 3D ResNet-50, $Z^i \in \mathbb{R}^{C\times T\times W\times H}$, as the input of the i-th spatio-based deep aggregation network, SDA computes the spatio attention matrix A_s^i via:

$$A_s^i = softmax(\alpha(Z^i) \otimes \beta(Z^i)) \tag{4}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ denote reshaping operations to obtain the feature map with the size of $(WH \otimes CT)$ and the size of $(CT \otimes WH)$, respectively. The elements in A_s^i indicate the correlation between features at each spatial position in Z^i and features at other spatial positions. In other words, positions with strong correlations to other spatial positions in the feature map have higher spatial attention scores, and vice versa.

Then, the spatially enhanced feature map $\bar{Z}_s^i$ is obtained from:

$$\bar{Z}_s^i = \pi(softmax(A_s^i \otimes \eta(Z^i))) \tag{5}$$

Where $\eta(\cdot)$ and $\pi(\cdot)$ denote reshaping operations to obtain the feature map with the size of $(WH \otimes CT)$ and the size of $C \times T \times W \times H$, respectively.

SDA unveils additional spatial correlation information within images, exploring intricate dependencies across different regions and consolidating spatial information. Notably, in each processing stage, the output from the upper part of the spatial aggregation module is seamlessly integrated into the main branch, fortifying spatial information in the primary pathway. The incorporation of the spatial aggregation module elevates the spatial modeling capabilities of the 3D ResNet. Moreover, the interaction between the two branches delves into a richer trove of information within the image. This synergistic approach enhances explorations in nuanced spatial relationships and dependencies.

3.4. The multi-level fusion loss of ST-MANet

Convolutional neural networks generate hierarchical feature maps through repeated pooling and subsampling operations. As the hierarchy goes deeper, the receptive field becomes larger, indicating stronger representational capabilities but leading to the loss of detailed spatial information. In other words, deep features in convolutional neural networks possess high-level visual perception but discard some crucial local information. Shallow layers, on the other hand, exhibit higher spatial resolution features, preserving fundamental

details of the image but lacking semantic understanding. To enhance the overall expressive power of the network, ST-MANet introduces temporal and spatial diversity constraints for different depths, which aims to encourage attention matrices at different depths to focus on distinct temporal and spatial information.

Considering that temporal and spatial attention matrices also have different dimensions, a downsampling operation is applied to the attention matrices of shallower layers to ensure that all attention matrices have the same size:

$$\bar{A}^i = S^i(A^i), i = 2, 3, 4, 5 \tag{6}$$

where i represents the i-th stage of ResNet-50 and $\bar{A}^i$ represents the attention matrix after downsampling for the i-th stage of 3D ResNet-50. $S^i(\cdot)$ is the downsampling function of the i-th stage of ResNet-50. Then, ST-MANet minimizes the cosine similarity between the aligned attention matrices to maximize the differences at different network depths in terms of time and space:

$$\begin{aligned} L_t &= \sum_{(i,j)\in\delta} f(\bar{A}_t^i, \bar{A}_t^j) \\ L_s &= \sum_{(i,j)\in\delta} f(\bar{A}_s^i, \bar{A}_s^j) \end{aligned} \tag{7}$$

where L_s and L_t represent the spatial and temporal diversity constraints, $\delta = \{(i, j) \mid i, j \in \mathbb{Z}, 2 \leq i, j \leq 5, i \neq j\}$, where i and j denote the stages of the network. This constraint is computed only for stages 2 and onwards. $f(\cdot)$ denotes the cosine similarity function. During training, the spatial and temporal diversity constraints tend to decrease, encouraging a reduction in the cosine similarity of the spatial and temporal attention maps across different stages. This reduction further contributes to compelling the network to focus on distinct the spatial and temporal locations at various stages. This enhances the diversity of the spatial and temporal information within multi-scale features.

Meanwhile, ST-MANet fuses four feature maps of different depths by cascading along the channel dimension to obtain a comprehensive and good fusion representation Z for the classification of action recognition;

$$Z = Con(Z^2; Z^3; Z^4; Z^5) \tag{8}$$

where $\bar{Z}^i = Z^i \oplus \bar{Z}_t^i \oplus \bar{Z}_s^i$ denotes final output of the i-th stage of 3D ResNet-50, which fully aggregates spatio-temporal information in each layer.

The multi-level fusion loss of ST-MANet is defined:

$$L = L_{ce} + \lambda_1 L_s + \lambda_2 L_t \tag{9}$$

where L_{ce} is the cross-entropy loss. L_s is the spatial diversity constraint. L_t is the temporal diversity constraint. λ_1 and λ_2 are the balance coefficients.

4. Experiments

4.1. Set up

Dataset and metric: The performance of the ST-MANet in action recognition is verified based on three popular datasets, i.e., UCF-101, HMDB-51, and Diving 48. The comprehensive statistical details for each dataset are provided in Table 2. Building upon prior

research, the evaluation of performance relies on accuracy as the primary metric, with higher values indicating superior performance.

- UCF-101 is a coarse-grained description of the action recognition. It contains 13320 action videos spanning 101 distinct classes.
- HMDB-51 is a coarse-grained description of the action recognition. It contains 6849 action videos spanning 51 distinct classes.
- Diving 48 is a fine-grained description of the action recognition. It contains 18404 action videos spanning 48 distinct classes.

Table 2. Statistical Details For Each Dataset

Dataset	Sample	Class	Type
Ucf-101	13320	101	Coarse-Grained Description
Hmdb-51	6849	51	Coarse-Grained Description
Diving 48	18404	48	Fine-Grained Description

Implementation Details: ST-MANet conducts preprocessing and data augmentation on the input video data. Initially, it randomly samples continuous video segments of 64 frames from the videos. Subsequently, the video segments undergo subsampling with a stride of 2, resulting in frame sequences of 32 frames. Next, random cropping is applied to the frames, with the length and width set between 0.5 and 1 times the length of the shorter side of the original video frame. Finally, a spatial scale transformation is applied to resize both the length and width of the frame sequences to 112. To increase training samples, random horizontal flipping is employed as a data augmentation technique. During the training process, the paper utilizes batch SGD to train the ST-MANet, employing cross-entropy loss function with a momentum coefficient of 0.9 and 32 batch size. The network parameters are initialized using the Kaiming initialization strategy. The initial learning rate is set at 0.01, and every 15 epochs, the learning rate is decayed by a factor of 0.1. The entire training process spans 90 epochs. In contrast to the preprocessing of data before training, during testing, the entire video is first divided into multiple continuous video segments of 64 frames each. These segments then undergo subsampling with a stride of 2, resulting in multiple video frame sequences of 32 frames. Subsequently, central cropping is applied to each frame sequence. Finally, multiple frame sequences of a video are input into the network.

4.2. Comparison with baselines

Comparison methods: Eight baseline methods are compared to evaluate the performance of ST-MANet in the experiments, including TCNet [19], LFSNet [23], TSNet [26], LTNet [24], MRNet[7], RMRNet [21], MVTNet [3], and A3DNet [29].

Comparison results: Table 3 presents a comparison of experiment outcomes regarding accuracy on the three few-shot datasets. Specifically, ST-MANet outperforms eight baseline methods across both settings, which proves effectiveness and superiority in the action recognition. The seasons are fourfold. (1) ST-MANet designs temporal-based deep

Table 3. Comparison of ST-MANet to top-performing methods on three datasets in terms of the accuracy

Method	UCF-101	HMDB-51	Diving 48
TCNet	0.8200	0.5940	0.5222
LFSNet	0.8580	0.5492	0.5645
TSNet	0.8640	0.5881	0.5987
LTNet	0.9270	0.6720	0.6719
MRNet	0.8440	0.5900	0.5904
RMRNet	0.9418	0.6275	0.6668
MVTNet	0.9040	0.5460	0.6048
A3DNet	0.9260	0.6126	0.6521
ST-MANet	**0.9605**	**0.6872**	**0.7215**

aggregation network to enhances the temporal information in the 3D RestNet-50 and incorporates a temporal aggregation branch for exploring and aggregate additional temporal information. (2) ST-MANet designs the spatial-based deep aggregation network to strengthen discriminative spatial information relevant to action recognition. (3) ST-MANet introduces a multi-level fusion strategy to model various visual rhythms and explore multi-scale temporal and spatial information concurrently. (4) By introducing temporal and spatial diversity constraints, ST-MANet encourages temporal and spatial attention at multiple scales to focus on different positions, ensuring that the ultimately fused multi-scale features contain more comprehensive and complete spatiotemporal information.

In addition, there are four observations. (1) TCNet and TSNet, while leveraging the powerful dual-stream architecture, struggle with capturing long-term temporal dynamics due to their focus on optical flow between adjacent frames. This limitation suggests the need for more advanced mechanisms that can handle temporal dependencies over larger time spans. Furthermore, their spatial stream networks, which are crucial for understanding the context and structure within each frame, could benefit from more innovative designs to better extract relevant spatial features. (2) LFSNet's adoption of the C3D network, although it offers a straightforward and efficient architecture, falls short in delving deeper into the intricate spatiotemporal relationships present in videos. LTNet's enhancement to include more frames is a step in the right direction, yet the network's overall simplicity may not be sufficient to fully harness the rich information encoded in video sequences. This points towards the necessity for more complex and sophisticated network structures that can integrate information from multiple frames effectively and robustly. (3) MRNet's focus on refining the GRU for temporal exploration is commendable, but the inherent challenges in extracting spatial information using RNN-based methods cannot be overlooked. ST-MANet's integration of a spatial attention module with 3D ResNet-50 represents a significant advancement, as it allows the network to adaptively focus on relevant regions within the frame, leading to a more nuanced understanding of the spatial content. This underscores the importance of attention mechanisms in enhancing the model's ability to discern and prioritize spatial information. (4) RMRNet's investigation into residual frames is an innovative approach to bolster the motion capture capabilities of 3D convolutional networks. However, the lack of structural innovation and the reliance on 3D

ResNet-18 for temporal feature extraction may limit the network's overall performance. This indicates that while residual frames can offer valuable insights into motion, the underlying model architecture must also be optimized to ensure comprehensive temporal analysis. In summary, these observations reveal that while current networks have made strides in video action recognition, there is still much room for improvement. The key lies in striking a balance between the complexity of the network, which allows for thorough spatiotemporal feature extraction, and the efficiency required for practical applications. Future research should focus on developing architectures that can effectively handle long-term temporal dependencies, innovate in spatial feature extraction, and integrate attention mechanisms to refine the model's focus on critical information. By addressing these aspects, we can expect significant advancements in the field of video action recognition, leading to more accurate and efficient models capable of understanding and analyzing the rich content of video data.

4.3. Ablation Study

Table 4. Ablation experiments of each component in ST-MANet

3D ResNet-50	TDA	SDA	MF	Accuracy
√				0.8317
√	√			0.8945
√		√		0.9012
√	√	√		0.9278
√			√	0.8416
√	√	√	√	0.9605

We conducts five ablation experiments about 3D ResNet-50, TDA (temporal-based deep aggregation network), SDA (spatial-based deep aggregation network), and MF (multi-level fusion strategy). Specifically, (1) ST-MANet utilizes 3D ResNet-50 to extract features for performing action recognition tasks. (2) ST-MANet intergrades 3D ResNet-50 and the temporal-based deep aggregation network to extract features for performing action recognition tasks. (3) ST-MANet intergrades 3D ResNet-50 and the spatial-based deep aggregation network to extract features for performing action recognition tasks. (4) ST-MANet intergrades 3D ResNet-50, the temporal-based deep aggregation network and the spatial-based deep aggregation network to extract features for performing action recognition tasks. (5) ST-MANet intergrades 3D ResNet-50 and the multi-level fusion strategy to extract features for performing action recognition tasks.

As shown in Table 4, there are six conclusions: (a) The Effectiveness of 3D ResNet-50 as a Basic Feature Extractor: Through experiment (1), we found that using 3D ResNet-50 alone for action recognition tasks, the model can effectively extract spatiotemporal features. This indicates that 3D ResNet-50, as a deep learning base network, has good performance in the field of action recognition. However, its performance may be limited by the rigidity of the network structure, which may not fully capture the complex spatiotemporal information in videos. (b)The Complementary Role of the Temporal-Based

Deep Aggregation Network (TDA): In experiment (2), by integrating 3D ResNet-50 with the Temporal-Based Deep Aggregation Network (TDA), the model's performance in action recognition tasks was improved compared to using 3D ResNet-50 alone. This suggests that TDA can effectively supplement the temporal feature extraction capabilities of 3D ResNet-50, enhancing the model's ability to capture the temporal sequence changes in actions. (c) The Integration Effect of the Spatial-Based Deep Aggregation Network (SDA): In experiment (3), the model's performance was also enhanced by combining 3D ResNet-50 with the Spatial-Based Deep Aggregation Network (SDA). This indicates that SDA effectively strengthens the feature extraction in the spatial dimension, enabling the model to better understand spatial relationships and layouts in videos. (d) Significant Improvement with Spatiotemporal Fusion Strategy: In experiment (4), by fusing both Temporal-Based and Spatial-Based Deep Aggregation Networks, the model's performance in action recognition tasks reached its peak. This result confirms the importance of joint optimization of spatiotemporal information for enhancing the accuracy of action recognition, where spatiotemporal fusion strategy can more comprehensively capture the complex spatiotemporal features in videos. (e) Further Optimization with Multi-Level Fusion Strategy (MF): in experiment (5), by introducing the Multi-Level Fusion Strategy (MF) in conjunction with 3D ResNet-50, the model's performance was further optimized in some cases. This indicates that the Multi-Level Fusion Strategy can more meticulously integrate spatiotemporal information at different levels, providing richer feature representations for action recognition tasks. (f) the combined effect of these modules transcends the sum of their individual contributions, highlighting the synergy and complementarity inherent in ST-MANet's design. This observation underscores the rationality and effectiveness of the modules within ST-MANet, demonstrating their collective power in enhancing the model's accuracy and performance. Such a finding not only validates the architectural choices made in ST-MANet but also offers insights into the importance of holistic system design in achieving superior results in action recognition tasks.

4.4. The effectiveness of multi-level fusion

To delve deeper into the contributions of each stage within ST-MANet to the overall performance of action recognition, we conducted a comparative analysis. This analysis involved directly classifying features extracted from individual stages and then comparing these results with the outcomes obtained from features that have undergone multi-level fusion. The detailed results of this comparison are presented in Table 4.

These are some observations: (1) The top-level features, which are the output of the deepest layers in the network, are found to contain rich semantic information that is crucial for understanding the actions within the videos. These features have undergone a series of transformations and aggregations that enable them to capture complex patterns and relationships present in the data. On the other hand, features from the shallower layers tend to lack the sophistication to adequately model semantic information. They exhibit limited reasoning capabilities when it comes to processing both temporal and spatial information, which is evident from the decreasing trend in recognition accuracy as we move towards shallower features. (2) The features extracted from Stage 5 represent a turning point in this analysis. At this stage, the features have been fused from both the main branch and the temporal aggregation branch. This fusion results in a rich amalgamation of temporal and spatial information, alongside semantic insights. As a result, Stage 5 features achieve the

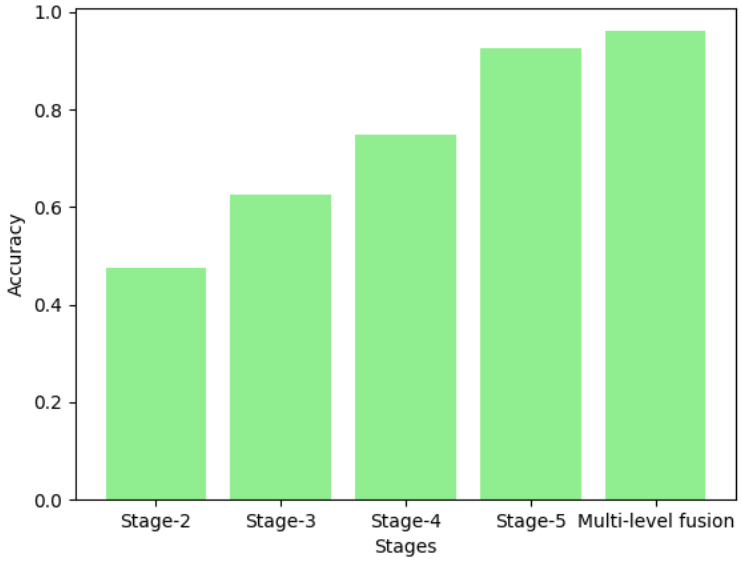
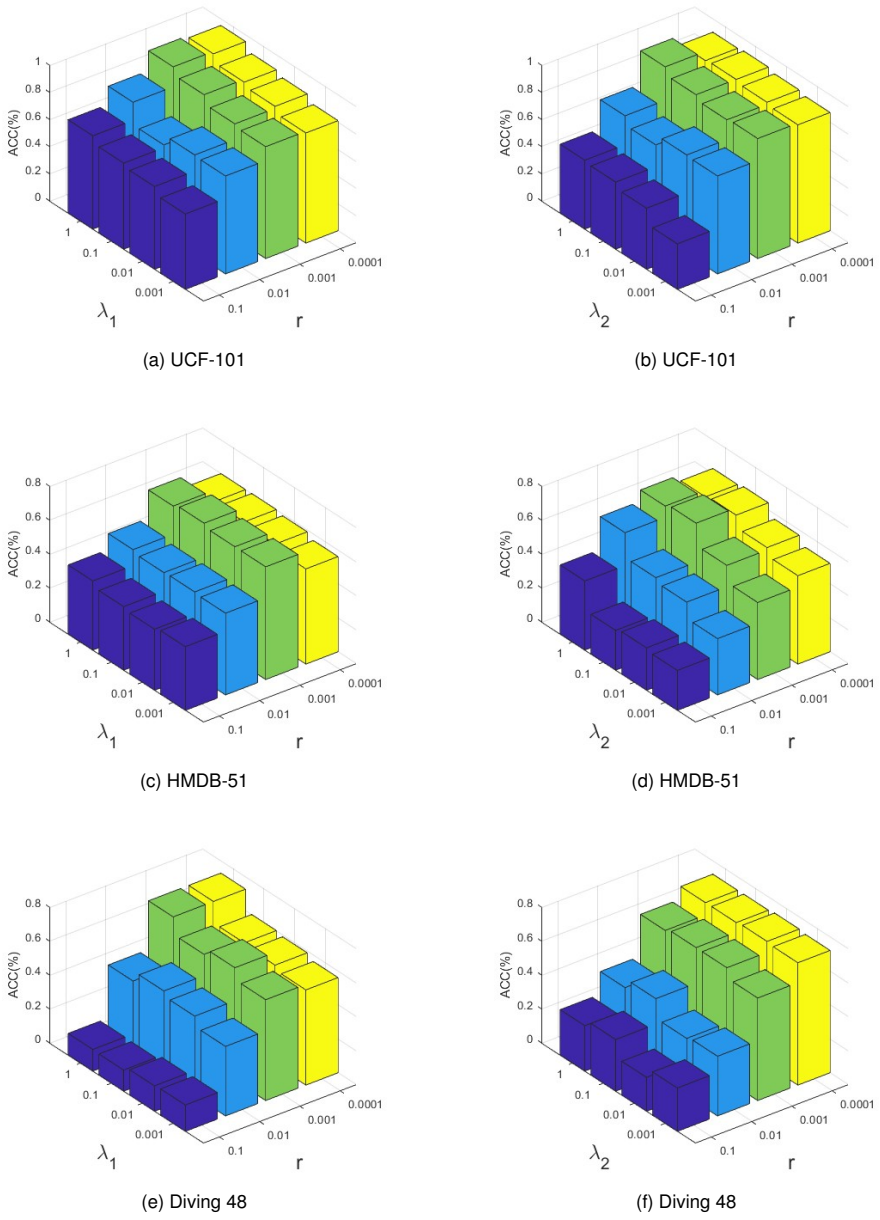


Fig. 4. Contribution ablation of each stage of in ST-MANet

highest accuracy among the single-scale features, indicating the importance of the temporal aggregation module in enhancing the network's temporal modeling capabilities. (3) The multi-level fusion strategy takes this a step further by integrating features from multiple stages of the network. Temporally, this strategy spans a wide range of information, capturing various visual rhythms and exploring multiple time scales. This comprehensive temporal coverage allows the network to recognize actions that may vary in speed and duration. Spatially, the fused features not only include deeply learned, highly discriminative information but also incorporate finer details from the shallower layers. This combination of deep and shallow features ensures that the model can leverage both the nuanced understanding of the scene's layout and the broader context of the action being performed. Consequently, the multi-level fusion strategy stands out as the most effective approach, achieving the highest accuracy in the action recognition task. This strategy exemplifies the power of combining diverse types of information at different levels of abstraction, which is a key principle in designing deep learning models for complex tasks such as video action recognition. By harnessing the collective strength of features from various stages, ST-MANet is able to outperform other approaches and set a new benchmark for accuracy in the field.

4.5. Parameter Analysis

Fig. 5 illustrates the results of the analysis experiments investigating the influence of trade-off parameters λ_1 and λ_2 and learning rate r on three datasets. Specifically, we



(a) UCF-101 (b) UCF-101

(c) HMDB-51 (d) HMDB-51

(e) Diving 48 (f) Diving 48

Fig. 5. The sensitivity analysis of parameters λ_1 and λ_2 and learning rate r on three datasets for ST-MANet

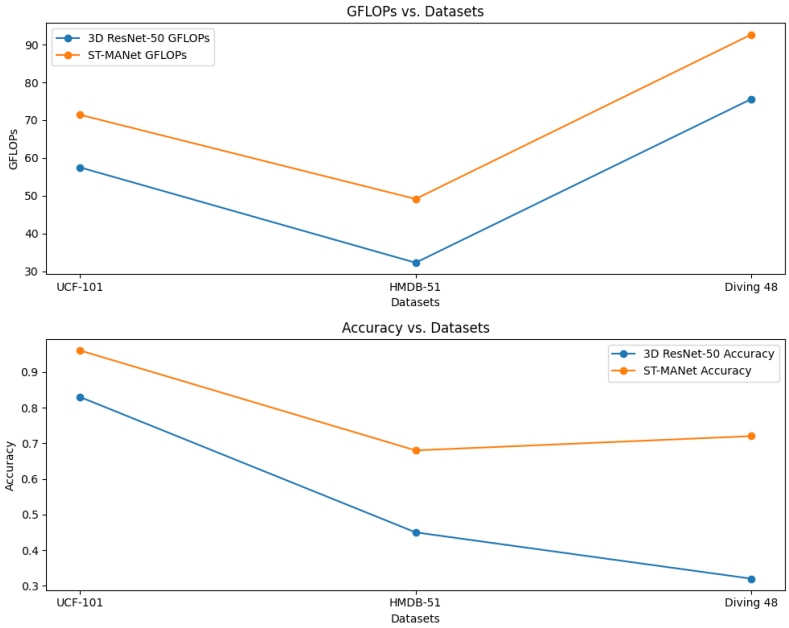


Fig. 6. Comparison of computational complexity and accuracy of ST-MANet and 3D ResNet-50 in terms of three datasets

constrained λ_1 and λ_2 within the set $\{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. Throughout the experiments, one parameter was maintained at a constant value while the other was varied. The results displayed in Fig. 5 demonstrate the robustness of ST-MANet to changes in λ_1 and λ_2. The performance consistently remains satisfactory, particularly when λ_1 is set to 0.01 and λ_2 is set to 0.001. As a result, for the three datasets, ST-MANet is configured with $\lambda_1 = 0.01$ and $\lambda_2 = 0.001$ in the experiments. This configuration has been empirically determined to yield a high level of accuracy, suggesting that it effectively weights the different components of the network's loss function in a manner that is conducive to learning robust action recognition features. The parameters λ_1 and λ_2 play a crucial role in the network by controlling the balance between certain regularization terms and the overall loss. The optimal values found in this study provide a valuable reference for future research and applications of ST-MANet, as they offer a blueprint for achieving high performance with a reasonable computational cost.

In addition to the qualitative assessment of the ST-MANet's performance, we have conducted a quantitative analysis to evaluate its computational efficiency and effectiveness. This analysis includes a direct comparison of the floating-point operations required by ST-MANet and the 3D-ResNet-50 architecture when applied to the UCF-101 dataset, a benchmark for action recognition research. As illustrated in Table 6, the implementation of ST-MANet results in an approximate 24% increase in floating-point operations when contrasted with the 3D ResNet-50. This increase is attributed to the additional temporal and spatial modules integrated into ST-MANet, which are designed to enhance the net-

work's ability to capture complex spatiotemporal features within videos. Despite this rise in computational demand, it is important to note that the computational footprint of both ST-MANet and 3D ResNet-50 remains relatively comparable. This suggests that while ST-MANet is more computationally intensive, the increase is not prohibitive and is justified by the performance gains. The accuracy improvements achieved by ST-MANet are particularly noteworthy. On the UCF-101 dataset, ST-MANet demonstrates a 12.88% increase in accuracy over the baseline 3D ResNet-50. This significant leap in performance underscores the value of the network's advanced feature extraction capabilities, particularly its ability to model the intricate temporal dynamics that are critical for accurate action recognition. In conclusion, while ST-MANet does require more computational resources than 3D ResNet-50, the increase is moderate and does not lead to a significant disparity in computational requirements. The substantial improvement in accuracy, especially on challenging datasets like UCF-101, is a testament to the network's advanced feature aggregation and modeling capabilities. This makes ST-MANet a compelling choice for applications where high accuracy in action recognition is paramount, justifying the additional computational investment.

5. Conclusion

This paper introduces ST-MANet, a novel spatio-temporal-based multi-level aggregation network for action recognition.Specifically, by leveraging correlations between different spatial and temporal positions on the feature map, ST-MANet effectively explores long-range spatial and temporal dependencies, generating spatial and temporal attention maps that assign varying weights to features at different locations. Moreover, the integration of a multi-scale approach enables the modeling of various visual rhythms and the capture of multi-scale spatiotemporal information. The proposed spatial diversity constraint further enhances the network's ability to focus on distinct spatial areas at different scales, thereby incorporating more diverse spatial information into multi-scale features. Experimental results demonstrate the effectiveness of ST-MANet compared to existing approaches, showcasing high accuracy across three datasets. In the future, there are several promising avenues for future research. Firstly, exploring more sophisticated attention mechanisms could further improve the network's ability to capture salient spatial and temporal features. Additionally, investigating ways to incorporate domain knowledge or context information may enhance the model's robustness in real-world scenarios. Furthermore, extending the application of ST-MANet to other related tasks beyond action recognition, such as anomaly detection or event localization, could offer valuable insights into its broader applicability and effectiveness. Overall, ST-MANet presents a strong foundation for advancing the state-of-the-art in spatio-temporal-based action recognition and opens up exciting possibilities for future research in the field.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)

2. Chai, W., Jiang, Z., Hwang, J.N., Wang, G.: Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In: ICCV. pp. 14655–14665 (2023)
3. Chen, J., Ho, C.M.: Mm-vit: Multi-modal video transformer for compressed video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1910–1921 (2022)
4. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: computer vision and pattern recognition. pp. 203–213 (2020)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: international conference on computer vision. pp. 6202–6211 (2019)
6. Gao, J., Liu, M., Li, P., Zhang, J., Chen, Z.: Deep multiview adaptive clustering with semantic invariance. IEEE Transactions on Neural Networks and Learning Systems (2023)
7. Guan, X., Yang, Y., Li, J., Xu, X., Shen, H.T.: Mind the remainder: taylor's theorem view on recurrent neural networks. IEEE Transactions on Neural Networks and Learning Systems 33(4), 1507–1519 (2021)
8. Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C.: Augfpn: Improving multi-scale feature learning for object detection. In: computer vision and pattern recognition. pp. 12595–12604 (2020)
9. Gupta, P., Thatipelli, A., Aggarwal, A., Maheshwari, S., Trivedi, N., Das, S., Sarvadevabhatla, R.K.: Quo vadis, skeleton action recognition? International Journal of Computer Vision 129(7), 2097–2112 (2021)
10. He, Z., Lv, J., Fang, S.: Representation modeling learning with multi-domain decoupling for unsupervised skeleton-based action recognition. Neurocomputing p. 127495 (2024)
11. Huo, J., Cai, H., Meng, Q.: Independent dual graph attention convolutional network for skeleton-based action recognition. Neurocomputing p. 127496 (2024)
12. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: Stm: Spatiotemporal and motion encoding for action recognition. In: the IEEE/CVF international conference on computer vision. pp. 2000–2009 (2019)
13. Karim, M., Khalid, S., Aleryani, A., Khan, J., Ullah, I., Ali, Z.: Human action recognition systems: A review of the trends and state-of-the-art. IEEE Access (2024)
14. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems 27 (2014)
15. Li, P., Gao, J., Zhang, J., Jin, S., Chen, Z.: Deep reinforcement clustering. IEEE Transactions on Multimedia (2022)
16. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: computer vision and pattern recognition. pp. 909–918 (2020)
17. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: international conference on computer vision. pp. 7083–7093 (2019)
18. Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T.: Teinet: Towards an efficient architecture for video recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11669–11676 (2020)
19. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems 27 (2014)
20. Sun, L., Jia, K., Chen, K., Yeung, D.Y., Shi, B.E., Savarese, S.: Lattice long short-term memory for human action recognition. In: IEEE international conference on computer vision. pp. 2147–2156 (2017)
21. Tao, L., Wang, X., Yamasaki, T.: Rethinking motion representation: Residual frames with 3d convnets. IEEE Transactions on Image Processing 30, 9231–9244 (2021)
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE international conference on computer vision. pp. 4489–4497 (2015)

23. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE international conference on computer vision. pp. 4489–4497 (2015)
24. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence 40(6), 1510–1517 (2017)
25. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: computer vision and pattern recognition. pp. 1895–1904 (2021)
26. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36 (2016)
27. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition. In: Computer Vision and Pattern Recognition. pp. 1529–1538 (2017)
28. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: computer vision and pattern recognition. pp. 591–600 (2020)
29. Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., Maybank, S.J.: Asymmetric 3d convolutional neural networks for action recognition. Pattern recognition 85, 1–12 (2019)
30. Zhang, D., Dai, X., Wang, Y.F.: Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In: Computer Vision. pp. 712–728 (2019)

Yuhang Wang is with the School of Physical Education, Harbin University. Research direction: image processing, sports analysis, education data analysis, artificial intelligence.

Received: April 18, 2024; Accepted: June 29, 2024.