IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 4, 2019, accepted August 18, 2019, date of publication August 21, 2019, date of current version September 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936628

Two-Level Attention Model Based Video Action Recognition Network

HAIFENG SANG[1], ZIYU ZHAO[1], AND DAKUO HE[2]

[1]School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China
[2]College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

Corresponding author: Ziyu Zhao (643306538@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773105 and Grant 61374147, in part by the Fundamental Research Funds for the Central Universities under Grant N182008004, in part by the Natural Science Foundation of Liaoning Province under Grant 20170540675, and in part by the Scientific Research Project of Liaoning Educational Department under Grant LQGD2017023.

ABSTRACT The complex environment background, lighting conditions, and other action-irrelevant visual information in the video frame bring a lot of redundancy and noise to the action spatial features, which seriously affects the accuracy of action recognition. Aiming at this point, we propose a recurrent region attention cell to capture the action-relevant regional visual information in the spatial feature, and according to the temporal sequential natures of the video, on the basis of the recurrent region attention cell, a Recurrent Region Attention model (RRA) is proposed. The recurrent region attention cell in the RRA iterates according to the temporal sequence of the video, so that the attention performance of the RRA is gradually improved. Secondly, we propose a Video Frame Attention model (VFA) that can highlight the more important frames in the whole action video sequence, so as to reduce the interference caused by the similarity between the heterogeneous action video sequences. Finally, we propose an end-to-end trainable network: Two-level Attention Model based video action recognition network (TAMNet). We experimented on two video action recognition benchmark datasets: UCF101 and HMDB51. Experiments show that our end-to-end TAMNet network can reliably focus on the more important video frames in the video sequence, and effectively capture the action-relevant regional visual information in the spatial features of each frame of the video sequence. Inspired by the two-stream structure, we construct a two-modalities TAMNet network. In the same training conditions, the two-modalities TAMNet network achieved optimal performance on both datasets.

INDEX TERMS Action recognition, LSTM, recurrent region attention, video frame attention.

I. INTRODUCTION

Video action recognition has always been a research hotspot in the field of computer vision, with the goal of analyzing the action which is ongoing in an unknown video or image sequence. Identifying the action in a video is one of the basic abilities of human beings. Humans can combine the action-relevant spatial images in the video and the context between the images to identify and infer the category to which the action belongs.

Convolutional neural networks can learn the discriminative spatial representation of raw visual data with the help of large-scale supervised datasets. In recent years, with its excellent modeling capabilities, it has achieved great success in the recognition and classification tasks in the field of still images [1]–[6], [42], and gradually introduced into the video field to solve video-based action recognition problem [7], [8], [10], [11]. However, the convolutional neural networks focus on local patterns, and its improvement in recognition accuracy for action video with temporal sequential natures is not as remarkable as image recognition. The long-term structure plays an important role in learning the continuity of action videos. Therefore, recurrent neural networks, especially LSTM networks, can be invoked to better capture long-range temporal patterns and contexts.

Combining the convolutional neural network with the LSTM network [15], the end-to-end joint training directly on the dataset can better learn the spatio-temporal information of the action video sequence. However, as shown in Figure 1, the environment where the action in video frame located in is complex, and the proportion of the action subject, lighting conditions change frequently, which brings redundancy

The associate editor coordinating the review of this article and approving it for publication was Jingchang Huang.

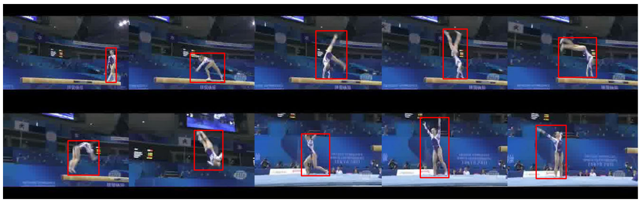


FIGURE 1. The environment where the action in video frame located in is complex, and the proportion of the action subject, lighting conditions change frequently.

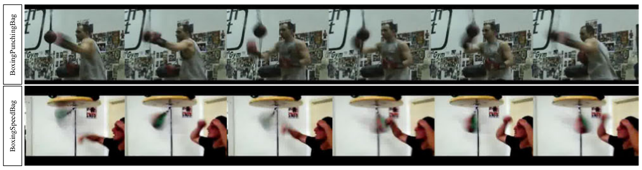


FIGURE 2. Different classes of action videos may have similar context in temporal sequence.

and noise to the spatial information. Secondly, as shown in Figure 2, different classes of action videos may have similar context in temporal sequence, making LSTM network prediction errors.

In order to address these challenges, in this paper, we have made the following contributions:

(1) We propose a recurrent region attention cell, which can effectively capture the action-relevant regional visual information in the spatial features of video frames in order to reduce the interference of redundant information and noise information on action spatial features. Then, according to the temporal sequential natures of the video, we propose a Recurrent Region Attention model (RRA) based on the recurrent region attention cell. The recurrent region attention cell in RRA iterates according to the temporal sequence of the video, so that the RRA can effectively capture the action-relevant regional visual information in the spatial features of each frame of the action video sequence.

(2) We propose a Video Frame Attention model (VFA) to highlight the more important frames in the entire video sequence to reduce the interference caused by the similarity between the heterogeneous action video sequences.

(3) We propose a Two-level Attention Model based video action recognition network (TAMNet) which can be end-to-end trained.

Sparse Temporal Sampling strategy [11] was adopted to obtain a subset of action video sequence as input to the TAMNet network, enabling TAMNet to model the long-range temporal pattern of the entire video sequence.

Inspired by the two-stream architecture, we fuse the video-level prediction of the RGB modality TAMNet with the video-level prediction of the optical flow modality TAMNet to produce the final action category prediction. The structure of the two-modalities TAMNet network is shown in Figure 3.

The remainder of the paper is organized as follows. In section 2, the popular video action recognition methods and the attention mechanisms used in action recognition and other fields are introduced. In section 3, we elaborate on the details of our proposed Two-level Attention Model based video action recognition network (TAMNet), and introduce our two-modalities fusion method. In Section 4, we discuss experiments that validate the effectiveness of our proposed method. In Section 5, we first summarize the work, then provide constructive comments and suggestions for future work.

II. RELATED WORKS

With the development of deep learning algorithms in recent years, convolutional neural networks have excellent feature extraction ability for unstructured data, and has achieved remarkable success in image recognition and classification [1]–[6], [42], it is also gradually being used in video action recognition tasks.

A. VIDEO ACTION RECOGNITION

Early video action recognition technology directly applied 2D CNN to RGB video frames. Karpathy *et al.* [7] divided the video into fixed-length segments, and designed several temporal sampling methods, including single-fusion, late-fusion, early-fusion, and slow-fusion, to pooling local spatio-temporal information of RGB video frame to expand the connectivity of CNN in temporal dimension. However, this method does not bring a remarkable promotion in recognition performance compared to using only the single frame method.

In order to overcome the shortcomings of 2D CNN in the temporal dimension, Simonyan and Zisserman [8] proposed a Two Stream Network. The spatial stream 2D CNN is applied to process the RGB video frames to obtain the spatial information of the action, the temporal stream 2D CNN processes the stacked optical flow field composed of consecutive multiple frames of dense optical flow graph [9] to obtain motion information, then fuse the probability scores of the two networks. The final recognition result is obtained using the method of averaging or SVM (the SVM is more accurate in the experiment). They show that even if the probability scores of the two networks are simply combined, the accuracy of action recognition is significantly improved, indicating that the optical flow provides high quality motion information.

On the basis of [8], Feichtenhofer *et al.* [10] studied the convolutional fusion method on spatial stream and temporal stream, and proposed a new convolutional neural network structure to make better use of spatio-temporal information. They show that the spatial stream network and the temporal stream network are fused into the last convolutional layer, and the abstract convolutional features are pooled in the spatio-temporal neighborhood, which not only reduces the number of the network parameters, but also further improves the performance of the network.

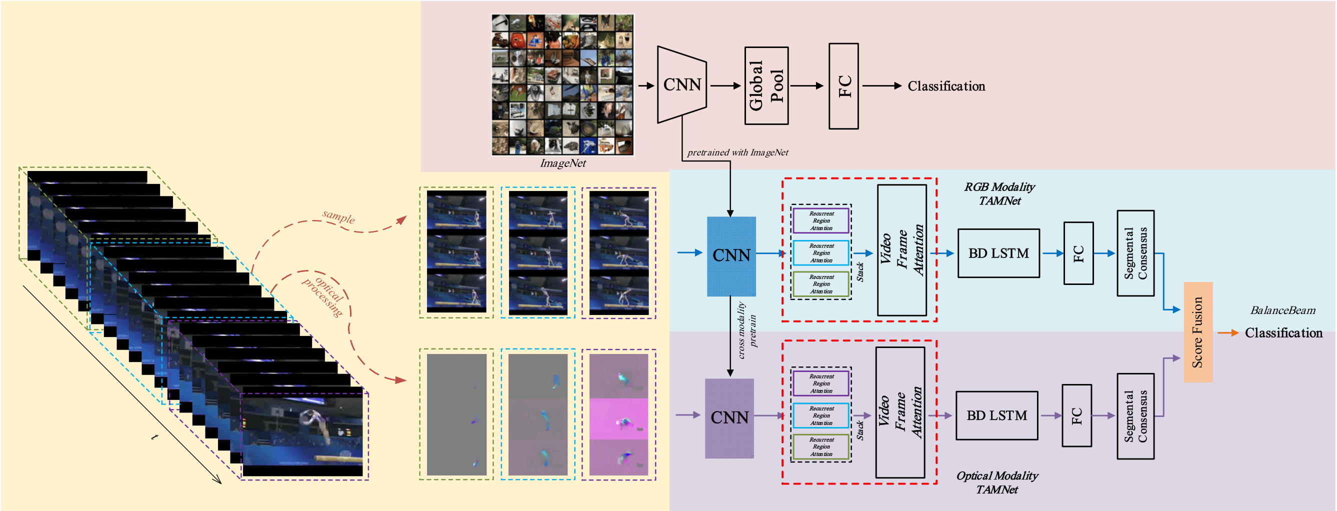


FIGURE 3. **Overview of two-modalities TAMNet network. The red dotted coil in the right half is the two-level attention model proposed by us. First, we use the Temporal Segment strategy to segment action video sequence, and multiple consecutive frames are sampled in each snippet that obtained after the Temporal Segment, to acquire a segmental-input sequence for each snippet. The set of all segmental-input sequences is the input sequence of the single-modality TAMNet network. The following operations are performed on the input sequence of each modality: Utilizing the convolutional neural network to obtain the spatial-feature sequence of each segmental-input sequence, and then our proposed Recurrent Region Attention model is used to capture the action-relevant regional visual information in each segmental-spatial-feature sequence. All of the segmental-spatial-feature sequences generated by the RRA are stacked and fed into the Video Frame Attention model. Feeding the action spatial-feature sequence generated by the VFA into the Bidirectional LSTM, making prediction at each moment with Bidirectional LSTM. Utilizing the Segmental Consensus Function to get the video-level prediction. Finally, the video-level predictions of the two modalities were fused to obtain the final prediction about the action category.**

In recent years, many methods based on two-stream 2D CNN have been proposed to improve the accuracy of video action recognition. However, these methods have limited access to the temporal context of video sequence, mainly because the spatial stream 2D CNN is only applied to single-frame RGB video frame, and the temporal stream 2D CNN is only applied to single-stack optical flow stack, ignoring the temporal sequential nature of video.

In response to the above problems, Wang *et al.* [11] proposed a Temporal Segment Network (TSN) based on long-range temporal structure and Sparse Temporal Sampling strategy. The TSN is also composed of a spatial stream network and a temporal stream network. But unlike previous input forms for two-stream networks [8], [10], the TSN performs Sparse Temporal Sampling strategy on the entire video sequence, and takes the sampled video snippets as inputs to the network. Each snippet will get its preliminary prediction about the action class through the network, and then get the video-level prediction of the whole video through the Segmental Consensus Function.

The input of the temporal stream network is the optical flow features obtained from the original action videos. The optical flow features provide high quality motion information, and the introduction of optical flow features brings a remarkable improvement in the accuracy of action recognition. Another way to obtain motion information is the C3D network proposed by Tran *et al.* [12]. The spatial features and temporal features of the video sequence are extracted by the 3D convolutional kernel to compensate for the lack of 2D CNN in the temporal dimension. This network can generate multi-channel information from consecutive video frames, and then perform convolution operation and down-sampling operation for each channel separately. Finally, all channel information is combined to obtain a final feature description. Instead of repeating the process for spatio-temporal models, Carreira and Zisserman [29] inflate all the filters and pooling kernels of 2D ConvNets, endow them with an additional temporal dimension. Designed an I3D network by inflating 2D ConvNets into 3D ConvNets. Wang *et al.* [40] believe convolutional and recurrent operations can only process one local neighborhood at a time. Inspired by the classical non-local means method in computer vision, they present a non-local operation as a generic family of building blocks that can capture long-range dependencies and compute the response at a position as a weighted sum of the features at all positions. Tran *et al.* [41] empirically demonstrate the accuracy advantages of 3D CNNs over 2D CNNs within the framework of residual learning and design a new spatio-temporal R(2+1)D convolutional block, this block divide the 3D convolution into 2D spatial convolution and 1D temporal convolution. Compared with 3D convolutional networks with the same structural parameters, R(2+1)D can achieve lower training error, which is easier to optimize.

The method of using convolutional neural networks to recognize action in video is mainly focused on the short-term pattern of the video, and it is difficult to directly capture the long-term pattern of the video. Recurrent neural networks, especially the LSTM [13] networks, are considered to be effective models for processing long-term sequence data. Ng *et al.* [14] designed several feature fusion methods to fuse the convolutional features of each frame in the video, and then used 5-layer LSTMs to extract the depth features of the video spatial features obtained after the fusion, finally, made predictions at each moment. In combination with CNN and LSTM, Donahue *et al.* [15] proposed an LRCN model capable of end-to-end training, using RGB modality and optical flow [16] modality as input to the LRCN. Finally, the outputs of RGB modality LRCN and optical flow modality LRCN are averaged to obtain the final classification result.

However, the method of simply combining convolutional neural network and LSTM network ignores the spatial redundancy and noise caused by action-irrelevant visual information, such as environment backgrounds, lighting conditions, etc. It also neglects the interference caused by the similar context of heterogeneous action videos. In view of these problems, we can focus on the spatial information and temporal information of action video sequence.

B. ATTENTION MECHANISM

When humans view visual images, the visual system does not process the entire image at the same time, but by quickly scanning the global image to obtain the target region which needs to be focused on. Then invest more attention resources in this region to get more details of the desired target and to suppress the impact of other useless information on current target. This visual attention mechanism of humans greatly improves the efficiency and accuracy of visual information processing.

Inspired by human visual attention mechanism, Xu *et al.*[17] introduced a soft attention mechanism in image caption. This soft attention mechanism is then applied to the video analysis tasks. Sharma *et al.* [18] proposed a soft attention LSTM model based on multi-layer recurrent neural networks, which selectively focuses on some of the video frames in the video sequence to improve the ability of the model to identify action in the video.

Liu *et al.* [19] proposed a Global Context-Aware Attention LSTM network. They obtain the initial global context memory by averaging the hidden representations of all spatio-temporal steps in the first LSTM layer. Then, the informativeness gate (score) for each spatio-temporal input is calculated by the input of each spatio-temporal step and the global context memory generated by the previous attention. The hidden state of the spatio-temporal LSTM unit in the second layer is updated by using the learned informativeness score. Finally, the output of the last spatio-temporal step in the second spatio-temporal LSTM layer is used to refine the global context memory. Through multiple iterations, the global context makes the classification more discriminative.

Yan *et al.* [20] proposed a novel Hierarchical Multi-scale Attention Network by combining Hierarchical Multi-scale RNN and attention mechanism. They used the newly proposed gradient estimation method for stochastic neurons, namely Gumbel-softmax, to implement temporal boundary detectors and the stochastic hard attention mechanism.

Wang *et al.* [21] proposed a Hierarchical Attention Network. They consider that although optical flow features and RGB features capture different aspects of a video frame, the attention location on the video is same. Moreover, the RGB features and optical flow features provide complementary information to each other, making predictions more accurate. Therefore, at each moment, they combine the hidden state of the RGB modality LSTM with the hidden state of the optical flow modality LSTM as an input to the hierarchical attention mechanism.

Yu *et al.* [22] proposed a novel high-level action representation using the joint spatial-temporal attention model. In spatial, inspired by ResNet, they built spatial convolution (2D) branch to obtain spatial attention guidance. Then, considering the temporal coherence in the short video clip, an extra temporal convolution (1D) branch is constructed. The two branches are integrated into a spatial-temporal unit, and a spatial attention gate is obtained by the softmax function. Finally, a two-level global attention branch is applied to get a better spatial attention guide. In temporal, they use a bidirectional LSTM to build a temporal attention model. Then use the sigmoid and softmax functions to convert the hidden state of the bidirectional LSTM into a temporal attention score.

However, these attention models are highly integrated with the recurrent neural networks, and the computational process is complicated, which brings expensive computational cost to the training process of the network. In order to avoid the heavy computational burden of the training process of the network, we respectively propose one novel, simple, and effective attention model for the spatial information and temporal information of the action video sequence.

III. PROPOSED METHOD

In this section, we describe our proposed Two-level Attention Model based video action recognition network (TAMNet) in detail, and introduce our two-modalities fusion method. Figure 4 shows the structure of our TAMNet model.

Our proposed TAMNet consists mainly of three parts: convolutional neural network, two-level attention, and Bidirectional LSTM. First, we use the Temporal Segment strategy to segment action video sequence, and multiple consecutive frames are sampled in each snippet that obtained after the Temporal Segment, to acquire a segmental-input sequence for each snippet. The set of all segmental-input sequences is the input sequence of the TAMNet network. Secondly, we utilize the convolutional neural network to obtain the spatial-feature sequence of each segmental-input sequence, and then our proposed Recurrent Region Attention model (RRA) is used to capture the action-relevant regional visual information in each segmental-spatial-feature sequence. All of the segmental-spatial-feature sequences generated by the RRA are stacked and fed into the Video Frame Attention model (VFA). Finally, we feed the action spatial-feature sequence generated by the VFA into the Bidirectional LSTM, making prediction at each moment with Bidirectional LSTM. Utilizing the Segmental Consensus Function to get the video-level prediction of action

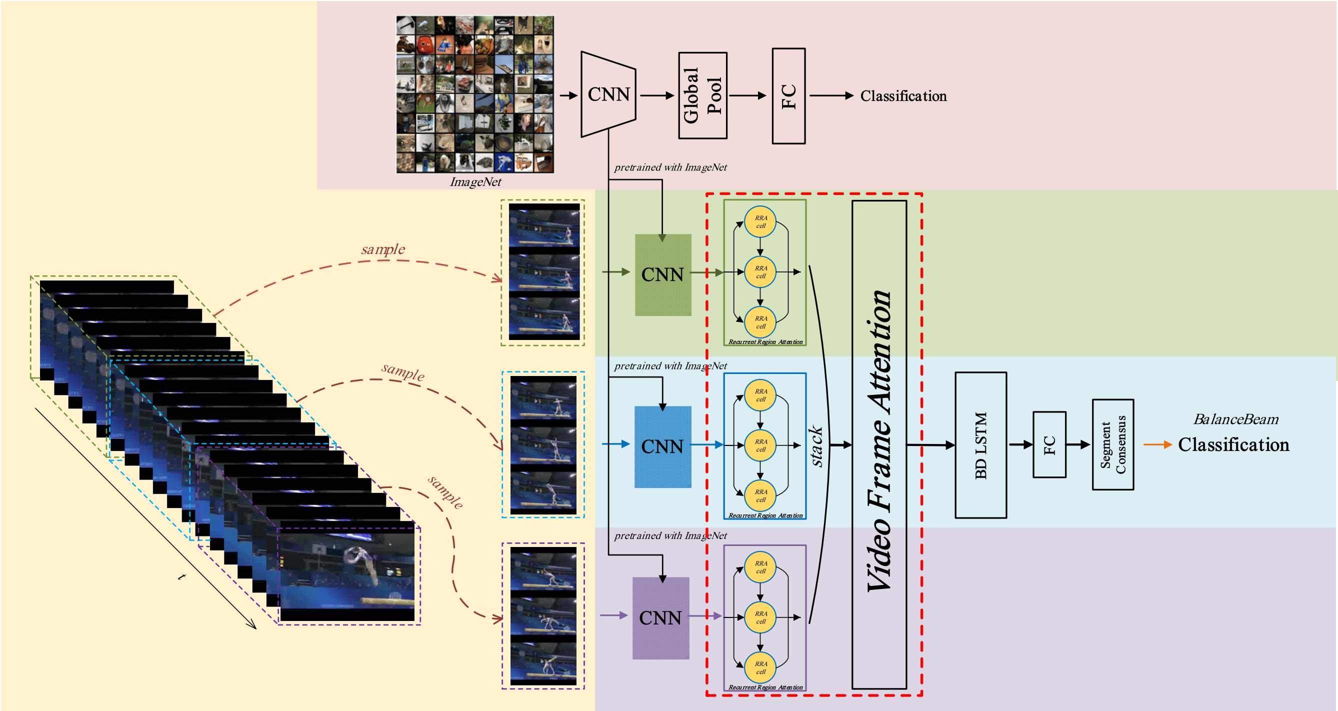


FIGURE 4. The structure diagram of TAMNet.

video sequence. Next, we will explain in detail each component of our proposed TAMNet network model in turn.

A. NETWORK INPUT

Assuming that a video sequence V has a total of T frames, we represent the video sequence as:

$$V = \{F_1, F_2, \ldots, F_t, \ldots, F_T\}, \quad t \in T$$

Then, using the Temporal Segment idea proposed by Wang *et al.* [11], the entire action video sequence is equally divided into I segments, and l consecutive video frames are randomly sampled in each snippet, and the sampled video sequence subset is used as the input to the TAMNet network. The input to the network is expressed as:

$$\begin{aligned} v = \{&P_1, P_2, \ldots, P_i, \ldots, P_l, \\ &P_{l+1}, P_{l+2}, \ldots, P_{l+i}, \ldots, P_{2l}, \ldots, P_{I\times l}\}, \\ v \in V, &\quad i \in l,\ I \times l \in T \end{aligned}$$

B. RECURRENT REGION ATTENTION MODEL

We first proposed a recurrent region attention cell, which can capture the action-relevant regional visual information in the spatial feature of the video frame, thereby reducing the influence of redundant information and noise information on the action spatial feature. It is worth noting that the internal structure of recurrent region attention cell is its characteristic. The output of this recurrent region attention cell is not only passed to the neural cells of the next layer, but also passed back to the recurrent region attention cell of this layer. This way of passing constitutes the loop of data. Figure 5 shows in detail the internal structure of the recurrent region attention cell and the flowing way of the data.

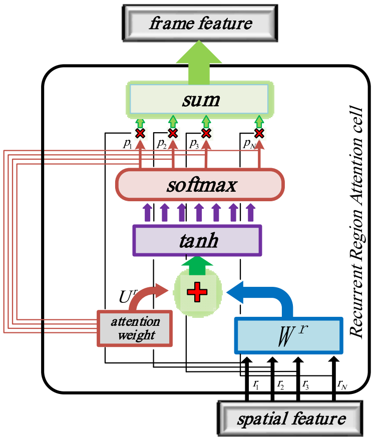


FIGURE 5. The internal structure diagram of the recurrent region attention cell.

Inspired by LSTM network, according to the temporal sequential natures of the video, on the basis of the recurrent region attention cell, we propose a Recurrent Region Attention model (RRA). It should be noted that different from the common existing spatial attention models at present, the RRA model proposed by us takes into account the temporal sequential characteristic of video and has temporal sequential nature. Therefore, it is possible to better capture action-relevant regional visual information through the temporal sequential affiliation between video frames. The recurrent region attention cell in the RRA iterates according to the temporal sequence of the video (the iterative approach of recurrent

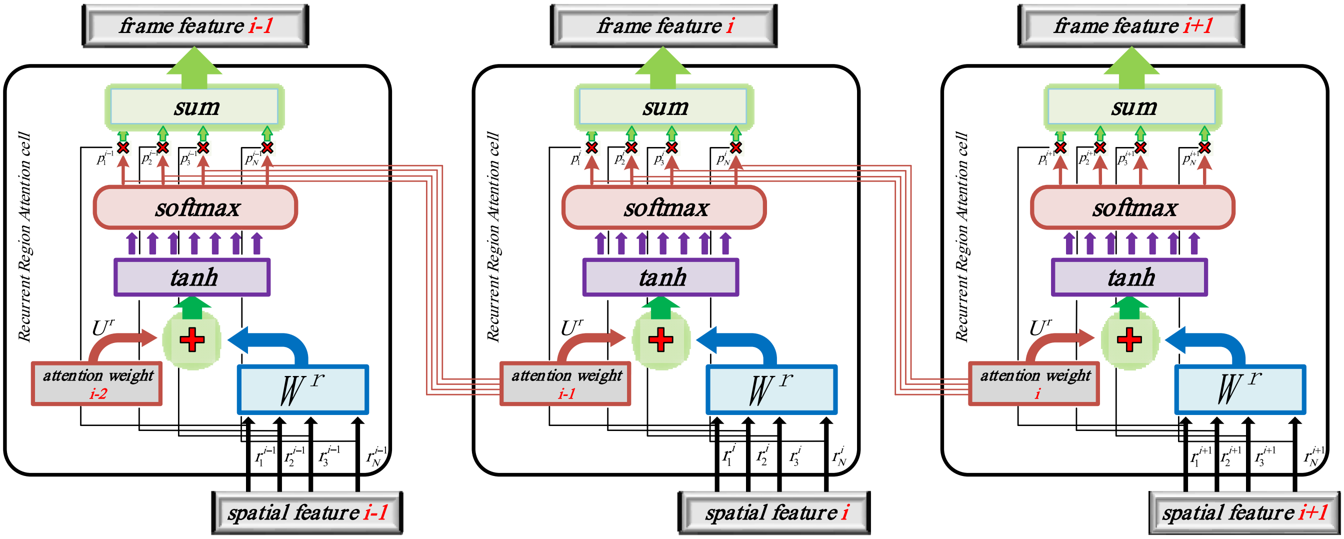

frame feature i-1
frame feature i
frame feature i+1
sum
softmax
tanh
Recurrent Region Attention cell
U^r
W^r
attention weight i-2
attention weight i-1
attention weight i
p_1^{i-1} p_2^{i-1} p_3^{i-1} p_N^{i-1}
p_1^{i} p_2^{i} p_3^{i} p_N^{i}
p_1^{i+1} p_2^{i+1} p_3^{i+1} p_N^{i+1}
r_1^{i-1} r_2^{i-1} r_3^{i-1} r_N^{i-1}
r_1^{i} r_2^{i} r_3^{i} r_N^{i}
r_1^{i+1} r_2^{i+1} r_3^{i+1} r_N^{i+1}
spatial feature i-1
spatial feature i
spatial feature i+1

FIGURE 6. The RRA structure diagram that unfolded according to the input sequence length. The spatial feature of the current frame (current moment) and the region attention weights of the previous frame (previous moment) are used as the input of the recurrent region attention cell of the current moment. The probabilities obtained by *softmax* function inside the recurrent region attention cell are the attention weights of region features of current frame. The weighted sum of the region features is the spatial feature that highlights the action visual information.

region attention cell is similar to the LSTM cell), so that the attention performance of the RRA on the action-relevant regional visual information is gradually improved. Different from the soft attention mechanisms [19], [21], [23], [35] used in action recognition, machine translation, and image caption, which takes the hidden state of the LSTM networks as one of inputs and integrates with the LSTM cell, and also different from the spatial attention mechanism of [39], which only uses the spatial feature of image as input, takes the convolutional layer with 1 × 1 kernel as calculate operation and does not consider temporal affiliation between video frames, however, the RRA proposed by us is a neural network model which takes into account the temporal sequential characteristic of video, has temporal sequential nature, and takes the spatial feature of the current frame (current moment) and the region attention weights of the previous frame (previous moment) as input, only relies on the recurrent region attention cell to form a directed annular connection between neural cells in the layer. The RRA does not belong to the feedforward neural network, mainly because the output of the recurrent region attention cell in the RRA is transmitted in both the depth direction of network and the temporal direction. In order to clearly express the recurrent characteristic of the RRA, in Figure 6, we unfold it according to the length of the input sequence. When unfolded, the RRA can be viewed as the array composed of the recurrent region attention cells. In this array, the recurrent region attention cells of the previous moment (previous frame) and the next moment (next frame) are connected to each other.

The spatial feature f_i of a certain video frame P_i in the video sequence subset v is extracted by using the convolutional neural network, the shape of which is $[height, width, channel]$. Where *height* represents the height of the spatial feature f_i, *width* represents the width of the spatial feature f_i, and *channel* represents the number of channels of the spatial feature f_i.

For a certain frame P_i in the video sequence subset v, we can get a spatial region feature set $\left\{r_1^i, r_2^i, \ldots, r_N^i\right\}$ for this frame, where N represents the total number of regions of the spatial feature of the video frame P_i. N can be calculated by the follow:

$$N = height \times width \tag{1}$$

Weighted summing these region features with the region attention weights, the spatial feature of the video frame generated by RRA can be obtained by the follow:

$$f_i^r = \sum_{j=1}^{N} p_j^i r_j^i \tag{2}$$

Then, the spatial feature sequence of the video sequence subset v is represented as $v^r = \left\{f_1^r, f_2^r, \ldots, f_i^r, \ldots, f_{I \times l}^r\right\}$, where r_j^i represents the j-th region feature of the i-th frame, p_j^i is the attention weight of the j-th region in the spatial feature of the i-th frame corresponding to r_j^i, f_i^r represents the spatial feature of video frame after capturing the action-relevant regional visual information through the RRA.p_j^i is calculated by the follows:

$$u_j^i = \omega^r \tanh\left(W^r r_j^i + U^r p_j^{i-1}\right) + b^r \tag{3}$$

$$p_j^i = \frac{\exp\left(u_j^i\right)}{\sum_{j=1}^{N} \exp\left(u_j^i\right)} \tag{4}$$

where W^r, U^r, ω^r, and b^r are shared parameters learned by the RRA. The action-relevant regional visual information in the *i*-th frame is captured by p_j^i. In this way, the interference brought by action-irrelevant visual information such as the complex environment background and lighting conditions can be reduced. p_j^{i-1} represents the region attention weight of the *j*-th region of the previous frame. Therefore, the region attention weights of the *i*-th frame are determined by the spatial feature of current frame and the region attention weights of previous frame. We hope the RRA proposed by us can reduce the impacts of spatial noisy and spatial redundancy on the accuracy of recognition. At the same time, the feature ambiguity caused by the spatial similarity between frames and frames can be solved, so that the spatial feature has the ability to correctly represent action. Algorithm 1 shows the working process of our RRA.

Algorithm 1 The Working Process of Our RRA

```
Input: Segmental spatial feature sequence $\{f_1, f_2, \ldots, f_i, \ldots, f_l\}$
Output: Segmental spatial feature sequence $\{f_1^r, f_2^r, \ldots, f_i^r, \ldots, f_l^r\}$ after capturing action-relevant regional visual information through RRA
  (1) Randomly initialize the region attention weights $p^0 = \{p_0^0, p_1^0, \ldots, p_N^0\}$
  (2) for $i = 1,2,\ldots,l$
  (3)   Calculate the region attention weights $p^i = \{p_0^i, p_1^i, \ldots, p_N^i\}$ of the $i$-th frame using equations (3)-(4)
  (4)   Weighted summing the spatial region features of the $i$-th frame with attention weights $p^i$ using equation (2), to obtain the spatial feature $f_i^r$ generated by recurrent region attention cell.
  (5)   $i$+1, repeat step (3) and step (4)
  (6) endfor
```

C. BIDIRECTIONAL RECURRENT NEURAL NETWORK

The architecture of TAMNet build upon the LSTM [13]. Figure 7 shows the internal structure of the LSTM cell. The output of the LSTM cell contains a memory vector. This memory vector represents the new memory obtained at the current moment by the LSTM cell after synthesizing the memory of the previous moment and the input of the current moment. Even if the length of the input sequence is long, the LSTM cell can always transmit the information of the past moments, avoiding the loss of the sequence information and solving the long-term dependency problem.

The input of the LSTM cell is the state vector h_{t-1} of the hidden layer of the previous moment, the memory vector c_{t-1} obtained at the previous moment, and the input vector x_t of the current moment. The output of the LSTM cell is the state vector h_t of the hidden layer of the current moment and the memory vector c_t of the current moment.

Given a feature sequence $\{x_1, x_2, \ldots, x_i, \ldots, x_{I\times l}\}$ as input to the LSTM cell, then the corresponding hidden states $\{h_1, h_2, \ldots, h_i, \ldots, h_{I\times l}\}$ can be obtained by repeating the follows:

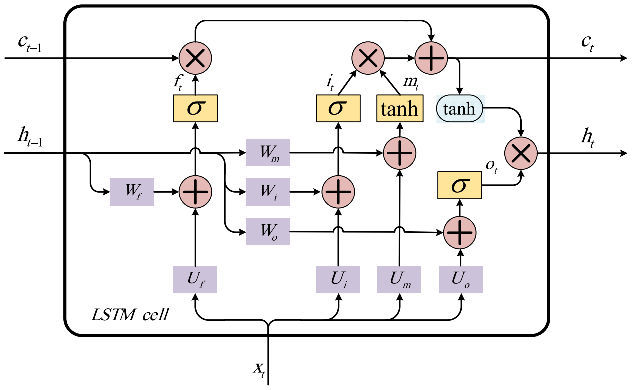


FIGURE 7. The internal structure diagram of the LSTM cell.

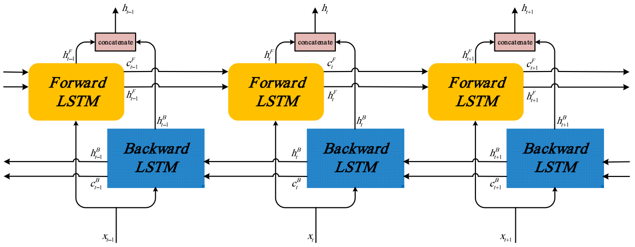


FIGURE 8. The Bidirectional LSTM structure diagram that unfolded according to the temporal.

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{5}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{6}$$

$$m_t = \tanh(W_m h_{t-1} + U_m x_t + b_m) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot m_t \tag{8}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{9}$$

$$h_t = o_t \odot \tanh(c_t) \tag{10}$$

where *Ws*, *Us*, and *bs* are shared parameters learned by the network, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes elementwise multiplication.

The action in the video sequence is coherent, so it is not rigorous to make prediction at the current moment based only on action information of the past moments. The LSTM cell only considers the effect of the inputs from the past moments on the output of the current moment, however the output we need is dependent on the entire video sequence. In order to consider the impact of future action information on the output of the current moment, we use the idea of Bidirectional RNN [24] to combine a LSTM cell that moves from the beginning of the sequence (Forward LSTM) with another LSTM cell that moves from the end of the sequence (Backward LSTM) to form a Bidirectional LSTM, making full use of the front and back dependencies in the video sequence. Figure 8 shows a Bidirectional LSTM structure that unfolded according to the temporal.

The input vector x_t of the current moment is input to the Forward LSTM and the Backward LSTM respectively, and

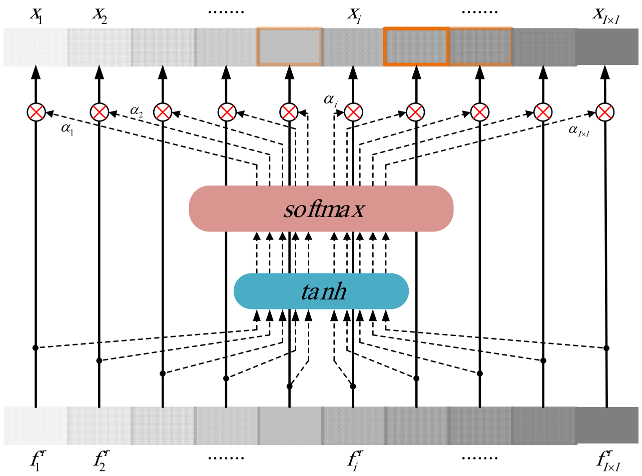


FIGURE 9. The structure diagram of VFA.

correspondingly generates a forward hidden state vector h_t^F and a backward hidden state vector h_t^B of the current time, then the output of the Bidirectional LSTM is represented as $h_t = [h_t^F, h_t^B]$, $[\cdot]$ means concatenating two hidden state vectors.

D. VIDEO FRAME ATTENTION MODEL

Although the Bidirectional LSTM can fully learn the contextual correlation of a certain class action video sequence, there may be a majority of consecutive similar frames in certain parts of the temporal sequence between heterogeneous action videos, this similar correlation in temporal will cause errors in predictions of the Bidirectional LSTM. Therefore, we propose a Video Frame Attention model (VFA) that highlights the more important frames in the whole video sequence to reduce the interference caused by the similar contexts between the different classes of action video sequences. It is worth noting that different from the temporal attention mechanism of [39], the temporal attention of [39] combined the hidden state of the LSTM cell, built upon the LSTM cell, and integrated with the LSTM cell, the attention mechanism of this structure such as [19], [21], [23], [35], [39] brings a lot of computation to the network and slows down the speed of the network, however, our VFA only takes the stacked output of RRA as input and separates from the LSTM cell. The VFA proposed by us not only reduces the amount of calculation, but also speeds up the calculation of the network and also improves the performance of the network. The data transfer process between RRA and VFA can be viewed in Figure 4. And the structure of Video Frame Attention model is shown in Figure 9.

Given the spatial feature sequence $v^r = \{f_1^r, f_2^r, \ldots, f_i^r, \ldots, f_{I\times l}^r\}$ of the video sequence subset v to the VFA, then the spatial feature sequence $x = \{x_1, x_2, \ldots, x_i, \ldots, x_{I\times l}\}$ generated by the VFA can be obtained by the follow:

$$x = \alpha^T v^r \tag{11}$$

where $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_{I\times l}\}$ is the frame weights of the spatial feature sequence v^r of the video sequence subset v, can be obtained by the follows:

$$q = u^{\mathrm{T}} \tanh(W^v v^r + b^v) \tag{12}$$

$$\alpha = soft\max(q) \tag{13}$$

where W^v, b^v, and u are shared parameters learned by the VFA.

We hope that the VFA can reduce the interference brought by the similar contexts of heterogeneous action video sequences in temporal. Algorithm 2 shows the working process of our VFA.

Algorithm 2 The Working Process of Our VFA

```
Input: Spatial feature sequence $\{f_1^r, f_2^r, \ldots, f_i^r, \ldots, f_{I\times l}^r\}$ for a video sequence subset
Output: Spatial feature sequence $\{x_1, x_2, \ldots, x_i, \ldots, x_{I\times l}\}$ obtained by VFA
  (1) Calculate the frame weights $\{\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_{I\times l}\}$ of the spatial feature
      sequence $\{f_1^r, f_2^r, \ldots, f_i^r, \ldots, f_{I\times l}^r\}$ using equations (12) and (13)
  (2) Calculate the action spatial feature sequence $\{x_1, x_2, \ldots, x_i, \ldots, x_{I\times l}\}$ generated by VFA using
      equation (11)
```

E. VIDEO LEVEL PREDICTION

Different from the past methods which used the output of the last moment of the LSTM for classification, we first pass the output of each moment of the Bidirectional LSTM to the fully connected layer to generate preliminary prediction of each moment about the action category. Then continue the Segmental Consensus Function proposed by Wang *et al.* [11], so that the preliminary prediction of each moment in the sequence can reach a consensus and generate a video-level prediction result. The Segmental Consensus Function is defined as follow:

$$\widehat{Y} = \text{Consensus}\left(\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_{I\times l}\right) \tag{14}$$

F. PARAMETERS UPDATE

Assume that there are K video sequence samples in the training set, we denote the authentic class label of the i-th video sequence sample as Y_i and its video-level prediction is expressed as $\widehat{Y}_i$. We can update the network parameters by minimizing the cost function (15):

$$J = \frac{1}{K}\sum_{i=1}^{K}\left[-Y_i^{\mathrm{T}}\log(\widehat{Y}_i) - (1 - Y_i)^{\mathrm{T}}\log(1 - \widehat{Y}_i)\right] \tag{15}$$

G. MODALITIES FUSION

Previously, we described the structure of single-modality TAMNet in detail. In order to make full use of the spatio-temporal information and motion information provided by the action video sequence, we consider fusing the video-level predictions of the RGB modality TAMNet and the optical flow modality TAMNet.

Since our TAMNet is an end-to-end network, we chose probability fusion method to fuse the video-level predictions of different modality TAMNet networks.

Probability fusion method is essentially an integrated approach. We first train a separate model for each modality input sequence. Then we perform probability fusion to fuse the video-level predictions of different modality models. We denote the video-level prediction of RGB modality TAMNet as $\widehat{Y}_{RGB}$ and the video-level prediction of optical flow modality TAMNet as $\widehat{Y}_{optical}$. Then the final prediction generated by the probability fusion method can be obtained by (16). Where λ represents the fusion weight of the RGB modality TAMNet video-level prediction $\widehat{Y}_{RGB}$.The specific probability fusion will be detailed in the experiment in Section 4.

$$\widehat{Y}_{Fusion} = \lambda * \widehat{Y}_{RGB} + (1 - \lambda) * \widehat{Y}_{optical} \tag{16}$$

IV. EXPERIMENTS

In this section, we first introduce the evaluation dataset and the implementation details of the proposed method, and then verify the effectiveness of our proposed two-level attention on improving the recognition accuracy and the recognition performance of the TAMNet network. Finally, the performance of the two-modalities TAMNet network is evaluated and compared with the SOTA methods.

A. DATASETS

We evaluated and compared the performance of our proposed TAMNet network on two popular video action recognition datasets.

UCF101[25] is a video action dataset with 101 categories collected from YouTube. It contains 13320 videos, and the average length of each video is 180 frames. UCF101 provides the greatest diversity in action categories, including daily activities in life and even extreme sports. And there is camera motion in the video, and the change of object scale, environment backgrounds, and the lighting conditions is also large, so UCF101 is a challenging dataset. Each video comes with a class label. We will report the average accuracy of the three train/test splits according to the original evaluation scheme of the dataset.

HMDB51 [26] is a video action dataset consisting of 6676 videos from various sources (such as movies and YouTube videos) with 51 action categories. We will follow the suggested evaluation scheme and report the average accuracy over the three train/test splits.

B. EXPERIMENTAL ENVIRONMENT

We used the Tensorflow framework to complete this work on a computer with two NVIDIA RTX2080Ti GPUs and 32G RAM.

C. EXPERIMENTAL DETAILS

Take the experiment of the UCF101 dataset as an example. In the experiment, different from the previous input-form of two-stream networks, we first convert the optical flow graphs [16] into a video file through OpenCV, and then convert the original video files and optical flow video files of UCF101 into a binary TFRecord format through Tensorflow and OpenCV. Due to the binary format of files, the training speed is greatly accelerated and the memory usage of the computer is reduced. The video sequence of each modality is sampled by the Sparse Temporal Sampling strategy [11], and the sampled video subsequence is used as the input of the single-modality TAMNet network. We use the convolutional neural network as feature extractor to extract the spatial features of the video frames and use the Momentum optimization algorithm to optimize the network parameters. The Batchsize is set to 32, the momentum is set to 0.9, the network input dimension is set to 15, and the snippet sequence length is set to 5. That means the video sequence is equally divided into 3 segments, and then 5 consecutive video frames are sampled in each snippet. The final video-level prediction of the RGB modality TAMNet or the optical flow modality TAMNet is obtained using the average Segmental Consensus Function [11].

For the RGB modality TAMNet network, we use the pre-trained model from ImageNet [27] to initialize the weights of the convolutional neural network. The initial learning rate is set to 0.001, and then judge whether the learning rate is reduced to its 1/10 according to the absolute value of the difference between the average loss of the previous epochs and the current epoch loss. The number of hidden units in the VFA is set to 256, the number of hidden units in the LSTM is set to 256, and the entire training process is stopped in 120 epochs.

For the optical flow modality TAMNet network, since the distribution of optical flow field is different from the RGB images, we use the linear transformation to discretize the optical flow field into the same 0-255 interval as the RGB images. Then use the Cross Modality Pretraining method to initialize the weights of the convolutional neural network in the optical flow modality TAMNet by utilizing the convolutional neural network in the RGB modality TAMNet, which not only reduces the training duration of the optical flow modality, but also appropriately avoid overfitting. The number of hidden units in the VFA is set to 256, the number of hidden units in LSTM is set to 256.The learning rate is initialized to 0.005, the learning rate attenuation strategy is consistent with the RGB modality TAMNet, and the maximum iteration is set to 250 epochs.

In order to avoid overfitting caused by the complexity of model during training, we use technology such as random cropping, horizontal flipping, corner cropping [11], and scale jittering [2] to augment data and add a Dropout layer after the bidirectional LSTM. The dropout rate of RGB modality TAMNet is set to 0.5, and the dropout rate of optical flow modality TAMNet is set to 0.7.

To speed up the training, we used a multi-GPU parallel strategy. The training time of the UCF101 dataset on the RGB modality TAMNet is about 50 hours, and the training time on the optical flow modality TAMNet is about 90 hours.

D. PERFORMANCE EVALUATION

1) TAMNET PERFORMANCE EVALUATION

In this section, we first make an experiment to choose which convolutional neural network can be used as a feature extractor for TAMNet, and then verify the effectiveness of our proposed RRA and VFA on improving network recognition performance, finally we verify the recognition performance of single-modality TAMNet network.

Since our RRA focuses on the spatial features extracted by the convolutional neural network (feature extractor), whether the spatial features extracted by the convolutional neural network have strong spatial representation ability directly affects whether RRA can be fully played its role. Therefore, we will experiment with which convolutional neural network can be chosen as the feature extractor for TAMNet. We use VGG16, ResNet50, and BN Inception as candidate feature extractor for TAMNet because the input sizes of these convolutional neural networks are 224 × 224, comparing the feature extraction ability of these convolutional neural networks with the same input size is more convincing. Secondly, we only use the RGB modality of UCF101 as input, and use the settings in the above experimental details to group experiments on the three feature extractors. We use the pre-trained model from ImageNet [27] to initialize the weights of the convolutional neural network. Each group of experiments consisted of the basic model: ConvNet + BDLSTM, the model with RRA: ConvNet + RRA + BDLSTM, the model with VFA: ConvNet + VFA + BDLSTM, and TAMNet (ConvNet + RRA + VFA + BDLSTM). The test results of each group on the RGB modality of UCF101 dataset are shown in Table 1. The recognition accuracy of each group is the average accuracy of the three train/test splits of the UCF101 dataset.

As shown in the results of Table 1, the addition of RRA and VFA improves the recognition performance of each group. Among them, by comparison, it can be found that the performance improvement provided by RRA and VFA for the BN Inception feature extractor group is far more than that of VGG16 feature extractor group and ResNet50 feature extractor group. Therefore, we believe that BN Inception can obtain more representative spatial features to fully exploit the recurrent region attention performance of RRA. In the end, we chose BN Inception as the feature extractor for TAMNet.

Through the ablation studies, it can be found that the addition of RRA component and VFA component brings different extents of performance improvement to the base model in each group. Among them, the performance improvement of ConvNet + RRA + BDLSTM compared to the basic model is much higher than that of ConvNet + VFA + BDLSTM. Meanwhile, the performance improvement of TAMNet (ConvNet + RRA + VFA + BDLSTM) compared with the recognition performance of the basic model is greater than the sum of the above two models. We believe that this situation is mainly caused by the RRA between the position of CNN feature extractor and that of the VFA in the network structure. Compared to focusing on the spatial features with action-irrelevant visual information, such as redundancy and noise, it is more meaningful for VFA to focus on the action-relevant regional visual information calculated by RRA in temporal. In this way, the RRA can also assist the VFA in reducing the prediction errors of BDLSTM caused by the similar context of heterogeneous action video in temporal. The ablation studies and the above analysis proved that both RRA and VFA are indispensable components in the TAMNet model.

Then we verify the effectiveness of our proposed two-level attention on improving the recognition accuracy and the recognition performance of the TAMNet network. We transferred the weights learned by BN Inception [5] on the ImageNet [27] dataset to the UCF101 dataset. For fair comparison, we used pre-trained method to experiment only on the RGB modality input of the UCF101 dataset to highlight the promotion in recognition performance by RRA and VFA.

The experiment results are summarized in Table 2. The recognition accuracy of each method is the average accuracy of the three train/test splits of the UCF101 dataset, and

TABLE 1. The average recognition accuracy of each group on the RGB modality of UCF101 dataset.

Method (**RGB modality**)	Pre-train dataset	Resolution	Backbone architecture	Accuracy
VGG16 + BDLSTM	ImageNet	224 × 224	VGG16	73.9%
VGG16 + RRA + BDLSTM				74.4%
VGG16 + VFA + BDLSTM				74.2%
TAMNet (VGG16)				75.0%
ResNet50 + BDLSTM	ImageNet	224 × 224	ResNet50	84.4%
ResNet50 + RRA + BDLSTM				86.1%
ResNet50 + VFA + BDLSTM				84.8%
TAMNet (ResNet50)				86.9%
BN Inception + BDLSTM	ImageNet	224 × 224	BN Inception	**86.1%**
BN Inception + RRA + BDLSTM				**88.5%**
BN Inception + VFA + BDLSTM				**86.8%**
TAMNet (BN Inception)				**89.6%**

TABLE 2. The average recognition accuracy of various methods on the UCF101 dataset.

Method	Pre-train dataset	Resolution	Backbone architecture	Accuracy
Two Stream(RGB) [8]	ImageNet	224 × 224	VGG-M	73.0%
Spatial Stream ResNet [28]	ImageNet	224 × 224	ResNet-50	82.3%
Conv Pooling Network(RGB) [10]	ImageNet	224 × 224	VGGNet-16	82.6%
RGB-I3D [29]	ImageNet	224 × 224	Inception V1	84.5%
TSN(RGB) [11]	ImageNet	224 × 224	BN Inception	85.7%
Multimodal Fusion Network(RGB) [34]	ImageNet	224 × 224	ResNet-152	86.2%
BN Inception + BDLSTM(RGB)	ImageNet	224 × 224	BN Inception	86.1%
BN Inception + RRA + BDLSTM(RGB)	ImageNet	224 × 224	BN Inception	88.5%
TAMNet(RGB)	ImageNet	224 × 224	BN Inception	**89.6%**

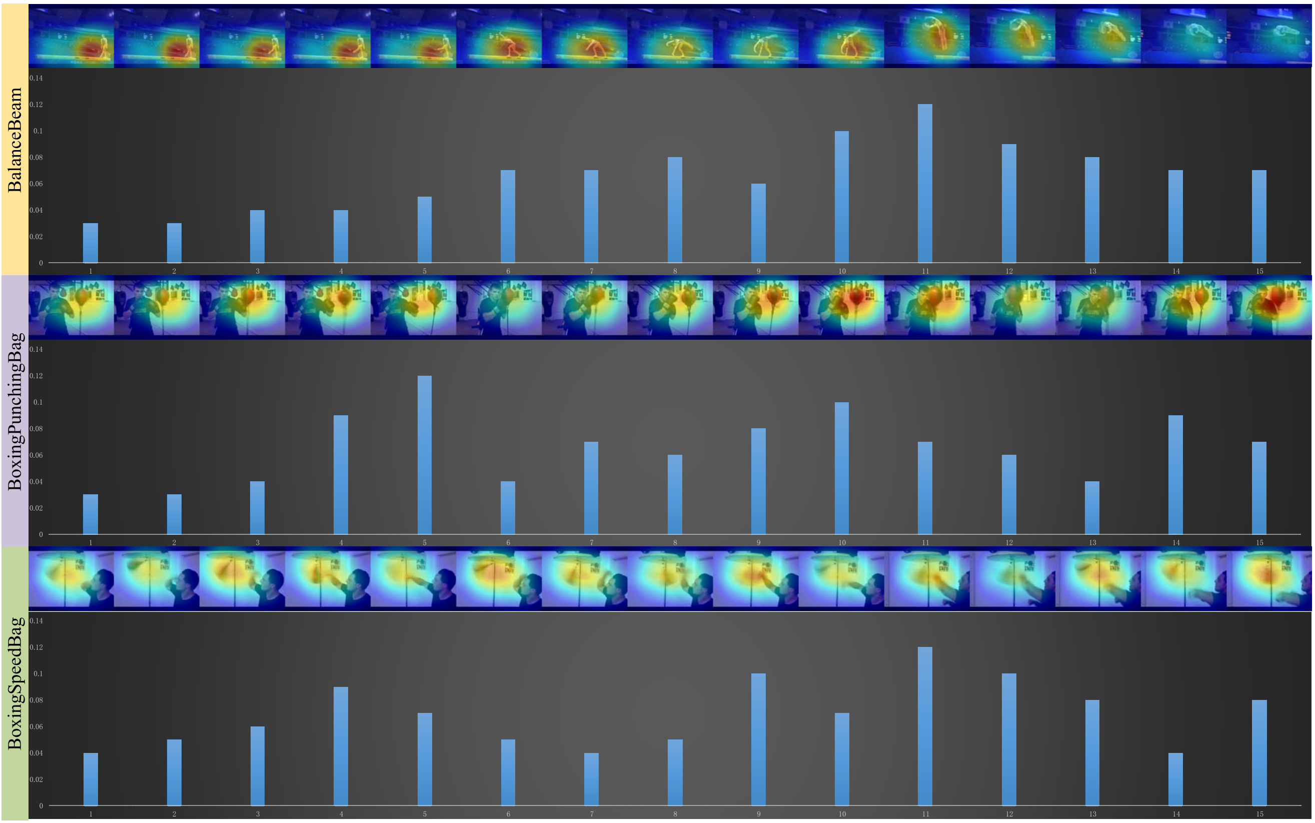


FIGURE 10. The visualized heat maps of RRA and the score bar chart of each frame calculated by VFA.

the network model is grouped according to the pre-trained dataset. The last three lines in the table are our basic model: BN Inception + BDLSTM, the network model with RRA: BN Inception + RRA + BDLSTM, and the final network model with two-level attention: TAMNet. Compared with our basic model, the recognition accuracy is improved by 2.4% with the addition of RRA, and the recognition accuracy is improved by 3.5% with the addition of two-level attention.

By comparing the recognition accuracy of the three models, we can see that the proposed RRA and VFA have brought significant promotion to the network recognition performance.

In order to more intuitively reflect the improvement that our RRA and VFA bring to the recognition performance of network, we have visualized the attention effect of RRA and graphicalized the score of each frame calculated by VFA. The visualized heat maps of RRA (upper part of each action sequence) and the score bar chart of each frame calculated by VFA (below part of each action sequence) are shown in Figure 10. It can be clearly observed that the attention resource of our RRA is mainly focused on the spatial regions associated with the action, thus reducing the interference caused by the action-irrelevant visual information such as the complex background and lighting conditions. And it also can be clearly observed that video frames which better reflect action trait get higher scores than other frames in the video sequence. Such as the frame in which the subject is swiveling jump in the BalanceBeam, the frame in which the subject is punching the bag in the BoxingPunchingBag, and the frame in which the subject is quickly raising his arm in the BoxingSpeedBag. The VFA catches the video frames which can highlight the action trait, thereby reducing the interference caused by similar contexts between heterogeneous action videos.

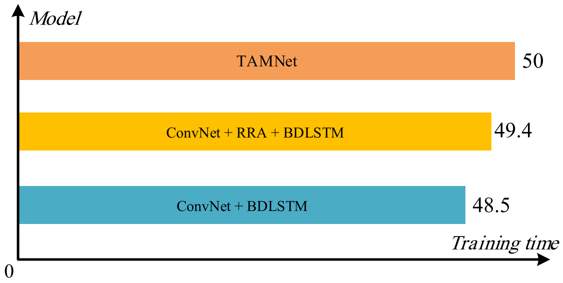


FIGURE 11. The training durations of the three models on split1 of the UCF101 dataset.

We also compared the training durations of our three network models on the split1 of the UCF101 dataset under the same number of training (epochs =120). The training durations of the three models on split1 of the UCF101 dataset is shown in Figure 11. Through the comparison of the training durations of these three models, it can be clearly reflected from the side that our proposed RRA and VFA only bring a small amount of computation to the network.

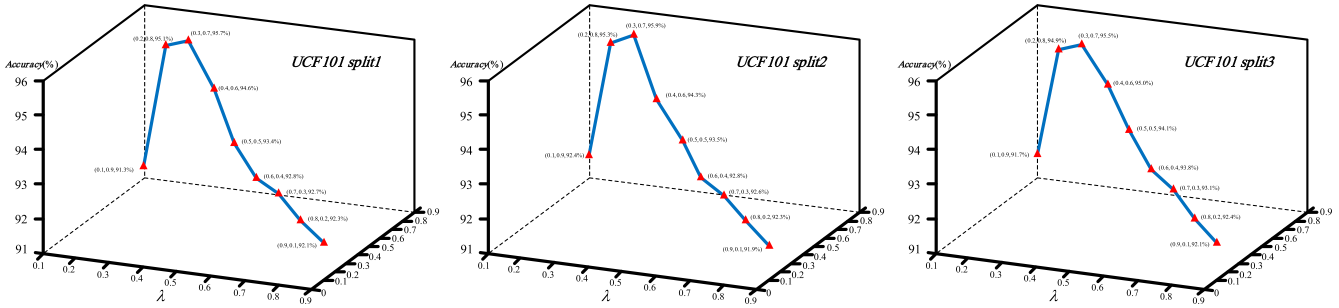


FIGURE 12. The 3D line charts of the two-modalities fusion results of UCF101 dataset in regard to different fusion weights.

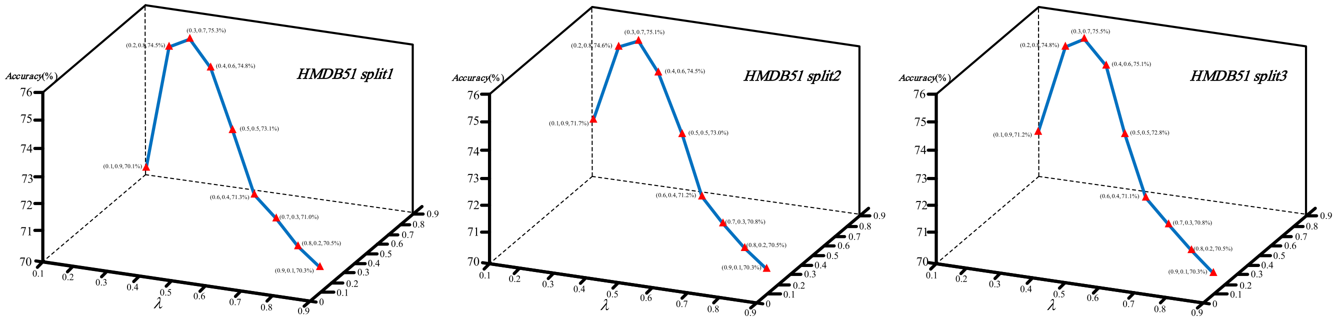


FIGURE 13. The 3D line charts of the two-modalities fusion results of HMDB51 dataset in regard to different fusion weights.

From what has been discussed above, the proposed RRA and VFA not only bring considerable promotion to the recognition performance of the network, but also do not bring an expensive computational burden to the network training process.

Then we compare the performance of TSN(RGB)-BN Inception [11], Multimodal Fusion Network(RGB)-ResNet152 [34], and TAMNet(RGB)-BN Inception in Table 2, all of these models are pre-trained on the ImageNet dataset. It can be seen that our TAMNet(RGB) outperforms TSN(RGB) by 3.9% and outperforms Multimodal Fusion Network(RGB) by 3.4% on the UCF101 dataset. This superior performance shows that the spatio-temporal representations learned by our TAMNet are more efficient than the TSN and Multimodal Fusion Network when using only RGB modality input for transfer learning. Our TAMNet network achieves the best performance with the same pre-trained settings.

TABLE 3. Comparison of two-modalities performance between TAMNet and SOTA methods.

Method	UCF101	HMDB51
DT + MVSV [30]	83.5%	55.9%
iDT + HSV [31]	87.9%	61.1%
Two Stream [8]	88.0%	59.4%
VideoLSTM [32]	89.2%	-
C3D [12]	85.2%	51.6%
Two Stream + LSTM [14]	88.6%	-
ConvNet + LSTM [15]	82.3%	-
TDD + FV [33]	90.3%	63.2%
HAN [21]	92.7%	64.3%
STAN [39]	93.6%	-
JSTA [22]	93.7%	65.3%
TSN (RGB+OF) [11]	94.0%	68.5%
Multimodal Fusion Network (RGB+OF) [34]	94.8%	-
ResNet + TSN [36]	94.8%	71.8%
TVNets + IDT (RGB+OF) [37]	95.4%	72.6%
Pillar Networks++ [38]	-	73.6%
TAMNet (RGB+OF)	**95.7%**	**75.3%**

2) TWO-MODALITIES TAMNET PERFORMANCE EVALUATION

In this section, we validated the two-modalities fusion performance of the TAMNet network on the UCF101 dataset and the HMDB51 dataset.

In this paper, we use the probability fusion method to fuse the video-level predictions of RGB modality TAMNet and optical flow modality TAMNet. We first research the fusion weights of probability fusion. The 3D line charts of the two-modalities fusion results of the 3 splits of UCF101 dataset and the HMDB51 dataset with respect to different fusion weights are shown in Figure 12 and Figure 13. The λ in the figure represents the fusion weight of the video-level prediction of the RGB modality TAMNet network. When λ is 0.3, the two-modalities TAMNet network achieves optimal recognition performance on both the three splits of the UCF101 dataset and the HMDB51 dataset. Therefore, we finally chose the fusion weight that $\lambda = 0.3$ to fuse the video-level predictions of the two-modalities TAMNet.

Finally, we compare the two-modalities fusion performance of our proposed TAMNet network with the current

SOTA methods on the UCF101 dataset and the HMDB51 dataset. Since our network adopts RGB modality and optical flow modality as input, in order to fairly compare the recognition performance of the network, for the network which adopts multi-modalities as input, we select the two-modalities fusion performance of it to compare the recognition performance with our network. The results are shown in Table 3. On the UCF101 dataset, the performance of our proposed TAMNet network outperformed TSN [11] by 1.7%, outperformed the Multimodal Fusion Network [34] and ResNet + TSN [36] by 0.9%, which is better than TVNets + IDT [37] by 0.3%, exceeding the performance of the current SOTA methods. The performance on the HMDB51 dataset outperformed TSN [11] by 6.8%, outperformed ResNet + TSN [36] by 3.5%, outperformed TVNets + IDT [37] by 2.7%, which is better than Pillar Networks ++ [38] by 1.7%. This indicates that the TAMNet proposed in this paper has good generalization ability.

V. CONCLUSION

In this work, our proposed Recurrent Region Attention and Video Frame Attention have brought significant improvements to the accuracy of video action recognition. The two-modalities recognition performance of the proposed TAMNet network has reached a new level of technology on both the UCF101 dataset and the HMDB51 dataset. Since our TAMNet network is an end-to-end network, we only use the probability fusion method to fuse the outputs of two-modalities TAMNet. In this fusion method, each modality model can only access the features of the current modality and cannot learn the interaction between different modalities. In the subsequent works, we will also try to use different fusion methods to conduct more comprehensive modality fusion at different locations of the model to achieve higher recognition performance.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2017, *arXiv:1602.07261*. [Online]. Available: https://arxiv.org/abs/1602.07261

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1725–1732.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 568–576.

[9] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L^1 optical flow," in *Proc. Joint Pattern Recognit. Symp.*, 2007, vol. 4713, no. 5, pp. 214–223.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1933–1941.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 20–36.

[12] 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)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] J. Y.-H. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4694–4702.

[15] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2625–2634.

[16] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Eur. Conf. Comput. Vis.*, Prague, Czech Republic, 2004, pp. 25–36.

[17] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[18] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," 2015, *arXiv:1511.04119*. [Online]. Available: https://arxiv.org/abs/1511.04119

[19] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1647–1656.

[20] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Hierarchical multi-scale attention networks for action recognition," *Signal Process. Image Commun.*, vol. 61, pp. 73–84, Feb. 2018.

[21] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li, "Hierarchical attention network for action recognition in videos," 2016, *arXiv:1607.06416*. [Online]. Available: https://arxiv.org/abs/1607.06416

[22] T. Yu, C. Guo, L. Wang, H. Gu, S. Xiang, and C. Pan, "Joint spatial-temporal attention for action recognition," *Pattern Recognit. Lett.*, vol. 112, pp. 226–233, Sep. 2018.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[24] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[25] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. –2563.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[28] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 3468–3476.

[29] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 6299–6308.

[30] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 596–603.

[31] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, vol. 150, pp. 109–125, Sep. 2016.

[32] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understand.*, vol. 166, pp. 41–50, Jan. 2018.

[33] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4305–4314.

[34] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal Keyless attention fusion for video classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Apr. 2018, pp. 7202–7209.

[35] H. Sang, C. Wang, D. He, and Q. Liu, "Multi-information flow CNN and attribute-aided reranking for person reidentification," *Comput. Intell. Neurosci.*, vol. 2019, Feb. 2019, Art. no. 7028107.

[36] Y. Yuan, D. Wang, and Q. Wang, "Memory-augmented temporal dynamic learning for action recognition," 2019, *arXiv:1904.13080*. [Online]. Available: https://arxiv.org/abs/1904.13080

[37] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6016–6025.

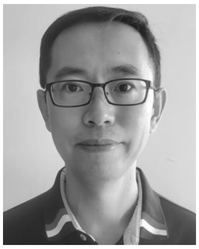
[38] B. Sengupta and Y. Qian, "Pillar networks++: Distributed non-parametric deep and wide networks," 2017, *arXiv:1708.06250*. [Online]. Available: https://arxiv.org/abs/1708.06250

[39] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified Spatio-Temporal attention networks for action recognition in videos," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019.

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6450–6459.

[42] B. Ma, Z. Liu, F. Jiang, Y. Yan, J. Yuan, and S. Bu, "Vehicle detection in aerial images using rotation-invariant cascaded forest," *IEEE Access*, vol. 7, pp. 59613–59623, 2019.



HAIFENG SANG received the B.S. and M.S. degrees from Northeast Normal University, Changchun, China, in 2000 and 2003, respectively, and the Ph.D. degree from Northeastern University, Shenyang, China, in 2006. He is currently a Professor with Shenyang University of Technology. His current research interests include machine vision detection technology and biometric identification technology research.

ZIYU ZHAO received the B.S. degree from the Shenyang University of Technology, Shenyang, China, in 2017, where he is currently pursuing the M.S. degree in instrument science and technology. His current research interest includes action recognition in computer vision.

DAKUO HE received the B.S. degree from the Harbin Information College, Harbin, China, in 1995, and the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 1988 and 2002, respectively, where he is currently a Professor and he is also with the Key Laboratory of Integrated Automation of Process Industry, Ministry of Education. His current research interests include modeling, control, and optimization in complex industrial systems.