Global and Local Knowledge-Aware Attention Network for Action Recognition

Zhenxing Zheng, Gaoyun An, *Member, IEEE*, Dapeng Wu, *Fellow, IEEE*, and Qiuqi Ruan, *Senior Member, IEEE*

***Abstract*— Convolutional neural networks (CNNs) have shown an effective way to learn spatiotemporal representation for action recognition in videos. However, most traditional action recognition algorithms do not employ the attention mechanism to focus on essential parts of video frames that are relevant to the action. In this article, we propose a novel global and local knowledge-aware attention network to address this challenge for action recognition. The proposed network incorporates two types of attention mechanism called statistic-based attention (SA) and learning-based attention (LA) to attach higher importance to the crucial elements in each video frame. As global pooling (GP) models capture global information, while attention models focus on the significant details to make full use of their implicit complementary advantages, our network adopts a three-stream architecture, including two attention streams and a GP stream. Each attention stream employs a fusion layer to combine global and local information and produces composite features. Furthermore, global-attention (GA) regularization is proposed to guide two attention streams to better model dynamics of composite features with the reference to the global information. Fusion at the softmax layer is adopted to make better use of the implicit complementary advantages between SA, LA, and GP streams and get the final comprehensive predictions. The proposed network is trained in an end-to-end fashion and learns efficient video-level features both spatially and temporally. Extensive experiments are conducted on three challenging benchmarks, Kinetics, HMDB51, and UCF101, and experimental results demonstrate that the proposed network outperforms most state-of-the-art methods.**

***Index Terms*— Action recognition, attention mechanism, convolutional neural networks-recurrent neural networks (CNNs-RNNs) framework, spatiotemporal feature.**

Manuscript received June 5, 2018; revised March 15, 2019 and October 31, 2019; accepted March 2, 2020. Date of publication March 30, 2020; date of current version January 5, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61772067, Grant 61472030, and Grant 61471032, in part by the Fundamental Research Funds for the Central Universities under Grant 2017JBZ108, and in part by the China Scholarship Council. *(Corresponding author: Gaoyun An.)*

Zhenxing Zheng and Qiuqi Ruan are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: zhxzheng@bjtu.edu.cn; qqruan@bjtu.edu.cn).

Gaoyun An is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: gyan@bjtu.edu.cn).

Dapeng Wu is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611-6130 USA (e-mail: dpwu@ieee.org).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2978613

I. Introduction

RECOGNIZING action is one of the most challenging problems in computer vision due to the complicated motion dynamics, cluttered background, viewpoint variations, and high computational complexity. To obtain robust video-level representation to these issues, previous studies made extensive explorations in how to fully exploit the spatial and temporal information to represent an action, such as improved dense trajectories (IDT) [1], space–time interest point (STIP) [2], and scale-invariant feature transform (SIFT) [3]. However, these methods achieved relatively marginal progress.

Recently, convolutional neural networks (CNNs) have attracted growing attention in computer vision for the superior properties providing support for the tasks of image classification [4], object segmentation [5], and object tracking [6]. Intensive interests applying CNNs to action recognition in videos have been raised. Previous model [7] used a CNN to extract the feature of each frame and pooled features of multiple frames belonging to one video for the video-level prediction, which failed to capture sufficient motion information for action recognition. The challenge of constructing effective spatiotemporal representation could be alleviated via fusing motion and appearance knowledge in the way of two streams. Although this model [8] achieved desirable results, it lacks the capacity of capturing long-term temporal dynamics. Recurrent neural networks (RNNs), especially long short-term memory (LSTM) [9], achieved impressive results in the sequence tasks due to the ability of long-term temporal modeling, so an alternative strategy is to adopt LSTM to model dynamics of frame-level features. However, most existing LSTM-based approaches do not make the distinction between various parts of video frames. In addition to 2-D CNNs used for image processing, 3-D CNNs [10] were proposed to process videos. They replaced 3×3 convolutional kernels with those of $3 \times 3 \times 3$ to perform 3-D convolutions over stacked frames. However, these methods usually have abundant parameters and need to be pretrained on a large-scale video data set, e.g., Kinetics [11].

As attention mechanisms can help models locate discriminative regions, it has been deployed in the image caption, machine translation, and fine-grained image recognition, achieving promising results. It is recorded in the cognitive psychology literature [12] that the attention is the cognitive process of selectively concentrating on a discrete aspect of information. The experimental results [13] showed that

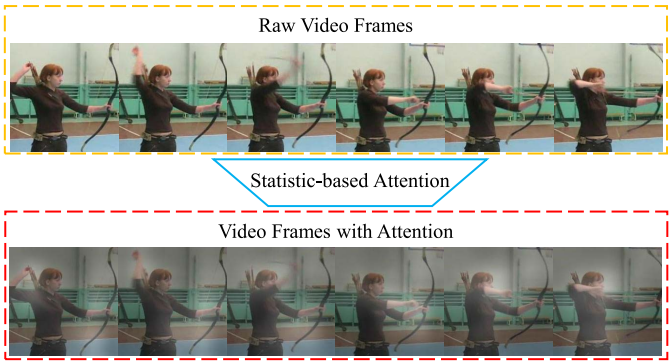


Fig. 1. Illustration of our SA for action Archery. We extract the features of video frames from a CNN and use a statistic-based method to map the resulted 3-D activation tensor to the 2-D attention map. Then, bilinear interpolation is used to up-sample the 2-D attention map to the original input size, referred to the attention image. Finally, we obtain these synthetic images by fusing raw video frames and corresponding attention images equally. This figure shows that most relevant parts for action Archery, such as the elbow, the hand, and the arrow, are attached with attention, which guides our model to make the correct prediction.

applying attention mechanisms can bring benefits to many tasks due to the allocation of limited processing resources reasonably. As depicted in Fig. 1, when raw video frames are fed into our attention model, most relevant parts for action Archery, such as the elbow, the hand, and the arrow, are attached with attention, which provides significant details for our model to make the correct classification. Thus, deploying attention mechanisms in a model properly would contribute to the task of action recognition.

Motivated by these facts and to tackle the abovementioned challenges, we propose a novel global and local knowledge-aware attention network for action recognition, which can make full use of the implicit complementary advantages of global and local information. More specifically, the proposed network consists of three streams: statistic-based attention (SA) stream, learning-based attention (LA) stream, and global pooling (GP) stream. The frame-level feature extracted from the base convolution layers is processed by three streams simultaneously, and then, each attention stream uses a fusion layer to aggregate outputs of the GP layer and the attention module, deriving the composite feature containing global and detailed local information. In addition, the global-attention (GA) regularization is proposed to guide two attention streams to better model dynamics with the reference to the global information. Finally, SA stream, LA stream, and GP stream are fused equally in the softmax layer to make comprehensive predictions. It is worth to note that three streams share the base convolution layers. The proposed network embedded with two types of attention can learn efficient video representation both spatially and temporally.

In summary, the main contributions of this article are summarized as follows.

1) We propose a novel global and local knowledge-aware attention network for action recognition. The proposed model adopts a three-stream architecture, including two attention streams and a GP stream, to exploit the implicit complementary advantages of global and local information for comprehensive predictions.
2) We propose to use the first attention mechanism named SA to focus on various parts of video frames. This attention uses the statistic of activation tensors across the channel dimension to locate the most discriminative region. It has no parameters and is nearly cost-free.
3) We propose to use the second attention mechanism named LA to make the distinction between different parts of each frame. It comprises two stacked fully connected layers and learns to pay attention to the essential regions precisely. This mechanism can be optimized in an end-to-end fashion by a standard backpropagation algorithm.
4) We propose the GA regularization to guide two attention streams to better model proper dynamics with the reference to the global information.

II. Related Work

There have been intensive studies on action recognition. In this section, we review the related works from three perspectives of CNNs, RNNs, and attention mechanisms.

A. CNNs for Action Recognition

Deep CNNs are witnessed significant advancements in numerous visual tasks since Krizhevsky *et al.* [14] proposed AlexNet. Motivated by these, early attempts used CNNs to learn spatiotemporal representation for action recognition through temporal information fusion strategies. Karpathy *et al.* [7] described several fusion methods and proposed a multiresolution approach such that high layers have access to the temporal structure across all input frames. To utilize motion information, an alternative way was to combine optical flow containing short-term motion information [8]. The two-stream network consists of two separate subnetworks, where one is for raw images and the other is for stacked optical flow, respectively, and captures spatiotemporal information by fusing the softmax scores of two streams. Wang *et al.* [15] introduced the spatiotemporal compact bilinear operator to process features at multiple abstraction levels to construct spatiotemporal pyramid representation. Fan *et al.* [16] proposed a TVNet to imitate the optimization iterations of TV-L1, which could be fine-tuned by specific tasks. In addition to RGB and optical flow information, audio, pose, and trajectory were also used to provide essential cues for action recognition [17], [18]. Wang *et al.* [19] concatenated dense trajectory descriptors with appearance features and achieved competitive performance. Choutas *et al.* [20] encoded the movement of human joints, and the resulted heatmaps were aggregated temporally, obtaining PoTion representation. The 3-D convolutional network [10] took multiple adjacent frames as inputs and constructed 3-D convolutional kernels to perform 3-D convolution operation over the stacked frames. To explore the benefits of deep 3-D features, Hara *et al.* [21] transferred 2-D residual connections to a 3-D structure and proposed a 101-layer residual 3-D CNN. In addition, by modeling the neural mechanism of the primary visual cortex and the middle temporal cortex, Liu *et al.* [22] proposed a bioinspired model to perform action recognition based on the spatiotemporal inseparability and center-surround

suppression of neurons. Wang *et al.* [23] stacked multiple SMART building blocks consisted of an appearance branch and a relation branch to model appearance and relations simultaneously.

B. RNNs for Action Recognition

Recurrent structures were also resorted to modeling the temporal dynamics due to the advances in sequence tasks. Ergen and Kozat [24] introduced a regression structure based on LSTM for efficient online learning. Donahue *et al.* [25] designed a recurrent convolutional architecture, which cascaded a CNN with a recurrent model into a unified model. CNN was used to extract features of each frame, and then, these features were fed into LSTM step by step for modeling dynamics of the feature sequence so that it could learn video-level representation in both spatial and temporal dimensions. Beyond short snippets, Ng *et al.* [26] combined the temporal feature pooling architecture with LSTM to allow the model to accept arbitrary-length frames. Wang *et al.* [27] utilized a deep 3-D-CNN to process salient-aware clips and fed the features extracted from the fully connected layer of a 3-D-CNN into LSTM for action recognition. According to the spatial–optical data organization, Yuan *et al.* [28] synthesized motion trajectories, optical, and video segmentation into spatial–optical data and used a two-stream 3-D CNN to process synthetic data and RGB data separately. Then, the resulted spatiotemporal features were fed into LSTM to mine their patterns. Sun *et al.* [29] proposed Lattice-LSTM to apply local space-variant superposition operations on the cell memory of LSTM, which enhanced the ability to capture various motion patterns.

C. Attention Mechanism

Recently, the attention mechanism has been widely used in machine translation [30] significantly. Xu *et al.* [31] incorporated two variant attention mechanism called soft attention and hard attention into a caption generation model. Zagoruyko and Komodakis [32] defined an activation-based attention map and a gradient-based attention map and transferred attention maps of a powerful network to a weak CNN to improve the performance of the latter for image classification. Inspired by Bahdanau *et al.* [30], Sharma *et al.* [33] proposed a recurrent network embedded with a soft attention mechanism that learned to attach higher importance to the elements relevant to actions. In addition to spatial attention, temporal attention was also integrated into a unified framework [34], [35] to consider both spatial and temporal cues. Zhang *et al.* [36] developed a variant LSTM incorporating an attention unit to explore the spatial–temporal relation between different parts. However, most attention mechanisms embedded in LSTM depend on inner hidden states of LSTM, which alleviates the speed of computation greatly. Unlike generating an attention map from LSTM, Li *et al.* [37] used a spatial attention neural cell to find the spatial regions where action appears, and a temporal attention neural cell to determine temporal segments containing the action. Long *et al.* [18] constructed RGB attention cluster, flow attention cluster, and audio attention cluster to integrate local features of multiple modalities. To exploit interactions among local features, Du *et al.* [38] proposed an interaction-aware attention network to extract features of different layers and used principal component analysis (PCA) to obtain the interaction information among these features.

Different from previous works, our goal is to make full use of both significant local details and global information. To address this issue, we propose a global and local knowledge-aware attention network for action recognition.

III. Proposed Network

The pipeline of the proposed network is shown in Fig. 2. Three streams from the top to the bottom of the network are named SA stream, GP stream, and LA stream, respectively. More specifically, each stream uses a shared residual network (ResNet) to extract spatial features. Then, two types of attention capture different detailed action information, while the GP stream learns global information from appearance features. Next, a fusion layer is designed to aggregate global and local details into a composite feature. Inspired by the rank loss between prediction probabilities used in fine-grained image recognition [13], a GA regularization is proposed to guide two attention streams to better model dynamics of composite features with the reference to the global information. Fusion at the softmax layer is adopted to make the final comprehensive predictions. Our network explores the implicit complementary advantages between global and local information at the feature level and the score level. Sections III-A–III-F will describe these parts in detail.

A. Feature Extraction

Recently, CNNs have shown an effective way to learn high-level semantic features. ResNet was initially introduced by He *et al.* [39]. It constructs the identity shortcut connections between the input and the output of each building block to mitigate the problems of vanishing and exploding gradient. Thus, in this article, considering the tradeoff between computation cost and performance, ResNet with 34 layers (ResNet-34) is used as the backbone network to extract appearance features and initialized by the parameters of ResNets-34 pretrained on ImageNet [40].

ResNet-34 consists of multiple layers, a convolutional layer ($Conv1$), four groups of building blocks ($Conv2\_x$, $Conv3\_x$, $Conv4\_x$, and $Conv5\_x$), an average pooling layer, and a classifier layer. Each group of blocks comprises a specific number of building blocks that perform convolution, batch normalization, and nonlinear activation. All convolutional layers have 3×3 kernels except that the first convolutional layer has 7×7 kernels. Raw video frames are fed into ResNet-34, and the output of $Conv5\_x$ is used as an appearance feature, called the activation tensor, $F^* \in R^{C \times H \times W}$, which is represented as

$$F_t^* = \left(f_{t,1}^*, f_{t,2}^*, \ldots, f_{t,H \times W}^*\right) \tag{1}$$

where $*$ stands for SA, LA, or GP, respectively. Thus, F_t^{SA} represents the feature of the tth frame for SA stream, F_t^{LA} is for LA stream, F_t^{GP} is for GP stream, and C, H, and W are the number of channels and height and width of the activation tensor, respectively. Each C-dimensional vector $f_{t,n}^* \in R^C$

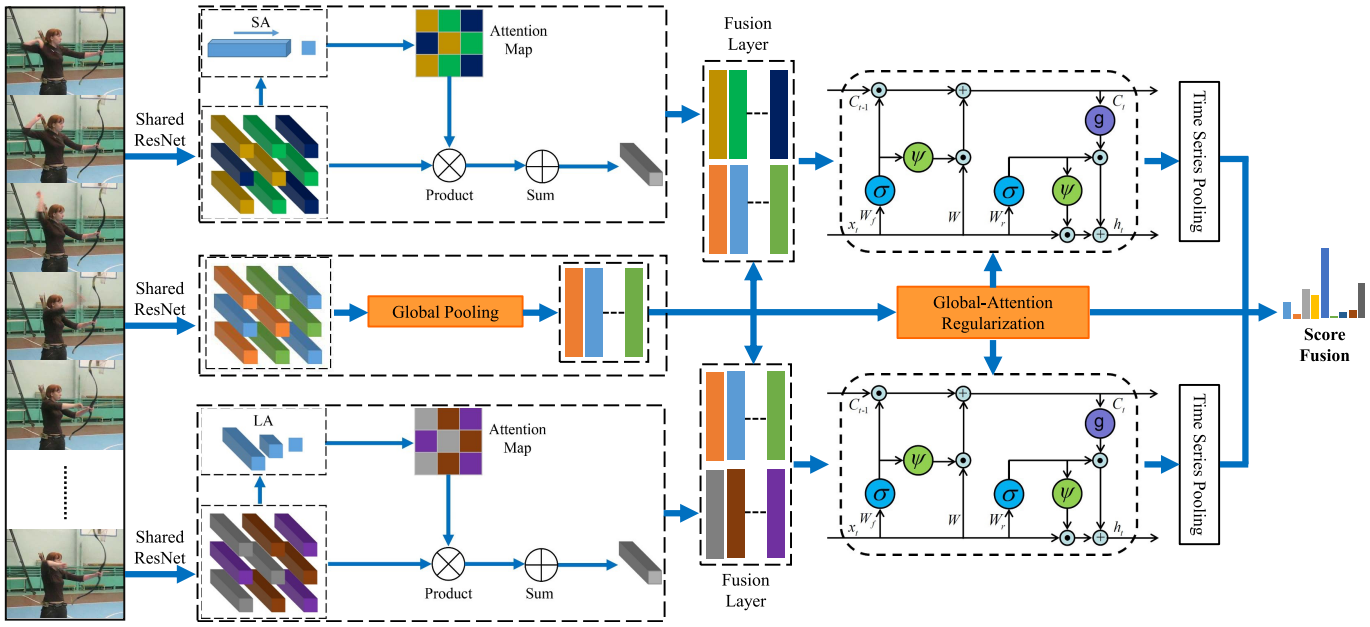


Fig. 2. Pipeline of the proposed network. Our network adopts a three-stream architecture, including two attention streams and a GP stream. The top is SA stream, the bottom is LA stream, and the middle is the GP stream. Shared ResNet is used to extract spatial features. Each attention stream employs a fusion layer to derive the composite features containing global and detailed local information. GA regularization is proposed to enable attention streams to better model dynamics with the reference to the GP stream. Fusion at the softmax layer is implemented to fully exploit the implicit complementary advantages of global and local knowledge for comprehensive predictions.

could be considered to correspond to a specified region of input frames. Consequently, the feature sequence F_t^* could be used to calculate the importance of input locations and find the most discriminative region.

B. Attention Mechanism

Most traditional algorithms do not take advantage of the complementary relation between global and local information for action recognition. These algorithms use the features from the fully connected layer to model temporal dynamics, which discard detailed action information. Some attention models operate on the features from convolutional layers that lack global information of action. Due to the impressive results of attention mechanism on various tasks, especially on natural language processing, we propose to use two types of attention to capture different significant detailed action information, which could be combined with global information to improve the performance of CNNs for action recognition.

Here, to capture global information being complementary to the detailed local information, we design the GP stream employing an average pooling layer to pool the activation tensor into a fixed-length feature vector averagely. The following formula is used to process activation tensors:

$$R_t^{\mathrm{GP}} = \frac{1}{H \times W} \sum_{n=1}^{H \times W} f_{t,n}^{\mathrm{GP}} \tag{2}$$

where $f_{t,n}^{\mathrm{GP}}$ denotes the nth vector of the activation tensor F^{GP} of the tth frame and R_t^{GP} denotes the frame-level feature vector of the tth frame. In the following, we discard the subscript t for brevity. After pooling the feature sequence of different appearance locations averagely, it can be considered as the global or general representation of an action. However, these features may be insufficient to model comprehensive dynamics of action due to the lack of details. Hence, two types of attention mechanisms are proposed to capture detailed information of actions, which is complementary to global representation.

1) Statistic-Based Attention: In this section, we introduce the first attention mechanism that focuses on the most discriminative parts of each frame and maps convolutional features to a fix-length vector. SA is parameter-free, which means that it needs no extra training data to train this attention specially. In this attention, we follow the assumption that the absolute value of each neuron in activation tensors could be used as an indicator of how important the corresponding image region is. Let us reconsider the activation tensor $F^{\mathrm{SA}} \in R^{C \times H \times W}$, and SA processes it according to three steps. First, a mapping function $\Phi(\cdot)$ is constructed to map activation tensors $F^{\mathrm{SA}} \in R^{C \times H \times W}$ to $M^{\mathrm{SA}} \in R^{H \times W}$

$$M_{h,w}^{\mathrm{SA}} = \Phi(F^{\mathrm{SA}}) = \sum_{c=1}^{C} \left| F_{c,h,w}^{\mathrm{SA}} \right|^2 \tag{3}$$

where $F_{c,h,w}^{\mathrm{SA}}$ denotes the value at the position (c, h, w) of F^{SA} and $M_{h,w}^{\mathrm{SA}}$ denotes the value at the position (h, w) of the 2-D output. Then, M^{SA} is normalized by a softmax function as weights, named the attention map

$$T_{h,w}^{\mathrm{SA}} = \frac{\exp\left(M_{h,w}^{\mathrm{SA}}\right)}{\sum_{h=1}^{H} \sum_{w=1}^{W} \exp\left(M_{h,w}^{\mathrm{SA}}\right)} \tag{4}$$

where the value of the attention map $T_{h,w}^{\mathrm{SA}}$ denotes the spatial importance of corresponding regions in the case of SA. Finally,

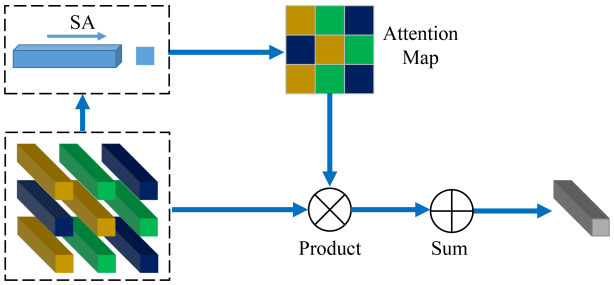


Fig. 3. Illustration of SA. SA uses an attention mapping function to map a 3-D activation tensor to a 2-D attention map based on the statistic of activation tensors across the channel dimension.

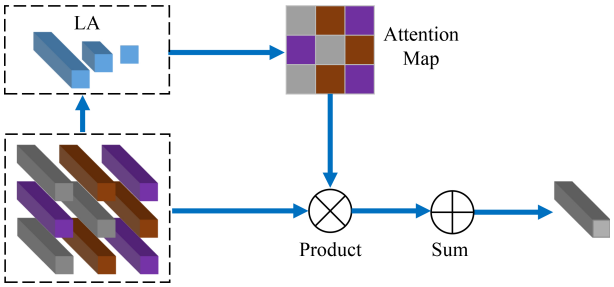


Fig. 4. Illustration of LA. LA employs two stacked fully connected layers to learn to focus on the most essential parts. This attention can be trained in an end-to-end fashion by a standard backpropagation algorithm.

we compute weighted summation between the attention map T^{SA} and feature sequences F^{SA}

$$R^{\mathrm{SA}} = \sum_{h=1}^{H}\sum_{w=1}^{W} T_{h,w}^{\mathrm{SA}} \times F_{h,w}^{\mathrm{SA}} \tag{5}$$

where R^{SA} is the frame-level feature vector produced by SA. Fig. 3 illustrates the pipeline of SA. Although this attention only uses the statistic of an activation tensor across the channel dimension, the model can focus on the most discriminative regions. Therefore, SA can capture detailed action information.

2) Learning-Based Attention: To offset the SA depending on the statistic of activation tensors, we propose to use the second attention mechanism, named LA, to learn to focus on the most important parts. Unlike SA that is parameter-free, LA has two stacked fully connected layers. The pipeline of LA is illustrated in Fig. 4.

The first fully connected layer has 128 neurons, while the second layer has one neuron whose output is used as the importance indicator of each input region. LA also consists of three steps that resemble the previous one. The first step of this attention uses two fully connected layers to map each C-dimensional vector f^{LA} to a scalar indicating the importance of different input regions, which can be described in the following formula:

$$M_{h,w}^{\mathrm{LA}} = \Psi(F^{\mathrm{LA}}) = W_1^{\mathrm{LA}} \times \tanh\left(W_0^{\mathrm{LA}} \times f_{h,w}^{\mathrm{LA}}\right) \tag{6}$$

where $W_0^{\mathrm{LA}} \in R^{128\times 512}$ is the parameter matrix of the first layer and $W_1^{\mathrm{LA}} \in R^{1\times 128}$ is the parameter matrix of the second layer. The bias term is omitted for simplicity. Similarity, $M_{h,w}^{\mathrm{LA}}$ is normalized by a softmax function

$$T_{h,w}^{\mathrm{LA}} = \frac{\exp\left(M_{h,w}^{\mathrm{LA}}\right)}{\sum_{h=1}^{H}\sum_{w=1}^{W}\exp\left(M_{h,w}^{\mathrm{LA}}\right)} \tag{7}$$

where $T_{h,w}^{\mathrm{LA}}$ represents the importance of input regions corresponding to the position (h, w) of the activation tensor in the case of LA. After obtaining the attention map, $T_{h,w}^{\mathrm{LA}}$ is used to aggregate feature sequence F^{LA} into a fix-length vector

$$R^{\mathrm{LA}} = \sum_{h=1}^{H}\sum_{w=1}^{W} T_{h,w}^{\mathrm{LA}} \times F_{h,w}^{\mathrm{LA}} \tag{8}$$

where R^{LA} denotes the frame-level feature vector produced by LA. LA compensates SA for the lack of learning ability.

Due to the LA mechanism is embedded in the network, the parameters (W_1^{LA} and W_0^{LA}) could be learned effectively by using a backpropagation algorithm in an end-to-end fashion. The main differences between our attention mechanism and others are twofold. First, two types of attention mechanisms described in this article determine the important parts without the guidance of the hidden state h_t in LSTM. Second, SA attends the crucial parts through the statistic of activation tensors directly.

C. Fusion of Global and Local Information

As described in Section III-B, the model with GP treats various parts of a frame equally, ignoring detailed action information, while the model with SA or LA focuses on the most discriminative local details. Thus, it is natural to construct practical spatiotemporal representation by fusing global and local information. We introduce the feature fusion layer that concatenates for GP representation with the statistic-based or LA representation, respectively, to obtain composite features

$$R^{\mathrm{GS}} = [R^{\mathrm{GP}}, R^{\mathrm{SA}}] \tag{9}$$

$$R^{\mathrm{GL}} = [R^{\mathrm{GP}}, R^{\mathrm{LA}}] \tag{10}$$

where R^{GS} represents the composite feature of GP and SA representation, R^{GL} represents the composite feature of GP and LA representation, and $[\cdot]$ denotes the feature concatenation operation. Then, composite features, R^{GS} and R^{GL}, are fed into the recurrent model to produce the video-level feature.

D. Efficient Temporal-Relation Modeling

Although LSTM has made significant advances in many fields as the superiority of modeling dynamics, the update of gate states in the recursion depends upon previous hidden states h_{t-1}, which dramatically restricts the speed of computation. Different from the previous recurrent models, simple recurrent unit (SRU) proposed by [41] breaks the dependence by completely dropping h_{t-1} in the recursion, which simplifies the state computation and discloses more parallelism while retaining the strong capability of representation. Thus, SRU is used to model temporal dynamics through feeding composite

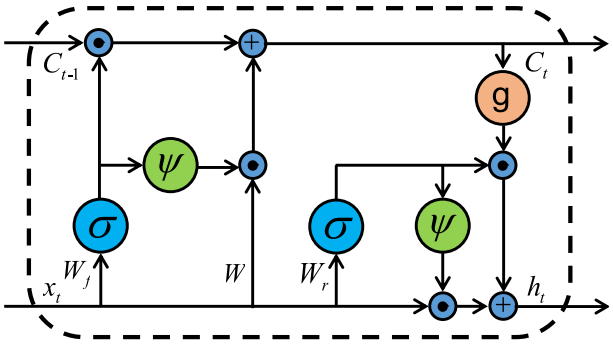


Fig. 5. Illustration of SRU. ψ represents the operation of 1-input, g is the hyperbolic tangent function, and σ is the sigmoid function. $\odot$ and $\oplus$ are elementwise multiplication and addition, respectively. We do not show the bias for simplicity.

features step by step

$$\widetilde{R}_t^\star = W^\star R_t^\star \tag{11}$$
$$f_t^\star = \sigma\left(W_f^\star R_t^\star + b_f^\star\right) \tag{12}$$
$$r_t^\star = \sigma\left(W_r^\star R_t^\star + b_r^\star\right) \tag{13}$$
$$c_t^\star = f_t^\star \odot c_{t-1}^\star + \left(1 - f_t^\star\right) \odot \widetilde{R}_t^\star \tag{14}$$
$$h_t^\star = r_t^\star \odot g(c_t^\star) + \left(1 - r_t^\star\right) \odot R_t^\star \tag{15}$$

where $\star$ stands for GS, GL, or GP, so R_t^{GS}, R_t^{GL}, and R_t^{GP} represent two composite features and global features, respectively, $f_t^\star$ and $r_t^\star$ are sigmoid gates referred to the forget gate and the reset gate, and $g(\cdot)$ is the hyperbolic tangent function, as shown in Fig. 5. When an input vector $R_t^\star$ is fed into this recurrent module, $\widetilde{R}_t^\star$, $f_t^\star$, and $r_t^\star$ at each time step can be computed simultaneously. Equations (14) and (15) are elementwise operations. A mean-pooling layer followed by a softmax layer is employed to compute the probability distribution on all action categories by the following formula:

$$P^\star(y = j) = \frac{\exp\left(\frac{1}{N}\sum_{t=1}^{N} W_{jt}^\star S_t^\star\right)}{\sum_{k=1}^{K}\exp\left(\frac{1}{N}\sum_{t=1}^{N} W_{jt}^\star S_t^\star\right)} \tag{16}$$

where $\star$ has the same meaning as in (11)–(15), $W_{jt}^\star$ represents the weight of the last fully connected layer which maps the output $S_t^\star$ of the recurrent model at the time step t to action j, K is the number of neurons in the softmax layer, and N is the number of time steps.

E. GA Regularization

Inspired by the rank loss used in image recognition [13], we propose the GA regularization to enable the recurrent network to model dynamics more reliable based on the global information. The GA regularization is defined as follows:

$$\mathrm{Loss}_{\mathrm{GA}}(\mathrm{Att}) = \max(0, P_{\mathrm{GP}} - P_{\mathrm{Att}} + \mathrm{margin}) \tag{17}$$

where P_{GP} is the prediction probability of the GP stream and P_{Att} is the prediction probability of the attention stream, so P_{SA} is for the prediction probability of SA and P_{LA} is for LA. We empirically set the margin to 0.1. Compared with the rank loss, the main difference is that GA regularization is used to optimize the whole model with the reference to GP representation. The overall loss is shown as follows:

$$\mathrm{Loss} = \mathrm{Loss}\, f(Y, \widetilde{Y}) + \mathrm{Loss}_{\mathrm{GA}}(\mathrm{SA}) + \mathrm{Loss}_{\mathrm{GA}}(\mathrm{LA}) \tag{18}$$

where Y is the predicted class vector, $\widetilde{Y}$ is the ground-truth vector denoted by a one-hot vector, and $\mathrm{Loss}\, f(\cdot)$ is the cross-entropy loss function. Based on global information, the GA regularization can guide two attention streams to generate more confident predictions.

F. Different Fusion Models

Finally, to strengthen the robustness of video-level representation, the following formula is used to combine different representation by averaging their softmax scores equally:

$$P_{\mathrm{early}} = a \times P_{\mathrm{GS}} + b \times P_{\mathrm{GL}} + c \times P_{\mathrm{GP}} \tag{19}$$

where $P_{(\cdot)}$ in the right-hand side the equation refers to the prediction of features R^{GS}, R^{GL}, or R^{GP} and P_{early} is the final prediction by combining three streams. We assign 1/3 to coefficients a, b, and c, respectively, in all experiments, unless stated otherwise. This model is termed the early fusion model because attention and global features are fused at the feature level early. With this fusion strategy, three streams are integrated into a unified network and make comprehensive predictions.

Instead of combining global information with local information at the feature level, we also propose an alternative network to fuse different types of information at the score level. As shown in Fig. 6, statistic-based representation, GP representation, and learning-based representation are fed into the recurrent model followed by a time-series pooling layer and a softmax layer to produce probabilities that are then fused to make the final classifications. Note that there is no GA regularization and a fusion layer in this late fusion network

$$P_{\mathrm{late}} = d \times P_{\mathrm{GP}} + e \times P_{\mathrm{SA}} + f \times P_{\mathrm{LA}} \tag{20}$$

where $P_{(\cdot)}$ in the right-hand side of this equation is the prediction of GP representation (P_{GP}), SA representation (P_{SA}), or LA representation (P_{LA}). P_{late} is the final prediction by fusing these three streams at the score level. We also empirically assign 1/3 to all coefficients d, e, and f.

G. Implementation Details

In this article, the off-the-shelf ResNets-34 pretrained on ImageNet is used as the appearance feature extractor. Depending on the sizes of GP features and composite features, the input sizes of corresponding recurrent models are set to 512 and 1024 neurons, respectively. The number of layers of recurrent models is kept the same as 3. The hidden units are set to 1024 in all experiments. In the LA mechanism, we experimentally find that the setting of 128 neurons for the first fully connected layer works best.

All videos are split into multiple 30-frame samples with a stride 15 for expanding the scale of samples. We impose the restriction on the maximum split number to 20. More specifically, to extract spatial features by ResNet, we first resize the

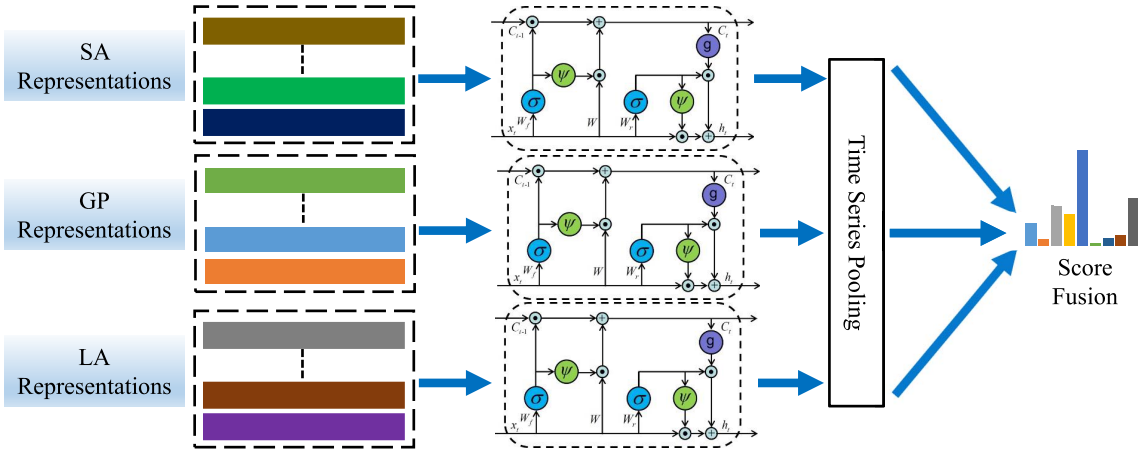


Fig. 6. Illustration of the late fusion network. There is no GA regularization and a fusion layer in it. SA representation, GP representation, and LA representation produced by well-trained CNNs are fed into the recurrent model followed by a time-series pooling layer and a softmax layer to produce probabilities that are then fused to make the final classification.

raw frame into 256×256. Second, we crop the rescaled frame with random size (0.08–1.0) and random aspect ratio (3/4–4/3) and then resize it to 224×224. Finally, frames are horizontally flipped with a probability of 0.5 for data augmentation. In addition, values of each frame in the range [0, 255] are linearly converted to a tensor in the range [0–1.0] and normalized with mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225) for RGB channels to be suitable for processing. The same preprocessing operation is adopted in training and testing phrases except for center crop and no flip in testing.

The weights of the recurrent model are initialized randomly, and all models are trained in an end-to-end fashion. The Adam optimization algorithm [42] with minibatch size 28 is used to optimize models. The learning rate starts from 1e-5 for the first eight epochs and changed to 1e-6 for the rest eight epochs. The dropout regularization ratio 0.6 is adopted in the linear transformation of the recurrent model. In testing, predictions of clips belonged to one video are averaged for the completed prediction.

Finally, after all hyperparameters are determined, we make experiments on the large-scale video data set, i.e., Kinetics [11]. Due to that Kinetics has about 500k videos and each video lasts around 10 s, it is necessary to reduce the computation burden by sampling. Inspired by the sampling strategy [43], each video is split into three segments evenly, and we sample four frames from each segment randomly, which can maintain vital information at every action intervals. Thus, with this strategy, videos with different temporal lengths are aligned and can be represented by 12 frames efficiently. Therefore, the batch size is changed to 64, the learning rate is changed to 0.001, and the training stops at 25 epochs. The rest details are the same as mentioned earlier.

IV. Experimental Evaluation

In this section, we first perform extensive experiments on two challenging benchmarks, i.e., HMDB51 [44] and UCF101 [45], to comprehensively demonstrate the effectiveness of various components in our network. Then, different fusion strategies are investigated to take advantage of different representations. Next, experiments on Kinetics [11] is conducted to validate the effectiveness on the large-scale data set, and we fine-tune our network trained on Kinetics to the experiments of UCF101 and HMDB51 to make comparisons with the state of the art. Finally, the attention regions located by our attention models are visualized to show superiority. The details will be discussed sufficiently in the following.

A. Experimental Setup

The HMDB51 data set is collected from various sources, mostly from commercial movies. This data set is composed of 3570 training clips and 1530 testing clips and organized as 51 distinct categories. There are 70 training videos and 30 testing videos for each action class. It provides three splits that divide all videos into different groups. We report the average accuracy over these three splits.

The UCF101 contains 13 320 realistic action videos from 101 action categories. It also provides three standard splits and aims at giving the largest diversity regarding actions. Similarly, we report the average accuracy over three splits. The action categories can be divided into five types: human–object interaction, body–motion only, human–human Interaction, playing musical instruments, and sports. These videos are also grouped into 25 groups that have a similar background.

Kinetics is a high-quality data set collected from YouTube. It consists of approximately 500k video clips and provides a training set, a validation set, and a testing set. In this article, we reported experimental results on a testing set of the latest version of Kinetics-600. To make a fair comparison with the previous methods, we also reported accuracies on Kinetics-400* that is constructed by selecting the overlapping videos between Kinetics-600 and the official released version Kinetics-400.

B. Analysis of Attention Mechanism

In this section, to demonstrate the effectiveness of two types of attention, we evaluate the performance of models with a GP layer or two types of attention individually, called the

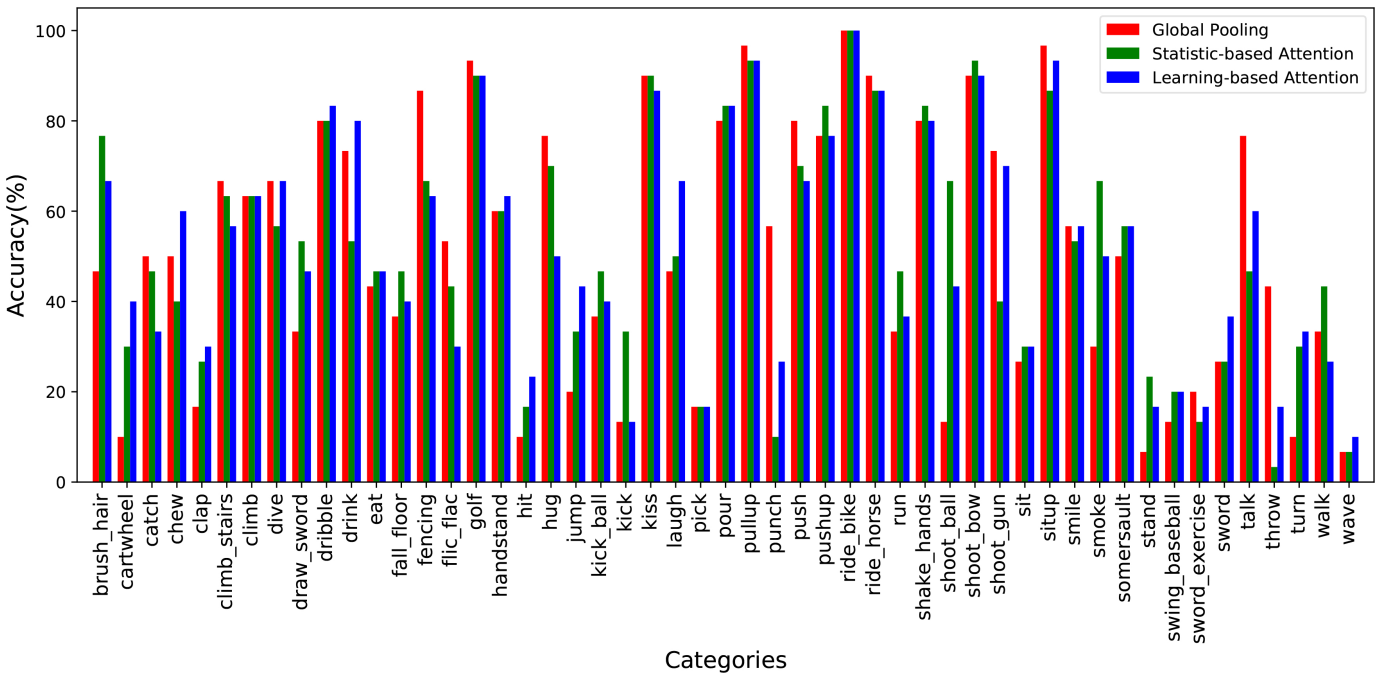


Fig. 7. Results of per-class actions on HMDB51. The lengths of red, green, and blue bars are confidences produced by the GP model, the SA model, and the LA model, respectively. This figure shows that different models work better on particular actions, such as GP model is good at action Walk, SA is good at action Hug, and LA is good at action Laugh.

TABLE I
ANALYSIS OF TWO TYPES OF ATTENTION MECHANISM

Model	HMDB51	UCF101
Global Pooling	51.11	81.06
Statistic-based Attention	52.22	81.65
Learning-based Attention	52.48	82.91

GP model, SA model, and LA model, respectively. The GP model and two attention models have the same structure except for the attention part. As the GP layer directly computes the average value of each feature map, it can be considered to learn global information.

Experimental results of three models on HMDB51 and UCF101 are listed in Table I. As shown in Table I, two attention models show significant improvements in both data sets compared with the GP model. The SA model improves the results by 1.11% and 0.59%, while the LA model improves the results by 1.37% and 1.85% on HMDB51 and UCF101, respectively, which demonstrates the effectiveness of two types of attention mechanism. It worth to mention that compared with the SA model, the LA model converges faster in training time. It suggests that the network with learning ability can quickly focus attention on various regions of video frames.

To further verify whether there is a complementary advantage between different models, Fig. 7 shows the results of per-class action. It can be seen that different models work better on particular actions, such as GP model is good at action Walk, SA is good at action Hug, and LA is good at action Laugh. We speculate that there may be implicit complementary advantages of combining global and detailed local information. In the following, global and local representations are investigated in various ways.

C. Analysis of Global and Local Information Fusion

In this section, we explore the complementary advantages of global and detailed local information. Intuitively, GP directly computes the average value of each feature map and treats various parts equally to make the prediction from the global perspective, while attention models attend the most discriminative parts and make the prediction depending on action details. Based on this, we make the combination to exploit abundant benefits brought by the fusion of global and local information. Here, we introduce two composite features, where GP representation is fused with SA representation (GS) or with LA representation (GL) at the feature level and the score level. Fusion at the feature level has four forms (max, multiply average, and concatenation), while score fusion has the average form.

Base convolutional layers and two attentions well-trained in Section IV-B are used to extract global and local representation. Their parameters are frozen in training time, and recurrent models are trained for modeling dynamics. Results in Tables II and III show that the performance is boosted by a large margin in two cases of feature concatenation and score fusion. GL using score fusion achieves the best results, where the maximum of improvement is 4.12% on HMDB51 and 2.73% on UCF101, which shows that there are implicit complementary advantages between GP representation and attention representation. Meanwhile, models using score fusion or feature concatenation achieve comparable results,

TABLE II
ANALYSIS OF GLOBAL AND SA INFORMATION FUSION

Methods	HMDB51	UCF101
Global Pooling	51.11	81.06
Statistic-based Attention	52.22	81.65
GS (feature max)	50.59	80.39
GS (feature multiply)	50.52	80.67
GS (feature average)	53.86	82.57
GS (feature concatenation)	54.25	82.65
GS (score fusion)	53.59	83.66

TABLE III
ANALYSIS OF GLOBAL AND LA INFORMATION FUSION

Methods	HMDB51	UCF101
Global Pooling	51.11	81.06
Learning-based Attention	52.48	82.91
GL (feature max)	50.39	81.22
GL (feature multiply)	49.93	81.22
GL (feature average)	52.88	82.28
GL (feature concatenation)	54.31	83.18
GL (score fusion)	55.23	83.79

TABLE IV
ANALYSIS OF GA REGULARIZATION

Methods	HMDB51	UCF101
Global Pooling	51.11	81.06
GS (concatenation) with GA Regularization	54.44	83.52
GL (concatenation) with GA Regularization	54.58	83.42

which justifies that the main improvements of feature concatenation may be derived from the fusion of two networks. It may also indicate that feature concatenation could exploit the advantages of fusion of attention and global information although they are interrelated with each other. As the model using feature concatenation has fewer parameters, therefore, our fusion strategy can efficiently exploit global knowledge and detailed local knowledge and obtain the effective video-level representation.

D. Analysis of GA Regularization

In this section, we evaluate the effect of GA regularization applied to two composite features and GP features. In this case, the GP model is used as the reference. Intuitively, with the help of global information, two attention streams would better model the dynamics of composite features. Similar to the previous experimental settings, all parameters of spatial feature extract parts and attention parts are frozen, and we only optimize recurrent models of two attention streams. Results in Table IV show that GA regularization contributes to improving performance on two data sets. On HMDB51, the best improvement is achieved by GL, and the best improvement is achieved by GS on UCF101, where improvement of the accuracy is 0.27% and 0.87% on HMDB51 and UCF101, respectively. It illustrates that GA regularization can guide our recurrent network to model dynamics more reliable with global information as a reference.

TABLE V
ANALYSIS OF DIFFERENT FUSION MODELS

Methods	HMDB51	UCF101
Global Pooling	51.11	81.06
Late fusion model (max)	54.12	83.95
Late fusion model (multiply)	56.14	85.32
Late fusion model (average)	55.82	85.03
Early fusion model (GS + GL)	54.51	84.29
Early fusion model (GS + GL + GP)	56.21	84.71

E. Analysis of Different Fusion Models

In this section, we investigate the different fusion methods for global and local information. GS and GL are the same as those of the three aforementioned cases, as described in Section IV-C. In the late fusion model, the softmax scores derived from different models, as illustrated in Fig. 6, are directly fused to make the final prediction by using three forms (max, multiply, and average). The symbol "+" represents the prediction score fusion of different features, including two composite features and global features, as illustrated in Fig. 2.

From Table V, we can see that the late fusion model and early fusion model achieve comparable results and improve the performance by a large margin on both data sets compared with a single model. In the case of the late fusion model, where different types of representations are fused directly without the concatenation layer and GA regularization, the max improvement is 5.03% and 4.26% on HMDB51 and UCF101, respectively, compared with GP model. Compared with the early fusion model (GS + GL), when incorporating global information, the early fusion model (GS + GL + GP) improves accuracies by 1.70% and 0.42% on HMDB51 and UCF101, respectively. Their cooperation outperforms either of them, which demonstrates that two attention representation and the GP representation may be complementary to each other in a way. In the following, the early fusion model (GS + GL + GP) is used to compare with the state-of-the-art models.

F. Comparison With the State of the Arts

Finally, to verify the effectiveness of our model on the large-scale data set of action recognition, we perform experiments on Kinetics and report top-1 and top-5 and average accuracies on Kinetics-600 and Kinetics-400*. The accuracies of other algorithms in Table VI are copied from original articles that mainly take RGB frames as inputs. 3-D ResNeXt-101 [21] designed a residual 3-D CNN to take advantage of deep 3-D features. Our model outperforms 3-D ResNeXt-101 by 4.1% at top-1 accuracy of the experiment on Kinetics-400*. Note that the accuracy of I3D-RGB [46] is lower than ours by 0.8% at the top 1 accuracy on Kinetics-400*, but when combined optical information, two-stream I3D achieves superior performance. This demonstrates that the optical flow can provide complementary advantages from the perspective of motion, and we speculate that the performance of our model would also be boosted when incorporating optical flow information.

TABLE VI

COMPARISON OF THE NETWORKS TAKING RGB FRAMES AS INPUTS ON KINETICS

Model	Top-1	Top-5	Average
CNN+LSTM [11]	57.0	79.0	68.0
Two-stream CNN [11]	61.0	81.3	71.2
3D ResNeXt-101 [21]	65.1	85.7	75.4
I3D-RGB [46]	68.4	88.0	78.2
ARTNet [23]	69.2	88.3	78.7
TSN (BNInception) [43]	69.1	88.7	78.9
ECO_{En} [47]	70.0	89.4	79.7
Two-Stream I3D [46]	71.6	90.0	80.8
Our model (Kinetics-400*)	69.2	88.5	78.9
Our model (Kinetics-600)	70.0	89.4	79.7

Meanwhile, we fine-tune our network trained on Kinetics to the experiments on UCF101 and HMDB51. The classification accuracies listed in original articles on HMDB51 and UCF101 are summarized in Table VII. First, compared with I3D, our model highlights the discriminative part of actions, which may be complementary to I3D with appearance and motion streams. When combined with I3D, our model achieves state-of-the-art performance on UCF101 and very competitive performance with the state-of-the-art model on HMDB51. This shows that the attention stream is effective for action recognition. Second, our model outperforms recent RNN-based approaches by a large margin. The most relevant work to ours is the spatiotemporal attention model, i.e., recurrent spatial-temporal attention network (RSTAN) [48]. It uses an attention-driven appearance-motion fusion strategy to integrate appearance and motion streams into a unified model, which achieves 70.5% and 94.6% on HMDB51 and UCF101, respectively, while our model outperforms RSTAN by 10.6% on HMDB51 and 3.6% on UCF101, which may demonstrate that our model can learn more discriminative features. PoTion [20] extracts pose motion from video frames to construct the third modality, which may be complementary to appearance and motion modalities. Compared with PoTion, our attention stream is more complementary to the traditional two-stream network, achieving a better result on HMDB51. Third, deep networks with temporal pyramid pooling (DTPP) [49] and PoTion [20] only got comparable performance on one data set, HMDB51 or UCF101, respectively. However, our model may get comparable performance on both data sets. In the future, a great effort may be devoted to investigating the impact of other modalities, e.g., pose or audio.

Table VIII displays the experimental results of the models employing attention mechanisms. Our model achieves the best results compared with other models. RSTAN [48] introduces a spatial–temporal attention module to identify which part or frame is important. Although it achieves the promising performance, our proposed model can make further improvements by 10.6% and 3.6% on both data sets, respectively. This may indicate that our model can focus on important regions relevant to actions. STAN [37] derives temporal attention from the holistic consensus across the frame, optical flow, and clip modalities, and it achieves 93.6% on UCF101. Our model outperforms it by 4.6% although we do not decide which frame is essential explicitly. S3D-G [68] replaces many of the 3-D convolutions by low-cost 2-D convolutions and introduces a feature gating mechanism to build an efficient network. Although it achieves the promising performance, our proposed model can make further improvements by 2.9% and 1.4% on both data sets, respectively.

TABLE VII

COMPARISONS WITH THE STATE OF THE ARTS ON HMDB51 AND UCF101

Model	Source	HMDB51	UCF101
C3D Model [10]	Sports-1M	56.8	82.3
Two-Stream [50]	ImageNet	59.4	88.0
Asymmetric 3D-CNN + iDT [51]	ImageNet	65.4	92.6
Grass Match Kernels [52]	ImageNet	65.4	93.4
Trajectory Pooling [53]	ImageNet	65.6	92.1
Sequential Framework [28]	Sports-1M	65.7	90.9
TDD+iDT [54]	ImageNet	65.9	91.5
Lattice LSTM [29]	ImageNet	66.2	93.6
Pyramid Net (ResNet-50) [15]	ImageNet	66.5	93.8
STDDCN [55]	ImageNet	66.9	93.8
LTC + iDT[56]	ImageNet	67.2	92.7
D^3-LND [57]	Sports-1M	67.8	92.8
Two-Stream TSN [43]	ImageNet	68.5	94.0
TCLSTA [58]	ImageNet	68.7	94.0
Pyramid Net (BN-Inception) [15]	ImageNet	68.9	94.6
Chained Network [17]	Sports-1M	69.7	91.1
ActionVLAD + iDT [59]	ImageNet	69.8	93.6
STMN + iDT [60]	Sports-1M	70.2	92.8
3D ResNeXt-101 [21]	Kinetics	70.2	94.5
RSTAN [48]	ImageNet	70.5	94.6
ARTNet with TSN [23]	Kinetics	70.9	94.3
Multiplier + iDT [61]	ImageNet	72.2	94.9
ECO_{En} [47]	Kinetics	72.4	94.8
PBNet + iDT [62]	ImageNet	72.5	95.4
TVNet + iDT [16]	KITTI	72.6	95.4
Deep Manifold Learning [63]	ImageNet	72.5	96.7
Motion Map 3D Network [64]	Sports-1M	73.7	91.9
OFFNet [65]	ImageNet	74.2	96.0
S-TPNet + iDT [66]	Kinetics	74.8	96.0
Dense Dilated Network [67]	Kinetics	74.5	96.9
S3D-G [68]	Kinetics	78.2	96.8
I3D [46]	Kinetics	80.2	97.9
PoTion + I3D [20]	Kinetics	80.9	**98.2**
Attribute Model [69]	-	81.1	95.1
DTPP [49]	Kinetics	**82.1**	98.0
Ours(early fusion) + I3D	Kinetics	81.1	**98.2**

TABLE VIII

COMPARISON OF NETWORKS WITH ATTENTION

Model	Source	HMDB51	UCF101
Recurrent Attention [36]	ImageNet	45.4	75.8
Hierarchical Attention [70]	ImageNet	64.3	92.7
STAN [37]	ImageNet	-	93.6
Attention Cluster [18]	Kinetics	69.2	94.6
RSTAN [48]	ImageNet	70.5	94.6
Pyramid Attention Network [38]	ImageNet	70.7	95.5
S3D-G [68]	Kinetics	78.2	96.8
Ours(early fusion) + I3D*	Kinetics	81.1	**98.2**

G. Spatial Attention Visualization

To better understand the effectiveness of our two types of attention, we visualized attention regions on which statistic-based and LA focus. As illustrated in Fig. 8, it is clear that two types of attention help the network focus on the most relevant parts. Both statistic-based and LA can locate the most relevant parts, the bike and body for action Biking, and the body for action SkyDiving, which contributes to making correct

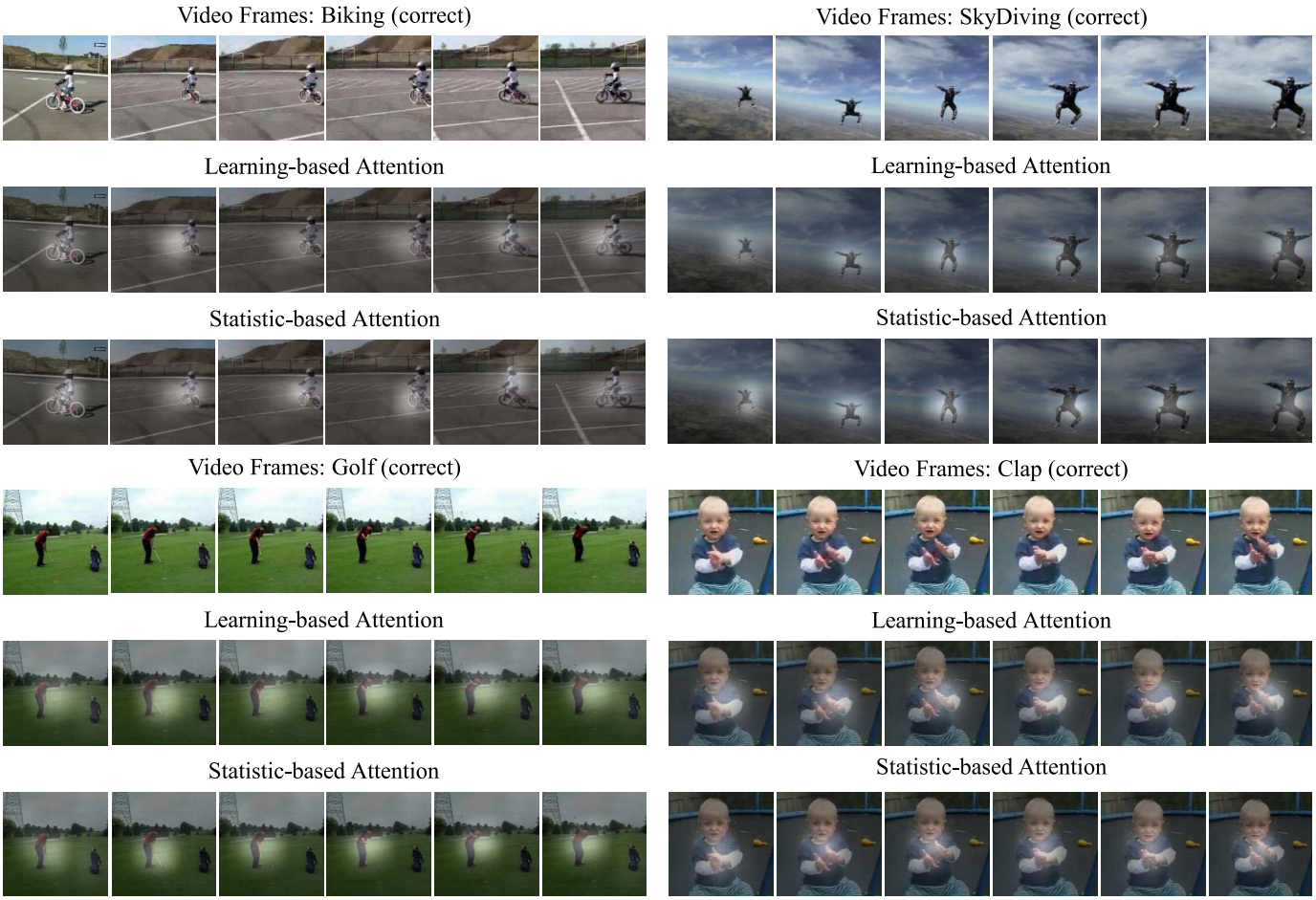


Fig. 8. Correct illustration of our SA and LA. These images are cropped from raw video frames centrally. Frames in the second and third rows of each action display the attention regions (white) produced by SA and LA, respectively. One can see that both statistic-based and LA can locate the most relevant parts, the bike and the body for action Biking, the body for action SkyDiving, the body and the golf club for action Golf, and hands for action Clap.

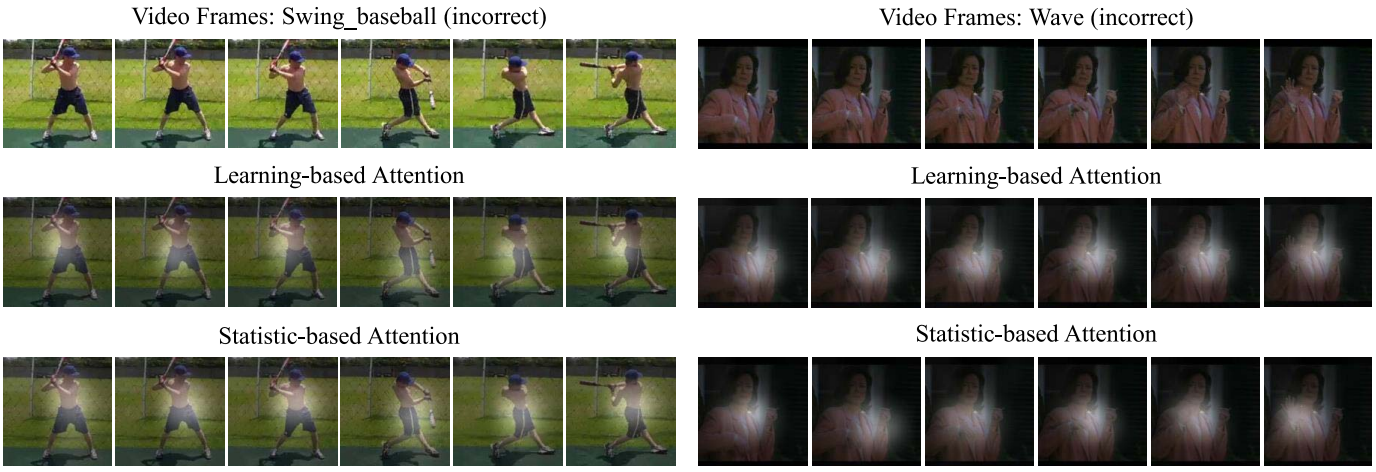


Fig. 9. Incorrect illustration of our SA and LA. One can see that attention attends the body and baseball bat for action Swing baseball and hands for action Wave, respectively. Our models are confused with these two action groups, Swing baseball and Golf, Wave, and Clap due to the similarities of them and make incorrect predictions.

predictions. However, when attention attends the body or the region between two hands that are irrelevant to action Swing baseball or Wave, as shown in Fig. 9, our model classifies the action Swing baseball as Golf and action Wave as Clap incorrectly. Because both actions Swing Baseball and Golf use a bat to hit a ball and both actions Wave and Clap are acted by swinging hands, substantial similarities between these actions perplex our models. In the future, we will address this issue by the auxiliary tasks of scene recognition, fine-grained detection, and localization for limbs and objects.

We also compare the attention visualization results of RSTAN [48] with those of our attention models in Fig. 10. It is

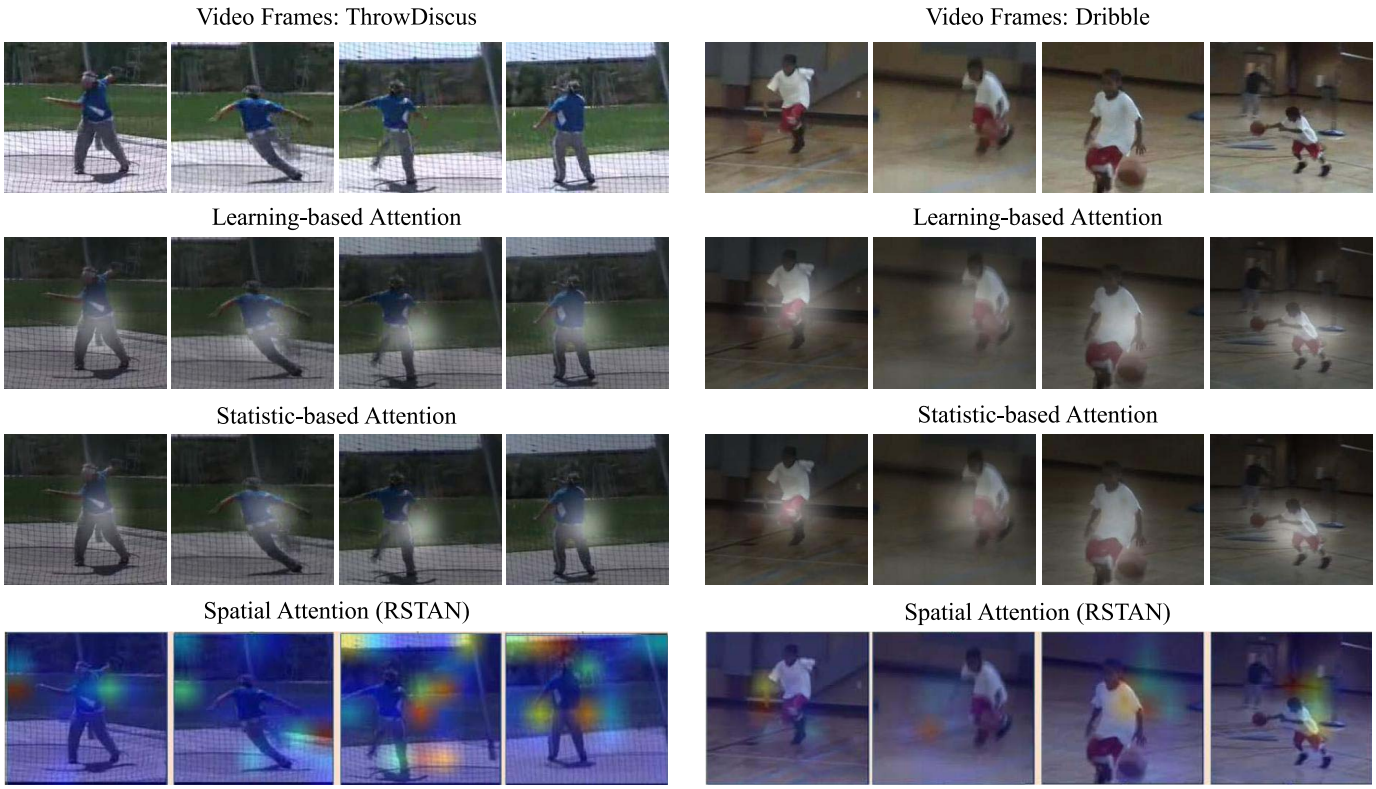


Fig. 10. Comparisons of visualization results produced by RSTAN [48] and ours. The first row shows raw video frames, the second and third rows display visualization results of our attention models, and the bottom row copied from the original article shows the spatial visualization results of RSTAN for given frames. The brightness of color indicates the importance of corresponding regions.

obvious that our attention models focus attention on the person as the action moves for ThrowDiscus, while RSTAN focuses on sky, grass, and ground that are intuitively irrelevant to this action. For action Dribble, our models pay attention to the basketball and person more precise compared with RSTAN. These illustrate that our model can capture more relevant information.

V. Conclusion

In this article, we propose a novel global and local knowledge-aware attention network for action recognition. The experimental results demonstrate that our proposed model can make full use of the implicit complementary advantages of global and local information and outperforms the state-of-the-art methods on three challenging benchmarks, i.e., Kinetics, UCF101, and HMDB51.

Acknowledgment

Code is available at https://github.com/ZhenxingZheng/attention-network.

References

[1] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[2] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.

[3] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 9–16.

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[6] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 408–417.

[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.*, Jun. 2014, pp. 1725–1732.

[8] 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)*, Jun. 2016, pp. 1933–1941.

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

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[11] W. Kay *et al.*, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[12] J. R. Anderson, *Cognitive Psychology and Its Implications*. New York, NY, USA: Worth, 1985.

[13] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, vol. 1, no. 6, pp. 1097–1105.

[15] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[16] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6016–6025.

[17] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2904–2913.

[18] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7834–7843.

[19] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2613–2622, Dec. 2017.

[20] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose MoTion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033.

[21] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[22] H. Liu, N. Shu, Q. Tang, and W. Zhang, "Computational model based on neural network of visual cortex for human action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1427–1440, May 2018.

[23] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.

[24] T. Ergen and S. Serdar Kozat, "Efficient online learning algorithms based on LSTM neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3772–3783, Aug. 2018.

[25] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[26] J. Yue-Hei Ng, M. 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)*, Jun. 2015, pp. 4694–4702.

[27] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 510–514, Apr. 2017.

[28] Y. Yuan, Y. Zhao, and Q. Wang, "Action recognition using spatial-optical data organization and sequential learning framework," *Neurocomputing*, vol. 315, pp. 221–233, Nov. 2018.

[29] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2166–2175.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[31] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[32] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[33] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[34] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[35] 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.

[36] M. Zhang, Y. Yang, Y. Ji, N. Xie, and F. Shen, "Recurrent attention network using spatial-temporal relations for action recognition," *Signal Process.*, vol. 145, pp. 137–145, Apr. 2018.

[37] 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.

[38] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 373–389.

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

[40] 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.*, Jun. 2009, pp. 248–255.

[41] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, "Simple recurrent units for highly parallelizable recurrence," in *Proc. Empirical Methods Nat. Lang. Process.*, 2018, pp. 4470–4481.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[43] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[44] 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.*, Nov. 2011, pp. 2556–2563.

[45] 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: http://arxiv.org/abs/1212.0402

[46] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[47] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 713–730.

[48] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.

[49] J. Zhu, Z. Zhu, and W. Zou, "End-to-end video-level representation learning for action recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 645–650.

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

[51] H. Yang *et al.*, "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognit.*, vol. 85, pp. 1–12, Jan. 2019.

[52] L. Zhang, X. Zhen, L. Shao, and J. Song, "Learning match kernels on Grassmann manifolds for action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 205–215, Jan. 2019.

[53] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu, and Q. Tian, "Pooling the convolutional layers in deep ConvNets for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1839–1849, Aug. 2018.

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

[55] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition," *Pattern Recognit.*, vol. 92, pp. 13–24, Aug. 2019.

[56] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[57] M. Tong, M. Zhao, Y. Chen, and H. Wang, "D3-LND: A two-stream framework with discriminant deep descriptor, linear CMDT and non-linear KCMDT descriptors for action recognition," *Neurocomputing*, vol. 325, pp. 90–100, Jan. 2019.

[58] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 773–786, Mar. 2019.

[59] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3165–3174.

[60] C. Li *et al.*, "Deep manifold structure transfer for action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4646–4658, Sep. 2019.

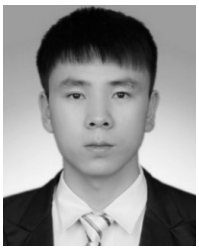
[61] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7445–7454.

[62] W. Huang *et al.*, "Toward efficient action recognition: Principal back-propagation for training two-stream networks," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1773–1782, Apr. 2019.

[63] X. Chen, J. Weng, W. Lu, J. Xu, and J. Weng, "Deep manifold learning combined with convolutional neural networks for action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 3938–3952, Sep. 2018.

[64] Y. Sun, X. Wu, W. Yu, and F. Yu, "Action recognition with motion map 3D network," *Neurocomputing*, vol. 297, pp. 33–39, Jul. 2018.

[65] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1390–1399.
[66] Z. Zheng, G. An, D. Wu, and Q. Ruan, "Spatial-temporal pyramid based convolutional neural network for action recognition," *Neurocomputing*, vol. 358, pp. 446–455, Sep. 2019.
[67] B. Xu, H. Ye, Y. Zheng, H. Wang, T. Luwang, and Y.-G. Jiang, "Dense dilated network for video action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4941–4953, Oct. 2019.
[68] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.
[69] D. Roy, K. S. R. Murty, and C. K. Mohan, "Unsupervised universal attribute modeling for action recognition," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1672–1680, Jul. 2019.
[70] 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: http://arxiv.org/abs/1607.06416



Zhenxing Zheng received the B.S. degree from the Hebei University of Science and Technology, Shijiazhuang, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China.

His main research interests are in computer vision, pattern recognition and image processing, in particular focusing on action recognition.

Gaoyun An (Member, IEEE) received the B.S. degree in biological engineering and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2003 and 2008, respectively.

He was a Visiting Scholar with the University of Florida, Gainesville, FL, USA, from 2017 to 2018. He is currently an Associate Professor and the Doctorate Supervisor with the Institute of Information Science, Beijing Jiaotong University. His research interests include image processing, deep learning, computer vision, and pattern recognition.

Dapeng Wu (Fellow, IEEE) received the B.E. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1990, the M.E. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1997, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2003.

He is currently a Professor with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. His research interests include networking, communications, signal processing, computer vision, machine learning, smart grid, and information and network security.

Qiuqi Ruan (Senior Member, IEEE) received the B.S. and M.S. degrees from Northern Jiaotong University, Beijing, China, in 1969 and 1981, respectively.

From January 1987 to May 1990, he was a Visiting Scholar with the University of Pittsburgh, Pittsburgh, PA, USA, and also with the University of Cincinnati, Cincinnati, OH, USA. He is currently a Professor and the Doctorate Supervisor with the Institute of Information Science, Beijing Jiaotong University, Beijing. He is the IEEE Beijing Section Chairman. His main research interests include digital signal processing, computer vision, pattern recognition, and virtual reality.