IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2019, accepted January 12, 2020, date of publication January 20, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968054

Context-Aware Cross-Attention for Skeleton-Based Human Action Recognition

YANBO FAN[1], SHUCHEN WENG[2], YONG ZHANG[1], BOXIN SHI[2], AND YI ZHANG[3]
[1]Tencent AI Lab, Shenzhen 518057, China
[2]National Engineering Laboratory for Video Technology, Department of Computer Science, Peking University, Beijing 100871, China
[3]College of Intelligence and Computing, Tianjin University, Tianjin 300072, China

Corresponding author: Yong Zhang (zhangyong201303@gmail.com)

This work was supported in part by the Tencent AI Lab, and in part by the National Science Foundation of China (NSFC) under Grant 61702359.

ABSTRACT Skeleton-based human action recognition is becoming popular due to its computational efficiency and robustness. Since not all skeleton joints are informative for action recognition, attention mechanisms are adopted to extract informative joints and suppress the influence of irrelevant ones. However, existing attention frameworks usually ignore helpful scenario context information. In this paper, we propose a cross-attention module that consists of a self-attention branch and a cross-attention branch for skeleton-based action recognition. It helps to extract joints that are not only more informative but also highly correlated to the corresponding scenario context information. Moreover, the cross-attention module maintains input variables' size and can be flexibly incorporated into many existing frameworks without breaking their behaviors. To facilitate end-to-end training, we further develop a scenario context information extraction branch to extract context information from raw RGB video directly. We conduct comprehensive experiments on the NTU RGB+D and the Kinetics databases, and experimental results demonstrate the correctness and effectiveness of the proposed model.

INDEX TERMS Action recognition, cross-attention, context information.

I. INTRODUCTION

Human action recognition is a fundamental and challenging research problem in computer vision [1]–[8]. The performance of human action recognition has an important influence on many other tasks like video understanding and video surveillance. Many works have been proposed with different input modalities, including RGB video [1], [2], [9], [10], optical flow [4], [11] and human 2D/3D skeletons [8], [12], [13] (the optical flow and human skeletons can be estimated directly from the RGB video). Comparing to RGB video and optical flow, skeleton data is computationally more efficient and is robust to the variations in clothing and illumination. With the development of depth sensors like Kinetic [14] and pose estimation technique [15], [16], skeleton-based human action recognition receives more and more attention recently [6], [17]–[19].

As shown in Fig. 1, human actions can be represented by a sequence of skeleton joints. It is well studied that for a certain action, different joints may contain different information and should have different influences for action recognition [3], [5], [6], [8]. For example, when performing *snatch weight lifting*, the movements of *arms* may be more informative than that of *feet*, the influence of joints w.r.t. *arms* is more important accordingly. Yet the informativeness degree of joints may also vary over frames for a certain sequence. To address these, attention mechanisms [20] have been incorporated into skeleton-based action recognition to adaptively assign different weights to different joints. For example, Yang *et al*. proposed a convolutional neural network (CNN) based attention architecture to focus on the informative joints and frames [8]. Liu *et al*. proposed a global context-aware framework to increase the attention capability of the LSTM model [3]. However, these existing works usually learn joints' attention from skeleton modality only. Seen from Fig. 1, being a high-level abstraction of human action, the skeleton data may ignore the helpful scenario context information, such as *the action is performed at a weightlifting venue with a barbell alongside*. Yet human actions are often closely related to such context information. For example, one is more likely to do *snatch weight lifting* than *playing football* at *a weightlifting venue and a barbell alongside*, and the skeleton

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

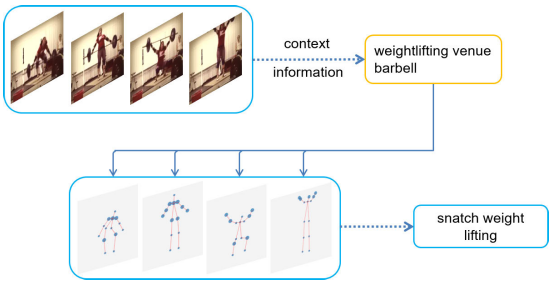


FIGURE 1. An illustration of the sequence of RGB frames and their corresponding skeleton joints for action *snatch weight lifting* (the sample is selected from the Kinetics [1] database). The circle sizes of joints indicate their attention weights for action recognition. The scenario context information extracted from RGB frames benefits the learning of attentions for skeleton joints.

joints w.r.t. *arms* should be more informative than that of *feet* with high probability accordingly. The scenario context information can help us to learn attentions for skeleton joints.

In this paper, we propose a cross-attention module for skeleton-based human action recognition. The joints attention in the proposed cross-attention module consists of two branches: self-attention branch and cross-attention branch. The self-attention branch measures the informativeness degree of each joint and is learned from the skeleton representations directly. The cross-attention branch measures the relevancy between joints and the scenario context information. By combining the self-attention branch and the cross-attention branch, our cross-attention module pursues to extract skeleton joints that are not only more informative but also highly related to its context information, which in consequence helps to learn more discriminative features for action recognition. Besides, the cross-attention module is designed to maintain the input variables' size such that it can be flexibly embedded into many existing frameworks without breaking their structures. We further develop two instantiations for the cross-attention branch and conduct a comprehensive ablation study to investigate their behaviors. To facilitate end-to-end training, we propose a flexible and lightweight scenario context extraction branch to learn context information from raw RGB video directly. Finally, the whole framework is evaluated on the NTU RGB+D and Kinetics databases and experimental results demonstrate its correctness and effectiveness.

The main contributions are three-fold. (1) We propose a context-aware cross-attention module that considers the helpful scenario context information. It helps to extract joints that are not only more informative but also highly related to the context information. (2) We develop a lightweight context information extraction branch that enables end-to-end training of the whole model. (3) We conduct experiments on the NTU RGB+D and the Kinetics databases and experimental results demonstrate the effectiveness of the proposed cross-attention module.

II. RELATED WORK

A. SKELETON-BASED ACTION RECOGNITION

Skeleton-based human action recognition has received more and more attention [5], [6], [12], [17], [21]–[25]. Various handcrafted features have been proposed, for example, the covariance matrices of joint trajectories [26], the pairwise relative position features [27] and the histograms of 3D joints locations [28]. Recently, the great success of deep neural networks has also encouraged its implementation in skeleton-based action recognition. For example, recurrent neural networks (RNN) are widely adopted due to its power in modeling the temporal relations [3], [5], [29]. Some works also transform the sequence of skeleton joints into 2D arrays and utilize the 2D convolutional kernels to extract the spatial and temporal relations [8], [17]. However, these works usually need predefined transformation orders that may break the graph properties of human skeletons.

Recently, Graph Convolutional Neural Networks (GCNs) are proposed to generalize the 2D convolutional kernel from grid-like structured data such as images to arbitrary graph structures [30]. While human skeleton data forming a natural graph with its node being joints and edges being body bones, Yan *et al.* [13] extended the ideas of GCNs to skeleton-based action recognition and propose the ST-GCN model. ST-GCN represents the sequence of skeleton joints into a spatio-temporal graph. By performing graph convolutions on the constructed spatio-temporal graphs directly, ST-GCN obtains better representational ability and achieves promising results for action recognition. Following [13], [31] further developed a graph edge convolutional neural networks to learn complementary information form body bones to the skeleton joints, and propose hybrid neural networks to combine graph node convolutional neural networks and graph edge convolutional neural networks. However, these works treat joints in an equal manner and ignore the fact that different joints and frames may have different influences for certain action recognition.

B. ATTENTION MECHANISMS

Our work is also motivated by the success of attention mechanisms [20]. The attention modules allow the networks to adaptively focus on informative response and have been successfully implemented in many applications, including machine translation [20], image caption [32] and action recognition [3], [8], [33]. Though human actions can be represented by a sequence of 2D/3D skeleton joints, the informativeness degree of different joints may vary for action recognition. For example, when performing waving hands, the skeleton joints with respect to hands may be more informative than that of legs. The attention strategy has also been implied in skeleton-based action recognition to help the network to focus on more informative joints. For example, Yang *et al.* transformed the sequences of skeleton joints into 2D arrays and developed a global long-sequence attention network based on 2D convolutional kernels to extract the key joints and frames [8]. Liu *et al.* proposed a global context-aware attention module to promote the selective attention of LSTM networks [3]. Song *et al.* proposed a spatial attention module to adaptively select

the informative joints and a temporal attention module for keyframes extraction [5].

Although the attention mechanisms have been used for skeleton-based action recognition, our cross-attention module is significantly different from existing works. In existing works, joints attention is usually learned from skeleton data only and ignores the scenario context information that is helpful to action recognition. For example, in an indoor situation, it is more likely to perform eating or reading rather than playing football. Thus the skeleton joints with respect to the upper body should be more informative with high probability. Our proposed cross-attention module consists of a self-attention branch that learned from skeleton representation and a cross-attention branch that incorporates the scenario context information. It helps to extract joints that are not only important to skeleton representations but also highly relevant to the context information.

C. 3D CONVOLUTIONAL NEURAL NETWORKS

Deep neural networks based on 2D convolutional kernels have achieved great success in learning discriminative features or encoding the context information from raw RGB images [34]. Yet it has limitations in extracting the temporal information in sequence data such as videos. Recently 3D convolutional kernels become more and more popular in action recognition. Comparing to 2D CNNs, 3D CNNs can extract the spatio-temporal context information directly from raw videos [1], [2], [7]. For example, [1] proposed the I3D model that inflates the pretrained 2D kernels into 3D kernels. Reference [2] further explored that the pretrained 3D CNNs models on the Kinetics database can help to finetune the models on relative small databases like HMDB [35] and UCF101 [36].

III. CROSS-ATTENTION MODULE

In this section, we introduce the proposed cross-attention module for skeleton-based action recognition. We first briefly review a base two-branch self-attention module, then we propose a flexible cross-attention module to embed the context information. Finally, we develop two instantiations for the proposed cross-attention branch.

A. SELF-ATTENTION MODULE

We first review a base two-branch self-attention module [37], which includes a mask branch and a trunk branch. The trunk branch performs necessarily transformations on the input features to further increase its representational ability. While the mask branch aims to learn a mask from input representations to weight the output of the trunk branch. These two branches usually have the same output size. For example, given a sequence of skeleton joints with T frames and each frame contains N joints, let $\mathbf{v} \in \mathbb{R}^{T\times N\times d}$ be the input representation of all joints for a certain layer, where d is the number of input channels. As shown in Fig. 2, the two-branch self-attention

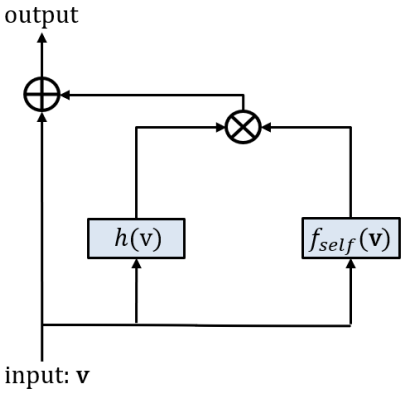


FIGURE 2. Self-attention module. ⊗ denotes element-wise product and ⊕ is element-wise sum. v represents the input skeleton representations. h(v) is the truck branch and set to be the identity mapping in our experiments. f_{self}(v) is the mask branch, implementation details can be found in section III-A.

module can be described as

$$\mathbf{y} = f_{self}(\mathbf{v}) * h(\mathbf{v}), \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{T\times N\times C}$ is the output representation and C is the number of output channels, $*$ denotes element-wise product. Function h represents the transformation operator in the trunk branch and $h(\mathbf{v}) \in \mathbb{R}^{T\times N\times C}$, for example, it can be linear transformations or just identity mapping. $f_{self}(\mathbf{v}) \in \mathbb{R}^{T\times N\times C}$ represents the mask branch that performs as a control gate for the output of the trunk branch, it is realized via 1×1 convolution in our experiments. A residual connection "$+\mathbf{v}$" can be added to $\mathbf{y}$ to maintain the original behaviors.

B. CROSS-ATTENTION MODULE

The mask branch of the base two-branch self-attention module in eq. (1) is learned from skeleton data only. However, as analyzed above, being a high-level abstraction of human actions, skeleton data usually ignores the helpful scenario context information. However, human actions usually have close relations to their scenario context information, for example, one is more likely to perform *snatch weight lifting* in *a weightlifting venue with a barbell alongside* than *playing football*. As different actions usually have different subsets of informative joints, the scenario context information can help to learn attentions for skeleton joints.

To utilize the complementary scenario context information, we propose a cross-attention module for skeleton-based action recognition. Let $\mathbf{a} \in \mathbb{R}^{k\times m}$ denote the context information w.r.t. t-th frame, and $\mathbf{A} \in \mathbb{R}^{T\times k\times m}$ be the context information of the video with T frames, the proposed cross-attention module is designed as

$$\mathbf{y} = g\left(f_{self}(\mathbf{v}), f_{cross}(\mathbf{v}, \mathbf{A})\right) * h(\mathbf{v}), \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^{T\times N\times C}$ is the output representation and C is the number of output channels. $f_{self}(\mathbf{v}) \in \mathbb{R}^{T\times N\times C}$ denotes the self-attention branch that learned from skeleton data only and $h(\mathbf{v})$ represents the necessary transformations in the trunk

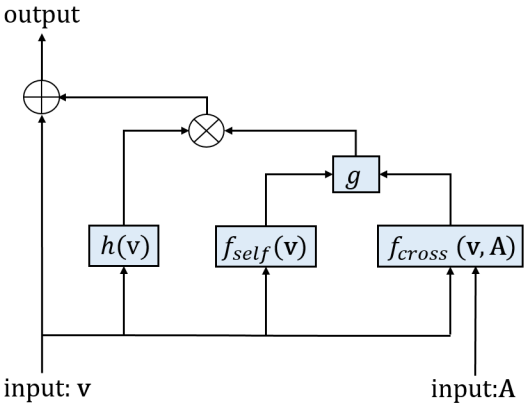


FIGURE 3. Cross-attention module. ⊗ denotes element-wise product and ⊕ is element-wise sum. **v** represents the input skeleton representations and **A** is the corresponding scenario context information representation. h(**v**) is the truck branch and set to be the identity mapping in our experiments. f_{self}(**v**) and f_{cross}(**v**, **A**) are self-attention branch and cross-attention branch, respectively. Implementation details can be found in section III-B.

branch. $f_{cross}(\mathbf{v}, \mathbf{A}) \in \mathbb{R}^{T\times N\times C}$ represents the cross-attention branch that measures the relevancy between joints representations **v** and scenario context information **A**. The joints that are closely related to the context information should be assigned with high scores of $f_{cross}(\mathbf{v}, \mathbf{A})$ and vice-versa. Operator g unifies these two types of attention, for example, through the element-wise average or product.

Comparing to the base two-branch self-attention module in eq. (1), by combining the self-attention branch $f_{self}(\mathbf{v})$ and the cross-attention branch $f_{cross}(\mathbf{v}, \mathbf{A})$, our cross-attention module (2) aims to put more attention to joints that are not only more informative in the perspective of skeleton representation but also has a close relation to the context information. It aims to suppress the influence of joints that are less relevant to the context information. Fig. 3 plots the structure of our cross-attention module, which contains a truck branch h, a self-attention branch f_{self}, and a cross-attention branch f_{cross}. The truck branch h is set to identity mapping in our experiments to reduce the complexity. We also add a residual connection such that it can be incorporated into many existing architectures without breaking their initial behaviors.

C. INSTANTIATIONS OF CROSS-ATTENTION BRANCH

The cross-attention branch $f_{cross}(\mathbf{v}, \mathbf{A})$ measures the relevancy between joints representation **v** and scenario context information **A**, it plays an important role in our cross-attention module. In this section, we describe two instantiations. For simplicity, we illustrate the calculation of relevancy w.r.t. the joints in the t-th frame.

1) DOT PRODUCT

In t-th frame, the relevancy of i-th joint $v \in \mathbb{R}^d$ to its scenario context information $\mathbf{a} \in \mathbb{R}^{k\times m}$ is calculated by

$$f_{cross}(v, \mathbf{a}) = \sigma\left(\frac{1}{k}\sum_{j=1}^{k}\theta(v)^{\top}\phi(\mathbf{a}_j)\right), \tag{3}$$

where $\mathbf{a}_j \in \mathbb{R}^m, j \in \{1, \cdots, k\}$ represents the j-th context information, $\theta : \mathbb{R}^d \to \mathbb{R}^z$ is the embedding operator on joints representation and $\phi : \mathbb{R}^m \to \mathbb{R}^z$ denotes the embedding operator on context information. The relevancy between the v-th joint and the j-th context representation is calculated by the dot-product in the embedded space. σ is designed to be *sigmoid* activation to normalize the relevancy value to [0, 1]. Besides, the *average* operator in eq. (3) can also be replaced by max, however, they perform similar according to our experiments. The embedding functions θ and ϕ are simply designed to be linear transformations, *i.e.*, 1×1 convolution in our experiments. The dimension z is set to $\lceil (d + m)/2 \rceil$ in our experiments.

2) CONCATENATION

Another realization is based on feed-forward network on feature concatenation, *i.e.*,

$$f_{cross}(v, \mathbf{a}) = \sigma\left(\frac{1}{k}\sum_{j=1}^{k}\psi([v, \mathbf{a}_j])\right), \tag{4}$$

where $[\cdot, \cdot]$ represents feature concatenation and $\psi : \mathbb{R}^{(d+m)} \to \mathbb{R}$ denotes the feed-forward network that learns the relevancy between joint v and context $\mathbf{a}_j$, σ is *sigmoid* activation. We didn't include the embedding functions θ and ϕ in eq. (4) to reduce its complexity. For the feed-forward network ψ, we experimentally try one-layer MLP and two-layer MLP (in which the number of hidden units is set to $\lceil (d + m)/2 \rceil$). Experimental results and detail comparisons can be found in section V.

IV. CROSS-ATTENTION BASED ACTION RECOGNITION

The proposed cross-attention module maintains the input variables' size and can be flexibly combined with many existing skeleton-based action recognition networks. To investigate its behaviors, in this section we describe a specific instantiation based on the recently proposed spatio-temporal graph convolution networks (ST-GCN) [13]. We also analyze the design of the scenario context information extraction branch to facilitate end-to-end training.

A. REVIEW OF ST-GCN

We first briefly review the framework of ST-GCN [13] for skeleton-based action recognition. Given a sequence of human skeleton data with T frames and N joints in each frame, a nature graph $G = (V, E)$ can be constructed with its vertices being the joints, *i.e.*, $V = \{v_{tj} | t = 1, \cdots, T, j = 1, \cdots, N\}$, v_{tj} denotes the representation of j-th joint in t-th frame. The edge set E usually contains two types of edges: the intra-frame edges that constructed based on natural connections between body joints in each frame, and the inter-frame edges that connect the same joints between consecutive frames. After constructing graph G, multiple spatio-temporal graph convolutional layers are applied in sequence to extract discriminative features for action recognition. Table 1 shows the overall network

TABLE 1. The baseline ST-GCN network [13] with 9 layers, where T is the number of frames for each sample and N is the number of skeleton joints in each frame. The input size is 16 × 18 with each joint represented by a 3-dimensional vector. The strides of temporal convolutions at gcn3 and gcn6 are set to be 2 for down-sampling. Fully connected layer and softmax are performed on the 256-d output of the global average pooling layer for classification.

layer		output size $(T \times N)$
gcn0-2	$\begin{bmatrix} 1 \times 3, & 64 \\ 9 \times 1, & 64 \end{bmatrix}$	16×18
gcn3-5	$\begin{bmatrix} 1 \times 3, & 128 \\ 9 \times 1, & 128 \end{bmatrix}$	8×18
gcn6-8	$\begin{bmatrix} 1 \times 3, & 256 \\ 9 \times 1, & 256 \end{bmatrix}$	4×18
global average pool		1×1

structures of ST-GCN, for each layer, the spatial convolution is done by graph convolution with kernel size 3×1 while that for temporal convolution is 1×9. We refer to [13] for implementation details about graph convolution.

As the spatial size (*i.e.*, the number of joints in each frame) keeps unchanged at different layers of ST-GCN, it is convenient for us to extract the learned representation of each joint at different layers. More importantly, the proposed cross-attention module in eq. (2) maintains the input variables' size and thus can be added between any layers of ST-GCN. We experimentally compare the performance while inserting the cross-attention module into different layers.

B. CONTEXT INFORMATION EXTRACTION

Another important part of the cross-attention module is the representation of scenario context information. To facilitate end-to-end training, in this work, we propose to extract scenario context information from raw RGB video directly. Specifically, let $I \in \mathbb{R}^{T \times w \times h \times 3}$ be the input of raw RGB video with T frames, $(w, h, 3)$ are the width, height and number of channels for each frame, respectively. The structure of the scenario context information extraction branch is given in Table 2. It consists of 3D convolution and residual connections. The learned context information is represented by a 512-d feature, it is further used to learn a relevancy score for different joints, as described in section III-B. The context information extraction branch, the cross-attention module as well as the graph convolution network are learned jointly in an end-to-end manner.

TABLE 2. Scenario context information extraction branch. The input size is 16 × 112 × 112. 3D convolutional kernels are adopted, and residual blocks are shown in brackets. The strides at layer2_1 are set to be (2, 2, 2) to perform down-sampling.

layer			output size $(T \times H \times W)$
layer0	$7 \times 7 \times 7$, 64, strides 1, 2, 2		$16 \times 56 \times 56$
pool	$3 \times 3 \times 3$, strides, 2, 2, 2		$8 \times 28 \times 28$
layer1_x	$\begin{bmatrix} 1 \times 1 \times 1, & 64 \\ 3 \times 3 \times 3, & 64 \\ 1 \times 1 \times 1, & 256 \end{bmatrix}$	$\times 3$	$8 \times 28 \times 28$
layer2_x	$\begin{bmatrix} 1 \times 1 \times 1, & 128 \\ 3 \times 3 \times 3, & 128 \\ 1 \times 1 \times 1, & 512 \end{bmatrix}$	$\times 4$	$4 \times 14 \times 14$

Besides, although we incorporate the RGB video into the learning of cross-attention module, our implementation is significantly different from existing RGB video-based action recognition methods: (1) we aim to extract context information with a lightweight and relatively shallow network (shown in Table 2) from RGB video to promote the cross-attention module. The extracted 512-d context representation from RGB video is only used to learn a scalar attention weight for each joint. In comparison, existing RGB-based action recognition methods [2], [10] focus on learning high-level representative features from RGB video, which is used directly for action recognition. For example, the C3D model in [10] extracted a 4096-d feature with a complicated 3D neural network, and directly fed the 4096-d feature into a softmax layer for action recognition. (2) Comparing to the existing RGB-based action recognition networks, our context information extraction branch is very shallow and lightweight. The number of model parameters of C3D [10] is 25 times larger than that of our context extraction branch. The performance of directly applying our context extraction branch for action recognition is very poor. We conduct a comprehensive ablation study to verify the effectiveness of the proposed cross-attention module in the experimental part.

V. EXPERIMENTS

In this section, we conduct experiments on the NTU RGB+D database [38] and the Kinetics database [1] to verify the effectiveness of the proposed model.

A. DATABASE

1) KINETICS [1]

The DeepMind Kinetics human action database contains 400 human action classes and around 300,000 videos in total, each action has at least 400 video clips. The video clips are taken from YouTube video and each clip lasts around 10s. We use the provided training set with 240,000 samples for training and evaluate model performance on the evaluation set with 20,000 samples. Besides, the Kinetics database only provides raw RGB video without skeleton data. To perform skeleton-based action recognition, [13] estimated its skeleton information through the Openpose [15] toolbox. For a fair comparison, here we also adopt their released skeleton data for training. Specifically, the skeleton of each person is represented by 18 joints, and each joint is encoded by (x, y, c) with (x, y) being its 2D coordinates and c being the confidence score generated by Openpose. For multiple person situations, the skeletons of two person with the highest average joint confidence scores are recorded.

2) KINETICS-MOTION

The Kinetics-Motion database is a subset of Kinetics [1] adopted by [13]. It contains 30 classes that strongly related to body motions. The selected classes are *belly dancing, punching bag, capoeira, squat, windsurfing, skipping rope, swimming backstroke, hammer throw, throwing discus, tobogganing, hopscotch, hitting baseball, roller skating, arm wrestling, snatch weight lifting, tai chi, riding mechanical bull, salsa dancing, hurling (sport), lunge, skateboarding, country line dancing, juggling balls, surfing crowd, dead lifting, clean and jerk, crawling baby, push up, front raises, pull ups*.

3) NTU RGB+D [38]

The NTU RGB+D database is captured in a controlled environment. It contains 56,000 videos in 60 categories, and both RGB video and 3D skeleton joints annotations are provided. It contains 25 major body joints. The skeleton joints information is collected by kinetic depth sensors. There are two standard benchmarks on this database: **cross-subject** and **cross-view**. For the cross-subject benchmark, the training and testing sets contain 40,320 and 16, 560 videos, respectively. For the cross-view benchmark, the training and testing sets contain 37,920 and 18,960 videos, respectively. Following common practice [13], we conduct experiments on these two benchmarks and report their top-1 classification accuracy.

B. IMPLEMENTATION

1) TRAINING

Our model is trained with input skeleton clips of 16 frames. For each sequence of skeleton joints with T frames in total, we first randomly select a temporal position and then crop 16 consensus frames around it for training. To extract context information from RGB video, we also generate the corresponding RGB clips with 16 frames for each skeleton clip. For each RGB clip, following [2], we further perform random cropping from the 4 corners or center position, it is then spatially resized to 112×112 and feed into networks. The whole model is trained end-to-end with standard cross-entropy loss and the stochastic gradient descent optimizer. The learning rate is set to be 0.1 and is reduced by a factor of 10 every 80 epochs. The model is trained for 260 epochs with a weight decay of 10^{-3} and a momentum of 0.9. The model is trained on an 8-GPU machine with batchsize of 128.

2) INFERENCE

Following [2], we perform sliding window to generate input clips from each test sample, *i.e.*, its skeleton data as well as RGB video are split into non-overlapped 16 frame clips. Each RGB clip is further spatially cropped around center position to a size of 112×112. The skeleton clips and its corresponding RGB data are feed into the network to generate the class scores. Fig. 4 illustrates the inference procedure of one input clip. The final prediction is the averaged class scores of all clips.

Input: a clip of 16-frames with $\mathbf{x}$ be the RGB frames and $\mathbf{s}$ be the corresponding skeleton information.

procedure INFERENCE($\mathbf{x}$, $\mathbf{s}$)
  $\mathbf{x}$ = Resize(Centercrop($\mathbf{x}$))
  $\mathbf{x}$ = ContextExtraction($\mathbf{x}$)
  $\mathbf{s}$ = $\text{GCN}_{0-6}(\mathbf{s})$
  $\mathbf{s}$ = CrossAttention($\mathbf{x}$, $\mathbf{s}$)
  $\mathbf{s}$ = $\text{GCN}_{7-8}(\mathbf{s})$
  $\mathbf{y}$ = FC(Pooling($\mathbf{s}$))
end procedure

Output: prediction $\mathbf{y}$

FIGURE 4. Inference procedure of the proposed model, where one cross-cross attention module is added after layer gcn6. The "GCN" network is given in Table 1 and the "ContextExtraction" network is given in Table 2.

C. ABLATION STUDY ON KINETICS-MOTION

We first conduct a comprehensive ablation study on the relatively small database Kinetics-Motion.

1) CROSS-ATTENTION BRANCH

We conduct experiments to verify the performance of different realizations of the cross-attention branch, including dot-product, one-layer MLP and two-layer MLP that described in Section III-C. For a fair comparison, we insert one cross-attention module after gcn6 of the ST-GCN network in Table 1, the operator g in eq. (2) is set to be *average* (*i.e.*, $g = (f_{self} + f_{cross})/2$).

TABLE 3. Classification accuracy (%) of different attention modules on the Kinetics-Motion database. The best results are shown in bold.

Method	top-1	top-5
Baseline	81.22	95.42
Self-Attention	82.34	95.32
Dot-Product	84.05	96.77
One-layer MLP	87.35	98.12
Two-layer MLP	**88.16**	**98.59**

The top-1 and top-5 classification accuracy is reported in Table 3. The "Baseline" refers to the ST-GCN model [13] without any attention module (its network structure is given in Table 1); the "Self-Attention" is realized by inserting one two-branch self-attention module in eq. (1) after gcn6 of ST-GCN; the "Dot-Product" denotes the variation that inserting one dot-product based cross-attention module in eq. (3) after gcn6 of ST-GCN; the "One-layer MLP" and "Two-layer MLP" denote the variations that inserting one concatenation based cross-attention module in eq. (4) after gcn6 of ST-GCN. Seen from Table 3, the cross-attention module with two-layer MLP achieves the best performance. Besides, all attention-based models obtain higher top-1 accuracy than the baseline, this illustrates that the attention mechanisms can help the skeleton-based action recognition. Comparing to the results of self-attention, our three cross-attention realizations

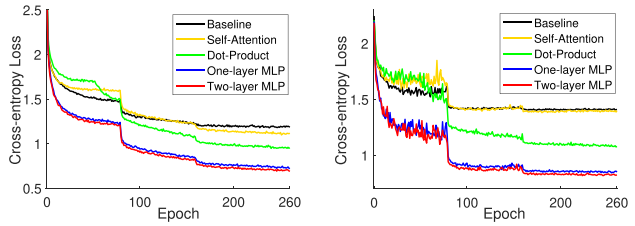


FIGURE 5. Tendency curves of training loss (left plot) and validation loss (right plot) for different models on the Kinetics-Motion database (best viewed in color).

all get better performance. Thus the scenario context information can help skeleton-based action recognition. The performance of two-layer MLP based realizations outperforms that of dot-product and one-layer MLP. Fig. 5 further plots the tendency curves of training loss and validation loss for different methods.

2) OPERATOR g IN CROSS-ATTENTION MODULE

We then investigate different ways to combine the self-attention branch and the cross-attention branch, *i.e.*, the design of operator g in eq. (2). We compare the performance of *average* (*i.e.*, $g = (f_{self} + f_{cross})/2$) and *multiplication* (*i.e.*, $g = \sqrt{f_{self} \times f_{cross}}$). We also report the performance of self-only (*i.e.*, $g = f_{self}$) and cross-only (*i.e.*, $g = f_{cross}$) to better understand their behaviors, note that self-only reduces to the base two-branch self-attention module in eq. (1). We fix the cross-attention branch to be two-layer MLP and insert one cross-attention module after layer gcn6. Table 4 tabulates their classification accuracy. We can observe that cross-only achieves better performance than self-only in terms of both top-1 and top-5 classification accuracy. This illustrates that the learned context information by f_{cross} can better help the model to focus on more informative joints. Combining the self-attention branch and cross-attention, we obtain better performance, and the *average* operator achieves higher classification accuracy than the *multiplication* operator.

TABLE 4. Classification accuracy (%) w.r.t. different operator g on the Kinetics-Motion database. The best results are shown in bold.

Method	top-1	top-5
Baseline	81.22	95.42
Self-only	82.34	95.32
Cross-only	87.82	98.45
Multiplication	87.95	97.98
Average	**88.16**	**98.59**

3) CROSS-ATTENTION AT DIFFERENT GCN LAYERS

By maintaining the input variables' size, the cross-attention module can be flexibly added into any layer of ST-GCN. To better investigate their behaviors, we experiment with one cross-attention module added after layer gcn0, gcn3, gcn6, and gcn8. We also try two cross-attention modules that inserted after layer gcn3 and gcn6 simultaneously (with one for each). To embed scenario context information into different gcn layers, we modify the context information extraction branch in Table 2 to make its output temporal size to be consistent with that of different gcn layers. For example, the temporal stride of layer2_1 is modified to 1 to generate output with $8 \times 14 \times 14$, while the temporal strides of both *pool* layer and layer2_1 are set to 1 to make the output temporal size to be 16.

TABLE 5. Classification accuracy (%) w.r.t. cross-attention module added after different stages of ST-GCN on the Kinetics-Motion database. The best results are shown in bold.

Method		top-1	top-5
Baseline		81.22	95.42
One cross attention module	gcn0	87.69	98.05
	gcn3	87.95	98.18
	gcn6	88.16	**98.59**
	gcn8	84.05	96.84
Two cross attention module	gcn3+gcn6	**88.49**	98.05

Table 5 reports their numerical results, in which the cross-attention branch is realized by the two-layer MLP. We observe that all cross-attention based realizations significant outperform the baseline ST-GCN model, this demonstrates the effectiveness of the proposed cross-attention module. When adding one cross-attention module, due to the powerful skeleton representations extracted by higher gcn layers, the performance consistently improves for gcn0, gcn3, and gcn6. Yet there is a relative performance drop for gcn8. The cross-attention module aims to learn an attention weight for each joint. When a cross-attention module is added after layer gcn8, the learned attention weight influences the following global average pooling layer and the fully connected layer. In comparison, when the cross-attention module is added after an earlier layer such as gcn6, the layer gcn7 and gcn8 will also benefit from the learned cross-attention weight. The implementation with two cross-attention modules achieves the highest top-1 accuracy. Due to time complexity, we didn't test implementations with more cross-attention modules.

To better investigate what we have learned from the cross-attention module, Fig. 6 visualizes the learned attention weights of different skeleton joints on eight classes, in which one cross-attention module is added after layer gcn6. The cross-attention branch is realized by the two-layer MLP. Seen from Fig. 6, the cross-attention module trends to assign more weight to skeleton joints that are highly related to the actions. For example, for the cases of *punching bag* and *hitting baseball* in Fig. 6, the movements of joints w.r.t. the upper body are more significant than that of feet, these joints also obtain more attention weights in our cross-attention module as expected.

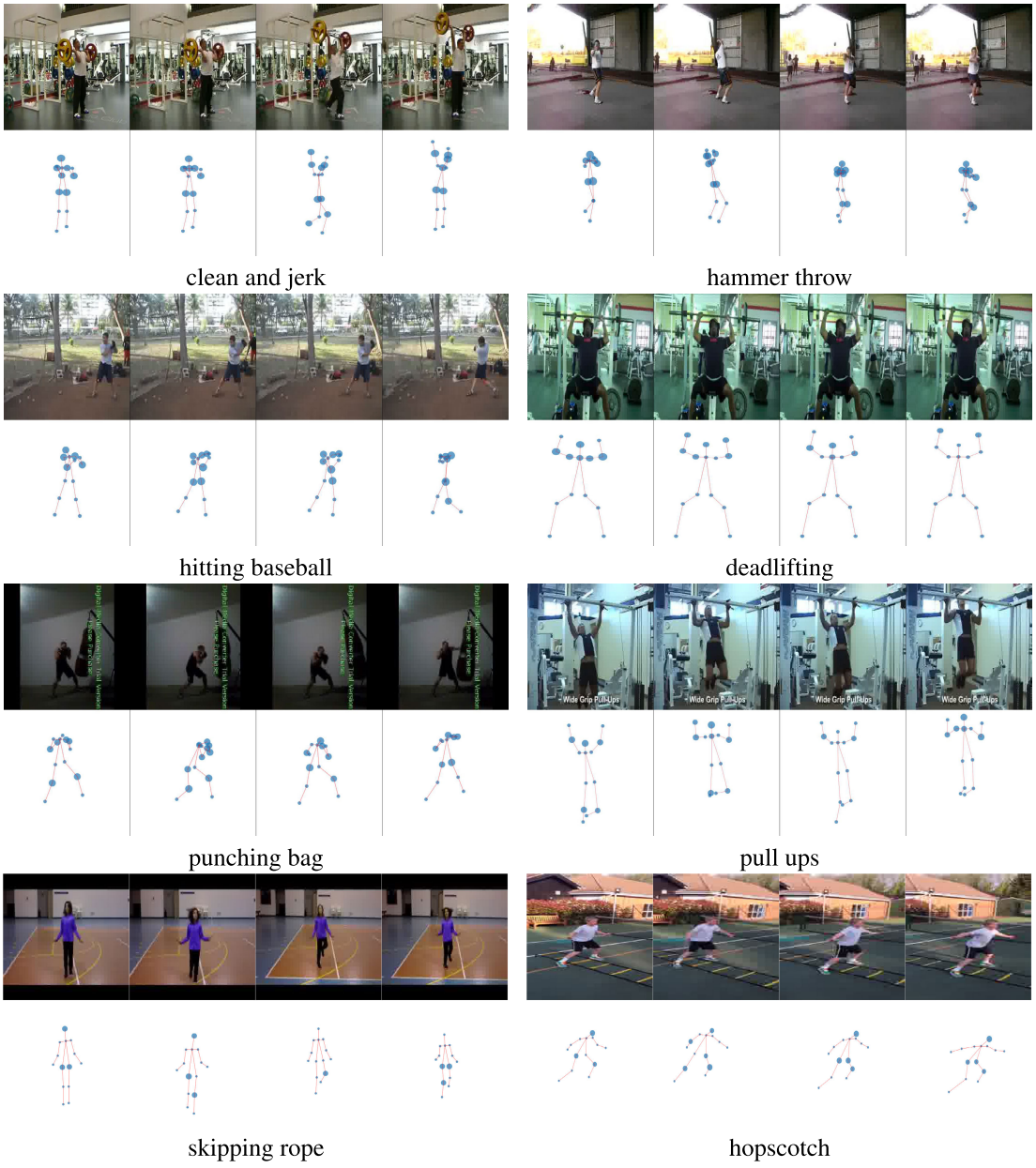


FIGURE 6. Visualization of learned attention weights from the cross-attention module on the Kinetics-Motion database. For each panel, the first row plots raw RGB frames and the second row plots the 18 skeleton joints for the corresponding person in each frame. The circle size around each joint indicates the magnitude of the attention weight learned from the cross-attention module.

TABLE 6. Classification accuracy (%) of different variations on the Kinetics-Motion database. The best results are shown in bold.

Method	Top-1	Top-5
RGB-only	82.77	96.36
Skeleton-only (ST-GCN)	81.22	95.42
RGB+Skeleton (concat)	86.00	97.51
RGB+Skeleton (cross-attention)	**88.16**	**98.59**

4) CONTEXT INFORMATION EXTRACTION

The proposed model consists of three main components, i.e., the context information extraction branch based on RGB video, the baseline ST-GCN branch based on skeleton and the cross-attention module. We conduct an ablation study on several variations to better explore the effectiveness of each component. Table 6 reports the numerical results of different variations. The "Skeleton-only" denotes the variation that only contains the ST-GCN branch (i.e., the baseline ST-GCN [13] given in Table 1). The "RGB-only" denotes the variation that only contains the context information extraction branch given in Table 2. The "RGB+Skeleton (concat)" denotes the variation that contains both branches. It concatenates the 512-d context feature learned from RGB video and the 256-d skeleton feature learned from the skeleton modality. The "RGB+Skeleton (cross-attention)" variation denotes the proposed model that fuses the two branches

TABLE 7. Top-1 classification accuracy (%) of different variations on the NTU RGB+D database. The best results are shown in bold.

Method	Cross-subject	Cross-view
RGB-only	76.30	79.71
Skeleton-only (ST-GCN)	80.05	83.26
RGB+Skeleton (concat)	81.54	84.61
RGB+Skeleton (cross-attention)	**84.23**	**89.27**

TABLE 8. Top-1 classification accuracy (%) on the NTU RGB+D database. The best results are shown in bold.

Method	Cross-subject	Cross-view
Deep LSTM [38]	60.7	67.3
GCA-LSTM network [3]	74.4	82.8
TCN [39]	74.3	83.1
ST-GCN [13]	81.5	88.3
TSSI+SSAN+GLAN [8]	82.4	89.1
Cross-Attention (ours)	**84.2**	**89.3**

via the proposed cross-attention module. Specifically, one two-layer MLP based cross-attention module is added after layer gcn6. A fully connected layer is adopted for each variation for action recognition. For all variations, the context information extraction branch (if included), the ST-GCN branch (if included), the cross-attention module (if included) and the fully connected layer are learned jointly in an end-to-end manner. From the table, the "RGB-only" obtains better performance than the "Skeleton-only". By introducing the RGB video information, the 'RGB+Skeleton (concat)" outperforms both the "RGB-only" and "Skeleton-only" variations. Thus, fusing these two modalities could improve the performance of action recognition. Comparing to the "RGB+Skeleton (concat)", our "RGB+Skeleton (cross-attention)" obtains better performance in terms of both top-1 and top-5 accuracy. The performance improvement is 2.16% and 1.08% in terms of top-1 and top-5 accuracy, respectively. This demonstrates the effectiveness of the proposed cross-attention module.

D. EXPERIMENTS ON NTU RGB+D

We now conduct experiments on the NTU RGB+D database. According to the ablation study in the previous section, we fix the cross-attention branch to be the two-layer MLP and the operator g in eq. (2) is set to be *average*. Due to the time complexity, we only add one cross-attention module after layer gcn6 of ST-GCN. To verify the effectiveness of the proposed cross-attention module, we first evaluate the performance of several variations of the proposed model. The experimental results are given in Table 7. The definition of each variation is given in "Section V-C: Context information extraction". Note that the "Skeleton-only" refers to the baseline ST-GCN model given in Table 1. From the table, we have the following observations: (1) comparing to the "RGB-only", the "Skeleton-only" obtains 3.75% and 3.55% performance improvements on cross-subject and cross-view, respectively. This verifies the good quality of the skeleton modality of the NTU RGB+D database; (2) by concatenating the context information and the skeleton information, the "RGB+Skeleton (concat)" obtains better performance than both "Skeleton-only" and "RGB-only"; (3) the "RGB+Skeleton (cross-attention)" significantly outperforms "RGB+Skeleton (concat)" on both tasks. Its performance improvement is 2.69% and 4.66% on cross-view and cross-subject, respectively. This demonstrates the effectiveness of the proposed cross-attention module. We also compare with several characteristic action recognition methods, their results are given in Table 8. Note that our baseline ST-GCN model and the ST-GCN [13] share the same network structures. The performance gap between our ST-GCN (*i.e.*, the Skeleton-only in Table 7) and ST-GCN in [13] (*i.e.*, the ST-GCN in Table 8) may because of different experimental settings. For example, we adopt a temporal size of 16, while the temporal size in [13] is 300. The results in Table 8 show the effectiveness of the proposed model.

E. EXPERIMENTS ON KINETICS

Now we evaluate the performance of the cross-attention module on the challenging Kinetics database. We fix the cross-attention branch to be the two-layer MLP and adopt the *average* operator for g in eq. (2). The number of model parameters is 8.4 million. The training on the Kinetics takes 66.8 hours for 260 epochs on an 8-GPU machine. We first conduct an ablation study to verify the effectiveness of each component of the proposed model. Table 9 reports classification accuracies of different variations. Please refers to "Section V-C: Context information extraction" for the definition of each variation. From the table, the performance of the "RGB-only" is significantly better than that of the "Skeleton-only". Similar to that on the NTU RGB+D database, by concatenating the learned context feature and the skeleton feature, the "RGB+Skeleton (concat)" outperforms both the "RGB-only" and "Skeleton-only". Moreover, the "RGB+Skeleton (cross-attention)" obtains better performance than the "RGB+Skeleton (concat)" in terms of both top-1 and top-5 accuracy. These demonstrate the effectiveness of the proposed cross-attention module.

TABLE 9. Classification accuracy (%) of different variations on the Kinetics database. The best results are shown in bold.

Method	Top-1	Top-5
RGB-only	32.95	59.85
Skeleton-only (ST-GCN)	25.50	46.20
RGB+Skeleton (concat)	38.43	64.98
RGB+Skeleton (cross-attention)	**39.90**	**66.70**

Table 10 reports the comparisons with several characteristic action recognition methods. As analyzed in Section V-D, the performance gap between our Skeleton-only in Table 9 and ST-GCN [13] in Table 10 is because of different experimental settings. Seen from Table 10, our cross-attention

TABLE 10. Classification accuracy (%) on the Kinetics database. The best results are shown in bold.

Method	top-1	top-5
Feature Enc. [40]	14.9	25.8
Deep LSTM [38]	16.4	35.3
TCN [39]	20.3	40.0
ST-GCN [13]	30.7	52.8
Cross-Attention (ours)	**39.9**	**66.7**

model obtains the best performance on the Kinetics database. These demonstrate the correctness and effectiveness of the proposed cross-attention module.

VI. DISCUSSION

In this paper, we proposed to exploit helpful scenario context information to benefit skeleton-based human action recognition. We presented a novel cross-attention module that helps to extract joints that are not only more informative but also highly related to the scenario context information. We also provided two instantiations of the cross-attention module. In the experiments, we developed a context information extraction branch to extract context information from raw RGB video directly. We conducted comprehensive experiments on the NTU RGB+D database and the Kinetics database. The experimental results demonstrated the effectiveness of the proposed cross-attention module.

There are several interesting questions regarding the cross-attention module. For example, the context information is not limited to the feature that learned from the RGB video. Some other side information, such as video caption, can also be applied to depict the context information. Meanwhile, except for the ST-GCN network, the cross-attention module maintains the input variables' size and can be combined with many other skeleton-based action recognition networks. It is interesting and helpful to investigate the influence of the cross-attention module with different realizations.

REFERENCES

[1] 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.

[2] 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. 18–22.

[3] 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)*, Jul. 2017, pp. 1647–1656.

[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS*, 2014, pp. 568–576.

[5] 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*, 2017, pp. 4263–4270.

[6] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5323–5332.

[7] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, 2018, pp. 7794–7803.

[8] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[9] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[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] D. Zhang, G. Guo, D. Huang, and J. Han, "PoseFlow: A deep motion representation for understanding human behaviors in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6762–6770.

[12] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. ECCV*, 2018, pp. 103–118.

[13] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018, pp. 7444–7452.

[14] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia Mag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[16] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[17] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," Nov. 2017, *arXiv:1711.05941*. [Online]. Available: https://arxiv.org/abs/1711.05941

[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.

[19] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[21] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," *Assoc. Adv. Artif. Intell.*, vol. 33, pp. 8561–8568, Aug. 2019.

[22] W. Li, L. Wen, M. C. Chuah, and S. Lyu, "Category-blind human action recognition: A practical recognition system," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4444–4452.

[23] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[24] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[25] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[26] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. IJCAI*, 2013, pp. 2466–2472.

[27] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.

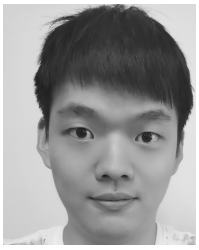
[28] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.

[29] W. Li, L. Wen, M.-C. Chang, S. N. Lim, and S. Lyu, "Adaptive RNN tree for large-scale human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1453–1461.

[30] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018, pp. 1–12.

[31] X. Zhang, C. Xu, and D. Tao, "Graph edge convolutional neural networks for skeleton based action recognition," May 2018, *arXiv:1805.06184*. [Online]. Available: https://arxiv.org/abs/1805.06184
[32] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
[33] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," Nov. 2015, *arXiv:1511.04119*. [Online]. Available: https://arxiv.org/abs/1511.04119
[34] 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.
[35] 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.
[36] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," Dec. 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402
[37] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[38] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.
[39] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1623–1631.
[40] B. Fernando, E. Gavves, M. Jose Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5378–5387.

YANBO FAN received the B.S. degree in computer science and technology from Hunan University, in 2013, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2018. He is currently a Senior Researcher with the Tencent AI Lab. His research interests are in computer vision and machine learning.



SHUCHEN WENG received the B.E. degree from Tianjin University, in 2019. He is currently pursuing the Ph.D. degree with the Camera Intelligence Group, Peking University, under the supervision of Prof. B. Shi. His research interests are in computer vision and machine learning.

YONG ZHANG received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2018. From 2015 to 2017, he was a Visiting Student with the Rensselaer Polytechnic Institute. He is currently with the Tencent AI Lab. His research interests include computer vision, machine learning, and probabilistic graphical models.

BOXIN SHI received the B.E. degree from the Beijing University of Posts and Telecommunications, in 2007, the M.E. degree from Peking University, in 2010, and the Ph.D. degree from the University of Tokyo, in 2013. He did Postdoctoral research at the MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, from 2013 to 2016, and worked as a Researcher at the National Institute of Advanced Industrial Science and Technology, from 2016 to 2017. He is currently a Boya Young Fellow Assistant Professor and a Research Professor with Peking University, where he leads the Camera Intelligence Group. He received the Best Paper Runner-up Award at the International Conference on Computational Photography in 2015. He has served as an Area Chair for ACCV 2018, BMVC 2019, and 3DV 2019.

YI ZHANG received the master's and Ph.D. degrees in computer science from Tianjin University, in 2006 and 2009, respectively. She is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University. Her research interests include object recognition and visual analysis.