Zoom Transformer for Skeleton-based Group Activity Recognition

Jiaxu Zhang, *Student Member, IEEE,* Yifan Jia, Wei Xie, *Member, IEEE,* Zhigang Tu, *Member, IEEE*

***Abstract*—Skeleton-based human action recognition has attracted increasing attention and many methods have been proposed to boost the performance. However, these methods still confront three main limitations: 1) Focusing on single-person action recognition while neglecting the group activity of multiple people (more than 5 people). In practice, multi-person group activity recognition via skeleton data is also a meaningful problem. 2) Unable to mine high-level semantic information from the skeleton data, such as interactions among multiple people and their positional relationships. 3) Existing datasets used for multi-person group activity recognition are all RGB videos involved, which cannot be directly applied to skeleton-based group activity analysis. To address these issues, we propose a novel Zoom Transformer to exploit both the low-level single-person motion information and the high-level multi-person interaction information in a uniform model structure with carefully designed Relation-aware Maps. Besides, we estimate the multi-person skeletons from the existing real-world video datasets *i.e.* Kinetics and Volleyball-Activity, and release two new benchmarks to verify the effectiveness of our Zoom Transfromer. Extensive experiments demonstrate that our model can effectively cope with the skeleton-based multi-person group activity. Additionally, experiments on the large-scale NTU-RGB+D dataset validate that our model also achieves remarkable performance for single-person action recognition. The code and the skeleton data are publicly available at https://github.com/Kebii/Zoom-Transformer.**

***Index Terms*—Activity recognition, Skeleton-based action, Visual transformer, Attention mechanism.**

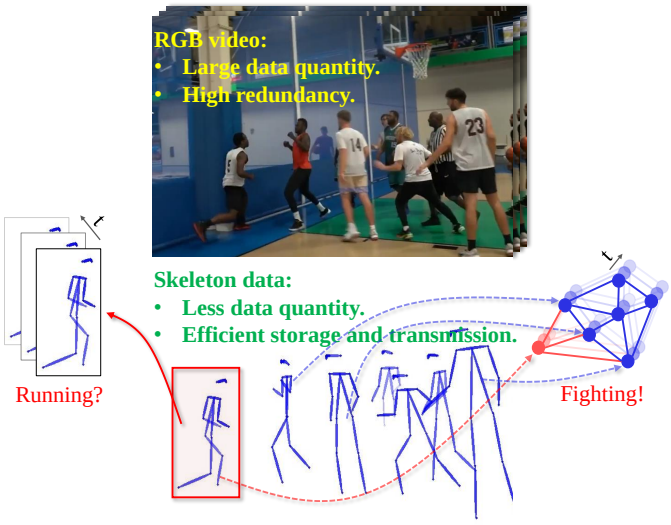

Fig. 1. Group activity in the data forms of RGB and skeleton. The independent action of each actor, the interactions of the actors and the positional relationships among the actors, constitute the complex multi-person group activity – fighting. Compared to the RGB data, the skeleton data is lightweight and can be processed more efficiently. Therefore, how to take full advantage of the skeleton data and explore multi-level relational information for group activity recognition is desired to be investigated.

I. Introduction

ACTION recognition is one of the fundamental problems in the field of computer vision [1], [2], and has a wide range of applications, *e.g.* human behavior analysis [3], intelligent video surveillance [4], [5], human-machine interaction [6], etc. However, as a common recording format of human action, videos have the disadvantages of large data quantity, high redundancy, and inefficient storage and transmission, which limits the application of video-based action recognition. Except for video-based action recognition, recent studies also pay great attention to skeleton-based action recognition as its robustness against changes in camera viewpoints and interference of cluttered backgrounds. The skeleton data is also compact, thus it is suitable for long-term storage and fast transmission. Accordingly, many advanced methods, which focus on skeleton-based action recognition, have been proposed recently [7]–[12]. These methods can extract the spatio-temporal motion information of the skeleton sequence, explore the physical structure of the skeleton graph, and significantly improve the accuracy of action classification. However, the existing works still have the following disadvantages:

(1) The prior researches mainly focus on the skeleton actions of one or two persons [13], but in real-world videos, human activities are often composed of multiple persons. In fact, each individual's independent action, the positional relationships among persons, and their interactions with each other constitute a complicated group activity. Taking a fighting incident from a surveillance video as an example (see Figure 1), in which some people are walking, some are running, and some are pushing. Only if we observe and consider both the multiple individual's actions and their interactions in the video, it can be recognized as the activity "fighting". Due to the lack of scene information in the skeleton data, it becomes more difficult to recognize group activities from the skeleton sequences.

Noticeably, identifying multi-person group activity is as important as recognizing single-person action. E.g., for video surveillance, if the system can discern groups of people

Jiaxu Zhang and Zhigang Tu are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China. (Jiaxu Zhang and Yifan Jia contributed equally and the co-first authors.) (Email: tuzhigang@whu.edu.cn).

Yifan Jia is with the Department of Pain, Renmin Hospital of Wuhan University, Wuhan 430060, China.

Wei Xie is with the School of Computer, Central China Normal University, Wuhan 430079, China.

fighting or being crowded and trampled, it will promptly notify the security to deal with the public safety incident in reality efficiently. Group activity recognition, a very meaningful visual task, is widely investigated in video-based action/event recognition but is rarely explored in skeleton-based action/event recognition. The long-sequence video, which is a primary type of video data, is not suitable for intelligent analysis of group activity because of the high computational complexity of the video-based method. Therefore, the mainstream video-based action recognition methods usually handle the video clips that are generally short in length [14]–[16]. For the long-sequence video clips, existing video-based methods usually only sample 8-64 frames from them. Furthermore, in practical applications, it is also impossible to store the long-sequence video data for a long time, since that requires a huge storage cost. Recently, in some scenarios, there has been a trend to use the skeleton data instead of the video data. For example, motion capture in animation making or human-computer interaction, the skeleton data estimated by the depth camera is more and more widely used than the video data, due to the skeleton data is more compressed and accurate to express human motion. Consequently, how to take full advantage of the lightweight and robust skeleton data is prospective for group activity recognition (see Figure 1).

(2) The existing skeleton-based action recognition methods lack an effective feature extraction model to learn the characteristic of multi-person group activity [17]. For example, they simply utilize a *global-pooling* layer to fuse the motion features of multiple persons. However, this strategy ignores the rich interaction information and positional relation among people. Taking a volleyball game video as an example. In this scenario, each individual may perform different actions. Some are running, some are walking, some are waving, and some are spiking. But their actions together form the activity of playing volleyball. Therefore, how to enable the model has the ability to extract a single person's action feature, as well as to exploit both the interactions and positional relationships among multiple persons, is a core issue for skeleton-based activity recognition.

(3) The current datasets, which were constructed for skeleton-based action recognition, contain two persons at most [8], [13], [15], [18], thus they are unsuitable for group activity recognition. Furthermore, the existing datasets, which are used for multi-person group activity recognition, only involve RGB videos [19], [20], thus they cannot be directly applied to analyze the skeleton-based group activity. Consequently, it is necessary to extract the skeleton data from real-world videos to promote the research and application of skeleton-based multi-person group activity analysis.

To address the above issues, we propose a novel Transformer-based model named Zoom Transformer, and design two skeleton-based group activity benchmarks (i.e. Kinetics-Skeleton-Activity (K-SA) and Volleyball-Skeleton-Activity (V-SA)) derived from the existing video-based datasets, for *skeleton-based multi-person group activity recognition*. Our Zoom Transformer, which consists of two parts, is valid to extract multi-level features of the skeleton sequence hierarchically. Specifically, the Zoom-in Transformer (ZiT) part, which takes the joint features of a single person as input to construct a graph of human body joints, uses the Multi-head Relation-aware Attention and the Multi-scale Temporal Convolution to learn the motion features of a single person in the spatial and temporal dimensions, respectively. The Zoom-out Transformer (ZoT) part, takes the features of multiple persons extracted by ZiT as input and builds a graph of the human group with the help of the powerful relation mining ability of the Relation-aware Attention to capture the spatio-temporal interaction information among multiple persons. Additionally, we carefully design a Relation-aware Attention mechanism, which comprehensively leverages the prior knowledge of the human body structure and the global characteristic of the human motion to fully exploit the multi-level features of group activities.

Our skeleton-based group activity benchmarks *i.e.* K-SA and V-SA aim to verify the effectiveness of the proposed Zoom Transformer model and promote future studies. For K-SA, we select 14 common group activities from the large-scale video-based human action dataset *i.e.* Kinetics [19] and employ the AlphaPose [21] to extract the skeleton data of at most 5 persons from the corresponding videos. For V-SA, we estimate the skeleton data of at most 12 persons from the popular video-based group activity dataset *i.e.* Volleyball Activity [20]. Apparently, the skeleton data of our K-SA and V-SA are both derived from real-world videos. Due to the limitation of the raw data quality and the ability of AlphaPose, K-SA and V-SA contain many noises, which makes group activity recognition more challenging and conformable to the real scenarios. Section Experiments describes how to acquire the group skeleton data in detail.

In summary, the main contributions of this work are three-fold:

- We extend the skeleton-based action recognition for single-person action to multi-person group activity, which has a wide range of applications in practice. We hope this work can attract more attention and promote the investigation of skeleton-based group activity recognition.
- A Zoom Transformer with Relation-aware Attention, which consists of a Zoom-in Transformer and a Zoom-out Transformer, is explored to extract the low-level motion information of a single person and the high-level interaction information of multiple persons from the skeleton sequence hierarchically.
- Two new skeleton-based group activity benchmarks (*i.e.* K-SA and V-SA) are designed and released publicly based on the existing real-world RGB video datasets, which can verify the effectiveness of our proposed model and facilitate the future study.

II. Related Work

A. Skeleton-based Action Recognition

Most of the conventional studies in skeleton-based action recognition relied on handcrafted features [22], [23], which cannot effectively extract the spatio-temporal correlation from the skeleton sequences. Thus, they have limited accuracy and weak robustness for sophisticated human actions. Recently,

deep learning based action recognition methods, e.g. RNN-based methods [11], [24]–[26] and CNN-based methods [27]–[31] focusing on the temporal information and spatial information of the skeleton sequence respectively, have achieved rapid progress. Yan *et al*. [8] firstly proposed a GCN-based method ST-GCN, which boosts the performance of the skeleton-based action recognition for the ST-GCN can explore the spatio-temporal features simultaneously and uniformly. Based on ST-GCN, many variants have been explored, which typically introduce some incremental modules, e.g. the attention module [32], the context-aware module [10], the semantics-guided module [33], and the class activation maps [34] to enhance the network capacity. Liu *et al*. [35] introduced a multi-scale 3D GCN that can disentangle and unify the dense cross-spacetime information. Cheng *et al*. [36] introduced a graph-based shift operation to provide flexible receptive fields and used the point-wise convolutions to lighten the computational complexity. Zhang *et al*. [7] explored a spatial attentive and temporal dilated GCN to extract the features of skeleton sequences with different spatial attention weights and temporal scales. Tu *et al*. [37] fused the motion features of the joints and the bones and designed a temporal prediction head for self-supervised skeleton feature mining. Furthermore, Chiara *et al*. [38] presented to use the Transformer to explore the spatio-temporal correlation of the skeleton graph sequences. These methods are effective to capture the motion information of a single person but without concerning the group activity of multiple persons. Usually, a multi-person group activity contains more expressive motion clues, such as the interactions and the relative spatial positions, which need to be deeply exploited. In contrast to the above methods, the proposed Zoom Transformer with Relation-aware Attention can extract both the low-level motion information of an individual and the high-level interactions of multiple people hierarchically, which opens an effective way to analyze skeleton-based group activities.

B. Group Activity Recognition

Group activity recognition has been extensively studied based on the RGB videos. The popular datasets for group activity recognition also only involve RGB videos, e.g. Volleyball Activity [20] and Collective Activity [39]. The earlier approaches mostly combined hand-crafted visual features with probability graphical models [40]–[42] to represent the motion features of multiple people. Recent deep model based researches usually use CNN to extract low-level video features, and utilize GCN or attention mechanism to exploit the high-level semantic relationship among actors [43], [44]. Ibrahim *et al*. [20] designed a two-stage deep temporal model, which builds an LSTM model to represent the action dynamics of individual people and uses another LSTM model to aggregate person-level information. Wu *et al*. [45] proposed to construct an Actor Relation Graph (ARG) to simultaneously capture the appearance and position relation between actors. Azar *et al*. [46] presented a Convolutional Relational Machine (CRM) to represent the spatial relations between individuals in the video. In general, video-based group activity recognition focuses on extracting the motion information of a single person and the interaction information among multiple people. Similarly, these two kinds of information are also the core clues for skeleton-based group activity recognition. Without the RGB-based scene appearance information, skeleton-based multi-level feature extraction is more challenging. Therefore, we present the ZiT and the ZoT with a Transformer-based structure and specifically designed Relation-aware Attention to fully explore the motion information and the interaction information of the skeleton sequence respectively.

C. Visual Transformer

Transformers [47], which have been widely used in the natural language processing task, are the models that rely on the multi-head self-attention mechanism to draw global correlations from the input features. Transformers have robust feature extraction ability and global feature perception field, so it is effective in various tasks. Recently, using Transformer in vision tasks becomes the trend, e.g. object detection [48], image enhancement [49], image segmentation [50], video processing [51], and 3D point cloud processing [52]. For image classification, Dosovitskiy *et al.*proposed a Vision Transformer (ViT) [53], which divides an image into 16×16 patches and feeds these patches into a standard Transformer, obtains remarkable performance. Wu *et al.*represented images as semantic visual tokens and ran Transformer to densely model token relationships [54]. For object detection, Carion *et al*. [55] combined the transformer framework with the CNN network and proposed a simple and fully end-to-end object detector named DEtection TRansformer (DETR). These works demonstrated that Transformer has a strong capability to extract low-level visual features and high-level correlative features. However, in the field of skeleton-based action recognition, there are few studies discuss the applicability of the Transformer in learning multi-level features. In this work, we designed a Zoom Transformer with Relation-aware Attention to extract multi-level spatio-temporal features from the skeletons of group people.

III. METHOD

In this section, we first elaborate the two important parts of our Zoom Transformer model *i.e.* Zoom-in Transformer (ZiT) and Zoom-out Transformer (ZoT) mathematically. Then, we define the Relation-aware Maps in the proposed Relation-aware Attention for ZiT and ZoT respectively. Finally, we introduce the structure of the Zoom Transformer model.

A. Zoom-in Transformer

The role of the ZiT is to capture the motion information of every single human body in group activities. Just like a zoom camera, ZiT focuses on each human body with a joint-level perception field and extracts the motion information of every single human body by exploring the spatial-temporal correlation among the joints. The detailed structure of a ZiT block can be referred to in the left part of Figure 2. The core components of the ZiT block are the Multi-head Relation-aware Attention and the Multi-scale Temporal Convolution

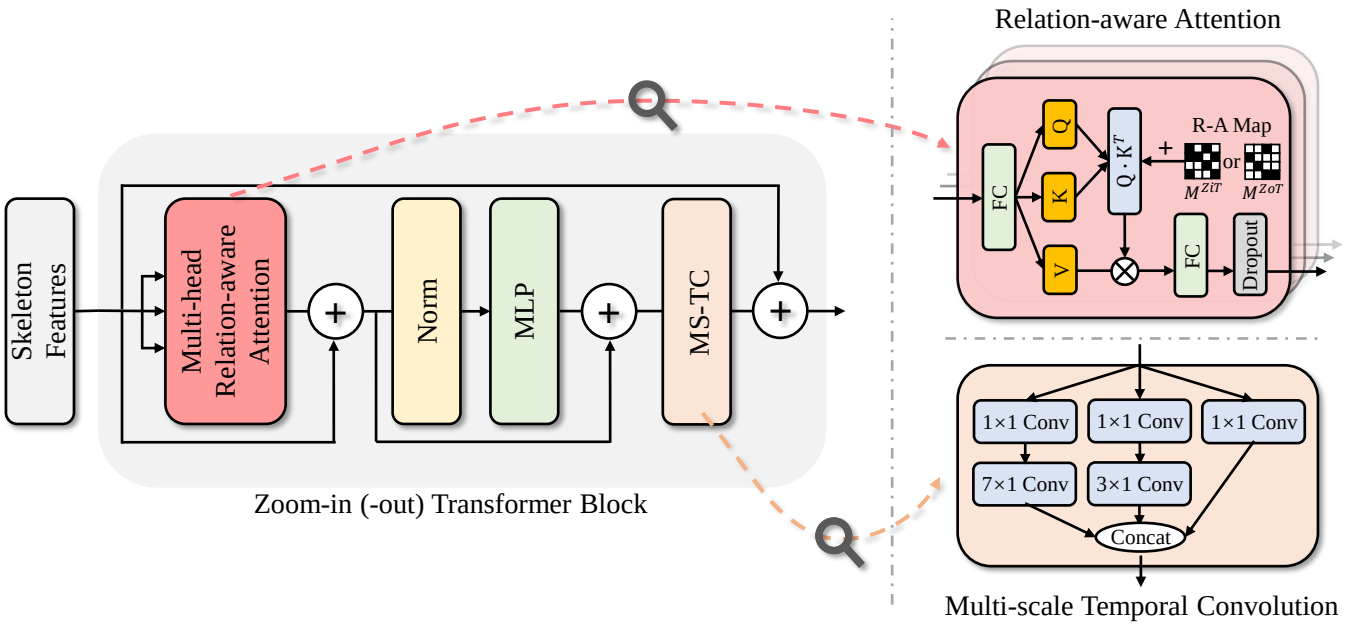

Fig. 2. The structure of a Zoom-in (-out) Transformer (ZiT/ZoT) block of the proposed Zoom Transformer model, which is suitable for extracting multi-level skeleton activity features. The core components of the ZiT/ZoT block are the Multi-head Relation-aware Attention and the Multi-scale Temporal Convolution (MS-TC), which can be effective to process spatial information and temporal information, respectively. The ZiT block and the ZoT block have different Relation-aware Maps. "R-A Map" denotes a Relation-aware Map. "Norm" denotes a normalization layer. "MLP" represents a Multi-layer Perceptron, which contains two fully connected layers and a dropout layer. The residual connections are also utilized in the ZiT/ZoT block. Note that the structures of the R-A Map in the ZiT block and in the ZoT block are different, which will be described in Secition III.C. Relation-aware Map.

(MS-TC), which can be effective to process spatial and temporal information, respectively. Let $X \in \mathbb{R}^{C\times T\times N\times M}$ be the N joint features across T frames of M persons, where C is the channel dimension. The initial value of C is 3, which represents the x-coordinate, the y-coordinate, and the confidence score of the joints.

The Multi-head Relation-aware Attention operates on the N joints, which can be described as mapping a query joint and a set of key-value joint pairs to an output, where the query (Q), key (K), value (V), and output are all feature vectors of the joints. The output is computed as a weighted sum of V, where the weight assigned to each V is computed by a correlation function Q with the corresponding K and the Relation-aware Map M_R. The Relation-aware Attention mechanism can be referred to in the right-top part of Figure 2. In practice, it is defined as:

$$Attention_R(Q,K,V,M_R) = \frac{1}{2}(softmax(\frac{QK^T}{\sqrt{d_k}}) + M_R)V, \tag{1}$$

where d_k is the dimension of K.

Integrating multiple Relation-aware Attention heads, the Multi-head Relation-aware Attention mechanism can be formulated as:

$$Multihead(Q,K,V,M_R) = Concat(head_1,...,head_h)W^O, \tag{2}$$

where $head_i = Attention_R(QW_i^Q, KW_i^K, VW_i^V, M_R)$. The projections are the parameter matrices $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W^O \in \mathbb{R}^{hd_v\times d}$. Multi-head Relation-aware Attention allows the model to jointly attend to the information from different motion representation subspaces and learn various features of the skeleton sequence. Using ZiT, we can capture the motion information of independent individuals, which is the basis for our method to further analyze the group activities.

The MS-TC operates on T frames with multi-scale temporal kernels. In our Zoom-in Transformer block, we use three branches with $1, 3, 7$ kernel sizes respectively. Each branch contains a 1×1 convolution to reduce the channel dimension. The results of the three branches are concatenated to obtain the output. The MS-TC can be referred to in the right-bottom part of Figure 2.

B. Zoom-out Transformer

The role of the ZoT is to exploit the interactions and positional relationships among multiple persons in group activities. If the ZiT is approximate to the macro mode of the Zoom Transformer, the ZoT can be considered as a panorama mode. ZoT focuses on the holistic scene with a group-level perception field. Thanks to the powerful feature mining capability of the Transformer, our ZoT block has the same structure as ZiT block with a different Relation-aware Map (see Figure 2), which is suitable for extracting the low-level motion feature of the human body joints and is also effective to learn high-level relation feature of multiple people. Let $X^{'} \in \mathbb{R}^{C\times T\times N\times M}$ denotes the output joint feature of ZiT. We use a Global Average Pooling (GAP) layer to squeeze the joint dimension, then we can get the feature $Y \in \mathbb{R}^{C\times T\times M}$ of M persons, which is the input feature of ZoT. The Relation-aware Attention of the ZoT that operates on the M actors enables to capture the information of interactions and positional relationships among them quite well.

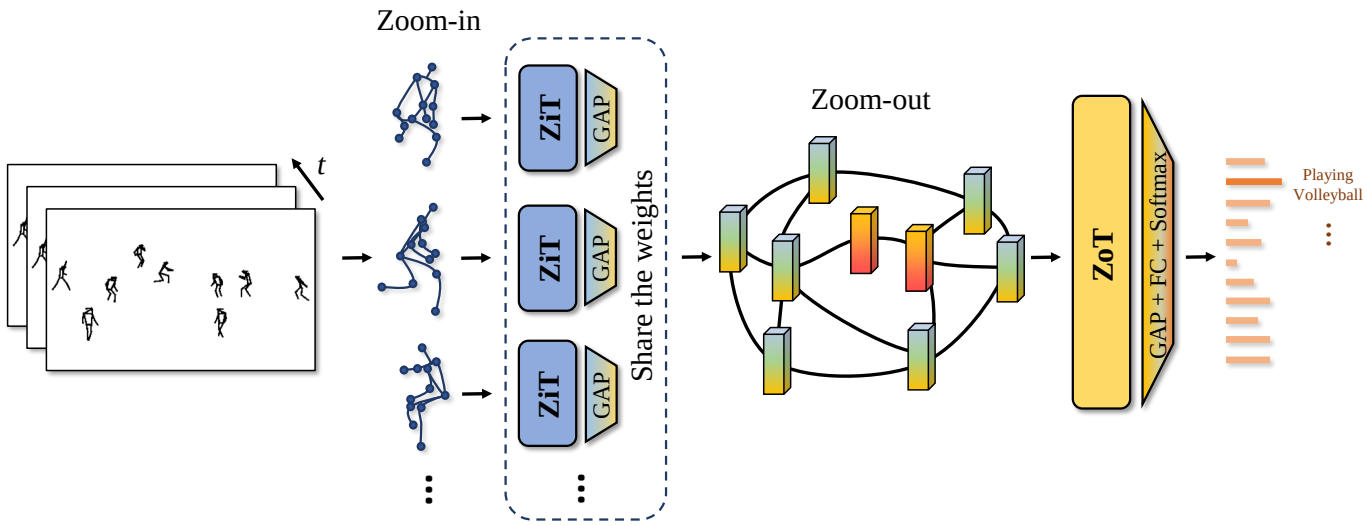

Fig. 3. The flow chart of our Zoom Transformer model, which consists of two parts – "ZiT" and "ZoT". "ZiT" in the zoom-in process can extract joint-level information, "ZoT" in the zoom-out process can exploit group-level information. ZiT and ZoT have the same structure with different Relation-aware Maps.

C. Relation-aware Map

The physical structure of the human body is an important clue for the extraction of the skeleton motion features. GCN-based models have achieved remarkable performance in skeleton-based action recognition because the adjacency matrix of GCNs can fully leverage the prior knowledge of body structure [7]. Inspired by this, we define a Relation-aware Map for the ZiT block, which enables the ZiT block to have the same ability as GCNs to use the prior knowledge of the skeleton graph for learning low-level features. We consider the Relation-aware Map for ZiT block as $M^{ZiT} \in \{0,1\}^{N\times N}$, where $M^{ZiT}_{i,j} = 1$ if the i-th and the j-th joints are physical connected, and $M^{ZiT}_{i,j} = 0$ otherwise. In the Relation-aware Attention, we make the M^{ZiT} normalized, accordingly, the M_R in ZiT block can be expressed as follows:

$$M_R = D^{-\frac{1}{2}} M^{ZiT} D^{-\frac{1}{2}}, \tag{3}$$

where $D \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix of M^{ZiT}. To make this prior Relation-aware Map more robust, we set the Relation-aware Maps of the last 5 ZiT blocks in our Zoom Transformer as the trainable parameters, which can be optimized during the training process.

For the Relation-aware Map in the ZoT block, unlike the ZiT which has an explicit physical graph structure, we use the feature output by ZiT to obtain an implicit head-specific Relation-aware Map that contains the global relation information. The $M^{ZoT} \in \mathbb{R}^{C\times M\times M}$ is formulated as:

$$M^{ZoT} = softmax(\psi(Y^{'}) - \phi(Y^{''})), \tag{4}$$

where $Y^{'} \in \mathbb{R}^{C\times T\times 1\times M}$ and $Y^{''} \in \mathbb{R}^{C\times T\times M\times 1}$ denotes the output features of ZiT with different tensor shape. $\psi(\cdot)$ and $\phi(\cdot)$ are the convolutional layers with 1×1 kernel. In general, the M_R in ZoT block can be expressed as follows:

$$M_R = M^{ZoT}. \tag{5}$$

It should be noticed that the channel mapping of the convolutional layers $\psi(\cdot)$ and $\phi(\cdot)$ is $C \rightarrow h$, which makes the Relation-aware Map in the ZoT block involve expressive global relation information for different attention heads.

D. Model Architecture

The flow chart of the deigned Zoom Transformer model is shown in Figure 3. As a baseline model of skeleton-based group activity recognition, our Zoom Transformer model is lightweight and effective, where we don't add any tricks and incremental modules to it. The input skeleton data is $X \in \mathbb{R}^{3\times T\times N\times M}$. In our Zoom Transformer, the ZiT has 7 blocks and each block has 6 attention heads. In the training process, we freeze the Relation-aware Maps of the first 2 ZiT blocks while setting the Relation-aware Maps of the last 5 ZiT blocks as the trainable parameters. In the zoom-in stage, we use the ZiT to parallelly process the skeleton data of M individual persons on the joint dimension N and the temporal dimension T. Therefore, for each person, the ZiT shares the weights. After the ZiT and the GAP layer, we can get the feature $Y \in \mathbb{R}^{C\times T\times M}$ of M persons, which is then used as the input of the ZoT. The ZoT has 2 blocks and each block also has 6 attention heads. In the zoom-out stage, the ZoT operates on the personal dimension M and the temporal dimension T. Finally, the output features of the ZoT are processed by a GAP layer and a Softmax classifier (FC + Softmax) to get the prediction score of the group activity. After the Zoom Transformer is processed, the channel dimension of the skeleton data is expanded from 2 to 276, so the final FC (*i.e.* Fully Connected) layer has 276 input channels and K output channels, where K is the total number of the activity categories.

IV. Experiments

A. Datasets

To test our skeleton-based multi-person group activity recognition method, we construct two new benchmarks *i.e.*

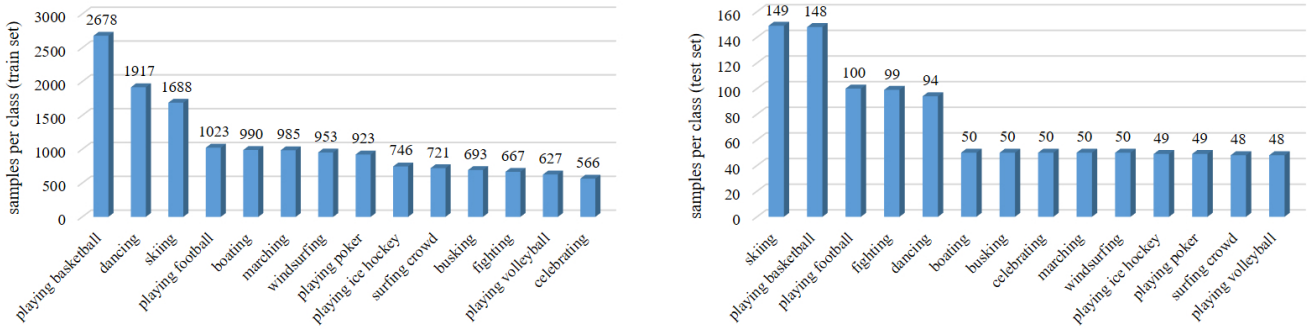

Fig. 4. The number of samples in each category of the training set (left) and the testing set (right) on the K-SA benchmark, which is derived from the Kinetics dataset.

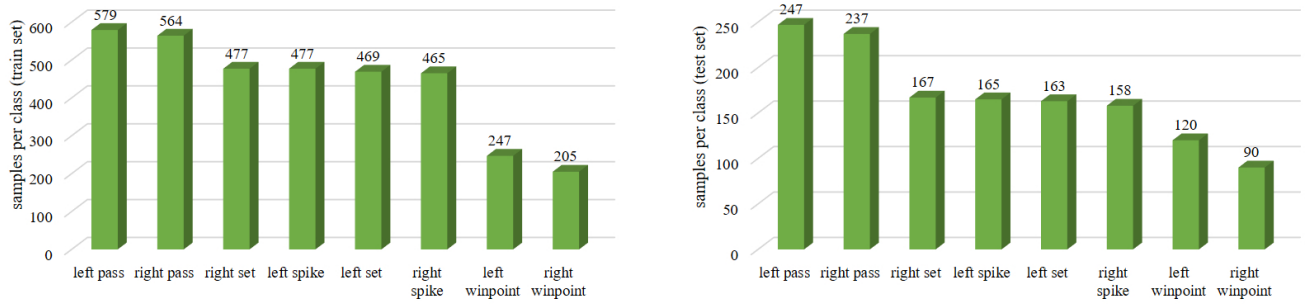

Fig. 5. The number of samples in each category of the training set (left) and the testing set (right) on the V-SA benchmark, which is derived from the Volleyball Activity dataset.

the Kinetics-Skeleton-Activity (K-SA) and the Volleyball-Skeleton-Activity (V-SA) based on the existing RGB video datasets Kinetics [19] and Volleyball Activity [20], respectively. Compared with the video data, the video-derived skeleton-based data we formed greatly reduce the data quantity. Take the Volleyball Activity dataset as an example. The original videos are about 59GB, but the corresponding skeleton data is only 473MB. The data size has been reduced by more than 120 times. Additionally, we also test the effectiveness of our model on the popular datasets NTU-RGB+D 60 [13] and NTU-RGB+D 120 [18] for single-person skeleton action recognition.

Kinetics-Skeleton-Activity (K-SA): Kinetics [19] consists of 300,000 video clips in 400 action classes. The video clips of Kinetics are sourced from YouTube and most of them are single-person actions. The action classes range from daily activities, sports scenes, to complex actions with interactions. In this work, we look at skeleton-based group activity recognition. Accordingly, based on the video content, we select 16,151 videos from Kinetics and divide them into 14 activity categories, including playing basketball, dancing, skiing, playing football, boating, marching, windsurfing, playing poker, playing ice hockey, surfing crowd, busking, fighting, playing volleyball and celebrating, to construct the K-SA benchmark. To obtain the joint locations, we first resize all videos to the resolution of 340×256 and convert the frame rate to 30 FPS. Then we use the public available AlphaPose [21] to estimate the location of 18 joints of each people on each frame. AlphaPose is a popular and advanced human pose estimation tool, which is widely used to extract human skeleton [56]–[58]. The 18 joints include 17 joints marked by the COCO dataset [59] and an additional "Neck" joint, which is the mean value of the left and right shoulders. The AlphaPose gives the 2D coordinates (X, Y) of the human body joints and the confidence scores C for the 18 joints. We represent each joint with a tuple (X, Y, C), and one person is recorded as an array of 18 tuples. For videos with more than 5 persons, we select 5 persons with the highest average joint confidence in each frame. In this way, one video with T frames is transformed into a skeleton sequence of multi-person. We represent the multi-person skeleton sequence with a tensor of $(3, T, 18, 5)$ dimensions. For simplicity, we pad every sequence by repeating the sequence from the start to ensure it contains $T = 300$ frames. We separate the selected 16,151 skeleton sequences into a training set that has 15,117 sequences and a testing set that has 1034 sequences associated with the Kinetics dataset. The number of samples in each category of the training set and the testing set can be seen in Figure 4. The top two rows of Figure 6 show two samples of the K-SA benchmark. We find that each individual can perform the same action or different actions in a group activity. The interaction among them is particularly important for recognizing group activities.

Volleyball-Skeleton-Activity (V-SA): The Volleyball Activity dataset [20] is composed of 4,830 video clips gathered from 55 volleyball games, with 3,493 training clips and 1,337 testing clips. Each clip is labeled with one of 8 group activity

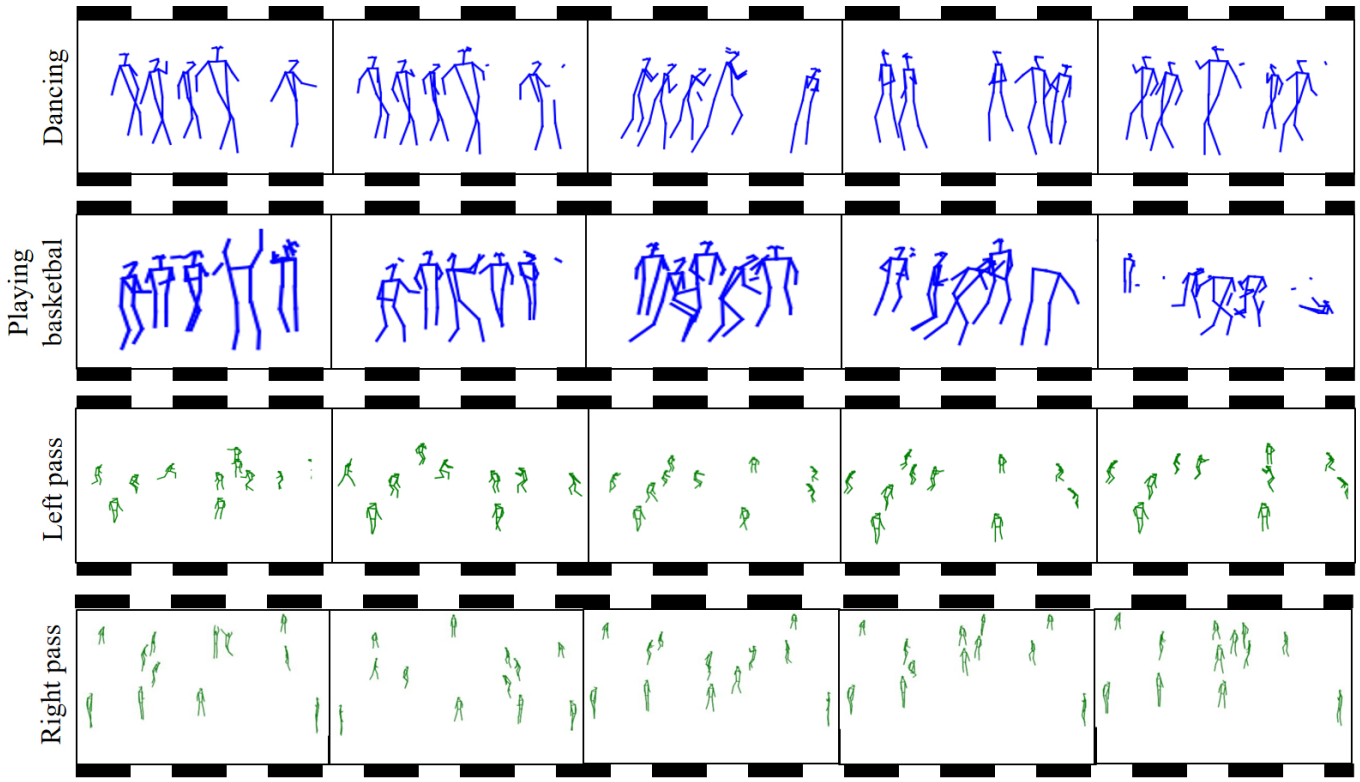

Fig. 6. Visualization of some samples of our released K-SA benchmark and V-SA benchmark. The K-SA benchmark is drawn in blue and the V-SA benchmark is drawn in green.

labels (right set, right spike, right pass, right winpoint, left set, left spike, left pass and left winpoint). Similar to the K-SA benchmark, we use the AlphaPose to estimate the 18 joints of each person on every frame of the clips. We select 12 persons with the highest average joint confidence in each frame and construct the skeleton sequence tensors with $(3, T, 18, 12)$ dimensions. Each video clip in the Volleyball Activity dataset contains 41 frames, thus we set $T = 41$. The number of samples in each category of the training set and the testing set can be seen in Figure 5. The bottom two rows of Figure 6 show two samples on the V-SA benchmark. Compared with the K-SA benchmark, the V-SA benchmark contains more persons in each sequence, and the relationship among them is more complicated.

NTU-RGB+D 60: NTU-RGB+D 60 [13] is a large in-door-captured dataset with annotated 3D joint coordinates for the human action recognition task. NTU-RGB+D contains 56,000 skeleton sequences in 60 action classes. There are 25 joints for each person in the skeleton sequences, while each sequence has no more than 2 persons. It includes two settings: (1) Cross-Subject (X-S) benchmark, in which the training set comes from one subset of 20 subjects and a model is validated on sequences from the remaining 19 subjects; (2) Cross-View (X-V) benchmark, in which the training samples come from camera views 2 and 3, and the evaluation samples are all from the camera view 1. We follow the conventional setting of [13] and report the top-1 accuracy on both sub-sets.

NTU-RGB+D 120: NTU-RGB+D 120 [18] is an extension of NTU-RGB+D 60, which adds 57,367 new skeleton sequences representing 60 new actions, for a total of 113,945 videos referring to 120 classes from 106 subjects under 32 camera setups. It includes two settings: (1) cross-subject (X-Sub) benchmark: the 106 subjects are split into training and testing groups. Each group contains 53 subjects. (2) cross-setup (X-Set) benchmark: the training data comes from samples with even setup IDs, and the testing data comes from samples with odd setup IDs.

B. Implementation Details

We implement our Zoom Transformer model based on the PyTorch deep learning framework [60]. The Zoom Transformer has a total of 9 blocks (7 ZiT blocks + 2 ZoT blocks), and the channel dimensions are respectively 48, 48, 48, 96, 96, 192, 192, 276 and 276 for each block (see Figure 7, ZiT+ZoT). We apply the stochastic gradient descent (SGD) algorithm with Nesterov momentum (0.9) as the optimizer. The weight decay is set to 0.0005. We use 4 GTX 2080Ti GPUs for model training, and set the batch size to 64 and the weighting factor to $\lambda = 0.1$. The initial learning rate is set as 0.05. For the K-SA benchmark, the number of training epochs is set as 50. The learning rate decay is set as 0.1 at the 20th epoch, 30th epoch, and 40th epoch. For the V-SA benchmark, the number of training epochs is set as 40. The learning rate decay is set as 0.1 at the 20th epoch and 30th epoch. For the NTU-RGB+D 60 and the NTU-RGB+D 120, the number of training epochs is set as 60. The learning rate decay is set as 0.1 at the 30th epoch, 40th epoch, and 50th epoch.

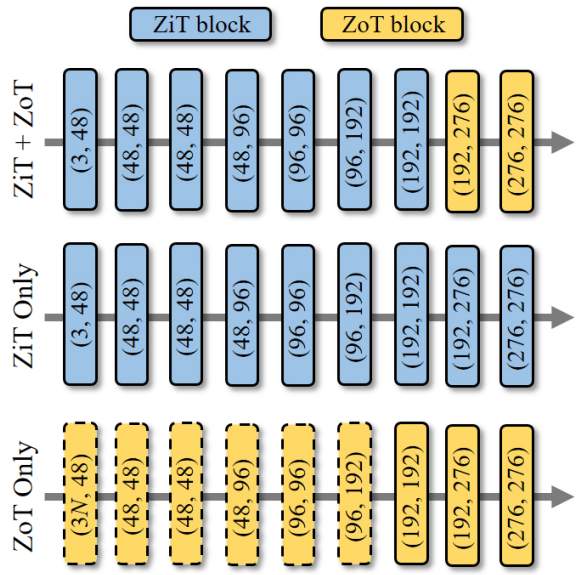

Fig. 7. The detailed structures of the ZiT+ZoT, the ZiT Only, and the ZoT Only. (3, 48) means the input channel of this block is 3, and the output channel of this block is 48. N is the number of the joints. The dotted line indicates the block without Relation-aware Map.

TABLE I
COMPARISON OF THE PARAMETERS AND ACCURACY ON THE K-SA BENCHMARK WITH DIFFERENT MODEL CONFIGURATIONS. * MEANS WITHOUT THE REALTION-AWARE MAP.

Model configs	Params	Top-1 (%)	Top-5 (%)
ST-GCN [8]	3.10M	64.60	94.10
AGCN [17]	3.44M	67.89	95.74
MS-G3D [35]	3.19M	68.08	96.03
S-TR [38]	3.07M	67.24	94.98
ZiT Only	2.93M	62.86	93.91
ZiT Only*	2.93M	62.37	93.42
ZoT Only	2.99M	67.41	94.58
ZoT Only*	2.71M	67.11	94.19
ZiT + ZoT*	2.71M	69.06	96.04
ZiT + ZoT	**2.93M**	**69.83**	**96.13**

C. Ablation Studies

We present an ablative study on the K-SA benchmark and the V-SA benchmark to evaluate the effectiveness of the proposed model. We analyze the effect of the ZiT, the ZoT, the Relation-aware Map, the temporal kernel size, and the number of attention heads. We also visualize the average Attention Maps in the ZiT and the ZoT to visually analyze the interpretability of the proposed Zoom Transformer model.

Model configurations. Experiments are conducted to test the performance of different model configurations. We take the popular GCN-based methods ST-GCN [8], AGCN [17], MS-G3D [35] and the Transformer-based method S-TR [38] as the baselines. Results on the K-SA benchmark and the V-SA benchmark are shown in Table I and Table II, respectively. "ZiT Only" means we only use the ZiT part and apply a GAP layer at the end of the model to pool the features of multi-person. "ZoT Only" means we only use the ZoT part and flatten the joint dimension N to the channel dimension C of the initial skeleton feature. The detailed structure of the ZiT Only and the ZoT Only can be seen in Fighre 7. Both the ZiT Only and the ZoT Only have 9 blocks, which are consistent with the Zoom Transformer. To balance the number of parameters, we only apply the Relation-aware Map in the last three blocks of the ZoT Only model. Table I shows the results on the K-SA benchmark, the performance of ZiT Only and ZoT Only is worse than AGCN and MS-G3D. But the ZiT + ZoT (i.e. the Zoom Transformer) outperforms AGCN by 1.94% (69.83% vs 67.89%) on the top-1 accuracy with less parameters (2.93M vs 3.44M) and outperforms MS-G3D by 1.75% (69.83% vs 68.08%) on the top-1 accuracy with less parameters (2.93M vs 3.19M). Table II shows the results on the V-SA benchmark, the performance of the three model configurations exceeds the baselines. ZiT + ZoT (the Zoom Transformer) performs better than ZiT Only and ZoT Only, where it surpasses AGCN by 4.79% (71.20% vs 66.41%), MS-G3D by 4.71% (71.20% vs 66.49%), and S-TR by 5.83% (71.20% vs 65.37%) on the top-1 accuracy. These results demonstrate that for skeleton-based group activity recognition, it is insufficient to extract only the low-level human joint motion information or only the high-level multi-person relation information. Only by fully mining the low-level information and the high-level information, and effectively integrating them can get remarkable performance. Besides, these results also prove that our Zoom Transformer model is more suitable than the existing methods to capture multi-level features hierarchically for skeleton-based group activity recognition.

Relation-aware Map. For the Relation-aware Map in the ZiT and ZoT, it can bring a suitable accuracy improvement on both the K-SA and V-SA benchmarks. As can be seen in Table I and Table II, the ZiT with Relation-aware Map obtains 0.49% and 0.23% improvements of top-1 accuracy on

TABLE II
COMPARISON OF THE TOP-1 AND TOP-5 ACCURACY ON THE V-SA BENCHMARK WITH DIFFERENT MODEL CONFIGURATIONS. * MEANS WITHOUT THE REALTION-AWARE MAP.

Model configs	Top-1 (%)	Top-5 (%)
ST-GCN [8]	64.92	97.08
AGCN [17]	66.41	98.07
MS-G3D [35]	66.49	98.35
S-TR [38]	65.37	97.46
ZiT Only	70.23	98.74
ZiT Only*	70.00	98.42
ZoT Only	69.04	98.59
ZoT Only*	68.81	98.42
ZiT + ZoT*	70.75	99.05
ZiT + ZoT	**71.20**	**99.18**

TABLE III
COMPARISON OF THE INPUT SEQUENCE LENGTH (FRAMES), GFLOPS AND TRAINING DATA SIZE AND WITH THE VIDEO-BASED METHODS. THE RESULTS OUTSIDE THE BRACKETS IS ON THE VOLLEYBALL ACTIVITY DATASET, AND THE RESULTS INSIDE THE BRACKETS IS ON THE KINETICS DATASET.

Methods	Seq len	GFLOPs	Data size
I3D [15]	8 (8)	108.0 (108.0)	43G (≈100G)
SlowFast [16]	8 (8)	65.7 (65.7)	43G (≈100G)
Zoom Transformer	**41 (150)**	**1.6 (5.6)**	**0.34G (4.8G)**

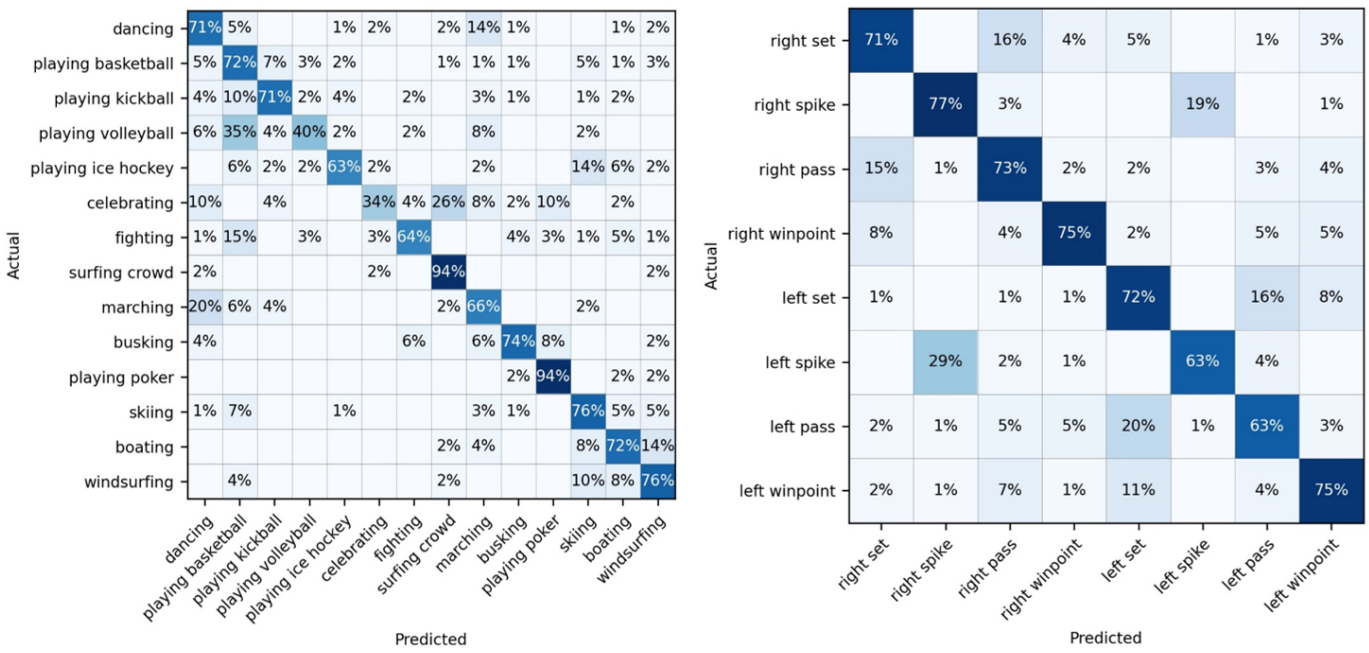

Fig. 8. The confusion matrices on our presented benchmarks K-SA (left) and V-SA (right).

TABLE IV
COMPARISON OF THE TOP-1 AND TOP-5 ACCURACY ON THE K-SA BENCHMARK WITH DIFFERENT NUMBER OF ATTENTION HEADS.

Attention heads	Top-1 (%)	Top-5 (%)
ZiT3 + ZoT3	66.25	95.74
ZiT3 + ZoT6	66.92	94.97
ZiT6 + ZoT3	67.50	95.36
ZiT6 + ZoT6	**69.83**	**96.13**
ZiT9 + ZoT9	68.38	96.03

TABLE V
COMPARISON OF THE TOP-1 AND TOP-5 ACCURACY ON THE K-SA BENCHMARK WITH DIFFERENT TEMPORAL KERNEL SIZES.

Kernel size	Top-1 (%)	Top-5 (%)
3	68.18	95.74
5	67.21	95.94
7	65.96	95.36
3 + 3 + 3	69.34	95.65
1 + 3 + 5	69.43	95.76
1 + 3 + 7	**69.83**	**96.13**

the K-SA and V-SA with almost no increase in the number of parameters, respectively. The ZoT with Relation-aware Map obtains 0.30% and 0.23% improvements of top-1 accuracy on the K-SA and V-SA, respectively. The Relation-aware Map also brings 0.77% and 0.45% top-1 accuracy improvements for the ZiT+ZoT on the benchmarks K-SA and V-SA, respectively. These results show that the Multi-head Attention with the Relation-aware Map can fully leverage the prior knowledge of the human body structure to extract motion features of individuals, and can effectively utilize global information of all actors to learn the multi-person interaction features.

Strength of the skeleton-based methods. Currently, on the video-based dataset, the activity recognition accuracy via utilizing the skeleton data (the skeleton data is directly estimated from videos), is lower than that by using RGB videos, where the accuracy gap is about 20% [8], [15]. The main reason is that a large amount of noise is generated during the process of extracting the skeleton data from RGB videos by pose estimation algorithms. Furthermore, the pose estimation precision is limited.

On the other hand, skeleton-based activity recognition has significant advantages over video-based activity recognition on other aspects. Table III shows the comparison of the input sequence length, GFLOPs, and training data size between the mainstream video-based methods and our skeleton-based Zoom Transformer model. It can be seen that our skeleton-based model requires fewer training data, can handle longer sequences, and the GFLOPs is much smaller than that of the video-based models (I3D [15]: 108.0 vs Ours: 1.6). Therefore, it is of great significance to explore group activity recognition based on the skeleton data. In fact, under the condition of accurate skeleton data estimated by the depth camera, the accuracy of skeleton-based action recognition will be superior to that of video-based action recognition [18].

Number of the attention heads. The number of attention heads affects the feature extraction ability of the Transformer critically. Table IV shows the comparison of the accuracies on the K-SA benchmark with different number of attention heads. The experimental results reveal that when both ZiT and ZoT have 6 attention heads, the performance becomes the best. When the number of the attention heads is 9 or 3, the performance of the model decreases. The reason for this phenomenon is that when the number of the feature channels is confirmed, increasing the number of attention heads appropriately can improve the ability of the attention mechanism to extract diverse features. However, too many

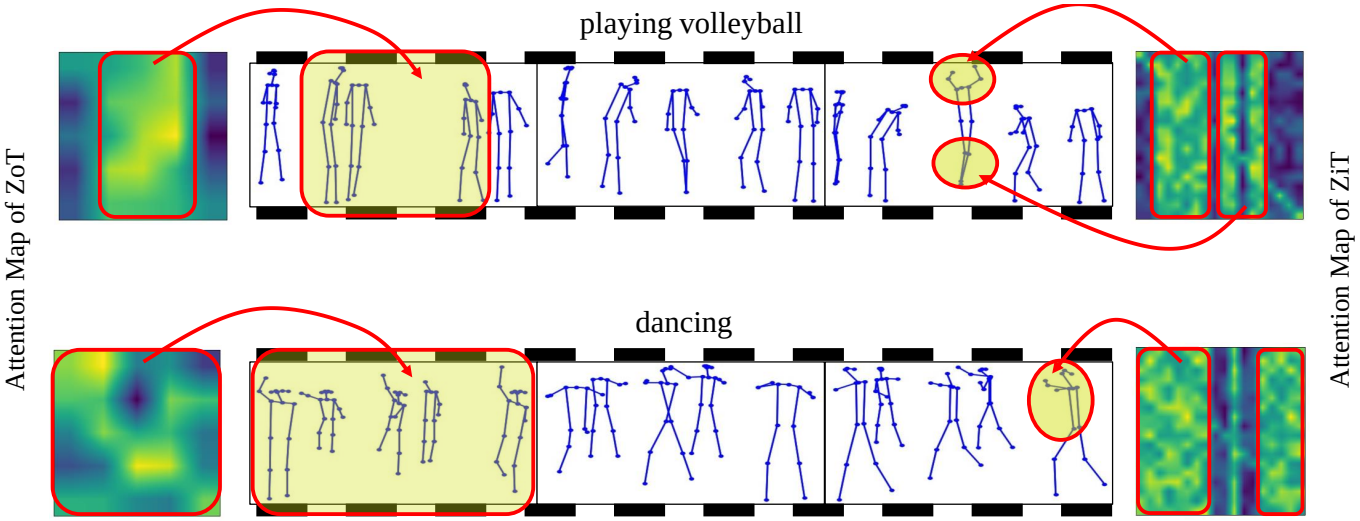

Fig. 9. Visualization of the average Attention Map in Zit and ZoT for different human action classes on the K-SA benchmark. The brighter area indicates that the weight of the Attention Map is larger there, where a key motion area for recognizing group activity.

TABLE VI
COMPARISON OF THE TOP-1 AND TOP-5 ACCURACY ON THE V-SA BENCHMARK WITH DIFFERENT NUMBERS OF PERSONS.

Persons	Top-1 (%)	Top-5 (%)
2	46.03	92.58
6	61.25	97.33
12	**71.20**	**99.18**

attention heads will lead to too few channels of a single attention head, which will reduce the informativeness of the features [47].

Temporal kernel size. To understand the function of the temporal kernel size of the MS-TC module in the Zoom-in (-out) Transformer block, some experiments are conducted in Table V. The first three rows show the results of the Zoom Transformer with a single temporal kernel and the last three rows show the results of the Zoom Transformer with a multi-scale temporal kernel. The results demonstrate that the multi-scale temporal kernel can significantly improve the accuracy. When the 3 branches of the temporal convolution have different kernel sizes, the performance becomes better, and the MS-TC with "1 + 3 + 7" kernel size obtains the best accuracy. Thus, multi-scale temporal information is very important for skeleton-based activity recognition.

Accuracy of each activity category. Figure 8 displays the confusion matrices of our model on the presented benchmarks K-SA (left) and V-SA (right). The results reveal that the main difficulty for the skeleton-based group activity recognition is to distinguish the confusing group actions (e.g. "celebrating" and "surfing crowd" in K-SA (left)), and to discern the individual's spatial position relationship of a group action (e.g. "left spike" and "right spike" in V-SA (right)). Which demonstrated our initial motivation that "exploring the interactions and positional relation among multiple people is a core issue".

Qualitative analysis of the Attention Map. Figure 9 shows the visualized average Attention Map in ZiT and ZoT for the activity "playing volleyball" and "dancing" in the K-SA benchmark. The average Attention Map is the mean value of the multi-head multi-block Relation-aware Attention Maps. The brighter area indicates that the weight of the Attention Map is larger there, where a key motion area for recognizing group activity. We can clearly see that for different activities, the Attention Map of ZiT can learn the obvious movement parts of the human body, and the Attention Map of ZoT can focus on the important actors in group activities. Taking the activity "playing volleyball" as an example. For the main actor's Attention Map of ZiT (right part of Figure 9), the activation values of the upper limb and lower limb are significantly higher than that of other regions, which means that the movement of the upper limb and lower limb is significant and is a crucial clue for the activity "playing volleyball". For the Attention Map of ZoT (left part of Figure 9), the activation values of the middle three people are obvious, which means that they have rich interaction information and the ZoT can fully learn this discriminative information.

Number of persons. The results in Table VI show the effect of different numbers of persons on the model performance. The experiments are conducted on the V-SA benchmark, and we retain the individuals with higher confidence for 2 and 6 persons. When using only the skeletons of 2 persons, the top-1 accuracy degrades by 25.17% compared to using the skeletons of 12 persons (46.03% vs 71.20%). When using the skeletons of 6 persons, the top-1 accuracy is 61.25%, which is also inferior to using the skeletons of 12 persons. These results reflect that the interaction information of multiple people is crucial for recognizing group activities. Notably, our Zoom Transformer with ZoT blocks can effectively mine the high-level interaction features.

D. Comparison with State-of-the-arts

We compare the proposed Zoom Transformer model with the state-of-the-art skeleton-based single-person action recognition methods on both the NTU-RGB+D 60 and NTU-

TABLE VII
COMPARISON OF THE TOP-1 ACCURACY WITH STATE-OF-THE-ART SINGLE-PERSON SKELETON-BASED ACTION RECOGNITION METHODS ON THE NTU-RGB+D 60 DATASET.

Methods	X-S (%)	X-V (%)
HBRNN (2015) [61]	59.1	64.0
Deep LSTM (2016) [13]	60.7	67.3
ST LSTM (2016) [24]	67.2	77.7
TCN (2017) [63]	74.3	83.1
S-CNN (2017) [28]	80.0	87.2
CNN+M+T (2017) [29]	83.2	89.3
ST-GCN (2018) [8]	81.5	88.3
M-GCNs+VTDB (2019) [65]	84.2	94.2
AS-GCN (2019) [66]	86.8	94.2
2s-AGCN (2019) [17]	88.2	94.9
CA-GCN (2020) [10]	83.5	91.4
SGN (2020) [33]	89.0	94.5
MS-G3D (2020) [35]	89.4	95.0
2s RA-GCN (2020) [34]	86.7	93.4
2s Shift GCN (2020) [36]	89.7	96.0
TS+SS+PS (2021) [67]	88.0	94.9
ST-TR-agcn (2021) [38]	**90.3**	**96.3**
Zoom Transformer	90.1	95.3

TABLE VIII
COMPARISON OF THE TOP-1 ACCURACY WITH THE STATE-OF-THE-ART SINGLE-PERSON SKELETON-BASED ACTION RECOGNITION METHODS ON THE NTU-RGB+D 120 DATASET.

Methods	X-Sub (%)	X-Set (%)
ST LSTM (2016) [24]	55.7	57.9
Clips+CNN+MTLN (2017) [27]	61.8	62.2
GCA-LSTM (2017) [62]	61.2	63.3
ST-GCN (2019) [8]	77.9	79.0
2s-AGCN (2019) [17]	82.9	84.6
SkeMotion (2019) [64]	66.9	67.7
2s RA-GCN (2020) [34]	81.0	82.5
MS-G3D (2020) [35]	**86.9**	**88.4**
2s Shift GCN (2020) [36]	85.3	86.6
AdaSGN (2021) [68]	85.9	86.8
ST-TR-agcn (2021) [38]	85.1	87.1
Zoom Transformer	84.8	86.5

RGB+D 120 datasets in Table VII and Table VIII, respectively. The methods which are selected for comparison include the RNN-based methods [13], [24], [61], [62], the CNN-based methods [27]–[29], [63], [64], the GCN-based methods [8], [10], [17], [33]–[36], [65]–[68], and the Transformer-based method [38]. Notably, our Zoom Transformer model performs better than most of the state-of-the-art methods. E.g., For NTU-RGB+D 60, compared with ST-GCN [8], which is one of the most influential GCN-based models, our results outperform it by 8.6% (90.1% vs 81.5%) on the X-S benchmark and 7.0% (95.3% vs 88.3%) on the X-V benchmark. For NTU-RGB+D 120, compared with ST-GCN [8], our results exceed it by 6.9% (84.8% vs 77.9%) on the X-Sub benchmark and 7.5% (86.5% vs 79.0%) on the X-Set benchmark. The results verify that our model has a strong capability to capture motion information for the skeleton data. Besides, the results also show that our model is not only suitable for group activity recognition but is also effective for single-person action recognition.

V. CONCLUSIONS

In this work, we proposed a novel Zoom Transformer model for the skeleton-based group activity recognition which has rarely been studied before. The Zoom Transformer model consists of two parts *i.e.* the Zoom-in Transformer (ZiT) and the Zoom-out Transformer (ZoT). Specifically, the ZiT can extract the low-level motion information of independent individuals from the skeleton sequence with the help of the body structure based Relation-aware Map. The ZoT is able to mine the high-level multi-person interaction information and reason their relations with the help of the global feature based Relation-aware Map. The ZiT and ZoT have the same structure, but process different levels of skeleton features. Combining ZiT and ZoT, our Zoom Transformer model can learn distinctive features from the multi-person skeleton sequence hierarchically and identify the skeleton-based multi-person group activity effectively. Moreover, to promote the study of skeleton-based group activity recognition, we present and release two new skeleton group activity benchmarks *i.e.* K-SA and V-SA, based on the existing video datasets. Extensive experiments are conducted on these two benchmarks to evaluate the performance of our model. The results demonstrate that the proposed model is valid to recognize multi-person group activities and exceeds the GCN baseline by a large margin with fewer parameters. In addition, experiments on the large-scale NTU-RGB+D dataset show that our Zoom Transformer model is also useful for single-person action recognition. We expect that this work can arouse more investigations about skeleton-based group activity recognition, and promote the development of high-level visual relationship understanding.

A. Limitations

- The newly constructed K-SA and V-SA benchmarks are completely based on the existing video datasets, and automatically estimate the human pose through the existing pose estimation algorithm AlphaPose [21]. Limited by the quality of the raw videos and the performance of the AlphaPose, the skeleton data contains a lot of noise, which affects the accuracy of group activity recognition.
- The proposed Zoom Transformer model adopts Multi-scale Temporal Convolution and we fully test the effect of the convolution kernel size, however, the Temporal Convolution does not have a global perception field like the spatial attention mechanism. Therefore, when dealing with long skeleton sequences, the global temporal information extraction ability of our Zoom Transformer model is still insufficient.
- Our Zoom Transformer model focuses on the multi-level feature extraction for multi-person group activities. Although the proposed Zoom Transformer can be used to recognize the actions of a single person, its performance for single-person action recognition is slightly inferior to the state-of-the-art methods (see Table VII and Table VIII). The Zoom Transformer is constrained to fully unify the skeleton-based group activity recognition and the single-person action recognition. The reason is that fine-grained motion features of human joints are crucial for single-person skeleton action recognition. However,

for group activities, the model generally without pay much attention to the subtle movement of the joints, but pays more attention to the interactive information among multiple persons.

B. Future Work

There are several ways to address these limitations, which are promising to enhance the practical application of our method for skeleton-based multi-person group activity recognition.

- First, to construct a more useful and advanced dataset for skeleton-based group activity recognition, we can manually label the human pose from the videos to obtain more accurate skeleton data. In addition, we can use the depth camera to collect 3D skeleton data and establish an effective dataset that contains the ground truth of the 3D human pose.
- Second, to handle the temporal information of the long skeleton sequence, we would design an adaptive temporal attention mechanism to exploit multi-scale temporal features. Furthermore, how to reduce the computational complexity of the attention mechanism for the long-sequence data is also an important issue that is worthy of exploring.
- Finally, it is meaningful to design a unified Transformer-based model that is suitable for both multi-person group activity recognition and single-person action recognition.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62106177 and the Joint Fund of the Ministry of Education of China under Grant 8091B032156. The numerical calculation was supported by the super-computing system in the Supercomputing Center of Wuhan University.

REFERENCES

[1] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, 2018.

[2] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal vlad for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2799–2812, 2019.

[3] Y. Chang, Z. Tu, W. Xie, and J. Yuan, "Clustering driven deep autoencoder for video anomaly detection," in *European conference on computer vision*. Springer, 2020, pp. 329–345.

[4] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, "Going deeper with two-stream convnets for action recognition in video surveillance," *Pattern Recognition Letters*, vol. 107, pp. 83–90, 2018.

[5] Z. Tu, W. Xie, J. Dauwels, B. Li, and J. Yuan, "Semantic cues enhanced multimodality multistream cnn for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1423–1437, 2018.

[6] N. Zengeler, T. Kopinski, and U. Handmann, "Hand gesture recognition in automotive human–machine interaction using depth cameras," *Sensors*, vol. 19, no. 1, p. 59, 2019.

[7] J. Zhang, G. Ye, Z. Tu, Y. Qin, J. Zhang, X. Liu, and S. Luo, "A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition," *CAAI Transactions on Intelligence Technology*, 2020.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[9] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 1227–1236.

[10] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 333–14 342.

[11] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, and Y. Zhang, "Skeleton-based action recognition with gated convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3247–3257, 2018.

[12] C. Wu, X.-J. Wu, and J. Kittler, "Graph2net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[13] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[14] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[15] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla, "Quo vadis, skeleton action recognition?" *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2097–2112, 2021.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[17] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 026–12 035.

[18] L. J, S. A, P. M, W. G, D. L, and K. A, C, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," in *CoRR, abs/1905.04757*, 2019.

[19] W. K, J. C, K. S, B. Z, C. H, S. V, F. V, T. G, T. B, P. N *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[20] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.

[21] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[22] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[23] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4471–4479.

[24] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[25] 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 *Thirty-first AAAI conference on artificial intelligence*, 2017.

[26] X. Jiang, K. Xu, and T. Sun, "Action recognition scheme based on skeleton representation with ds-lstm network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2129–2140, 2019.

[27] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[28] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 597–600.

[29] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017, pp. 601–604.

[30] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio–temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2018.

[31] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2206–2216, 2020.

[32] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 536–553.

[33] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1112–1121.

[34] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1915–1925, 2020.

[35] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[36] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192.

[37] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Transactions on Multimedia*, 2022.

[38] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[39] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE, 2009, pp. 1282–1289.

[40] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: Hierarchical random field for collective activity recognition in videos," in *European Conference on Computer Vision*. Springer, 2014, pp. 572–585.

[41] W. Choi and S. Savarese, "Understanding collective activitiesof people from videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1242–1257, 2013.

[42] W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A heat-map-based algorithm for recognizing group activities in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1980–1992, 2013.

[43] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "stagnet: an attentive semantic rnn for group activity and individual action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 549–565, 2019.

[44] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[45] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9964–9974.

[46] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7892–7901.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[48] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[49] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[50] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[51] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[52] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," *arXiv preprint arXiv:2012.09164*, 2020.

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[54] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020.

[55] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[56] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the invisible visible: Action recognition through walls and occlusions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[57] T.-W. Chen and W.-L. Lin, "3d human motion reconstruction in unity with monocular camera," in *2020 International SoC Design Conference (ISOCC)*. IEEE, 2020, pp. 191–192.

[58] Y. Yoon, J. Yu, and M. Jeon, "Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition," *Applied Intelligence*, pp. 1–15, 2021.

[59] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[61] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[62] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[63] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[64] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz, "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[65] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph cnns with motif and variable temporal block for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8989–8996.

[66] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[67] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3d action representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 423–13 433.

[68] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 413–13 422.

Jiaxu Zhang received the B.S. degree from Southeast University, Nanjing, China, in 2020. He is currently working toward the M.S. degree at the LIESMARS (State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing), Wuhan University, China. His research interests include computer vision, deep learning, and action recognition.

Jia Yifan received the M.S. degree from the school of clinical medicine of Wuhan University in 2013, and is purchasing the Ph.D. degree in Huazhong Agricultural University. He is currently the deputy director and deputy chief physician of the Department of Pain of Renmin Hospital of Wuhan University. He has co-/authored more than 30 academic journal papers. His research interests include pain treatment, and artificial intelligence in medical applications.

Wei Xie received the B.E. degree in electronic information engineering and the Ph.D. degree in communication and information system from Wuhan University, China, in 2004 and 2010, respectively. From 2010 to 2013, he was an Assistant Professor with the Computer School, Wuhan University. He is currently a Professor with the Computer School, Central China Normal University, China. His research interests include motion estimation, super-resolution reconstruction, image fusion, and image enhancement.

Zhigang Tu started his Master Degree in image processing at Wuhan University, China, 2008. In 2015, he received the Ph.D. degree in Computer Science from Utrecht University, Netherlands. From 2015 to 2016, he was a postdoctoral researcher at Arizona State University, US. Then from 2016 to 2018, he was a research fellow at Nanyang Technological University, Singapore. He is currently a professor at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote sensing, Wuhan University. His research interests include computer vision, image processing, video analytics, and machine learning. Special for motion estimation, video super-resolution, object segmentation, action recognition and localization, and anomaly event detection.

He has co-/authored more than 50 articles on international SCI-indexed journals and conferences. He is an Associate Editor of the SCI-indexed journal *The Visual Computer* (IF=2.601), a Guest Editor of *Journal of Visual Communications and Image Representation* (IF=2.678) and *CC&HTS* (IF=1.339). He is the first organizer of the ACCV2020 Workshop on MMHAU (Japan). He received the "Best Student Paper" Award in the 4^{th} Asian Conference on Artificial Intelligence Technology.