IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 October 2023, accepted 9 November 2023, date of publication 14 November 2023,
date of current version 22 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332651

RESEARCH ARTICLE

Attention Relational Network for Skeleton-Based Group Activity Recognition

CHUANCHUAN WANG[1,2] **AND AHMAD SUFRIL AZLAN MOHAMED**[1]
[1]School of Computer Sciences, Universiti Sains Malaysia, George Town, Penang 11800, Malaysia
[2]School of Engineering, Guangzhou College of Technology and Business, Guangzhou 510850, China

Corresponding author: Ahmad Sufril Azlan Mohamed (sufril@usm.my)

This work was supported in part by the Natural Science Foundation of Guangdong Provincial Higher Education Characteristic Innovation Research Foundation under Grant 2019KTSCX258, in part by the Natural Science Foundation of the Guangzhou College of Technology and Business, and in part by the Research on Group Activity Recognition and Video Action Understanding under Grant KYYB202231.

ABSTRACT Group activity recognition is a significant and challenging task in computer vision. The solution of group activity prediction can be classified with traditional hand-crafted features, RGB video features, and skeleton data-based deep learning architectures, such as Graph Convolutional Networks (GCNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTMs). However, they rarely explore pose information and rarely use relational networks to reason about group activity behavior. In this work, we leverage minimal prior knowledge about the skeleton information to reason about the interactions from group activity. The objective is to obtain discriminative representations and filter out some ambiguous actions to enhance the performance of group activity recognition. Our contribution is a proposed Attention Relation Network (ARN) that fuses the attention mechanisms and joint vector sequences into the relation network. The skeleton joints vector sequences are previously unexplored pose information and assign greater significance attributed to individuals who are more relevant for distinguishing the group activity behavior. First, our model focuses on the specified edge-level information (encompassing both edge and edge motion data) within the skeleton dataset, considering directionality, to analyze the spatiotemporal aspects of the action. Second, recognizing the inherent motion directionality, we establish diverse directions for skeleton edges and extract distinct motion features (including translation and rotation information) aligned with these various orientations, thereby augmenting the utilization of motion attributes related to the action. We also introduce a representation of human motion achieved by combining relational networks and examining their integrated characteristics. Extensive experiments were tested in the Hockey and UT-interaction datasets to evaluate our method, obtaining competitive performance to the state-of-the-art. Results demonstrate the modeling potential of a skeleton-based method for group activity recognition.

INDEX TERMS Group activity recognition, attention mechanism, relational network, skeleton joint director sequences.

I. INTRODUCTION

Group activity recognition is a burgeoning area of interest among researchers in computer vision, given its vast array of potential applications, including but not limited to human-computer interaction, security monitoring. References [1] and [2], sports posture correction. Reference [3], automatic driving [4], elderly person fall detection [5], medical and healthcare [6], smart cities [7], and so on. With the advancement of deep learning technology, prior researchers have extensively investigated group behavior recognition, mainly focusing on RGB videos and skeletal data-based approaches. These efforts have yielded numerous outstanding research outcomes [8], [9], [10], [11]. However, recognizing human behaviors amidst the complexities of visual scenes remains challenging.

Given that the human skeleton in the video is primarily depicted as a sequence of joint coordinate lists obtained through pose estimator data. Reference [12], the skeleton sequence can solely capture action information.

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

It only encompasses pose details, excluding background interference, such as background alterations and illumination changes [13]. As a result, many activity recognition algorithms based on Convolutional Neural Networks (CNNs), RNNs, LSTM, and GCNs [14], [15], [16], and even more derivative algorithms are springing up, such as TA-CNN [17] and ST-GCN [9], and so on. Typically, these solutions entail manually designed features and a classification technique that theoretically incorporates the interdependencies among various body parts. Subsequently, they generated a rigid architecture heavily reliant on presumed prior knowledge of the human skeleton structure.

Substantial existing works address various issues of human skeletal joints in action recognition applications. For instance, Abdullahi and Chamnongthai [18], [19] improved attributes extracted from 3D skeletal videos acquired via a Leap Motion controller are employed as a state transition pattern input to a classifier for sign word classification, demonstrating state-of-the-art performance on human action recognition for skeletal 3D data sets.

Additionally, existing algorithms integrate human gesture information with neural network models or spatial and temporal data fusion to better achieve group activity recognition. These approaches have demonstrated remarkable achievements in terms of enhanced performance and accuracy. However, these solutions seldom leverage the information concerning the interplay between individual poses and the interconnection between the postures of two individuals engaged in interactions. Besides, most existing approaches primarily concentrate on modeling information at the joint level, neglecting the skeleton edge size and direction information. However, combining these edge attributes, such as the directionality of human motion, to portray the motion variations information of the action plays a crucial role in action recognition. From a modeling perspective, considering the direction of the human body movement and the size of the skeleton is more natural and logically sound. Not viewing them may lead to suboptimal results.

In this work, we proposed an Attention Relational Network (ARN). Starting from the input layer, the direction of the human body's movement individually pairs up the joints of both individuals. Subsequently, fuse the skeleton joint and temporal streams into independent relation modules. Then, the pair-wise inferred individual's relationships in the final stage of human interaction recognition. The overview of ARN is summarized in Figure 1. The solution is to improve the performance of group activity recognition based on Interaction Relational Network (IRN) but consider more cues of inter-individuals and intra- individuals, such as the individual's inward and outward motion edge attributes. In the initial phase, re-structuring the poses across the video with inward vector sequences and outward vector sequences as input. The features representing the joints as independent objects as the IRN [20] did, comprising the coordinates from multiple frames for the joint and temporal streams separately. Then, we pair-wise inferred individual relationships in the final stage of the attention relational network.

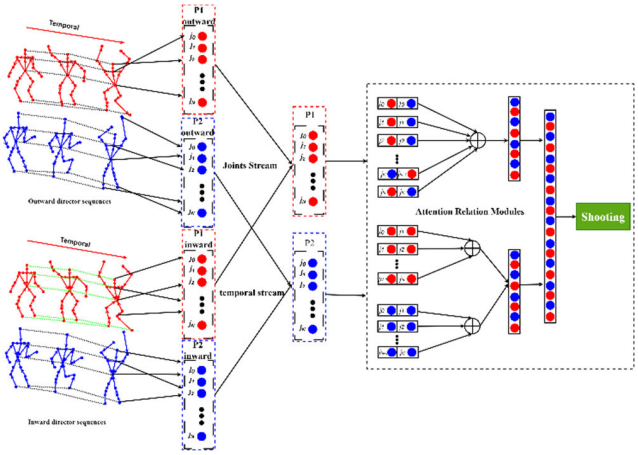

FIGURE 1. Overview of the proposed attention relational network (ARN) architecture.

We propose two relationship mappings for our specific designing modeling problem: an inward vector sequence of inter-person mapping, which pairs joints from one body with joints from the other body, and an outward vector sequence of intra-person mapping, where joints from the same body. Remarkably, the Relation Modules within the same relationship mapping share weights, enabling them to learn about the prevailing relations between joints and discern those pivotal for interaction recognition, all based on the provided data.

Subsequently, the descriptions generated by all the Relation Modules are pooled and directed to a module responsible for interpreting this arrangement of relations and performing appropriate classification. Moreover, we have devised various fusion techniques to combine the two defined types of relationships. By leveraging both forms of information, we achieve enhanced recognition accuracy. Our research underscores the significance of selecting an appropriate fusion architecture and initializing it with suitable models before commencing training.

We validate our approach through experiments on a traditional human interaction recognition dataset: The hockey [21] and UT-interaction [22] datasets. Our proposed solution demonstrates competitive performance to the state-of-the-art on these mutual action datasets.

Our contribution can be briefly summarized as follows:

1. We propose a novel solution to skeleton-based group activity recognition, combining the individual motion edge attributes, treating individual body parts derived from pose information as independent objects, and establishing pair-wise relationships between them.

2. We assess and devise effective fusion methods for diverse relationships, harnessing their complementary aspects to enhance overall performance. Extensive experiments were evaluated on The Hockey and UT-interactions datasets, which obtained promising performance.

3. We further enhance the relational network formulation and expand design ideas for group activity recognition. Following the attention mechanism, our solution enables it to autonomously augment its input with pertinent information extracted from the input pair.

Recent developments in skeleton-based Group Activity Recognition have highlighted the effectiveness of [23] as a potent feature extraction method, surpassing RGB frameworks in terms of efficiency and robustness. However, the potential loss of crucial cues, such as contextual information, when utilizing skeleton data can hinder the distinction of ambiguous actions, resulting in misclassification. Prior research in group activity recognition predominantly relied on RGB videos [8] and diverse feature sets, overlooking an essential aspect: the motion relationships [24] between interacting body parts. Understanding these motion relationships is imperative for comprehensive group activity recognition. Notably, incorporating joint size and motion direction information through additional channels or fusion techniques can significantly enhance the network's ability to capture individual characteristics, discriminate between activities, and generalize across diverse group dynamics, ultimately leading to improved performance and more accurate group activity recognition. The contributions collectively form the novel Attention Relational Network architecture, which is simple yet highly effective and efficient. By conducting extensive validation on substantial datasets, we attain state-of-the-art performance, demonstrating the resilience and effectiveness of the proposed solution.

In this novel endeavor, we introduce an enhanced architecture for relationship fusion, capitalizing on higher-level inferred connections. Another significant addition in this manuscript involves the integration of the Attention mechanism into our framework, facilitating temporal relational reasoning and the capacity to reason across the entire interaction sequence. To comprehensively evaluate our proposed approach, we conducted experiments on two additional datasets with demanding characteristics: UT-Interaction, where pose estimation was required, and Hockey dataset, featuring numerous diverse classes. Lastly, in this recent research, we conducted a more comprehensive qualitative analysis, utilizing confusion matrices and a bar chart to assess performance across interaction classes.

To the best of our knowledge, this work first attempts to study the effectiveness of skeleton joint vector sequences with the ARN for group activity recognition. Our paper is structured as follows. Part. 2 reviews the related work of the topic. Part. 3 presents the overview of the interaction relational network. Part. 4 elaborates on the experiment datasets, details, results, and ablation studies. Finally, Part 5 contains a conclusion, discussion, and future work.

II. RELATED WORK

A. SKELETON-BASED GROUP ACTIVITY RECOGNITION

The primary objective of skeleton-based group activity recognition is to classify action categories utilizing 3D coordinate data of the human body. With the development of deep learning and neural networks, current group activity recognition methods mainly include RNN. References [25] and [26], CNN [27], [28], and GCN-based models [29], [30], [31], [32]. Among these methods, RNN [14] has proven effective in handling time series data by establishing recursive connections and enabling feature extraction and classification after converting the skeleton sequence into a one-dimensional time series. To enhance the temporal context learning ability, improvements in standard RNNs have emerged, such as LSTM [18], [33], [34], [35], [36] and Gated Recurrent Unit (GRU) [34]. However, RNNs suffer from poor spatial modeling ability, difficulties dealing with long-term dependencies, and issues with exploding or vanishing gradients, leading to unsatisfactory recognition accuracy. On the other hand, CNNs excel in spatial feature extraction but are primarily used for image-based tasks. While successful, CNN-based methods encounter challenges like high computational requirements and parameter count. To fully utilize the topological graph structure of the human skeleton and capture spatial dependency between joints, GCNs [37] were introduced. Li et al. [38] proposes a Graph Diffusion Convolutional Network(GDCN) approach that integrates graph diffusion and GCNs for enhanced two-person action recognition. Yan et al. [9] proposed the ST-GCN architectures to represent the skeleton sequence as a spatial-temporal graph, enabling comprehensive capture of human behavior's spatial-temporal change relationship and achieving unprecedented recognition accuracy. However, those methods rarely explore pose information and rarely use relational networks to reason about group activity behavior. They disregard the magnitude and orientation details of the skeletal edges, crucial for action recognition, potentially leading to suboptimal outcomes in these methodologies. Furthermore, integrating the directional aspects of human motion to depict variations in action dynamics, a more intrinsic and rational approach to modeling action sequences, remains largely overlooked in current methodologies.

B. RELATION NETWORK

The architecture of the Relational Network was initially designed by Santoro et al. [39]. The network's ability to reason about the relationships between entities and their properties is crucial for achieving generally intelligent behavior [40]. The network solves the problems for neural networks to generally intelligent behavior. The authors evaluated that it covers not only distinct purposes, question-answering, and physical modeling but also different types of input data, such as visual, textual, and spatial state-based [20] information. Which through three tasks: First, they experimented with the CLEVR dataset to achieve visible question answering (QA) super-human performance; Second, test-based question answering through the bAbI suit; Finally, complex reasoning about dynamic physical systems. Some previous researchers expanded the RNs to Video QA. Such as Hierarchical

Relation Attention (HRA) [41] joins attention modules, and Ibrahim and Mori [42] proposed a hierarchical relational network for group activity recognition.

Benefit from the ability of the Relation Networks to handle relational reasoning, reduce overall network complexity, and gain a general power to reason about the relations between entities and their properties, the RNs are receiving increasing attention from researchers. Perez et al. [43] The proposed Interaction Relation Network(IRN) for mutual action recognition uses pose information for Human Interaction Recognition. Most importantly, it is the first model that extends the RNs for reasoning domain and application. Besides, for group activity recognition, to directly reason about the person interactions, Perez et al. [44] designed Group Interaction Relational Network(GIRN). The authors leverage the skeleton information, including joint relations that were previously not considered, to learn the interactions between the individuals and obtain competitive performance results. However, not only are those solutions usually designed to be complex, but they also rarely explore pose information and rarely use relational networks to reason about group activity behavior. They disregard the magnitude and orientation details of the skeletal edges, crucial for action recognition, potentially leading to suboptimal outcomes in these methodologies.

To our knowledge, no study is available on motion direction and joint size fusion with Attention Relation Networks. Figure 2 shows the ARN architecture for skeleton-based group activity recognition. Our contribution is to leverage minimal prior knowledge about the skeleton information to reason about the interactions from group activity.

III. OVERVIEW OF THE INTERACTION RELATIONAL NETWORK

The Interaction Relational Network(IRN) proposed by Perez et al. [20], [43], which adapted to action recognition for processing directly the joints' information. They drive the relation network to identify how to relate the body parts of the individuals interacting. To better understand the IRN, let's look at the Relation Network(RN) [39].

The Relation Network proposed by Santoro et al. [39], its simplest form is the following equation below:

$$RN(O) = f_\Phi\left(\sum_{i,j} g_\theta(O_i, O_j)\right) \tag{1}$$

Equation (1) lets O be a set of objects where each i^{th} object is represented by an arbitrary $\mathbb{R}^m$ vector, containing its properties. The function g, with learnable parameters θ, serves as the Relational Module, modeling relationships between input object pairs. Meanwhile, function f, with trainable parameters φ, performs reasoning based on the combined relationships inferred by g_θ.

The RN architecture is a versatile formulation for various data types, enabling easy modifications to incorporate extra input information and relationship types as long as objects are paired with shared weights. Perez et al. [20], [43] inspired by RN architecture, they derived three relationships: Inter-person Relationships, Intra-person Relationships, and Fusing Relations—corresponding to equations (2), (3), and (4), respectively.

$$IRN_{inter}(P_1, P_2) = f_\emptyset\left(\sum_{i,k} g_\theta\left(j_i^1, j_k^2\right) \oplus \sum_{i,k} g_\theta\left(j_i^2, j_k^1\right)\right) \tag{2}$$

In theory, $f_\emptyset$ and g_θ can represent Multi-Layer Perceptrons (MLPs) characterized by trainable parameters φ and θ, respectively. Notably, it $\oplus$ can encompass various pooling operations, including summation, maximization, averaging, or concatenation [43]. However, based on our experimental findings, we have opted to employ the averaging operation due to its superior performance.

$$IRN_{intra}(P_1, P_2) = f_{\emptyset'}\Big(\sum_{i=1}^{N}\sum_{K=i+1}^{N} g_\Theta(j_i^1, j_k^1) \frown \sum_{i=1}^{N}\sum_{K=i+1}^{N} g_\Theta\left(j_i^2, j_k^2\right)\Big) \tag{3}$$

Given that the intra-personal relationships among the joints can yield valuable information, we introduce an alternative architecture in which the joints of each individual are paired with the corresponding joints from the same individual. In this scenario, bidirectional pairing is unnecessary, as the paired joints originate from the same individual. Our preliminary experiments have shown that such bidirectional pairing can introduce unnecessary redundancy into our model and, in some instances, may even contribute to overfitting. The aggregated output from each individual is concatenated ($\frown$) before being processed through function f, characterized by its trainable parameters $\emptyset$.

$$IRN_{inter+intra}(P_1, P_2) = f_{\emptyset''}\Big(\sum_{i,k} g_\theta\left(j_i^1, j_k^2\right) \oplus \sum_{i,k} g_\theta\left(j_i^2, j_k^1\right) \sum_{i=1}^{N}\sum_{K=i+1}^{N} g_\Theta\left(j_i^1, j_k^1\right) \frown \sum_{i=1}^{N}\sum_{K=i+1}^{N} g_\Theta\left(j_i^2, j_k^2\right)\Big) \tag{4}$$

Conclusively, we propose an architecture that amalgamates both categories of relationships within a unified function f (parametrically defined by $\emptyset$), achieved through concatenating the pooled information from each function g, each governed by its distinct parameters θ and Θ.

Initially, Perez et al. [20], [43] extract information from each joint separately (j_n) across frames. Subsequently, the set of joints from both individuals (P_n) serves as input to our architectures, IRN_{inter} and IRN_{intra}. Each architecture models different relationships among the joints and can independently predict the action. Additionally, the models can be fused as $IRN_{inter+intra}$, leveraging both relationship types for improved prediction accuracy.

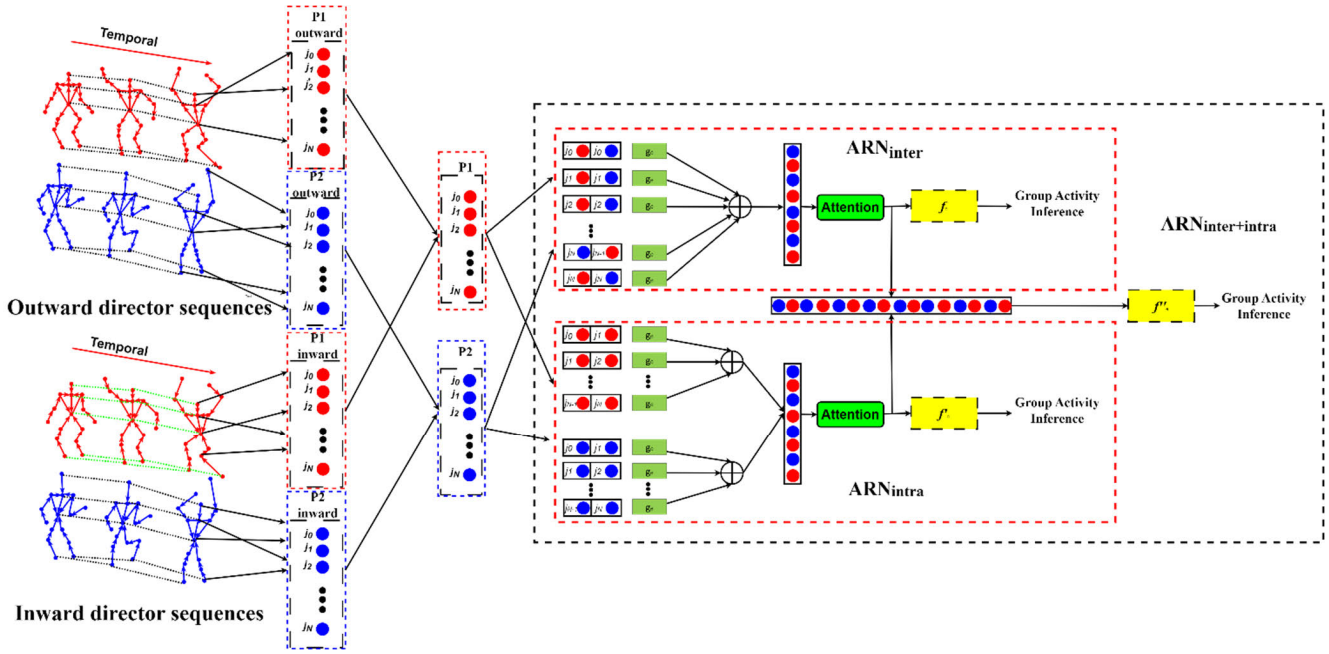

FIGURE 2. Illustration of the ARN architecture for skeleton-based group activity recognition. First, in the input layer, human body joints move inward and outward, and vector direction sequences were pre-extracted. Then, the direction of the human body's movement individually pairs up the joints of both individuals. Subsequently, fuse the skeleton joint and temporal streams into independent relation modules with an attention mechanism. Furthermore, the pair-wise inferred individual's relationships in the final stage of human activity recognition.

IV. ATTENTION RELATIONAL NETWORK

A. REBUILT EDGE-LEVEL INFORMATION

Given that the skeletal edges, formed by contiguous joints of the human physique, adhere to the anatomical framework of the human form, employing directional edge-level data to portray the spatio-temporal attributes of the action holds greater validity. Figure 3 shows that we leverage the inward and outward direction sequences to rebuild the skeleton edge-level information to represent activity motion features.

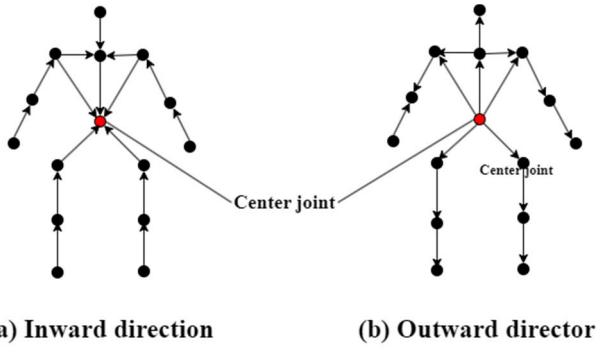

FIGURE 3. Illustration of the human body 15 skeleton joints with direction. The center joint is the red one. (a) human body skeleton joints with an inward direction, the arrow pointing to the center joint, and (b) human body skeleton joints with an outward dir ection, the arrow pointing to opposite the center joint.

B. GENERATE MOTION DIRECTION INFORMATION

Since the human body motion is closely related to the edge vector itself, in this paper, we define the angle between pose before (Figure 4 (a)) and pose after (Figure 4 (b)) two edges in Euclidean n-space [45] as $A_n(a, b)$, can be simple drive the following equation [46], [47]:

$$A_n(a, b) = \cos^{-1}\left(\frac{a \bullet b}{\|a\|\|b\|}\right) \tag{5}$$

where a is the pose before the human body moves edge vector, and b is the pose after the human body moves edge vector.

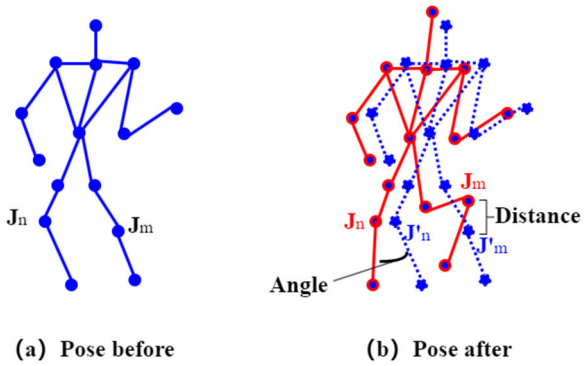

FIGURE 4. Illustration of skeleton motion information in different directions. (a) pose before the human move, remarked blue color, and (b) pose after the human move, remarked red color.

Also, we define the joint moved distance as D_n, the equation is defined as follows:

$$D_n\left(J_m, J'_m\right) = \sqrt{(J_1 - J'_1)^2 + (J_2 - J'_2)^2 + \ldots + (J_m - J'_m)^2} \tag{6}$$

where J_m is one of the skeleton joints, which pose move before, and J'_m is the same joint move after.

C. ATTENTION MECHANISM

While engaging in the same group activity, each individual may assume distinct roles during execution [44]. Some joints can be more crucial or discriminatory in determining the specific activity. Hence, instead of naively averaging the relations from all players. We applied an attention mechanism [48] to our architecture to assign greater weight to potentially significant individuals. This approach enhances the model's ability to focus on key contributors within the group activity recognition process.

V. EXRIMENTS

A. DATASETS

1) THE HOCKEY PENALTY DATASET

Reference [21] includes five Slashing, Holding, Tripping, Hooking, and No Penalty classes, with 76, 80, and 98 clips, respectively. The multi-person videos in this study depict intricate interactions between players within a non-laboratory recording setup. This dataset consists of multi-person videos capturing complex player interactions in a real-world recording setup. Collected from National Hockey League (NHL) broadcasts, it comprises three classes: No Penalty, Tripping, and Slashing, with 98, 80, and 76 videos, respectively. The clips are two to six seconds long, recorded at 30 fps, and presented in actual speed or slow-motion replays. Each penalty is entirely encompassed within the clip's duration, with the clip starting before and ending after the penalty. The dataset poses challenges like view variation, camera motion, occlusions, blurry frames, and complex interactions. It offers ground-truth pose annotations for all players in each clip, including 14 body key points and two hockey stick endpoints [49].

2) UT-INTERACTION DATASET

Reference [22] is a six-type, two-person interactions human action dataset. The six classes of interactions comprise 10 non-periodic atomic-level actions, such as Shaking hands, pointing, hugging, pushing, kicking, and punching are included. The dataset consists of 10 atomic actions, including stretch arm, withdraw arm, stretch leg, lower leg, and shift forward of left and right directions, forming the interactions. It comprises 10 sets, each featuring videos of different pairs of individuals engaged in all six interactions. Sets 1 to 4 show two interacting persons, while sets 5 to 8 involve interacting persons and pedestrians. Sets 9 and 10 depict several interacting persons performing activities simultaneously. Each set presents distinct backgrounds, scales, and illuminations. Across the dataset, 6 participants executed activities under 10 different clothing conditions, resulting in 60 interactions and over 180 atomic actions.

B. IMPLEMENTATION DETAILS

1) INPUT LAYERS

This work uses the Hockey and UT-interactions datasets as data sources. We used the skeleton information extracted by the OpenPose [50] tool. Then, we generated them to our project's sequences format to facilitate the next phase, classifying them as inward and outward vectors, respectively.

2) MLPS CONFIGURATION

For the MLPs configuration, we have fine-tuned the hyper-parameters described here during initial tests. In this work, we built the ARN as an MLP, with the first three layers having 1000 units each and the least with 500. Meanwhile, f_Φ contains a dropout layer for the input, with a dropout rate of 0.10 for ARN_{inter} and 0.25 for ARN_{intra} and $ARN_{inter+intra}$. Then, five fully connected layers with 100, 150, 200, 500, and 200 units are connected to a Softmax layer to perform the video classification. Moreover, training was carried out with the Adam optimizer, a learning rate setting 1e-4, and weight initialization using a truncated normal distribution with a zero mean and 0.045 standard deviations.

TABLE 1. Our result compares with previous methods on the hockey dataset.

Methods	Accuracy (%)
LRCN [51]	63.64
ST-GCN [52]	67.35
PoseC3D [13]	81.63
Askari et al.[21]	93.93
ARN*inter*	92.14
ARN*intra*	95.36
ARN*inter+intra*	**94.12**

3) MODELING

During the modeling phase, we also trained and randomly switched the input order between the joints of the people to help with generalization, which was significantly beneficial for the ARN_{intra} architecture to avoid bias on the order of the concatenated feature generated after the g_Θ. The settings for $ARN_{inter+intra}$ parameters θ and Θ are adjusted and based on the weights obtained previously by training ARN_{inter} and ARN_{intra} respectively. Meanwhile, the parameter φ is initialized at random.

4) JOINTS AND FRAMES SAMPLING

For the Hockey dataset, we sampled only 15 of them, analogous to what is provided by the UT-interactions data. For the Hockey dataset, since the videos are shorter and the frame rate is lower (15 FPS), we used the central 10 consecutive frames as a sampling for our input feature. For the UT-interactions dataset, we first sample half of the frames alternately, then sample the central 32 frames. Our input now has a wider temporal range. Since they are likely to include the more pertinent elements of the encounter, we have opted to sample the center frames.

5) OUTPUT LAYERS

In the output CSV file, we list the results of the training model, including the accuracy, precision, recall values, etc.

TABLE 2. Result from our methods on the UT-interaction dataset. Experiments were conducted on sub-datasets UT-1 and UT-2 to evaluate the network performance and average accuracy.

Methods	Accuracy (%)		Average acc(%)
	UT-1	UT-2	
Kong and Fu [53]	93.33	91.67	92.50
Lan et al.[54]	88.33	83.33	85.83
T Lan et al.[55]	83.33	81.67	82.50
Berlin et al.[56]	95.00	88.00	91.50
Fadime et al.[57]	95.00	92.00	93.50
Mohamed et al.[58]	82.00	86.00	84.00
Ke et al. [59]	93.33	91.67	92.50
IRN*inter+intra*[20]	96.08	96.10	96.09
LSTM-*IRN'inter+intra*[43]	98.27	96.65	97.47
ARN*inter*	96.17	95.34	95.76
ARNintra	95.10	92.33	93.72
ARN*inter+intra*	**97.77**	**97.12**	**97.45**

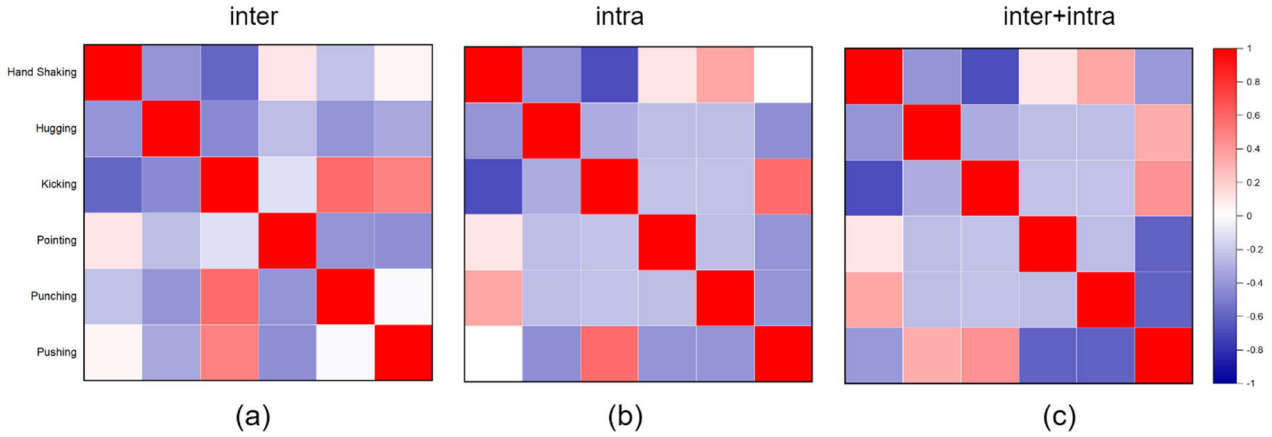

FIGURE 5. Confusion matrices for UT-interaction dataset with methods:(a) ARN_{inter}, (b) ARN_{intra}, (c) $ARN_{(inter+intra)}$.

This layer makes observing experimental outcomes and statistical information more accessible for us.

C. EXPERIMENTAL RESULTS

To evaluate our proposed methodology with the accuracy matric, we first report our best results on the Hockey dataset, shown in Table 1. Here, we compared previous work results to our completed method.

In this experiment, our baseline architectures, ARN_{inter} and ARN_{intra}, achieved accuracies of 92.14% and 95.36%, respectively, demonstrating the successful mapping of various relationships in the skeleton-based group activity recognition problem. Attempting to fuse the two models by simply averaging their scores proved ineffective, yielding lower performance than only ARN_{inter}. Conversely, our approach of integrating the models into a single architecture ($ARN_{inter+intra}$) exhibited better correlation among distinct types of relationships, resulting in a slight performance improvement.

Comparing our best results for the Hockey dataset in Table 1 with those of previous works, it becomes evident that our approach, $ARN_{inter+intra}$ has achieved markedly superior accuracy compared to the previous methods. It is crucial to emphasize that our approach currently relies on a fixed number of sampled frames. Specifically, we utilized only 10 frames for the SBU dataset. In contrast, most other methods can use all frames within the videos, thus incorporating more information than our proposed solution.

Table 2 presents the results of our experiments on the UT-interactions dataset. In contrast to the previous approaches, our ARN_{inter} architecture exhibited superior performance compared to previous methods, with a notable difference of

1-3% higher accuracy. Nevertheless, the fusion of architectures remains advantageous, and the $ARN_{inter+intra}$ fusion architecture achieves equivalent performance. Moreover, our approach outperforms previous studies on the UT-2 dataset. It demonstrates that our solution possesses certain advantages.

D. ABLATION STUDY

In this section, we evaluate the performance of our algorithms from a quantitative analysis perspective for a more comprehensive performance analysis of our method and to visually identify variations in results between our implementations with the UT-dataset, we present the ARN_{inter}, ARN_{intra}, and $ARN_{inter+intra}$ confusion matrices in Figure 5, respectively.

The figure contains the confusion matrices for all ARN architectures. Focusing first on the three-confusion matrix, consistent confusion, including Handing Shaking, Hugging, kicking, Pointing, Punching, and Pushing, can be seen for the same type of activities. Notably, the two relationship models exhibit confusion in distinct interaction cases. For instance, the ''Inter'' (ARN_{inter}) model displays confusion between ''Pushing'' and ''Handing Shaking,'' whereas the ''Intra'' ARN_{intra} model does not. This confusion is significantly reduced when both models are combined in ''Inter+Intra.'' ($ARN_{inter+intra}$). Moreover, interaction classes that were previously confused by both models, such as ''Pushing'' and ''Shaking Hands,'' are nearly entirely distinguishable from ''Inter+Intra.'' This qualitative analysis indicates the efficacy of our proposed architecture in preserving the strengths of both relationship models and harnessing their complementary attributes to differentiate even more challenging cases.

VI. CONCLUSION, DISCUSSION, AND FUTURE WORK

A. CONCLUSION

In this work, we proposed a novel Attention Relational Network (ARN) architecture for skeleton-based group activity recognition. We demonstrated its substantial value for two-person activity recognition, leveraging pose information to analyze the relationships among different body parts during the actions. Our proposed solution achieved promising performance on the conventional interaction dataset UT-interactions, and it also outperformed other approaches on the Hockey dataset subset involving mutual activities exclusively.

B. DISCUSSION

Table 2 shows that our method performs equivalent to the state-of-the-art (SOTA) on the UT-interactions dataset and achieves SOTA results on the subset UT-2. Perez et al. [43] presents 98.27% and 96.65% accuracy on datasets UT-1 and UT-2, respectively. What's more, the average accuracy is up to 97.47%. In this paper, we demonstrated how to adapt the skeleton-based Attention Relation Network for interaction recognition from group activity. Specifically, our proposed method fuses the attention mechanisms and joint vector sequences into the relation network. The skeleton joints vector sequences are previously unexplored pose information and assign greater significance attributed to individuals who are more relevant for distinguishing the group activity behavior. Evaluate the Hockey dataset, Askari et al. [21] reports 93.93% accuracy using RNN equipped with a time-varying attention mechanism, our baseline architectures, ARN_{inter} and ARN_{intra}, achieved accuracies of 92.14% and 95.36% respectively, demonstrating the successful mapping of various relationships in the skeleton-based group activity recognition problem. Then, we attempted to fuse the two models by simply averaging their scores, but proved ineffective, yielding lower performance than only ARN_{inter}. Conversely, our approach of integrating the models into a single architecture ($ARN_{inter+intra}$) exhibited better correlation among distinct types of relationships, with an outcome of 94.12% accuracy a slight performance improvement.

C. FUTURE WORK

We will extend our achievements, providing the ARN with higher-level information, such as features derived from an LSTM or Graph Convolutional Network (GCN) approach, which may yield improved results. Our method still has room to improve by emphasizing commonality over individuation for group activity recognition.

DECLARATION OF COMPETING INTEREST

The authors declared no competing interests.

ACKNOWLEDGMENT

The authors appreciate IEEE Access reviewers' prospective evaluations and valuable insights.

REFERENCES

[1] I. R. Dave, C. Chen, and M. Shah, ''SPAct: Self-supervised privacy preservation for action recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20132–20141.

[2] W. Liu, G. Kang, P.-Y. Huang, X. Chang, L. Yu, Y. Qian, J. Liang, L. Gui, J. Wen, P. Chen, and A. G. Hauptmann, ''Argus: Efficient activity detection system for extended video analysis,'' in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 126–133.

[3] Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang, ''MultiSports: A multi-person video dataset of spatio-temporally localized sports actions,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13516–13525.

[4] A. J. Siddiqui and A. Boukerche, ''A novel lightweight defense method against adversarial patches-based attacks on automated vehicle make and model recognition systems,'' *J. Netw. Syst. Manage.*, vol. 29, no. 4, Oct. 2021.

[5] J. Liu, R. Tan, G. Han, N. Sun, and S. Kwong, ''Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding,'' *IEEE Trans. Multimedia*, vol. 23, pp. 3684–3699, 2021.

[6] R. Presotto, G. Civitarese, and C. Bettini, ''FedCLAR: Federated clustering for personalized sensor-based human activity recognition,'' in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2022, pp. 227–236.

[7] M. H. Sharif, L. Jiao, and C. W. Omlin, ''CNN-ViT supported weakly-supervised video segment level anomaly detection,'' *Sensors*, vol. 23, no. 18, p. 7734, Sep. 2023.

[8] X. Wang and Q. Ji, "Hierarchical context modeling for video event recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1770–1782, Sep. 2017.
[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.
[10] J. Rajasegaran, G. Pavlakos, A. Kanazawa, C. Feichtenhofer, and J. Malik, "On the benefits of 3D pose and tracking for human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–9.
[11] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10608–10617.
[12] Y. Liu, P. Xue, H. Li, and C. Wang, "A review of action recognition using joints based on deep learning," *J. Electron. Inf. Technol.*, vol. 43, no. 6, pp. 1789–1802, 2021.
[13] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.
[14] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.
[15] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.
[16] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, Apr. 2021.
[17] K. Xu, F. Ye, and Q. Zhong, "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1–9.
[18] S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition of skeletal videos using processed video driven multi-stacked deep LSTM," *Sensors*, vol. 22, no. 4, p. 1406, Feb. 2022.
[19] S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach," *IEEE Access*, vol. 10, pp. 15911–15923, 2022.
[20] M. Perez, J. Liu, and A. C. Kot, "Interaction recognition through body parts relation reasoning," in *Pattern Recognition*. Auckland, New Zealand, Cham, Switzerland: Springer, 2020.
[21] F. Askari, R. Ramaprasad, J. J. Clark, and M. D. Levine, "Interaction classification with key actor detection in multi-person sports videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3579–3587.
[22] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1593–1600.
[23] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang, "GAIM: Graph attention interaction model for collective activity recognition," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 524–539, Feb. 2020.
[24] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowl.-Based Syst.*, vol. 158, pp. 43–53, Oct. 2018.
[25] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.
[26] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–10.
[27] 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.
[28] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1971–1980.
[29] X. Wang, X. Xu, and Y. Mu, "Neural Koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10597–10607.
[30] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13339–13348.
[31] 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. 12018–12027.
[32] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1–10.
[33] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*, 2012, pp. 37–45.
[34] S. Farah, W. David A, N. Humaira, Z. Aneela, and E. Steffen, "Short-term multi-hour ahead country-wide wind power prediction for Germany using gated recurrent unit deep learning," *Renew. Sustain. Energy Rev.*, vol. 167, Oct. 2022, Art. no. 112700.
[35] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.
[36] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 636–647, Feb. 2022.
[37] R. Li, S. Wang, and F. Zhu, "Adaptive graph convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.
[38] S. Li, X. He, W. Song, A. Hao, and H. Qin, "Graph diffusion convolutional network for skeleton based semantic recognition of two-person actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8477–8493, Jul. 2023.
[39] A. Santoro, D. Raposo, and D. G. Barrett, "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[40] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1988–1997.
[41] M. I. Hasan Chowdhury, K. Nguyen, S. Sridharan, and C. Fookes, "Hierarchical relational attention for video question answering," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 599–603.
[42] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 1–16.
[43] M. Perez, J. Liu, and A. C. Kot, "Interaction relational network for mutual action recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 366–376, 2022.
[44] M. Perez, J. Liu, and A. C. Kot, "Skeleton-based relational reasoning for group activity analysis," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108360.
[45] R. Lyons, "Distance covariance in metric spaces," *Ann. Probab.*, vol. 41, no. 5, pp. 3284–3305, Sep. 2013.
[46] Y. Sohn and N. S. Rebello, "Supervised and unsupervised spectral angle classifiers," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 12, pp. 1271–1282, 2002.
[47] A. Rosenfeld and J. S. Weszka, "An improved method of angle detection on digital curves," *IEEE Trans. Comput.*, vol. C-24, no. 9, pp. 940–941, Sep. 1975.
[48] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[49] F. Askari, R. Jiang, Z. Li, J. Niu, Y. Shi, and J. J. Clark, "Self-supervised video interaction classification using image representation of skeleton data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 5229–5238.
[50] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[51] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
[52] M. Jiang, J. Dong, D. Ma, J. Sun, J. He, and L. Lang, "Inception spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Int. Symp. Control Eng. Robot. (ISCER)*, Feb. 2022, pp. 208–213.
[53] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 167–178, Jan. 2016.
[54] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2658–2665.
[55] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *Computer Vision—ECCV 2014*, Zurich, Switzerland, Cham, Switzerland: Springer, 2014.
[56] S. J. Berlin and M. John, "Human interaction recognition through deep learning network," in *Proc. IEEE Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2016, pp. 1–4.
[57] F. Sener and N. Ikizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 63–73, Oct. 2015.
[58] M. R. Amer and S. Todorovic, "Sum product networks for activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 800–813, Apr. 2016.
[59] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Leveraging structural context models and ranking score fusion for human interaction prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1712–1723, Jul. 2018.

CHUANCHUAN WANG received the B.S. degree in electronics and information engineering from Xidian University, Xi'an, China, in 2013, and the master's degree from the School of Computer Science, Central South University, Changsha, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia. From 2016 to 2022, he was a Software Engineer, working in internet software development, including games and educational platforms. He is actively exploring research interests, including computer vision and image processing.

AHMAD SUFRIL AZLAN MOHAMED received the B.I.T. degree (Hons.) from Multimedia University, Malaysia, and the M.Sc. degree from The University of Manchester, U.K., and the Ph.D. degree from the University of Salford, U.K. He is currently an Associate Professor with the School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia. His research interests include image processing, video tracking, facial recognition, virtual reality, motion capture, and medical imaging.