IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 August 2022, accepted 12 October 2022, date of publication 14 October 2022, date of current version 27 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3214812

RESEARCH ARTICLE

Pose-Guided Graph Convolutional Networks for Skeleton-Based Action Recognition

HAN CHEN, (Graduate Student Member, IEEE), YIFAN JIANG, AND HANSEOK KO, (Senior Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Hanseok Ko (hsko@korea.ac.kr)

This work was supported by the Major Project of the Korea Institute of Civil Engineering and Building Technology (KICT) [Development of Modular Integrated Equipment System for Emergency Response to Virus Disaster_Development of Modular Facilities and Quarantine System for Preventing Spread of Infectious Diseases and Rapid Response (grant number 20220238-001)].

ABSTRACT Graph convolutional networks (GCN), which can model the human body skeletons as spatial and temporal graphs, have shown remarkable potential in skeleton-based action recognition. However, in the existing GCN-based methods, graph-structured representation of the human skeleton makes it difficult to be fused with other modalities, especially in the early stages. This may limit their scalability and performance in action recognition tasks. In addition, the pose information, which naturally contains informative and discriminative clues for action recognition, is rarely explored together with skeleton data in existing methods. In this work, we proposed pose-guided GCN (PG-GCN), a multi-modal framework for high-performance human action recognition. In particular, a multi-stream network is constructed to simultaneously explore the robust features from both the pose and skeleton data, while a dynamic attention module is designed for early-stage feature fusion. The core idea of this module is to utilize a trainable graph to aggregate features from the skeleton stream with that of the pose stream, which leads to a network with more robust feature representation ability. Extensive experiments show that the proposed PG-GCN can achieve state-of-the-art performance on the NTU RGB+D 60 and NTU RGB+D 120 datasets.

INDEX TERMS Action recognition, attention mechanism, feature fusion, graph convolutional networks, human skeleton, pose information.

I. INTRODUCTION

Human action recognition is crucial in various applications ranging from video surveillance and human-computer interaction to video understanding [1], [2], [3], [4], [5]. In recent years, skeleton-based human action recognition has attracted significant research attention due to the development of low-cost motion sensors and their robustness when faced with complicated environments such as background clutter and changes in illumination.

Skeleton data for human recognition are made up of time sequences of 3D coordinates for human joints derived from pose estimation methods or the direct measurement by sensors, e.g., Kinect and wearable inertial measurement units [6]. Early deep-learning-based action recognition methods feed the skeleton sequences into a recurrent neural network (RNN) [7], [8] or employ them as a pseudo-image input for a convolutional neural network (CNN) [9], [10], [11], [12], [13] to classify the action labels. To further explore the inherent correlations between human joints, graph convolutional networks (GCN)-based methods [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] have been proposed to model the natural topological structure of the human body and have achieved promising results in the human action recognition tasks.

Despite the encouraging results achieved by previous work, state-of-the-art GCN-based methods are limited in the following aspects: 1) Flexibility: Existing GCN-based methods mainly employ manually pre-defined graph topologies to model the natural connections between human joints [14], [15], [16], [17]. However, this ignores the relationship between unconnected joints such as the hands and legs,

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

which may losses implicit joint correlations, especially for the higher-level features, and limits the representation ability of GCN. 2) Scalability: The graphical form of the skeleton representation limits the fusion with other modalities [18], [19], [20], [21], especially in the early or low-level stages, thus making it difficult to learn the features from one data stream under the supervision of the other data stream, which restricts the recognition performance. 3) Ignorance of pose information: [22], [23], [24]. Human pose data carry rich information on the spatial and temporal dynamics of human joints and have been proved to be of great help for action recognition tasks [29], [30], [31], [32]. However, few studies have considered utilizing pose data to enhance skeleton-based model performance, especially with the GCN-based models.

In this work, a novel pose-guided graph convolutional network (PG-GCN) is proposed. Our motivation is derived primarily from the fact that human actions are made up of motions that can be represented by skeleton sequence or pose sequence. Namely, a possible approach to overcome limitations on recognition potential is to use pose data along with skeleton data in order to get richer information about the object. Given this, we explore a dynamic skeleton graph guided by pose data to resolve the above-mentioned issues. Specifically, (1) To enhance the scalability of the network, instead of solely employing skeleton data as input for the GCN, we develop a multi-stream architecture suitable for multi-modal inputs (i.e., pose and skeleton data). (2) To improve the flexibility of the network, a dynamic attention module is proposed for feature fusion across different streams in the early stages. It is achieved by employing a shared graph that bridges and refines the learned features from the skeleton data with those from the pose data. This module is trained and updated jointly with other graph convolutions within the model and used for dynamically adjusting the skeleton graph, thus enhancing the flexibility in constructing the graph for the skeleton. Through the multi-stream architecture and dynamic attention module, the features from the skeleton stream are aggregated with the pose information, and the robust pose-guided skeleton graph features are then used for classification, which enhances the generalization and representation ability of our proposed model.

To verify the superiority of our PG-GCN, extensive experiments were conducted on two challenging datasets: NTU RGB+D 60 and NTU RGB+D 120. The experimental results show that our model outperforms most state-of-the-art approaches. Ablation analysis of the proposed method confirms the effectiveness of the dynamic pose-guided module. The main contributions of this work are summarized below:

- We proposed the PG-GCN, a multi-modal framework for human action recognition that can effectively fuse pose information with skeleton data and be trained end-to-end.
- We proposed a dynamic pose-guided attention module (PG-AM) that employs a trainable shared graph to extract and fuse features across multi-stream inputs, providing more powerful graph modeling capabilities and generalization.
- We conducted extensive experiments to show that the proposed PG-GCN outperforms state-of-the-art methods on two skeleton-based action recognition benchmarks, NTU RGB+D 60 and NTU RGB+D 120.

II. RELATED WORK

A. SKELETON-BASED ACTION RECOGNITION

With the development of deep learning, data-driven methods have become widespread for human action recognition [33], [34], [35]. Some early studies utilized RNNs or CNNs to learn the temporal dynamics of skeleton sequences [7], [8], [9], [10], [11], [12], [13]. However, these methods failed to represent the structure of the skeleton data, which are naturally embedded in graphs. Recently, GCN have been widely adopted for skeleton-based action recognition [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] due to their ability to represent relationships between human body joints. The first attempt at a GCN for action recognition was ST-GCN [14], in which the spatial graph convolutions and temporal convolutions were combined for spatio-temporal modelling. Following this work, Liu et al. [24] proposed MS-G3D and explored the effects of a multi-adjacency GCN for action recognition. However, despite the success of GCN in skeleton-based action recognition, most current methods employ a topology that is pre-defined according to the human body structure, and this topology is fixed in both the training and testing phases, which limits the generalization ability of the model.

To further explore discriminative features and boost the performance of the skeleton-based action recognition models, efforts have been made to extract patterns from other modalities. However, most methods focus on the fusion with modalities such as RGB and depth, and few of them have considered the pose information, which carries rich information on the spatial and temporal dynamics of human joints. A recent work [32] took advantage of pose estimation results and verified their effectiveness in action recognition. This method embedded temporal pose estimation results as a 3D feature representation and then sent it to a 3D CNN to learn the spatio-temporal features. However, this model failed to explicitly exploit the relationship between the pose information and skeleton sequences.

B. POSE FOR ACTION RECOGNITION

Pose coordinates and skeleton data are closely related information because both are concerned with the understanding of human motion. Recently, some research has proven the effectiveness of pose information for action recognition. Yan et al. [14] used OpenPose [36] to extract the pose from each frame and then tested their skeleton-based action recognition. Liu et al. [37] proposed the utilization of pose heatmaps estimated from RGB images input to enhance the skeleton-based action recognition. Some studies [29], [30], [31] attempted to solve both action recognition and pose

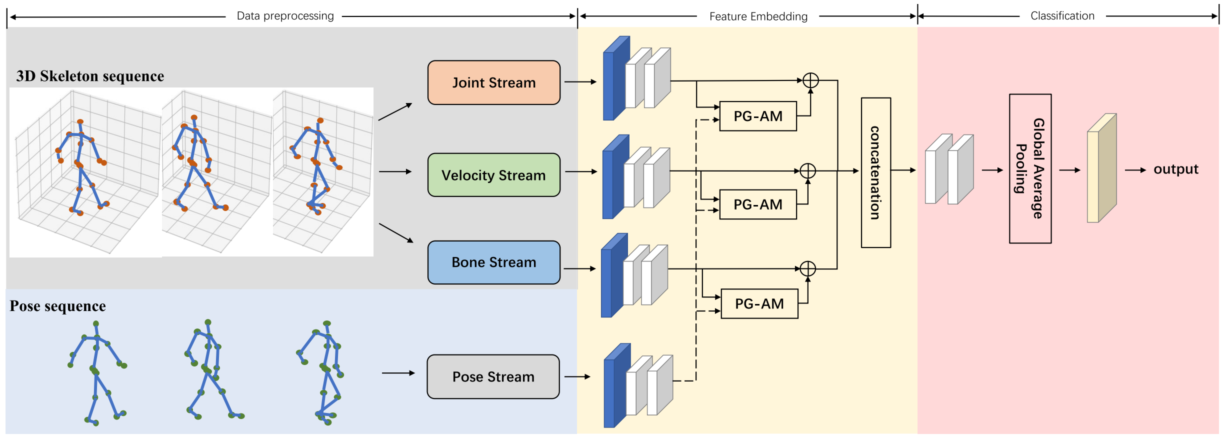


FIGURE 1. Overview of the proposed PG-GCN model. A pair of skeleton sequence and pose sequence from the same action fragment $\{I_s, I_p\}$ are first pre-processed and then fed into the feature embedding module to obtain the feature representations $\{F_s, F_p\}$. Then, the pose-guided attention module (PG-AM) computes the attention summaries that encode the correlations between F_s and F_p. Finally, the skeleton graph representation encoded with pose information will be handed over to the classification module to produce the action classification predictions.

estimation at the same time with a multi-task framework, further confirming that pose features can be used for action recognition. Despite the great help of pose information for action recognition, recently proposed pose-based action recognition algorithms are less powerful and not sufficiently robust against the noise present during the pose estimation process. Additionally, how to incorporate the pose data and skeleton data to best take advantage of the relationship of these two types of data for action recognition remains a problem to be solved. Based on the above findings, in this work, we proposed to utilize the pose data as the guide information for updating our GCN, thus avoiding the instability of pose data and enhancing the performance of GCN-based action recognition.

III. PROPOSED ALGORITHM

Our PG-GCN formulates action recognition as a pose-guided graph representation learning process. The pose-guided attention module (PG-AM) learns to explicitly encode correlations between the pose and skeleton from the same sequence, enabling PG-GCN to fuse multi-stream inputs, thus further helping to discover the generalized features and producing more robust recognition results. Specifically, during training, the pose-guided procedure can be decomposed into correlation learning between the learned graph feature pairs from the same sequence (Fig. 1). During testing, the PG-GCN takes advantage of the pose-guided attention information between the pose and skeleton input. We elaborate on the pose-guided attention mechanism in Section III-A and detail the overall PG-GCN architecture in Section III-B.

A. POSE-GUIDED ATTENTION MODULE IN THE PG-GCN

1) VANILLA POSE-GUIDED ATTENTION

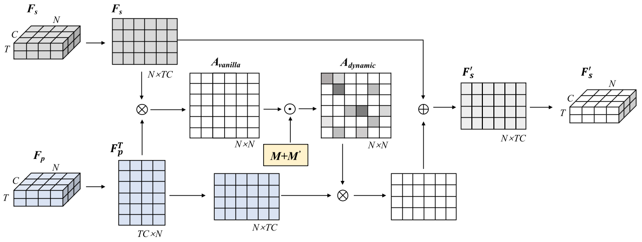


FIGURE 2. Illustration of our PG-AM. The yellow box indicates that the parameter is learnable. ⊗ denotes the matrix multiplication. ⊙ denotes element-wise multiplication. ⊕ denotes the element-wise summation.

As shown in Fig. 2, the two types of inputs are 2D pose sequence I_p and 3D skeleton sequence I_s from the same action fragment. $F_p \in \mathbb{R}^{T \times N \times C}$ and $F_s \in \mathbb{R}^{T \times N \times C}$ denote the corresponding feature representations from the feature-embedding network. F_p and F_s are 3D tensors with C channels, T frames and N joints. The proposed pose-guided attention mines the correlations between I_p and I_s in the feature-embedding space. This is achieved using $A \in [0, 1]$ to convert features from one input stream to another. For example, the features of the hand joint in the pose feature map can potentially guide the feature learning for the arm joint in the skeleton graph. To achieve pose-guided attention from I_p to I_s, we first compute the affinity matrix between F_p and F_s,

$$A_{vanilla} = softmax(F_s F_p^T) \in \mathbb{R}^{N \times N}, \tag{1}$$

where $F_p \in \mathbb{R}^{N \times (TC)}$ and $F_s \in \mathbb{R}^{N \times (TC)}$ are flattened into a matrix representation. As a result, each entry for $A_{vanilla}$ reflects the similarity between the features of each joint in F_p and F_s.

2) DYNAMIC POSE-GUIDED ATTENTION

Furthermore, we proposed the trainable affinity matrix $A_{dynamic}$, which is also an $N \times N$ matrix. In contrast to the vanilla affinity matrix in Eq. 1, the elements of the dynamic affinity matrix are parameterized and optimized together with the other parameters in the training process. $A_{dynamic}$ is first initialized by $A_{vanilla}$, modeling a prior for the correlation between the features of the pose stream and skeleton streams. Using this adjustable affinity matrix, the model can explore the most beneficial features of the recognition task. Dynamic pose-guided attention is formulated as follows,

$$A_{dynamic} = A_{vanilla} \odot (M + M'), \tag{2}$$

where $M, M' \in \mathbb{R}^{N\times N}$ denotes the trainable parameters and all of their elements are initialized with $1, 0$ respectively. In addition, $\odot$ denotes element-wise multiplication. Thus, during the training process, each element in $A_{dynamic}$ is adaptively tuned to capture a flexible correlation between the pose feature and skeleton features.

After obtaining the affinity matrix $A_{dynamic}$, we use it to fuse the pose features to the skeleton features. Given the pose feature F_p from one action sequence, the skeleton feature is updated as,

$$F'_s = A_{dynamic}F_p + F_s \in \mathbb{R}^{T\times N\times C}, \tag{3}$$

Thus, the feature of each joint in the skeleton stream adaptively absorbs detailed information from F_p. The fused F'_s is fed into the mainstream to produce a final action classification result.

B. FULL PG-GCN ARCHITECTURE

The pipeline of our proposed PG-GCN is presented in Fig. 1. The PG-GCN is fundamentally a framework that consists of three cascaded components: an ST-GCN-based [14] feature-embedding module, a pose-guided attention module (detailed in Section III-A), and a classification module. Inspired by [38], in which the joint positions, motion velocities, and bone features (i.e., lengths and angles) are considered, we employ the same data preprocessing for the skeleton data to produce the input for three skeleton sub-streams to fully exploit the skeleton information. The learned features of each of the three sub-stream are sent to the attention module and the correlation with the pose features is calculated. Finally, the three sub-streams are fused and passed through the classification module.

The feature embedding module is formed by orderly stacking a batch normalization layer for fast convergence, a block implemented by the ST-GCN layer, and two GCN blocks for informative feature extraction. After this module, the pose-guided module is employed to fuse features from the pose and skeleton streams. The pose-encoded skeleton feature maps are then sent into the classification module, which consists of two GCN blocks, a global average pooling layer, and a fully connected layer.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We conducted experiments on two large-scale datasets for action recognition: NTU RGB+D 60 [39] and NTU RGB+D 120 [40]. Ablation analysis was also employed to evaluate the contribution of each component in our PG-GCN.

1) NTU RGB+D 60

This large-scale dataset has been widely used to evaluate action recognition models. It contains 56,000 action clips for 60 action classes. The clips feature 40 volunteers ranging from 10 to 35 years old. Each clip is captured by three Kinect cameras from different angles: −45°, 0°, and 45° angles simultaneously. This dataset provides 3D skeleton sequences and corresponding 2D pose coordinates. A total of 25 human joints are captured. The author of this dataset recommends two benchmarks: 1) Cross-Subject (X-Sub), where half of the 40 volunteers are used for training (40,320 videos) and the rest for testing (16,560 videos); and 2) Cross-View (X-View), in which the sequences captured by cameras 2 and 3 are used for training (37,920 videos), and those captured by camera 1 are used for testing (18,960 videos).

2) NTU RGB+D 120

This is an extension of NTU RGB+D 60 and is currently the largest indoor action recognition dataset. It contains 114,480 action clips for 120 classes. The clips feature 106 volunteers. It also provides 3D skeleton sequences and corresponding 2D pose coordinates. Similarly, two benchmarks are suggested for this dataset: 1) Cross-Subject (X-Sub120), in which half of the 106 subjects are used for training (63,026 videos) and the rest for testing (50,922 videos); and 2) Cross-Setup (X-Setup120), in which the training (54,471 videos) and testing (59,477 videos) sets are split based on the parity of the camera setup IDs.

3) IMPLEMENTATION DETAILS

The batch size was set at 16. Stochastic gradient descent (SGD) was applied as the optimization strategy with the initial learning rate of 0.1 and the weight decay of 0.0001. Cross-entropy was employed as the classification loss function. The libraries involved in our work include PyTorch 1.10.1 [41] for network construction and Scikit-learn 1.0.2 [42] for confusion matrix visualization. The network training was accelerated with an NVIDIA RTX 3090 and an Intel(R) Core i7-9700K CPU.

In the NTU RGB+D 60 and NTU RGB+D 120 datasets, there are at most two people in each clip. We padded the data for the second person with 0 if there are fewer than 2 people in the clip. The maximum number of frames in each clip is 200. Sequences with fewer than 200 frames were padded with 0 at the end. In the experiments for X-View, a transformation [43] was conducted for view alignment.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

We compared our proposed method with the state-of-the-art skeleton-based action recognition methods on both the NTU RGB+D 60 dataset and NTU RGB+D 120 dataset. The comparison methods include RNN-based [7], [8], CNN-based [9], [10], [11], [12], [32], and GCN-based methods [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28].

1) NTU RGB+D 60

Table 1 presents a summary of the comparisons between the proposed method and other approaches. Compared with

TABLE 1. Comparison with state-of-the-art methods on NTU RGB+D 60 dataset with Top-1 accuracy (%). The first section shows RNN or CNN-based methods, while the second section includes GCN-based models.

Method	Year	X-Sub	X-View
STA-LSTM [7]	2017	73.4	81.2
ARRN-LSTM [8]	2018	81.8	89.6
HCN [9]	2018	86.5	91.1
3SCNN [10]	2019	88.6	93.7
ST-GCN [14]	2018	81.5	88.3
AR-GCN [18]	2019	85.1	93.2
PR-GCN [15]	2020	85.2	91.7
PB-GCN [19]	2018	87.5	93.2
GCN-NAS [20]	2019	89.4	95.7
JB-AAGCN [21]	2019	89.4	96.0
EfficientGCN-B0 [16]	2021	89.9	94.7
DGNN [22]	2019	89.9	96.1
4s Shift-GCN [17]	2020	90.7	**96.5**
Dynamic GCN [23]	2020	91.5	96.0
MS-G3D Net [24]	2020	91.5	96.2
PG-GCN (Ours)	2022	**91.8**	95.8

ST-GCN [7], which is currently the most widely used backbone model for skeleton-based action recognition, our PG-GCN exhibits an improvement of over 10% on X-Sub and 7% on X-View. PR-GCN [15] also utilizes pose information to enhance a skeleton-based action recognition model, with the pose data treated as prior information in refining the input skeleton information to reduce the impact of noise. The proposed method also outperforms PR-GCN for both benchmarks. In addition, AR-GCN [18], JB-AAGCN [21], and Dynamic GCN [23] also employ the attention mechanism, but there are obvious differences between these models and our PG-GCN, e.g., our attention is achieved through pose guidance, while these models focus on selecting key joints or introducing semantic information. JB-AAGCN attempts to learn a dynamic graph topology in a data-driven manner, leading it to perform better than our PG-GCN on X-View but significantly worse on X-Sub.

Overall, our model achieves more competitive results than the state-of-the-art models, confirming the superiority of our model. Notably, our method is the first to utilize pose data to guide and train a skeleton-based action recognition model, effectively maximizing the use of the pose information and benefiting the action recognition performance.

2) NTU RGB+D 120

Table 2 presents the experimental results for our proposed model and state-of-the-art methods. Of these methods, DSTA-Net [26], PA-ResGCN-B19 [27], and DualHead-Net [28] are enhanced by an attention mechanism, with the first two models exploring the dependencies between different joints in the skeleton sequence and the third utilizing attention to allow communication between coarse and fine-grained skeleton streams. Our proposed method outperforms DualHead-Net by 0.2% on X-Sub120 and achieves competitive performance compared with the other models, which can be attributed to the utilization of pose information and the pose-guided fusion strategy.

TABLE 2. Comparison with state-of-the-art methods on the NTU RGB+D 120 dataset with Top-1 accuracy (%). The first section shows RNN or CNN-based methods, while the second section includes GCN-based models.

Method	Year	X-Sub120	X-Set120
SkeleMotion [11]	2019	62.9	63.0
TSRJI [12]	2019	82.7	85.0
PoseC3D [32]	2021	86.9	90.3
ST-TR-agcn [25]	2020	65.5	59.7
EfficientGCN-B0 [16]	2021	85.9	84.3
4s Shift-GCN [17]	2020	85.9	87.6
DSTA-Net [26]	2020	86.6	89.0
MS-G3D Net [24]	2020	86.9	88.4
PA-ResGCN-B19 [27]	2020	87.3	88.3
DualHead-Net [28]	2021	88.2	**89.3**
PG-GCN (Ours)	2022	**88.4**	88.8

C. ABLATION ANALYSIS

In this section, we focus on exploration analysis of the PG-GCN components and verify the necessity of our dynamic attention strategy. The experiments were performed on the test set of NTU RGB+D 60 and NTU RGB+D 120. The evaluation criterion is the Top-1 accuracy.

1) EFFECTIVENESS OF INTRODUCING POSE FOR SKELETON-BASED ACTION RECOGNITION

We first studied the effect of the different data modalities for action recognition. In Table 3, we showed the results when using the pose, skeleton, or pose+skeleton data. For pose+skeleton (w/o attention), the learned features from the pose stream were directly concatenated with the skeleton feature maps, serving as a separate stream in the feature-embedding module. The results show that using only pose data can lead to successful action recognition, but the performance is less competitive than using only skeleton data for training. When using both the pose and skeleton data but without attention, we observe a significant drop in performance compared with the skeleton-only model. It indicates that, even though the pose information can contribute to action recognition, simply employing it with the skeleton data as input does not lead to better performance. In contrast, our proposed pose+skeleton (Dynamic attention) outperforms the other approaches across the four benchmarks, which confirms the importance of our proposed pose-guided fusion strategy. We attribute the success of to method to the dynamic pose-guided attention mechanism, which reduces the feature redundancy of the pose data while preserving the discriminative features, which benefits the learning of the features from the skeleton data.

2) EFFECTIVENESS OF THE POSE-GUIDED ATTENTION MECHANISM

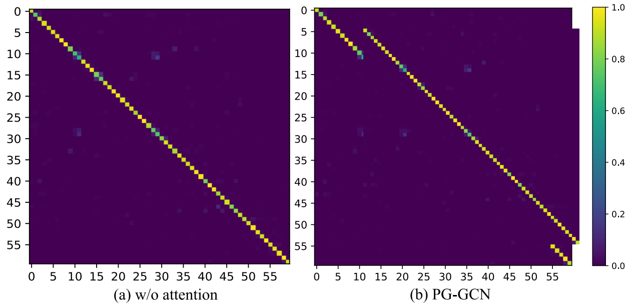
We also studied the effect of different pose-guided attention mechanisms in the PG-GCN, i.e., vanilla pose-guided attention (Eq. 1) and dynamic pose-guided attention (Eq. 3). As shown in Table 4, the dynamic attention achieves better performance than the vanilla attention mechanism. This confirms the importance of the learnable affinity matrix in dynamic attention. Furthermore, we observe a significant

TABLE 3. Comparison of different inputs on NTU RGB+D 60 and NTU RGB+D 120 with Top-1 accuracy (%). This experiment evaluates the effectiveness of pose data for action recognition and emphasizes the need to utilize pose data to improve recognition performance.

Method	X-Sub	X-View	X-Sub120	X-Set120
Pose-only	88.1	90.9	82.7	84.2
Skeleton-only	91.4	95.6	88.2	88.4
Pose+Skeleton (w/o attention)	90.1	94.3	85.5	85.5
Pose+Skeleton (Dynamic attention)	**91.8**	**95.8**	**88.4**	**88.8**

TABLE 4. Comparison of different pose-guided attention mechanisms on NTU RGB+D 60 and NTU RGB+D 120 with Top-1 accuracy (%). We also report the performance when excluding the attention module in our network.

Method	X-Sub	X-View	X-Sub120	X-Set120
w/o attention	90.1	94.3	85.5	85.5
Vanilla attention	91.6	**95.9**	88.3	88.6
Dynamic attention	**91.8**	95.8	**88.4**	**88.8**



(a) w/o attention (b) PG-GCN

FIGURE 3. Confusion matrix for (a) our network without the use of attention and (b) our PG-GCN with dynamic attention on the X-Sub benchmark of NTU RGB+D 60 dataset. The x-axis (true class) and y-axis (predicted class) are associated through the indices of action classes.

reduction in performance when excluding the attention module and simply concatenating the feature maps from the pose and skeleton streams (X-Sub: 91.8 → 90.1). The results clearly verify the effectiveness of our strategy, which employs the attention mechanism to incorporate pose information in the skeleton-based model and allows the model to learn more distinguishable features.

3) ANALYSIS OF CLASSIFICATION CONFUSION MATRIX

To further explore the performance of our proposed method for each action class and evaluate the effectiveness of our proposed pose-guided attention mechanism, we visualized the confusion matrix on NTU RGB+D 60 (Fig. 3). The diagonal represents the correct classification for each action class. The non-diagonal presents the misclassification results across different action classes. Compared with the results for our network without the use of attention, the confusion matrix of our proposed method is cleaner. In other words, our proposed method can achieve more accurate predictions and fewer misclassifications. This success can be attributed to our pose-guided attention mechanism, through which our network can utilize pose information to guide the robust feature learning of the skeleton. However, there are still some cases of failure in our results. For instance, the reading action (11) is often classified as playing with a phone (29), which can be attributed to the fact that these two actions include similar movements and are often confused when using sparse skeleton information. We plan to consider RGB data as complementary information to resolve these types of misclassification in future work.

V. CONCLUSION

In this paper, we proposed PG-GCN, a novel pose-guided multi-model framework for skeleton-based action recognition. We novelly employed the pose information as part of the input to the GCN. To fuse the features of the pose and skeleton streams, we proposed a pose-guided attention module to capture the correlations of the joints in the pose feature map and skeleton feature map as a dynamical guide for learning the graph features. The pose-guided attention module helps the network learn the most discriminating features from the skeleton sequence and improves the overall modeling capability. The proposed method achieved competitive performance on two large-scale action recognition datasets. The experimental results confirmed that our proposed method could effectively leverage pose information to improve action recognition accuracy.

REFERENCES

[1] A. K.-F. Lui, Y.-H. Chan, and M.-F. Leung, "Modelling of pedestrian movements near an amenity in walkways of public buildings," in *Proc. 8th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2022, pp. 394–400.

[2] W. Cao, Z. Zhang, C. Liu, R. Li, Q. Jiao, Z. Yu, and H.-S. Wong, "Unsupervised discriminative feature learning via finding a clustering-friendly embedding space," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108768.

[3] Y. Jiang, D. K. Han, and H. Ko, "Relay dueling network for visual tracking with broad field-of-view," *IET Comput. Vis.*, vol. 13, no. 7, pp. 615–622, Oct. 2019.

[4] Y. Jin, J. Hong, D. Han, and H. Ko, "CPNet: Cross-parallel network for efficient anomaly detection," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2021, pp. 1–8.

[5] Y.-F. Jiang, H. Shin, J. Ju, and H. Ko, "Online pedestrian tracking with multi-stage re-identification," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[6] R. Hou and Z. Wang, "Self-attention based anchor proposal for skeleton-based action recognition," 2021, *arXiv:2112.09413*.

[7] 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.*, vol. 31, 2017, pp. 1–8.

[8] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational network for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 826–831.

[9] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, *arXiv:1804.06055*.

[10] D. Liang, G. Fan, G. Lin, W. Chen, X. Pan, and H. Zhu, "Three-stream convolutional neural network with multi-task and ensemble learning for 3D action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–7.

[11] C. Caetano, J. Sena, F. Bremond, J. A. D. Santos, and W. R. Schwartz, ''SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition,'' in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

[12] C. Caetano, F. Bremond, and W. R. Schwartz, ''Skeleton image representation for 3D action recognition based on tree structure and reference joints,'' in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2019, pp. 16–23.

[13] H. Chen, Y. Jiang, and H. Ko, ''Action recognition with domain invariant features of skeleton image,'' in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2021, pp. 1–7.

[14] S. Yan, Y. Xiong, and D. Lin, ''Spatial temporal graph convolutional networks for skeleton-based action recognition,'' in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–10.

[15] S. Li, J. Yi, Y. A. Farha, and J. Gall, ''Pose refinement graph convolutional network for skeleton-based action recognition,'' *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1028–1035, Apr. 2021.

[16] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, ''Constructing stronger and faster baselines for skeleton-based action recognition,'' 2021, *arXiv:2106.15125*.

[17] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, ''Skeleton-based action recognition with shift graph convolutional network,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 183–192.

[18] X. Ding, K. Yang, and W. Chen, ''An attention-enhanced recurrent graph convolutional network for skeleton-based action recognition,'' in *Proc. 2nd Int. Conf. Signal Process. Mach. Learn.*, Nov. 2019, pp. 79–84.

[19] K. Thakkar and P. J. Narayanan, ''Part-based graph convolutional network for action recognition,'' 2018, *arXiv:1809.04983*.

[20] W. Peng, X. Hong, H. Chen, and G. Zhao, ''Learning graph convolutional network for skeleton-based human action recognition by neural searching,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 2669–2676.

[21] L. Shi, Y. Zhang, J. Cheng, and H. Lu, ''Skeleton-based action recognition with multi-stream adaptive graph convolutional networks,'' *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

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

[23] 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. 55–63.

[24] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, ''Disentangling and unifying graph convolutions for skeleton-based action recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 143–152.

[25] C. Plizzari, M. Cannici, and M. Matteucci, ''Skeleton-based action recognition via spatial and temporal transformer networks,'' *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[26] L. Shi, Y. Zhang, J. Cheng, and H. Lu, ''Decoupled spatial–temporal attention network for skeleton-based action recognition,'' 2020, *arXiv:2007.03263*.

[27] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, ''Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1625–1633.

[28] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, and E. Ding, ''Learning multi-granular spatio–temporal graph network for skeleton-based action recognition,'' in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4334–4342.

[29] D. C. Luvizon, D. Picard, and H. Tabia, ''2D/3D pose estimation and action recognition using multitask deep learning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5137–5146.

[30] M. Liu, F. Meng, C. Chen, and S. Wu, ''Joint dynamic pose image and space time reversal for human action recognition from videos,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8762–8769.

[31] A. Yan, Y. Wang, Z. Li, and Y. Qiao, ''PA3D: Pose-action 3D machine for video recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7922–7931.

[32] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, ''Revisiting skeleton-based action recognition,'' 2021, *arXiv:2104.13586*.

[33] M. Ronald, A. Poulose, and D. S. Han, ''ISPLInception: An inception-ResNet deep learning architecture for human activity recognition,'' *IEEE Access*, vol. 9, pp. 68985–69001, 2021.

[34] T. Wang, S. Lei, Y. Jiang, C. Chang, H. Snoussi, G. Shan, and Y. Fu, ''Accelerating temporal action proposal generation via high performance computing,'' *Frontiers Comput. Sci.*, vol. 16, no. 4, pp. 1–10, Aug. 2022.

[35] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan, and H. Snoussi, ''AED-Net: An abnormal event detection network,'' *Engineering*, vol. 5, no. 5, pp. 930–939, Oct. 2019.

[36] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, ''Realtime multi-person 2D pose estimation using part affinity fields,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[37] M. Liu and J. Yuan, ''Recognizing human actions as the evolution of pose estimation maps,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1159–1168.

[38] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, ''Constructing stronger and faster baselines for skeleton-based action recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 7, 2022, doi: 10.1109/TPAMI.2022.3157033.

[39] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, ''NTU RGB+D: A large scale dataset for 3D human activity analysis,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[40] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, ''NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[41] A. Paszke et al., ''PyTorch: An imperative style, high-performance deep learning library,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and M. Blondel, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

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

HAN CHEN (Graduate Student Member, IEEE) was born in Heilongjiang, China, in 1993. She received the master's degree in communication engineering from Harbin Engineering University, Harbin, China. She is currently pursuing the Ph.D. degree with the Department of Electronics and Computer Engineering, Korea University. Her research interests include intelligent signal process, image segmentation, and human action recognition.

YIFAN JIANG was born in Guangxi, China. He received the Ph.D. degree in electrical engineering from Korea University. He is currently a Research Professor with the Department of Electronics and Computer Engineering, Korea University. His research interests include image synthesis techniques that allow COVID-19 diagnostic approaches and other deep-learning based models to alleviate their dependency on high-quality data while maintaining advanced performance.

HANSEOK KO (Senior Member, IEEE) is a Professor of electrical and computer engineering with Korea University, Seoul, South Korea, and the Director of the Intelligent Signal Processing Laboratory. He has been actively engaged in the research efforts developing solutions addressing the multimodal-based technology issues, including human–machine interaction problems. He is the General Chair of IEEE ICASSP 2024 and Interspeech 2022.

• • •