International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01770-5

Learning Enriched Hop-Aware Correlation for Robust 3D Human Pose Estimation

Shengping Zhang[1] · **Chenyang Wang**[1] · **Liqiang Nie**[2] · **Hongxun Yao**[3] · **Qingming Huang**[4] · **Qi Tian**[5]

Received: 31 July 2022 / Accepted: 9 February 2023

Abstract
Graph convolution networks (GCNs) based methods for 3D human pose estimation usually aggregate immediate features of single-hop nodes, which are unaware of the correlation of multi-hop nodes and therefore neglect long-range dependency for predicting complex poses. In addition, they typically operate either on single-scale or sequential down-sampled multi-scale graph representations, resulting in the loss of contextual information or spatial details. To address these problems, this paper proposes a parallel hop-aware graph attention network (PHGANet) for 3D human pose estimation, which learns enriched hop-aware correlation of the skeleton joints while maintaining the spatially-precise representations of the human graph. Specifically, we propose a hop-aware skeletal graph attention (HSGAT) module to capture the semantic correlation of multi-hop nodes, which first calculates skeleton-based 1-hop attention and then disseminates it to arbitrary hops via graph connectivity. To alleviate the redundant noise introduced by the interactions with distant nodes, HSGAT uses an attenuation strategy to separate attention from distinct hops and assign them learnable attenuation weights according to their distances adaptively. Upon HSGAT, we further build PHGANet with multiple parallel branches of stacked HSGAT modules to learn the enriched hop-aware correlation of human skeletal structures at different scales. In addition, a joint centrality encoding scheme is proposed to introduce node importance as a bias in the learned graph representation, which makes the core joints (e.g., neck and pelvis) more influential during node aggregation. Experimental results indicate that PHGANet performs favorably against state-of-the-art methods on the Human3.6M and MPI-INF-3DHP benchmarks. Models and code are available at https://github.com/ChenyangWang95/PHGANet/.

Keywords 3D human pose estimation · Graph attention network · Multi-scale graph representation

1 Introduction

3D human pose estimation aims to obtain 3D coordinates of human joints from monocular images or 2D keypoints, which has a wide range of applications such as human–computer interaction (Chen et al., 2019b; Pustejovsky & Krishnaswamy, 2021), motion analysis (Liu et al., 2017; Liu & Yuan, 2018), and virtual reality (Mehta et al., 2017b). Although significant efforts have been devoted to developing effective 3D human pose estimation methods, it is still a challenging task since one 2D pose can be mapped to multiple 3D body configurations.

With recent advances in deep learning, a large number of deep neural networks based methods have been proposed, which can be roughly divided into two categories: one-stage

Communicated by Wenjun Kevin Zeng.

✉ Shengping Zhang
s.zhang@hit.edu.cn

Chenyang Wang
c.wang@stu.hit.edu.cn

Liqiang Nie
nieliqiang@gmail.com

Hongxun Yao
h.yao@hit.edu.cn

Qingming Huang
qmhuang@ucas.ac.cn

Qi Tian
wywqtian@gmail.com

[1] Harbin Institute of Technology, Weihai, China

[2] Harbin Institute of Technology, Shenzhen, China

[3] Harbin Institute of Technology, Harbin, China

[4] University of Chinese Academy of Sciences, Beijing, China

[5] Huawei Cloud & AI, Shenzhen, China

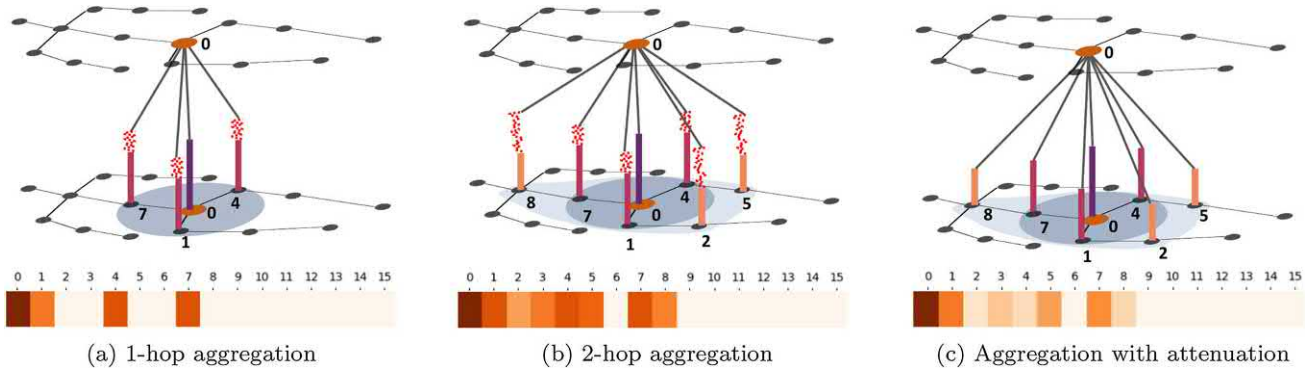

Fig. 1 Illustration of the hop-aware (2-hop as an example) node aggregation. Regard the pelvis (node 0) as the center node. The bars represent the correlation between nodes. The scatters and their heights are noise and noise degrees at different hops. The bottom heatmaps show the attention values for node aggregation. **a** Only 1-hop nodes attend to node aggregation. **b** 2-hop nodes are considered for node aggregation but introduce more noise. **c** Aggregation of 2-hop nodes with attenuation

methods (Chen & Ramanan, 2017; Li & Chan, 2014; Li et al., 2017; Moon & Lee, 2020; Sun et al., 2017) and two-stage methods (Chen et al., 2021; Li et al., 2020; Sharma et al., 2019; Wang et al., 2020; Xie et al., 2021; Zheng et al., 2021). Theformer predicts 3D heatmaps or coordinates of human joints from monocular images directly. However, the complex mapping relationship between monocular images and 3D coordinates limits the performance of these models. Benefiting from the high accuracy and good generalization ability of 2D human joint detection methods (Chen et al., 2018; Newell et al., 2016; Wang et al., 2021b), the two-stage methods use the detected 2D keypoints as the intermediate input to estimate 3D coordinates. Recently, because the detected 2D keypoints can be easily processed as a graph structure, graph convolution network (GCN) has been applied to the two-stage methods. GCNs-based methods (Ci et al., 2019; Hu et al., 2021; Liu et al., 2020a; Xu & Takano, 2021; Zhao et al., 2019; Zou & Tang, 2021) aggregate node semantic features by exploiting adjacent relationships of joints, which learns a strong human graph representation to improve the performance of 3D human pose estimation.

However, existing GCNs-based methods suffer from two main weaknesses. Firstly, as illustrated in Fig. 1a, these methods usually aggregate features from single-hop nodes, which are unaware of the correlation of multi-hop neighbors. The neglect of long-range dependency among distant nodes degrades the performance of estimating complex poses since the change of some joints may affect other joints that are not directly connected. As shown in Fig. 2, the limbs are prone to be wrongly estimated if only single-hop nodes are considered since non-directly connected torso joints affect the movement of the limbs. ModulatedGCN (Zou & Tang, 2021) and GraFormer (Zhao et al., 2022) make attempts to aggregate features among the high-order neighbors of each node. However, ModulatedGCN (Zou & Tang, 2021)

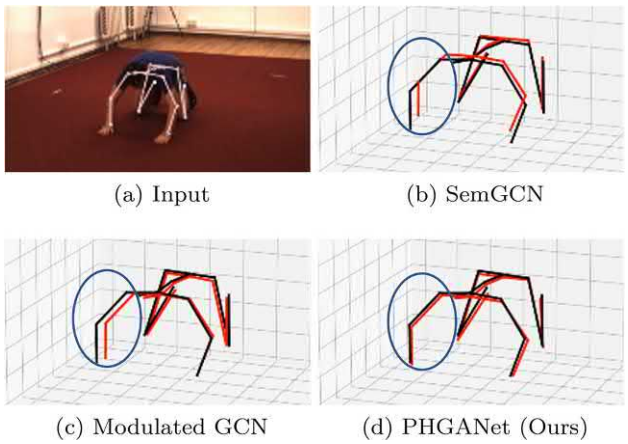

Fig. 2 Comparison results of estimating a complex pose between two 1-hop GCNs-based methods (SemGCN Zhao et al. 2019 and Modulated GCN Zou and Tang 2021) and our PHGANet. The estimation results and the corresponding ground truth are in red and black, respectively (Color figure online)

aggregates all hops neighbor features regardless of their distance, which may introduce useless information even noises. GraFormer (Zhao et al., 2022) can not capture the semantic correlation between nodes, making the model can not adjust the aggregating weights for the same hop.

In addition, existing GCNs-based methods typically operate on single-scale (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021) or sequential down-sampled multi-scale (Xu & Takano, 2021) graph representations. The former obtains 3D human poses from the input without downsampling and thereby maintains spatial details. However, such a single-scale design is less effective in encoding contextual information from different skeletal structures of the human body due to the limited receptive field of GCN. On the other hand, the sequential multi-scale framework downsamples the input to a subgraph for feature extraction and then recovers it

to the original graph structure to obtain the final representation. Although it captures a broad context from the subgraph, the original spatial details are lost due to the downsampling, which makes it difficult to recover the original graph structure accurately. HGN (Li et al., 2021) adopts a bottom-to-up structure to upsample the human graph to a denser graph for capturing multi-scale contexts. However, the upsampled graphs are optimized by the surface vertices of the SMPLify-X (Bogo et al., 2016), which may cause semantic ambiguity and also introduce redundant noises.

To address these problems, we propose a Parallel Hop-aware Graph Attention Network (PHGANet) for 3D human pose estimation, which learns enriched hop-aware correlation of the skeleton joints while maintaining the spatially-precise representations of the human graph. Specifically, we propose a Hop-aware Skeletal Graph Attention (HSGAT) module to capture the semantic correlation of multi-hop nodes, which first calculates skeleton-based 1-hop attention and then disseminates it to arbitrary given hops via graph connectivity. However, attention dissemination is usually accompanied by irrelevant noise from interactions with distant nodes. To alleviate the redundant noise, HSGAT uses an attenuation strategy to separate semantic correlation from distinct hops and adaptively assign them learnable weights according to their distances. Figure 1b, c illustrate that the noise from each hop is alleviated through attention attenuation (the scatters diminish and the corresponding heatmaps lighten in color). Therefore, compared to existing multi-hop methods (Zhao et al., 2022; Zou & Tang, 2021), HSGAT can capture the hop-aware correlation of distant node pairs while alleviating the negative effects introduced by the noise. Based on HSGAT, we further build parallel hop-aware graph attention network (PHGANet) with multiple parallel branches of stacked HSGAT modules to learn the enriched hop-aware correlation of human skeletal structures at different scales. Unlike HGN (Li et al., 2021), PHGANet continuously downsamples the original human graph to refine the node features, which maintains the original graph structure to learn spatially-precise representations of the human skeleton while parallel extracting enriched context from multi-scale subgraphs. In addition, we present a joint centrality encoding scheme to encode node importance as the bias in the input representation, which makes the core joints (e.g., neck and pelvis) more influential during node aggregation.

The contributions of this paper are summarized as follows:

- We propose a hop-aware skeletal graph attention (HSGAT) module to capture semantic correlation from arbitrary hop neighbors while efficiently alleviating undesired noise during the attention dissemination.
- We build parallel hop-aware graph attention network (PHGANet) upon HSGAT with a parallel multi-branch architecture, which maintains the original graph structure while learning enriched hop-aware correlation from skeleton parts at different scales for more robust estimations.
- Extensive experimental results on the Human3.6M and MPI-INF-3DHP datasets demenstrate that PHGANet outperforms the state-of-the-art methods.
  Meanwhile, HSGAT can be used as a plug-and-play component to improve the performance of existing GCNs-based methods.

2 Related Work

2.1 3D Human Pose Estimation

3D human pose estimation has drawn increasing attention during the past decade. In the early days, handcrafted features and geometric constraints are designed to regress 3D human pose (Agarwal & Triggs, 2006; Takano & Nakamura, 2015). Recent progress in 3D human pose estimation is mostly driven by the various deep neural network models. They are roughly divided into two categories: one-stage methods and two-stage methods.

One-stage methods (Chen et al., 2020; Fang et al., 2021; Pavlakos et al., 2017; Zhou et al., 2019) regress heatmaps or coordinates directly from monocular images through an end-to-end framework. Pavlakos et al. (2017) progressively extend the volumetric direction of the predicted heatmap for a fine-gained result. Zhou et al. (2019) regress a volumetric joint heatmap to predict 3D human poses from a single image. Chen et al. (2020) build a part-aware 3D pose estimator to select a suitable architecture to estimate each body part. Fang et al. (2021) exploit mirror symmetry constraints to optimize the performance of human pose estimation. However, since there is no intermediate supervision process, one-stage methods are vulnerable to the background and the lighting of the image. Moreover, the complex features make the learning process of a single model extremely difficult.

The second group of methods (Sharma et al., 2019; Wang et al., 2020; Li et al., 2020; Xie et al., 2021; Chen et al., 2021; Zheng et al., 2021; Luvizon et al., 2022) divide the work into two parts, which first predicts 2D keypoints from 2D detectors and then regresses the 3D pose with the knowledge from the 2D keypoints. Martinez et al. (2017) propose a simple baseline method Simple-3D, to estimate each joint from 2D keypoints. Hossain and Little (2018) utilize temporal features from 2D coordinate information to regress 3D coordinates. Chen et al. (2019a) use an unsupervised method to utilize cycle consistency between 2D pose and 3D projection to estimate 3D joints. Liu et al. (2021) leverage attention-based neural network with dilated convolutions to learn information from videos frames to estimate 3D human poses. Benefiting from the excellent performance of 2D detectors (Chen et al.,

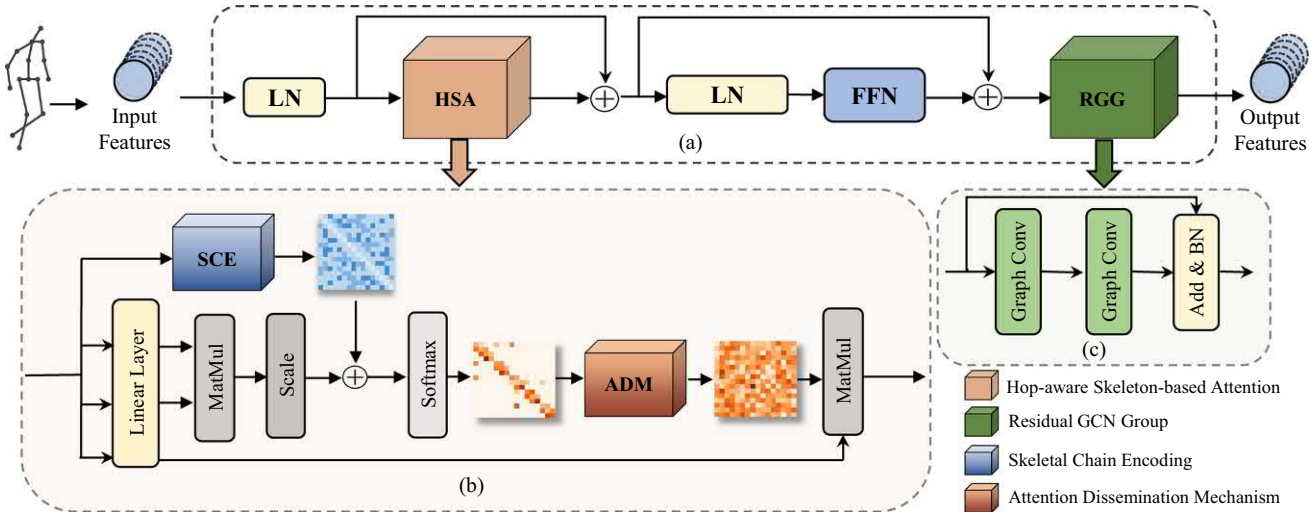

Fig. 3 Illustration of the proposed hop-aware skeletal graph attention (HSGAT) module. **a** The overview of the HSGAT module. **b** The details of Hop-aware Skeleton-based Attention. **c** The structure of Residual GCN Group

2018; Newell et al., 2016; Wang et al., 2021b), the two-stage methods commonly work better than the one-stage methods.

2.2 Graph Neural Networks

Recently, graph neural networks (GNNs) (Duvenaud et al., 2015; Hamilton et al., 2017; Velickovic et al., 2018) show a strong ability on representation learning of graph-structure data. For 3D human pose estimation, GNNs, especially graph convolution networks (GCNs), have become increasingly prevalent since the human skeleton can be easily regarded as a graph. GCN is a convolutional neural network that acts on graph-structure data. It can be grouped into two categories: spectral perspective and spatial perspective. The first category methods (Defferrard et al., 2016; Henaff et al., 2015; Kipf & Welling, 2017) transform the graph data into the Fourier domain for convolution calculations. Existing GCNs-based methods (Ci et al., 2019; Hu et al., 2021; Liu et al., 2020a; Xu & Takano, 2021; Zhao et al., 2019; Zou & Tang, 2021) for 3D human pose estimation fall in the second category, which directly acts convolution operations on graph to regress 3D poses from detected 2D keypoints. Zhao et al. (2019) present SemGCN to learn semantic relationships between human joints through learnable adjacency matrix parameters. Liu et al. (2020a) discuss the effect of different weight sharing methods for GCNs. Zou and Tang (2021) introduce a Modulated GCN to learn different modulation vectors for joints. Xu and Takano (2021) propose a sequential encoder-decoder architecture to explore both local and global semantic features. However, these GCNs-based methods only focus on neighboring nodes, ignoring the influence of non-directly connected nodes, which results in performance degradation.

2.3 Multi-hop Graph Neural Networks

Although GNNs (Duvenaud et al., 2015; Hamilton et al., 2017; Velickovic et al., 2018) improve the efficiency in learning node representations, aggregating information from local neighborhoods limits the performance. To solve the problem, early works (Li et al., 2019; Xu et al., 2018) simply construct deeper networks to extract features from long-range nodes, which causes over-fitting issues. Recently, Abu-El-Haija et al. (2019) design MixHop, which enlarges the receptive field by repeating feature representations of neighbors at various distances. However, with the increasing complexity of the graph, the computation increases significantly. Wang et al. (2021a) propose a multi-hop context-dependent self-attention GNN to solve the over-smoothing, while fixed decay weight makes it not flexible for different nodes at the same distance. For 3D human pose estimation, there are some researchers enlarging the receptive field of GCNs. Zou et al. (2020) propose an implicit fairing method to capture long-range dependencies among nodes. Quan and Hamza (2021) utilize high-order adjacency matrix to extend the scope of GCNs. Zhao et al. (2022) utilize ChebGCN to iterate GCN to improve aggregation distance of interaction joints. However, these methods may introduce noise from distant nodes since they lack the handling of redundant information of less-related nodes.

3 Hop-Aware Skeletal Graph Attention

In this section, we propose a novel graph attention module named hop-aware skeletal graph attention (HSGAT) for 3D human pose estimation, as shown in Fig. 3. HSGAT captures

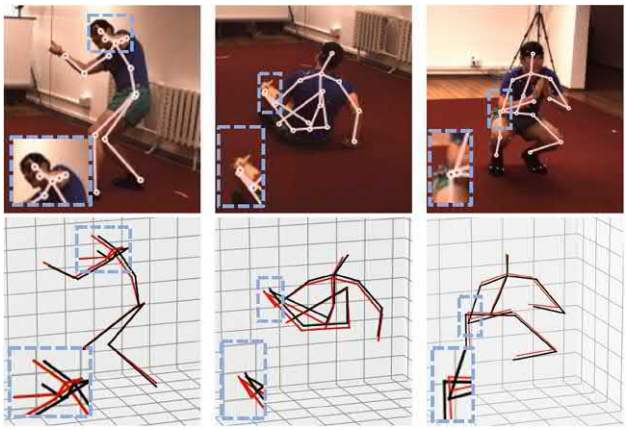

Fig. 4 Examples of complex poses. The upper row is the 2D keypoint input, the bottom row is the comparison between ground truth and predictions from the vanilla GAT. The ground truth are in black and the GAT predictions are in red (Color figure online)

hop-aware correlation with an embedded graph structure prior and alleviates the noise from distant neighbors. The details are as follows.

3.1 Vanilla Graph Attention Network

Graph attention network (GAT) (Velickovic et al., 2018) is a representation of spatial GNNs, which is inspired by the self-attention mechanism in Vaswani et al. (2017). GAT aggregates neighbor features by computing correlation of each node in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges connecting between nodes. In each attention layer, the input is node features $H = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N\}$, where $\boldsymbol{h}_i \in \mathbb{R}^F$ is the feature of each node, N is the number of nodes, and F is the hidden dimension of each node. After parametrized by a weight matrix $W \in \mathbb{R}^{F' \times F}$, the attention value of node j to i is expressed as

$$\alpha_{i,j} = LeakeyReLU\left(\boldsymbol{a}^T \left[W\boldsymbol{h}_i \| W\boldsymbol{h}_j\right]\right) \tag{1}$$

where $\boldsymbol{a} \in \mathbb{R}^{2F'}$ is the weight vector (of potentially different dimension F') and $\|$ is the concatenation operation. The $\alpha_{i,j}$ is computed only for nodes $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the 1-hop neighbors set of node i. After obtaining the attention matrix, the final output feature $\boldsymbol{h}'_i$ of node i is obtained from a linear combination between the neighboring features and the corresponding attention values

$$\boldsymbol{h}'_i = LeakeyReLU\left(\sum_{j \in \mathcal{N}_i} \alpha_{i,j} W\boldsymbol{h}_j\right) \tag{2}$$

where $W \in \mathbb{R}^{F' \times F}$ is the corresponding linear transformation weight vector of the input.

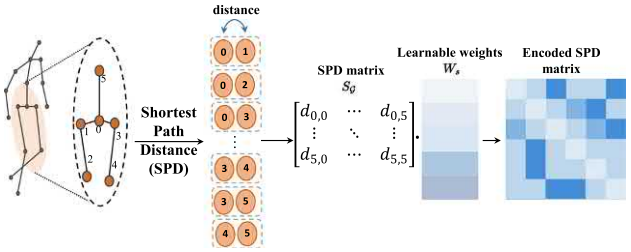

Fig. 5 The details of the skeletal chain encoding method. We use the subgraph with six joints (0–5) as an example to illustrate the process

3.2 Skeleton-Based Attention

Following the vanilla GAT, given a graph representation $H \in \mathbb{R}^{N \times F}$, we first calculate 1-hop correlation $q_{i,j}$ from node j to node i by Eq. (1). Then, according to the graph adjacency matrix $A^\mathcal{G} \in \mathbb{R}^{N \times N}$, the 1-hop correlation matrix $Q \in \mathbb{R}^{N \times N}$ can be expressed as

$$Q_{ij} = \begin{cases} q_{i,j}, & \text{if } A^\mathcal{G}_{ij} = 1 \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

Algorithm 1 Procedure of calculating the shortest path distance matrix.

Input: Human graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Initialize $S_\mathcal{G}$ to be a $|V| \times |V|$ array and initialize all elements in $S_\mathcal{G}$ to be $+\infty$
for each edge (u, v) in $\mathcal{E}$ **do**
  $S_\mathcal{G}[u][v] = 1$;
end for
for each node v in $\mathcal{V}$ **do**
  $S_\mathcal{G}[v][v] = 0$;
end for
for $k = 1$ to $|V|$ **do**
  for $j = 1$ to $|V|$ **do**
    for $j = 1$ to $|V|$ **do**
      if $S_\mathcal{G}[i][j] > S_\mathcal{G}[i][k] + S_\mathcal{G}[k][j]$ **then**
        $S_\mathcal{G}[i][j] = S_\mathcal{G}[i][k] + S_\mathcal{G}[k][j]$
      end if
    end for
  end for
end for
return $S_\mathcal{G}$;

However, the vanilla GAT suffers from performance degradation under some complex actions (e.g. pose or sitting) that have similar 2D coordinates for different joints. This is because the attention mechanism is unable to encode the positional dependency for different joints. For example, when coordinates overlap and depth ambiguity occur in some poses (the upper row in Fig. 4), GAT is confused to distinguish the position and the role of joints, which leads to the offsets of predictions (the bottom row in Fig. 4). To integrate the positional information, attention mechanisms in Computer

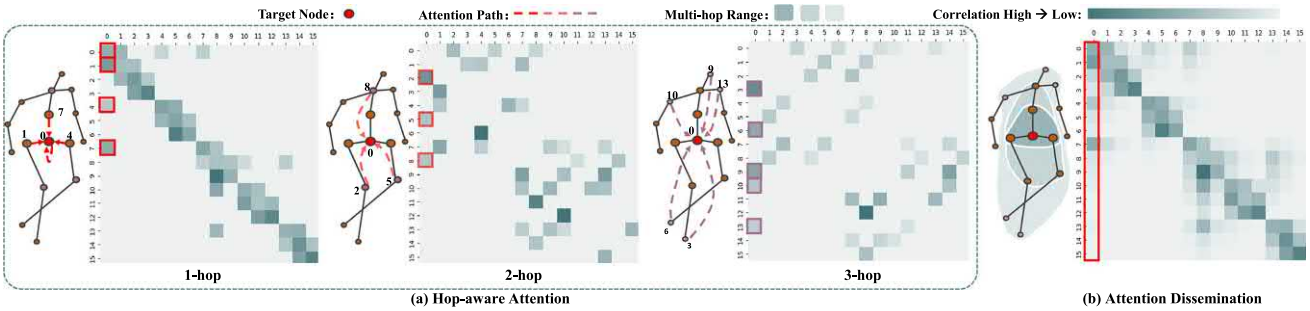

Fig. 6 Visualization of hop-aware attention. **a** Illustration of the attention value at different hops k ($k = 1, 2, 3$ in the figure). **b** The result of attention dissemination from all value in hop-aware attentions

Vision or Natural Language Process leverage the sinusoidal functions, but it is not suitable for graph-structure data since nodes of a graph are not arranged as a sequence. Therefore, we present a Skeletal Chain Encoding (SCE) method to encode the structure information of a human graph, which measures the spatial relation between node pairs. In general, the connectivity between nodes in a human graph is fixed and independent of the spatial position of each joint. From this perspective, we introduce a Shortest Path Distance (SPD) matrix $S_{\mathcal{G}} \in \mathbb{R}^{N\times N}$ to measure the relative position of each joint. Each element $d_{i,j}$ of the SPD matrix $S_{\mathcal{G}}$ represents the number of edges from node i to node j on the shortest path. The details are shown in Fig. 5. Given the human graph $\mathcal{G}$, $S_{\mathcal{G}}$ is calculated by the Floyd–Warshall algorithm. The algorithm compares all possible paths between each pair of nodes until obtaining the shortest path between two nodes. The detailed procedure is described in Algorithm 1. After obtaining $S_{\mathcal{G}}$, we assign the matrix with a learnable weight matrix $W_s \in \mathbb{R}^{N\times N}$ to make each node can adaptively attend to all other neighbors according to the spatial position. With this encoding method, the 1-hop attention matrix Q' of the graph $\mathcal{G}$ can be expressed as

$$Q' = Q + W_s S_{\mathcal{G}} \tag{4}$$

Combined with the skeletal chains encoding, the attention matrix adjusts the 1-hop correlation equipped with the joint position, which preserves the relative spatial position of nodes to handle confused joint entanglements. Then, we apply a softmax operation on Q' to acquire the attention matrix $A_1 \in \mathbb{R}^{N\times N}$ with the skeletal structure information. The skeletal chain encoding helps GAT incorporate the positional dependency into the 1-hop attention matrix, which is named skeleton-based attention.

3.3 Attention Dissemination

However, the skeleton-based attention only considers 1-hop node pairs and ignores non-directly connected nodes, which are essential for a human skeletal structure. As shown in Fig. 6, the 1-hop attention matrix owns a narrow receptive field, which hampers the ability of the network to gather long-range node features.

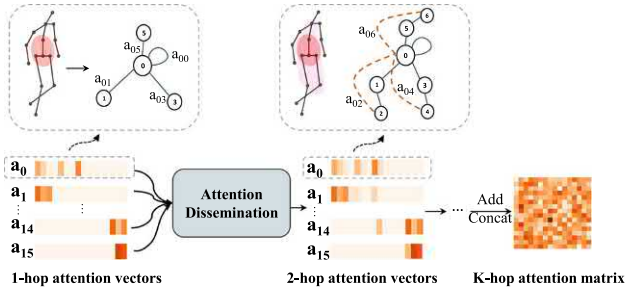

Fig. 7 The details of the attention dissemination

Although there are some multi-hop designs (Quan & Hamza, 2021; Zhao et al., 2022; Zou et al., 2020) for 3D human pose estimation, they are specifically designed for aggregating semantic features in GCNs, which can not reflect the correlations between node pairs at arbitrary hops. To address this issue, we design a novel k-order attention dissemination mechanism to increase the receptive field of the skeleton-based attention for effectively extracting relevant information from neighboring nodes. The details are shown in Fig. 7.

In the dissemination process, we borrow the property of the high-order adjacency matrix in the graph theory: the k-hop connection relationship of each node pair can be illustrated by the k-order adjacency matrix. Improving this principle to adjust the attention matrix, we disseminate the 1-hop skeleton-based attention A_1 to multi-hop attention matrix A_K based on the power operation to obtain the hop-aware attention value, which is shown in Fig. 6. Meanwhile, as the distance between nodes increases, the correlation between nodes gradually decays but brings more noise, which influences the accuracy of predictions. Different from existing methods (Quan & Hamza, 2021; Zhao et al., 2022; Zou et al., 2020) that ignore this key barrier and aggregate long-range semantic information with irrelevant noise, an adaptive attenuation strategy is presented to

overcome this obstacle. Specifically, when disseminating the 1-hop correlation to K-hop, HSGAT separates semantic correlation from distinct hops and assigns learnable attenuation weights $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_K\}$ to attention scores at each hop to further recalibrate the correlation of distant nodes. Such a design facilitates adaptively adjusting the extent of attenuation for different distances, which improves the effectiveness of feature aggregation for multi-hop nodes while reducing undesired noise. Following the above strategy, the K-hop attention matrix can be revised as

$$A_K = \sum_{k=0}^{K} \delta_k A_1^k \tag{5}$$

Compared with 1-hop attention, Fig. 6 shows that the dissemination mechanism generates a denser matrix containing node pairs at various hops to capture enriched semantic features. Then, the updated graph representation H' for the next layer is defined as

$$H' = A_K H \tag{6}$$

After adaptively disseminating nodes correlation from 1-hop to K-hop, attention matrix A_K is capable of expressing node correlation at arbitrary distances. It enlarges the receptive field of the attention layer and allows the network to capture long-range context between nodes while reducing the redundant information from distant nodes.

3.4 Hop-Aware Skeletal Graph Attention Module

Based on the above core components, we build a feature extraction module, which is named hop-aware skeletal graph attention (HSGAT) module, as shown in Fig. 3. Each block is divided into two parts. In the first part, we follow Xiong et al. (2020) and perform a layer normalization layer before calculating multi-hop skeleton-based attention to improve the training stability. After that, the graph representation goes through the multi-hop skeleton-based attention layer to obtain the hop-aware skeleton-based attention matrix with a residual connection. The updated features H_m is formulated as

$$H_m = H_{in} + HSA(LN(H_{in})) \tag{7}$$

where the $LN(\cdot)$ is the layer normalization function and the $HSA(\cdot)$ is the integrated process of aggregating node features using the proposed hop-aware attention dissemination method. After obtaining H_m, we develop a layer normalization layer, a fully connected feed-forward network in a residual connection manner. So far, the output of the first part is $H_m + FFN(LN(H_m))$, where $FFN(\cdot)$ is the feed-forward network.

Although the hop-aware attention aggregates information across long-range connected node pairs, we also hope to fuse local semantic information, which is beneficial for the network to model the human graph. To this end, we employ a residual graph convolution block in the second part to enrich the local feature of human graphs. The overall process of HSGAT is summarized as

$$H_{out} = RGG(H_m + FFN(LN(H_m))) \tag{8}$$

where $RGG(\cdot)$ is the residual graph convolution block as shown in Fig. 3. In this paper, we regard Modulated GCN (Zou & Tang, 2021) as the graph convolution layer and discuss the effect of different GCN layers in Section 5.5.

4 Parallel Hop-Aware Graph Attention Network

To encode the human topology, existing GCNs-based methods typically employ the following architecture designs: (a) single-scale residual architecture (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021) whose receptive field of neurons is fixed in each layer/stage. (b) the multi-scale encoder-decoder (Xu & Takano, 2021) that a complete graph representation is gradually recovered from sub-graph representations. However, these methods typically operate either on single-scale or sequential down-sampled multi-scale graph representations, resulting in the loss of contextual information or spatial details. Motivated by this issue, we build parallel hop-aware graph attention network (PHGANet) upon the HSGAT module in a parallel multi-scale manner. PHGANet maintains the original graph structure and learns enriched hop-aware correlation from the subgraphs at different scales, making the representation spatially precise and semantically enriched. As shown in Fig. 8, the framework contains the joint centrality encoding and parallel multi-scale branches. Next, we give the detail of each component of PHGANet.

4.1 Joint centrality encoding

The previous GCNs-based methods (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021) usually exploit the semantic features between joints without differentiating the importance of different nodes. In fact, joint centrality measures the importance of joints in the human topology, which plays a significant role in graph understanding. For example, more joints are affected to move when the position of the neck changes compared to a change in the wrist under some actions, such as "Photo" and "Sitting Down". Therefore, this paper introduces a simple but effective joint centrality encoding (JCE) method to incorporate this valuable signal into PHGANet. Concretely, we indicate the joint centrality by

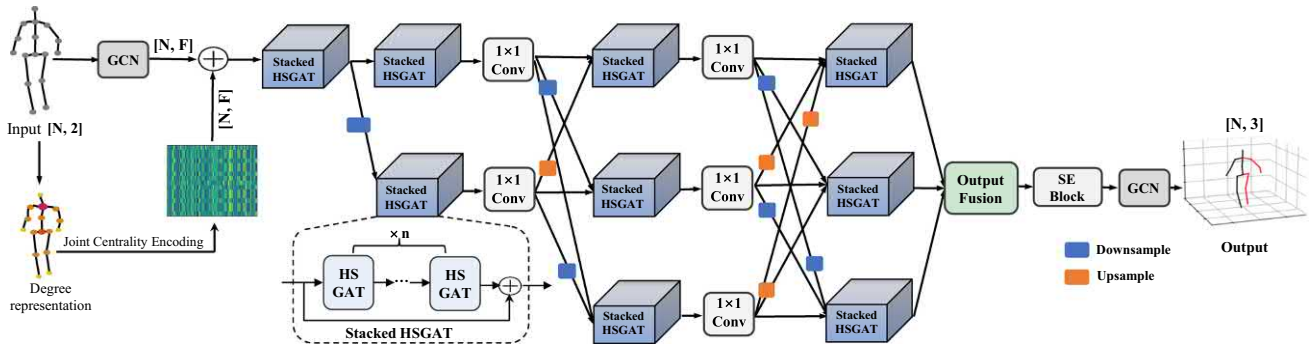

Fig. 8 Illustration of the network architecture of the proposed PHGANet with three parallel branches. Each branch is built upon the stacked HSGAT modules and 1×1 convolution layers. Features across branches are fused and exchanged to obtain semantically strong and spatially precise graph representations for accurate pose predictions

introducing the concept of degree, which can be explained as the number of edges connected to each node in the literature. Since the human body is an undirected graph, the degree of a node equals the number of its 1-hop neighbors. Then, we use an embedding layer to encode the degree and the output of JCE is an $N \times C$ matrix that concatenates N embedded degree vectors together. The embedding layer maps the degree of each node to high dimension vector for better integrating it into input feature while maintaining the original node importance in the graph (i.e., nodes with same degree own same embedding result.). After the joint centrality encoding, we add the encoded joint centrality to input latent features $X \in \mathbb{R}^{N \times F}$ where F is the hidden dimension. After that, the first layer representation $H^{(0)} \in \mathbb{R}^{N \times F}$ is formalized as

$$H^{(0)} = X + f(D_{\mathcal{G}}) \tag{9}$$

where $D_{\mathcal{G}} \in \mathbb{R}^{N \times 1}$ is the degree matrix of the graph, $f(\cdot)$ is the embedding operation achieved by a learnable vector. The joint centrality encoding assists PHGANet in capturing the node importance, making core joints (e.g., neck and pelvis) more influential when updating node features.

4.2 Parallel Multi-Scale Branches

After obtaining features with joint centrality, PHGANet generates output representations through a parallel multi-scale branches architecture. The architecture consists of multiple stacked HSGAT streams in parallel to operate on graphs at different scales. It maintains a complete human topology and receives complementary contextual information from multi-scale skeletal sub-structures to obtain spatial-precise and semantic-enrich graph representations. The essential feature extractor of each stream is Stacked HSGAT, which contains n HSGAT modules with a residual connection. Specifically, the proposed network first extracts coarse context and spatial features from initial latent representation. Then, multi-scale branches are gradually formed and connected in a parallel manner. For the same scale, the extracted features are fed into the stacked HSGAT blocks to explore hop-aware correlation and semantic features of multi-hop node pairs. Next, the network aligns and reduces the channel of each branch through an 1×1 convolution layer, followed by the skeletal sampling for feature fusion. We utilize the graph pooling method proposed in Xu and Takano (2021) to achieve transformation across different scales. Note that nodes at each scale bring various contextual information with a distinct degree of noise since different scales represent the human body from coarse to fine. Therefore, a branch-wise hop value is applied to each attention layer of skeletal structures at different branches, which is defined as

$$k^p = \begin{cases} k - (2^p - 1), & \text{if } (2^p - 1) < k \\ 1, & \text{otherwise} \end{cases} \tag{10}$$

where k^p represent the hop value in pth branch and k is the initial hop value. Combining branch-wise hop values with adaptive attenuation weights, PHGANet enables the attention matrix better capture the semantic correlation between nodes according to subgraph structures.

Several repetitions of feature extraction and exchanging strengthen the original graph features with the help of the subgraph features and vice versa. These features are incorporated into the output fusion layer to combine accurate spatial features and enriched hop-aware semantic correlation. To explore the effect of the features from different scales, we design three output fusion layers as shown in Fig. 9 and discuss the difference among them in Sect. 5.5. In this paper, we concatenate the (upsampled) features from all branches and reduce dimension through convolution layers to obtain the fused graph features. Finally, to enhance important semantic features, the fused features access the Squeeze-and-Excitation block (SE block) (Hu et al., 2018), which

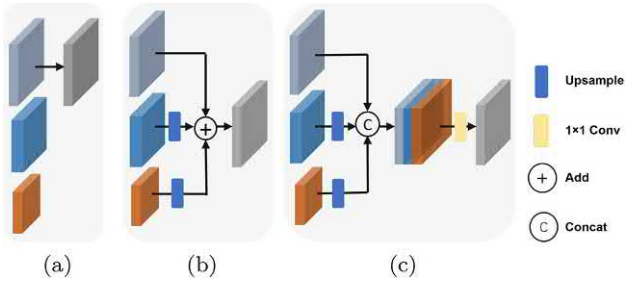

Fig. 9 Illustration of three output fusion layers. **a** Fusion Layer ♯1: only output the original graph topology representation. **b** Fusion Layer ♯2: rescale subgraph representations and add all of them as output. **c** Fusion Layer ♯3: concatenate the (upsampled) representations from all branches and reduce dimension through convolution layers

computes channel-wise weights for all features. At last, the 3D coordinates $\tilde{J}$ are regressed through an output graph convolution layer. Note that any graph convolution methods are easily embedded in the parallel multi-scale framework, which are verified in Sect. 5.5. In conclusion, the parallel framework is capable of generating a spatially-precise output by maintaining a complete skeleton structure while learning enriched semantic correlation from multi-scale subgraphs.

4.3 Loss Function

To obtain the most accurate 3D human pose, we leverage two loss functions to minimize the difference between the predicted and the corresponding ground truth joints. We predict 3D human poses from a set of 2D keypoints and design two loss functions: joints variance loss $\mathcal{L}_j$ and kinematic chain loss $\mathcal{L}_k$. These two losses are combined into a total loss $\mathcal{L}$ by a ratio λ, which is expressed as

$$\mathcal{L} = \lambda \times \mathcal{L}_j + (1 - \lambda) \times \mathcal{L}_k \tag{11}$$

Joints Variance Loss Different from previous methods (Liu et al., 2020a; Zhao et al., 2019), we employ Mean Absolute Errors (MAE) to compute the error between the predicted and the corresponding ground truth joint coordinates, which is defined as

$$\mathcal{L}_j = \frac{1}{M} \sum_{i=1}^{M} \left\| \tilde{J}_i - J_i \right\| \tag{12}$$

where M is the number of samples, $\tilde{J}_i \in \mathbb{R}^{N\times 3}$ is the predicted 3D coordinates consisting of N nodes, and $J_i \in \mathbb{R}^{N\times 3}$ is the corresponding ground truth in the dataset. Compared with the commonly used mean squared errors (MSE), MAE is more robust and stable to handle extreme poses.

Kinematic Chain Loss To ensure the plausibility of the predicted poses, we optimize PHGANet by formulating a kinematic chain loss, which measures the error of properties in human poses such as kinematic chains, symmetry, and bone angles. Inspired by Wandt et al. (2018), we introduce kinematic chain space (KCS) and develop it as a loss function. KCS represents the correlation between bone angles and bone lengths, which can be calculated from the human bone matrix. Specifically, given two nodes $\boldsymbol{j}_r$ and $\boldsymbol{j}_e$, a bone $\boldsymbol{b}_i$ between them can be formulated as

$$\boldsymbol{b}_i = \boldsymbol{j}_r - \boldsymbol{j}_e = J \times \boldsymbol{c}_{re} \tag{13}$$

where J is the 3D pose and $\boldsymbol{c}_{re}$ is the node relation vector with 1 at position r and -1 at position e, i.e,

$$\boldsymbol{c}_{re} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T \tag{14}$$

The bone matrix $B \in \mathbb{R}^{3\times b}$ can be defined as

$$B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \ldots, \boldsymbol{b}_b) \tag{15}$$

where b is the number of bones. After obtaining bone matrix, we multiply B with its transpose to calculate KCS matrix

$$\Psi = B^T B = \begin{pmatrix} l_1^2 & \cdot & \cdot & \cdot \\ \cdot & l_2^2 & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & l_b^2 \end{pmatrix} \tag{16}$$

Due to the inner product, the diagonal elements and other entries of the KCS matrix Ψ represent the squared bone lengths and angle relation between every two bones, respectively. Then, we formulate the kinematic chain loss as

$$\mathcal{L}_k = \frac{1}{M} \sum_{i=1}^{M} \left\| \tilde{\Psi}_i - \Psi_i \right\| \tag{17}$$

Accordingly, the kinematic chain loss effectively measures the lengths and angles of the predicted human structure, which helps produce robust results in various poses.

5 Results and Analysis

5.1 Implementation Details

Datasets We present extensive experimental evaluations of PHGANet on the Human3.6M (Ionescu et al., 2014) and MPI-INF-3DHP (Mehta et al., 2017a) datasets. Human3.6M is the largest dataset in 3D human pose estimation, including 3.6 million images taken by four cameras from 4 different views in an indoor environment. Specifically, it contains 11 professional actors (6 male, 5 female) who perform 15 daily activities such as eating, walking, and sitting. Following the

Table 1 Quantitative comparison on the Human3.6M dataset under Protocol #1

Method	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Lee et al. (2018)†	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Hossain and Little (2018)†	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2	59.9	65.6	55.8	50.4	52.3	43.5	45.1	51.9
Cai et al. (2019)†	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Pavllo et al. (2019)†	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Liu et al. (2020c)†	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.3	45.1
Wang et al. (2020)†	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
Hu et al. (2021)†	38.0	43.3	39.1	39.4	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Yang et al. (2018)	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	**43.6**	60.1	47.7	58.6
Fang et al. (2018)	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Zhao et al. (2019)	48.2	60.8	51.8	64.0	64.6	**53.6**	51.1	67.4	88.7	**57.7**	73.2	65.6	48.9	64.8	51.9	60.8
Sharma et al. (2019)	48.6	54.5	54.2	55.7	62.2	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
Pavllo et al. (2019) (single frame)	47.1	50.6	49.0	51.8	53.6	61.4	49.4	47.4	59.3	67.4	52.4	49.5	55.3	39.5	42.7	51.8
Ci et al. (2019)	46.8	52.3	**44.7**	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
Cai et al. (2019) (single frame)	46.5	**48.8**	47.6	50.9	52.9	61.3	48.3	45.8	59.2	64.4	51.2	48.4	53.5	39.2	41.2	50.6
Quan and Hamza (2021)	47.0	53.7	50.9	52.4	57.8	71.3	50.2	49.1	63.5	76.3	54.1	51.6	56.5	41.7	45.3	54.8
Zou et al. (2020)	49.0	54.5	52.3	53.6	59.2	71.6	49.6	49.8	66.0	75.5	55.1	53.8	58.5	40.9	45.4	55.6
Liu et al. (2020b)	48.4	53.6	49.6	53.6	57.3	70.6	51.8	50.7	62.8	74.1	54.1	52.6	58.2	41.5	45.0	54.9
Liu et al. (2020a)	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu and Takano (2021)	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Zhao et al. (2022)	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	39.7	43.1	51.8
Zou and Tang (2021)	45.4	49.2	45.7	49.4	**50.4**	58.2	47.9	46.0	57.5	63.0	**49.7**	**46.6**	52.2	38.9	40.8	49.4
Ours	**44.8**	49.1	45.6	**49.3**	51.0	57.5	**46.7**	**45.3**	**55.8**	62.5	51.0	46.7	52.8	**38.1**	**40.6**	**49.1**

CPN (Chen et al., 2018) donates detected 2D keypoints as input. Errors are in millimeters (mm). † means using multi-frames. The best and the second-best scores of single-frame methods are shown in bold and underlined, respectively

Table 2 Comparison of the Human3.6M dataset from ground truth 2D keypoints

Method	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch	Sitting	SittingD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Martinez et al. (2017)	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Zhao et al. (2019)*	37.8	49.4	37.6	40.9	45.1	**41.4**	40.0	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
Wang et al. (2019)	35.6	41.3	39.4	40.0	44.2	51.7	39.8	40.2	50.9	55.4	43.1	42.9	45.1	33.1	37.8	42.0
Ci et al. (2019)*	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	**36.2**	**39.5**	34.4	38.4	38.2	31.3	34.2	36.3
Liu et al. (2020b)*	36.2	40.8	33.9	36.4	38.3	47.3	39.9	34.5	41.3	50.8	38.1	40.1	40.0	30.3	33.0	38.7
Liu et al. (2020a)*	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
Zou and Tang (2021)*	37.3	40.7	33.1	36.8	37.2	44.4	40.1	36.1	41.8	47.5	37.8	**30.0**	39.2	30.2	32.8	38.3
Xu and Takano (2021)*	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
Zhao et al. (2022)*	**32.0**	38.0	30.4	34.4	**34.7**	43.3	**35.2**	31.4	38.0	46.2	34.2	35.7	36.1	**27.4**	30.6	35.2
Ours	32.4	**36.5**	**30.1**	**33.3**	36.3	43.5	36.1	**30.5**	37.5	45.3	**33.8**	35.1	**35.3**	27.5	**30.2**	**34.9**

Errors are in millimeters (mm). * means GCNs-based methods. The best and the second-best results are shown in bold and underlined, respectively

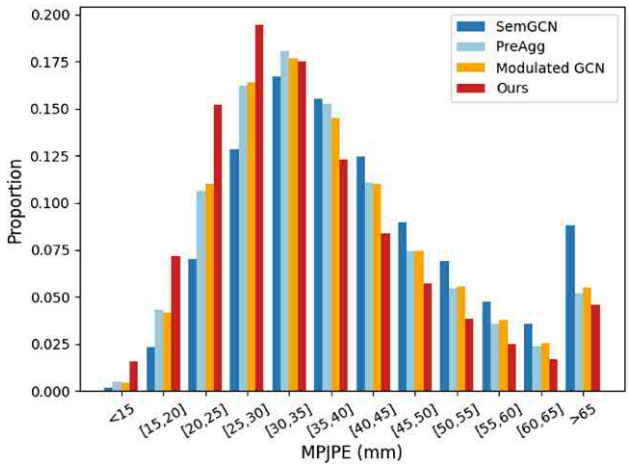

Fig. 10 MPJPE distribution on the testset of Human3.6M

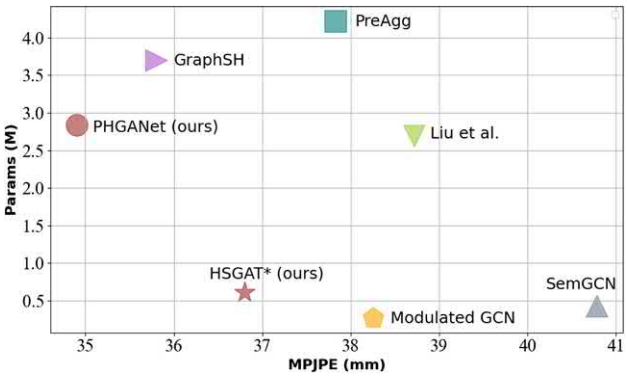

Fig. 11 MPJPE and parameters comparisons with different GCNs-based methods on Human3.6M regarding ground truth 2D keypoints as input

previous works (Xu & Takano, 2021; Zou & Tang, 2021), we use five subjects (S1, S5, S6, S7, S8) for training and two subjects (S9, S11) for evaluation. Additionally, MPI-INF-3DHP (Mehta et al., 2017a) is another 3D pose dataset that presents both indoor and outdoor scenes. We evaluate the generalization of PHGANet on its testset.

Evaluation Protocols Protocol #1 and Protocol #2 are commonly used (Li et al., 2020; Sharma et al., 2019; Wang et al., 2020; Xie et al., 2021; Zhou et al., 2019) as evaluation metrics on Human3.6M. Protocol #1 is Mean Per Joint Position Error (MPJPE), which means the average Euclidean distance between the ground truth and prediction. Protocol #2, Procrustes analysis MPJPE (P-MPJPE), calculates the error after performing rigid transformations (translation, rotation, and scaling). For the MPI-INF-3DHP dataset, we evaluate the proposed method by utilizing Percentage of Correct Keypoints (PCK) with a threshold of 150 mm and the Area Under Curve (AUC) for a range of PCK thresholds. Contrary to MPJPE and P-MPJPE, a higher PCK and AUC means better performance.

Training Setting We implement PHGANet with Pytorch and MindSpore and all experiments are conducted on a single Nvidia RTX 3090 GPU. The repeat time n in the stacked HSGAT block in Fig. 8 is set to 4 and the number of branches is set to 3. For HSGAT, we set 3 as the initial hop number k and use 128 as the hidden dimension. We choose Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.0001 and decay of 0.96 per 20,000 iterations. The batch size is set to 256. Moreover, to prevent overfitting, we apply Dropout (Srivastava et al., 2014) with the dropout rate of 0.25 to all graph convolution layers and 0.1 to graph attention layers. For training, we empirically set the loss ratio $\lambda = 0.9$.

5.2 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of the proposed PHGANet, we conduct comparison experiments on the Human3.6M and the MPI-INF-3D datasets. Then, we further elaborate on the improvement of complex actions made by the proposed model in detail.

Results on Human3.6M We evaluate PHGANet on the Human3.6M dataset and compare our result with other state-of-art methods.

Following previous works (Cai et al., 2019; Xu & Takano, 2021; Zhao et al., 2019; Zou & Tang, 2021), we regard 2D keypoints detected by the cascaded pyramid network (CPN) (Lin et al., 2017) as input. The results are reported in Table 1. Compared with the single-frame methods, PHGANet outperforms others and achieves state-of-the-art performance. Although Zou et al. (2020), Quan and Hamza (2021), and Zhao et al. (2022) also consider long-range nodes to make an improvement, they ignore uncertain noise from distant nodes, which leads to performance degradation. The proposed PHGANet addresses this issue by introducing learnable attenuation coefficients for hop-aware attention at various distances. Therefore, compared with the above multi-hop methods, PHGANet reduces the errors by 6.5, 5.7, and 2.7 mm (relative 11.7, 10.4, and 5.2%) respectively. Moreover, although some approaches (Cai et al., 2019; Hossain & Little, 2018; Hu et al., 2021; Lee et al., 2018; Liu et al., 2020c; Pavllo et al., 2019; Wang et al., 2020) extend multiple video frames as additional input († in Table 1), PHGANet is still comparable, especially in some complex action (e.g. Sitting).

We further compare our method with existing GCNs-based methods using 2D ground truth keypoints as input to eliminate the offsets from the 2D pose detector. As shown in Table 2, PHGANet surpasses all GCNs-based methods, which indicates its superiority. Meanwhile, Fig. 10 provides the error distributions comparison of test sets (S9, S11)

Table 3 Quantitative comparison of the MPI-INF-3DHP dataset

Method	Training data	GS	no GS	Outdoor	All (PCK)	All (AUC)
Zhou et al. (2017)	H36M+MPII	71.1	64.7	72.7	69.2	32.5
Luo et al. (2018)	H36M	71.3	59.4	65.7	65.6	33.2
Yang et al. (2018)	H36M+MPII	–	–	–	69.0	32.0
Zhou et al. (2019)	H36M+MPII	75.6	71.3	80.3	75.3	38.0
Pavllo et al. (2019)	H36M	76.5	63.1	77.5	71.9	35.3
Ci et al. (2019)	H36M	74.8	70.8	77.3	74.0	36.7
Liu et al. (2020b)	H36M	79.0	79.3	79.8	79.3	45.9
Liu et al. (2020a)	H36M	77.6	80.5	80.1	79.3	47.6
Xu and Takano (2021)	H36M	81.5	81.7	75.2	80.1	45.8
Zou and Tang (2021)	H36M	86.4	86.0	**85.7**	86.1	53.7
Zhao et al. (2022)	H36M	80.1	77.9	74.1	79.0	43.8
Ours	H36M	**88.7**	**88.5**	81.49	**86.9**	**55.0**

The best result is shown in bold

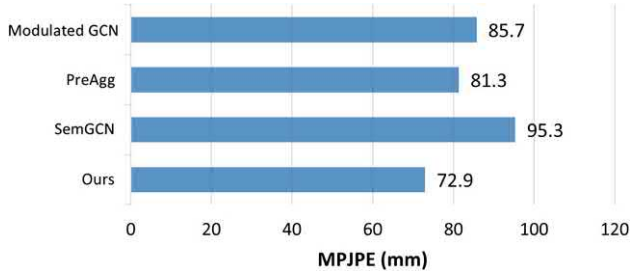

Fig. 12 Mean-Error comparison of the 5% hardest poses

among PHGANet and three previous methods (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021). There are more poses with minor prediction errors and fewer poses with high errors in the results from the proposed method. Specifically, the proportion of cases with errors below 30 mm is consistently higher, and the number of cases with errors above 35 mm is consistently fewer with our solution than with other methods. Moreover, the proportion of cases with errors below 15 mm in PHGANet is 15.9%, which is more than three times than that of the second one (4.9%) (Liu et al., 2020a). Meanwhile, we improve the performance in the poses with high errors from 5.5% (Zou & Tang, 2021) to 4.6% (relative 16.3%). The result further illustrates that PHGANet is stable for predicting common actions while better estimating complex poses than existing methods.

Results on MPI-INF-3D Dataset To verify the generalization of PHGANet, we evaluate the model on the testset of the MPI-INF-3D dataset following Zou and Tang (2021). In particular, we train the proposed method only on the Human3.6M dataset. The results in Table 3 indicate that PHGANet has a better generalization ability than other methods.

Parameters and Performance In this paper, we propose an HSGAT module and construct a parallel multi-scale graph attention network (PHGANet) based on HSGAT. To evaluate the effectiveness of HSGAT and PHGANet and compare parameters among existing graph-based methods, we conducted experiments on Human3.6M with ground truth 2D keypoints as input. Note that the HSGAT network is constructed under the same architecture as other GCNs-based methods (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021) and we denote it as HSGAT*. The results are shown in Fig. 11. Under the same residual architecture, HSGAT* uses 14% parameters of the SOTA method (Liu et al., 2020a) and reduces the errors from 37.8 to 36.8 mm. In addition, compared with the multi-scale method (Xu & Takano, 2021), PHGANet adopts a parallel framework to maintain spatial details and extract multi-scale semantic features at the same time. The comparison results in Fig. 11 indicate that the parameters of PHGANet is fewer than Xu and Takano (2021) (23% decrease) and MPJPE is reduced from 35.8 to 34.9 mm. In a nutshell, HSGAT is a robust graph attention module that can be embedded into various architectures and performs better than other GCNs-based methods. At the same time, both HSGAT and PHGANet achieve smaller prediction errors with fewer parameters.

Improvement on Complex Actions As discussed early, there are many complex poses (e.g. depth ambiguity and joints overlap) leading to model degradation. The proposed PHGANet employs several strategies to improve performance under hard conditions. In Fig. 12, we report the average errors of top 5%*het* hardest cases from four methods to validate the effectiveness of PHGANet. The mean errors of the proposed method is 72.9 mm, which is 22.4, 8.4, and 12.8 mm smaller than other GCNs-based methods (Liu et al., 2020a; Zhao et al., 2019; Zou & Tang, 2021), respectively. Meanwhile, we visualize some predictions of complex poses from the above methods to further evaluate the robustness of the proposed model as shown in Fig. 13. The visualization compares four different methods using some realistic but challenging instances with ambiguous depth and obscured 2D coordi-

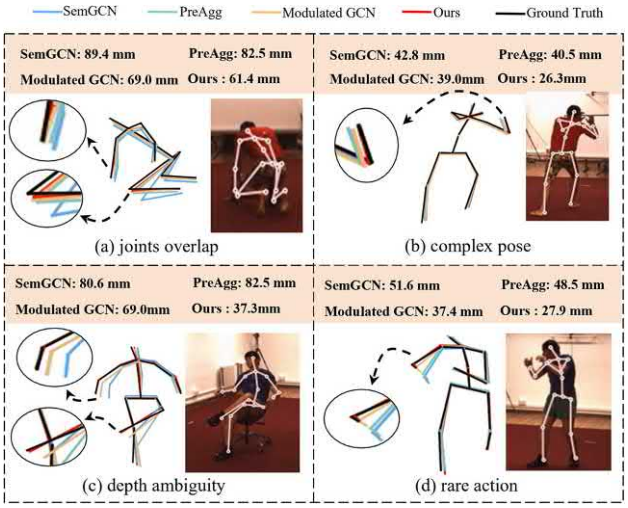

Fig. 13 Visualization comparison on hard poses

nates. We observe that PHGANet acquires the minimum MPJPE error among these methods, demonstrating the superior of the proposed method. Meanwhile, ground truth (black line) and the proposed method (red line) diverge less than others in the details of confused joints, which further illustrates that PHGANet is more effective in handling complex actions.

5.3 Qualitative Results

To qualitatively evaluate the proposed PHGANet, we show a visual comparison of the Human3.6M dataset for more cases in Fig. 14. The results show that PHGANet delivers reasonable 3D predictions for a variety of input situations. For instance, under simple actions (the first and second row in Fig. 14), the proposed method is almost identical to ground truth and the other methods deviate from it, which proves the stability of PHGANet. Meanwhile, benefiting from multi-scale semantic features and the structural information of the original graph, PHGANet is still robust when handling rare actions. For instance, we achieve accurate estimations for the head entanglement (the first case of the third row in Fig. 14) and the elbow node overlap (the second case of the third row in Fig. 14).

Table 4 Quantitative comparison on various hand pose estimation datasets

Method	ObMan	FHAD	GHD
Linear (Martinez et al., 2017)	23.64	26.15	39.25
Graph U-Net (Doosti et al., 2020)	7.63	13.82	8.45
GraFormer (Zhao et al., 2022)	3.29	11.68	4.25
PHGANet (ours)	3.06	11.32	4.34

5.4 Generalization

To verify the generalization capability of the proposed method, we conduct experiments on hand pose estimation task. Following GraFormer (Zhao et al., 2022), we use the ObMan (Hasson et al., 2019), FHAD (Garcia-Hernando et al., 2018), and GHD (Mueller et al., 2018) datasets under Protocol #1 (MPJPE) and compare the results with Linear model (Martinez et al., 2017), Graph U-Net (Doosti et al., 2020), and GraFormer (Zhao et al., 2022). Note that all methods estimate 3D poses by taking 2D ground truth keypoints as inputs.

As shown in Table 4, the proposed method achieves the best performance compared to other methods on the ObMan (Hasson et al., 2019) and FHAD (Garcia-Hernando et al., 2018) datasets and obtains comparable results that is slightly lower than GraFormer (Zhao et al., 2022) on GHD (Mueller et al., 2018), which verifies the generalization capability of the proposed method on hand pose estimation. The qualitative results in Fig. 15 also show that the proposed method is able to obtain desired performance on hand pose estimation.

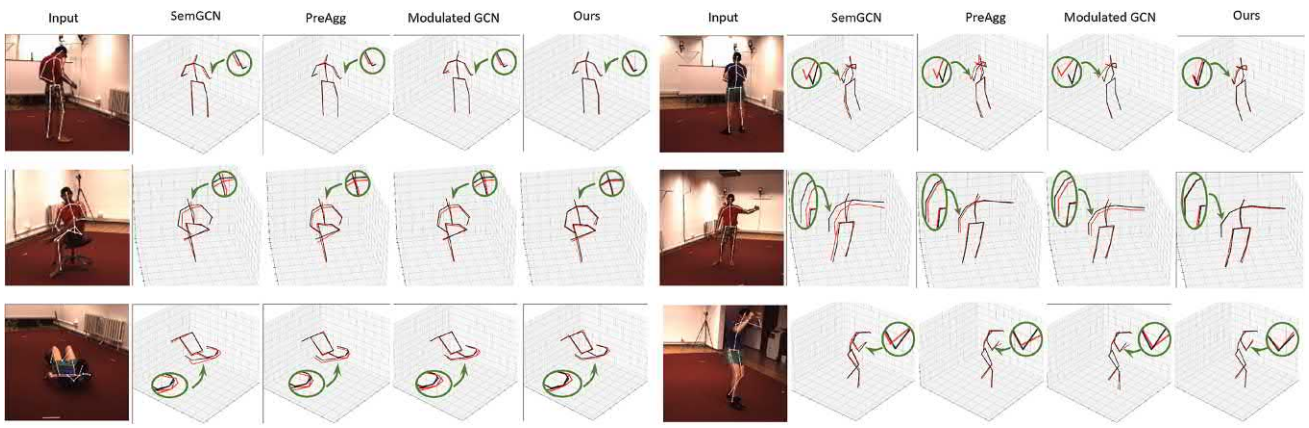

Fig. 14 Qualitative comparison on the Human3.6M dataset. The ground truth are in black and the predictions are in red

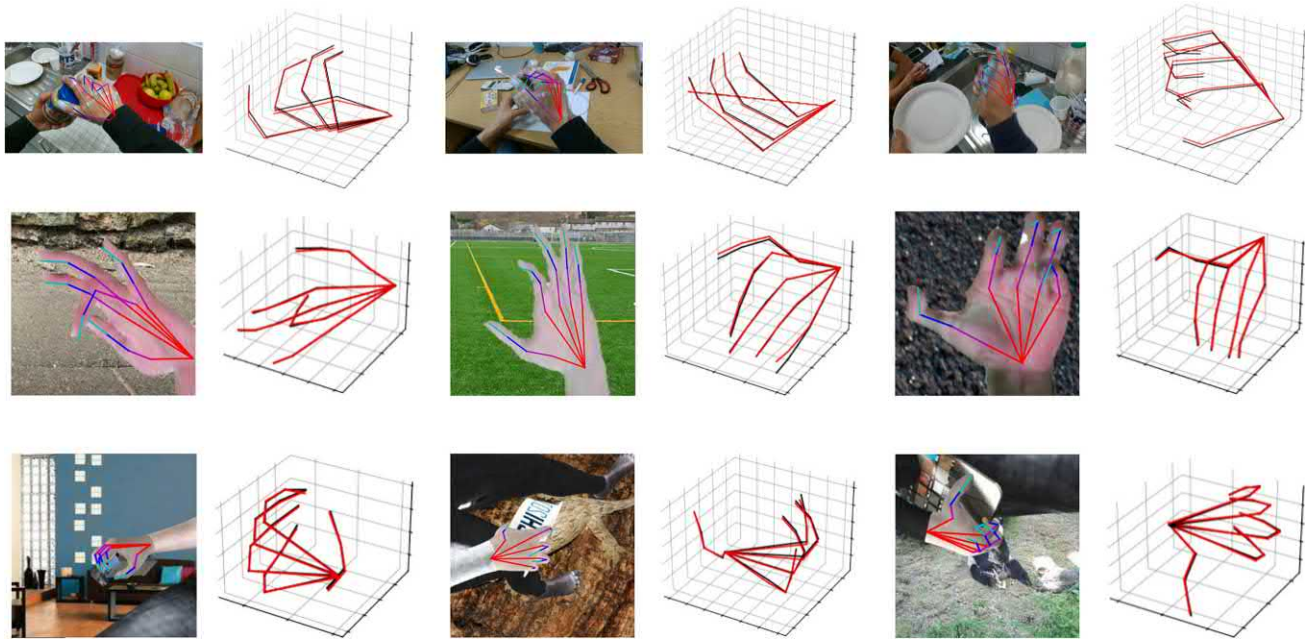

Fig. 15 Qualitative results on the FHAD (Garcia-Hernando et al., 2018) (top row), GHD (Mueller et al., 2018) (middle row), and ObMan (Hasson et al., 2019) (bottom row) datasets. The ground truth are in black and the predictions are in red

Table 5 The effectiveness of HSGAT on the Human3.6M dataset

Method		Params	MPJPE (mm)
SemGCN		0.27 M	43.8
	+ HSGAT	0.60 M	**38.6**
Modulated GCN		0.29 M	38.3
	+ HSGAT	0.61 M	**36.8**

Two different graph convolution layers are used. The best result is shown in bold

Table 6 Comparison of three different frameworks, including sequential residual blocks (SeqRes), graph stacked hourglass (GraphSH), and the proposed parallel multi-scale framework

Method	Frameworks	Params	MPJPE (mm)
SemGCN	SeqRes	0.27 M	52.5
	GraphSH	0.44 M	39.2
	Parallel multi-scale	0.79 M	**38.7**
Modulated GCN	SeqRes	0.29 M	38.3
	GraphSH	–	–
	Parallel multi-scale	2.94 M	**36.4**

Two different GCNs are used. The best result is shown in bold

5.5 Ablation Study

In this section, a series of ablation studies are provided to understand better how each component affects the performance. Note that all experiments are conducted on the Human3.6M dataset with 2D ground truth keypoints as input. In addition, we choose Modulated GCN (Zou & Tang, 2021) as the graph convolution layer in PHGANet.

Hop-Aware Skeletal Graph Attention HSGAT is the core component of PHGANet, which is easy to plug into existing GCNs-based methods. To verify its effectiveness for different GCNs-based methods, we remove the parallel multi-scale branches and adopt the sequential residual blocks proposed in Zhao et al. (2019) as the framework. Specifically, we integrate HSGAT into each layer of SemGCN (Zhao et al., 2019) and Modulated GCN (Zou & Tang, 2021), respectively, and compare errors to the original models. To obtain a fair comparison, we keep the hidden dimension and the number of blocks as the same as the compared methods. As shown in Table 5, HSGAT reduces the errors by 5.2 and 1.5 mm for exisiting methods, respectively. Therefore, HSGAT is a plug-and-play component that is capable of enhancing the model ability to learn accuracy graph representation through capturing hop-aware correlation.

Parallel Multi-Scale Framework To compare the proposed parallel multi-scale framework with both the sequential residual blocks framework (Zhao et al., 2019) and encoder-decoder framework (Xu & Takano, 2021), we remove the HSGAT unit and only apply the parallel multi-scale framework on SemGCN (Zhao et al., 2019) and Modulated GCN (Zou & Tang, 2021). To ensure fair comparisons, we set the hidden dimension to 64 to maintain the setting of GrahpSH (Xu & Takano, 2021) when training parallel multi-scale framework for the SemGCN (Zhao et al., 2019). For Modulated GCN, we maintain the original setting in Zou and Tang

Table 7 Effect of attenuation weights under the fixed or the adaptive strategy

δ	1	0.9	0.8	0.7	0.6	0.5	0.4	Adaptive
MPJPE	38.1	37.3	37.4	36.6	36.4	37.6	39.3	34.9

Table 8 Ablation study of skeletal chain encoding and joint centrality encoding

Method	Params	MPJPE (mm)
PHGANet w/o SCE/JCE	2.90 M	37.8
PHGANet w/o SCE	2.92 M	37.3
PHGANet w/o JCE	2.92 M	36.9
PHGANet	2.94 M	**34.9**

The best result is shown in bold

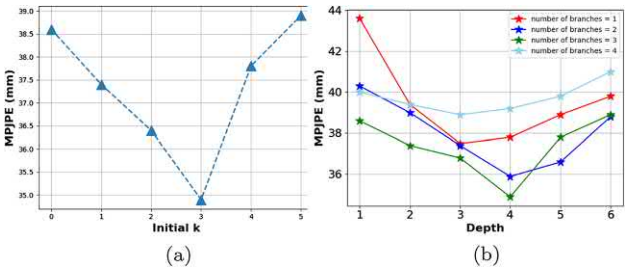

Fig. 16 Analysis of factors in PHGANet. **a** Effect of the initial hop k. **b** Effect of the number of depth and stage

Table 9 Ablation study of different output fusion methods in PHGANet

Method	Params	MPJPE (mm)
Fusion Layer ♯1	2.92 M	38.7
Fusion Layer ♯2	2.92 M	37.1
Fusion Layer ♯3	2.94 M	**34.9**

The best result is shown in bold

(2021). Table 6 shows that the proposed framework improves significantly than other frameworks, indicating that combining complete spatial information and multi-scale semantic features is beneficial for pose predictions. It also proves that the parallel network framework is general and flexible where arbitrary GCNs layers can be embedded.

Attention Attenuation Weights Because the noise from distant nodes may lead to performance degradation, we design an adaptive attenuation strategy to learn unshared weights for different hops. In Table 7, we verify the effectiveness of adaptive attenuation weights and compare it with the fixed coefficient strategy. Obviously, although fixed attenuation weight boosts the model, the value significantly influences the performance. When the weight becomes larger (> 0.6), the model focuses more on long-range nodes with redundant information, making more mistake predictions. Meanwhile, MPJPE increases as the attenuation coefficient decreases (< 0.5) since the high-order information is wasted. In contrast, the adaptive strategy is able to adjust the attenuation parameters adaptively according to different hops. The result shows that it achieves a great improvement on the MPJPE (from 36.4 to 34.9 mm).

Skeletal Chain Encoding and Joint Centrality Encoding This paper proposes two encoding methods to utilize the geometrical property of the graph: SCE and JCE. To verify the effect of the encoding methods, we conduct four different experiment settings: (1) Without neither SCE nor JCE. (2) Without SCE in HSGAT. (3) Without JCE at the beginning of PHGANet. (4) The proposed PHGANet. The results in Table 8 show that encoding skeletal prior boosts the performance of PHGANet. On the other hand, joint centrality is indispensable to yielding an improvement. Meanwhile, both SCE and JCE have parameters of 0.02 M, which provide benefits at extremely low computation.

Initial Hop Number k Figure 16a reports the effect of initial hop k. Note that in multi-scale branches, k is set to the longest path distance when k is larger than the maximum distance between node pairs. Obviously, the MPJPE significantly decreases when the network considers multi-hop neighbors ($k > 1$). The minimum error is generated at $k = 3$, where a better balance between long-range semantic information and noise can be achieved. However, the performance degradation occurs with the increase of k, especially when $k > 5$. The reason we analyze is that the noise is much more than useful information when neighbors are far away.

Depth and Number of Branches Here we conduct various experiments on depth (n in Fig. 8) and the number of branches. The depth is set from 1 to 6 and the number of branches is set from 1 to 4. Figure 16b shows that MPJPE is the smallest when the depth is 4 and the number of branches is 3. Considering the human graph is relatively simple and uniform compared with large-scale graph data such as society or medicine, the deeper network falls into the over-smoothing problem. Meanwhile, when the scale of the human subgraph is too small (e.g., containing only two nodes), the semantic information of individual joints is confusing, which has a negative impact on the final result.

Different Fusion Methods As shown in Fig. 8, multi-scale features are fed into the output fusion block. Inspired by HRNet (Wang et al., 2021b), we design three kinds of fusion layers to aggregate features from all branches as shown in Fig. 9. We investigate how to combine these features to improve performance. Table 9 shows that concatenating features from all branches (Fusion Layer ♯3) performs best. This is because each branch donates the context of different human body parts. For example, the subgraph containing 4 nodes represents the most important torso in human topology. Concatenating these features effectively integrates informa-

tion under different human structures, which enhances the ability of the network.

6 Conclusion

In this paper, we propose the parallel hop-aware graph attention network (PHGANet) for 3D human pose estimation, which learns enriched hop-aware correlation of the skeleton joints while maintaining the spatially-precise representations of the human graph. The proposed hop-aware skeletal graph attention (HSGAT) module obtains the skeleton-based 1-hop attention and disseminates it to arbitrary hops with an adaptive attenuation strategy. It is capable of capturing semantic correlation from multi-hop joints while alleviating undesired noise from them. Upon HSGAT, we build PHGANet with a parallel multi-branch of stacked HSGAT modules, which maintains the original graph structure while learning enriched hop-aware correlation from skeleton parts at different scales for more robust estimations under complex activities. Experimental results on the Human3.6M and MPI-INF-3DHP datasets show that the proposed method achieves state-of-the-art performance. In the future, we expect to make further improvements by exploring temporal information from multi-frames.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Nos. 62272134, 62236003 and 62072141), in part by the Taishan Scholars Program of Shandong Province (No. tsqn201812106), in part by the Shenzhen Colleges and Universities Stable Support Program (No. GXWD20220817144428005), in part by the National Key R&D Program of China (No. 2021ZD0110901) and in part by CAAI-Huawei MindSpore Open Fund.

Data Availability All datasets generated or analysed during the current study are included in the published articles (Ionescu et al. 2014; Mehta et al. 2017a). These datasets can be derived from the following public domain resources: https://github.com/jfzhang95/PoseAug/blob/main/DATASETS.md.

References

Abu-El-Haija, S., Perozzi, B., Kapoor, A., Alipourfard, N., Lerman, K., Harutyunyan, H., Steeg, G. V., & Galstyan, A. (2019). Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *ICML*.

Agarwal, A., & Triggs, B. (2006). Recovering 3D human pose from monocular images. *TPAMI, 28*(1), 44–58.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P. V., Romero, J., & Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*.

Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T., Yuan, J., & Magnenat-Thalmann, N. (2019). Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*.

Chen, C., & Ramanan, D. (2017). 3D human pose estimation = 2D pose estimation + matching. In *CVPR*.

Chen, C., Tyagi, A., Agrawal, A., Drover, D., MV, R., Stojanov, S., & Rehg, J. M. (2019a). Unsupervised 3d pose estimation with geometric self-supervision. In *CVPR*.

Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., & Luo, J. (2021). Anatomy-aware 3D human pose estimation in videos. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(1), 198–209.

Chen, Y., Huang, S., Yuan, T., Zhu, Y., Qi, S., & Zhu, S. (2019b) Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *ICCV*.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018) Cascaded pyramid network for multi-person pose estimation. In *CVPR*.

Chen, Z., Huang, Y., Yu, H., Xue, B., Han, K., Guo, Y., & Wang, L. (2020). Towards part-aware monocular 3D human pose estimation: An architecture search approach. In *ECCV*.

Ci, H., Wang, C., Ma, X., & Wang, Y. (2019). Optimizing network structure for 3D human pose estimation. In *ICCV*.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*.

Doosti, B., Naha, S., Mirbagheri, M., & Crandall, D. J. (2020) Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*.

Fang, H., Xu, Y., Wang, W., Liu, X., & Zhu, S. (2018). Learning pose grammar to encode human body configuration for 3D pose estimation. In *AAAI*.

Fang, Q., Shuai, Q., Dong, J., Bao, H., & Zhou, X. (2021). Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*.

Garcia-Hernando, G., Yuan, S., Baek, S., & Kim, T. (2018). First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*.

Hamilton, W. L, Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *NIPS*.

Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M. J., Laptev, I., & Schmid, C. (2019). Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Henaff, M., Bruna, J., & LeCun, Y. (2015). Deep convolutional networks on graph-structured data. arXiv:1506.05163

Hossain, M. R. I., & Little, J. J. (2018). Exploiting temporal information for 3D human pose estimation. In *ECCV*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *CVPR*.

Hu, W., Zhang, C., Zhan, F., Zhang, L., & Wong, T. (2021). Conditional directed graph convolution for 3d human pose estimation. In *ACM MM*.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI, 36*(7), 1325–1339.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *ICLR*.

Lee, K., Lee, I., & Lee, S. (2018). Propagating LSTM: 3D pose estimation based on joint interdependency. In *ECCV*.

Li, G., Müller, M., Thabet, A. K., & Ghanem, B. (2019). Deepgcns: Can GCNs go as deep as CNNs? In *ICCV*.

Li, H., Shi, B., Dai, W., Chen, Y., Wang, B., Sun, Y., Guo, M., Li, C., Zou, J., & Xiong, H. (2021). Hierarchical graph networks for 3D human pose estimation. In *BMVC*.

Li, S., & Chan, A. B. (2014). 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*.
Li, S., Zhang, W., Chan, A. B. (2017). Maximum-margin structured learning with deep networks for 3D human pose estimation. *IJCV*.
Li, S., Ke, L., Pratama, K., Tai, Y., Tang, C., & Cheng, K. (2020). Cascaded deep monocular 3D human pose estimation with evolutionary training data. In *CVPR*.
Lin, T., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. In *CVPR*.
Liu, K., Ding, R., Zou, Z., Wang, L., & Tang, W. (2020a). A comprehensive study of weight sharing in graph networks for 3D human pose estimation. In *ECCV*.
Liu, K., Zou, Z., & Tang, W. (2020b). Learning global pose features in graph convolutional networks for 3D human pose estimation. In *ACCV*.
Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition, 68*, 346–362.
Liu, M., & Yuan, J. (2018). Recognizing human actions as the evolution of pose estimation maps. In *CVPR*.
Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S. S., & Asari, V. K. (2020c) Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *CVPR*.
Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S. S., & Asari, V. K. (2021) Enhanced 3D human pose estimation from videos by using attention-based neural network with dilated convolutions. *IJCV*.
Luo, C., Chu, X., & Yuille, A. L. (2018). Orinet: A fully convolutional network for 3D human pose estimation. In *BMVC*.
Luvizon, D. C., Picard, D., & Tabia, H. (2022). Consensus-based optimization for 3D human pose estimation in camera coordinates. *IJCV*.
Martinez, J., Hossain, R., Romero, J., & Little, J. J. (2017). A simple yet effective baseline for 3D human pose estimation. In *ICCV*.
Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017a). Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*.
Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H., Xu, W., Casas, D., & Theobalt, C. (2017b). Vnect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics, 36*(4), 44:1–44:14.
Moon, G., & Lee, K. M. (2020). I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*.
Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., & Theobalt, C. (2018). Ganerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*.
Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *ECCV*.
Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*.
Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*.
Pustejovsky, J., & Krishnaswamy, N. (2021). Embodied human computer interaction. *Künstliche Intell, 35*(3), 307–327.
Quan, J., & Hamza, A. B. (2021). Higher-order implicit fairing networks for 3D human pose estimation. In *BMVC*.
Sharma, S., Varigonda, P. T., Bindal, P., Sharma, A., & Jain, A. (2019). Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*.
Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.
Sun, X., Shang, J., Liang, S., & Wei, Y. (2017). Compositional human pose regression. In *ICCV*.
Takano, W., & Nakamura, Y. (2015). Action database for categorizing and inferring human poses from video sequences. *Robotics and Autonomous Systems, 70*, 116–125.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *NIPS*.
Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *ICLR*.
Wandt B, Ackermann, H., & Rosenhahn, B. (2018). A kinematic chain space for monocular motion capture. In *ECCV*.
Wang, G., Ying, R., Huang, J., & Leskovec, J. (2021a). Multi-hop attention graph neural networks. In *IJCAI*.
Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., & Xiao, B. (2021). Deep high-resolution representation learning for visual recognition. *TPAMI, 43*(10), 3349–3364.
Wang, J., Yan, S., Xiong, Y., & Lin, D. (2020). Motion guided 3D pose estimation from videos. In *ECCV*.
Wang, L., Chen, Y., Guo, Z., Qian, K., Lin, M., Li, H., & Ren, J. S. J. (2019). Generalizing monocular 3D human pose estimation in the wild. In *ICCV*.
Xie, K., Wang, T., Iqbal, U., Guo, Y., Fidler, S., & Shkurti, F. (2021). Physics-based human motion estimation and synthesis from videos. In *ICCV*.
Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. (2020). On layer normalization in the transformer architecture. In *ICML*.
Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., & Jegelka, S. (2018). Representation learning on graphs with jumping knowledge networks. In *ICML*.
Xu, T., & Takano, W. (2021). Graph stacked hourglass networks for 3D human pose estimation. In *CVPR*.
Yang, W., Ouyang, W., Wang, X., Ren, J. S. J., Li, H., & Wang, X. (2018). 3D human pose estimation in the wild by adversarial learning. In *CVPR*.
Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. N. (2019). Semantic graph convolutional networks for 3D human pose regression. In *CVPR*.
Zhao, W., Tian, Y., Ye, Q., Jiao, J., & Wang, W. (2022). Graformer: Graph convolution transformer for 3D pose estimation
Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., & Ding, Z. (2021). 3D human pose estimation with spatial and temporal transformers. In *ICCV*.
Zhou, K., Han, X., Jiang, N., Jia, K., & Lu, J. (2019). Hemlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In *ICCV*.
Zhou, X., Huang, Q., Sun, X., Xue, X., & Wei, Y. (2017). Towards 3D human pose estimation in the wild: A weakly-supervised approach. In *ICCV*.
Zou, Z., & Tang, W. (2021). Modulated graph convolutional network for 3D human pose estimation. In *ICCV*.
Zou, Z., Liu, K., Wang, L., & Tang, W. (2020). High-order graph convolutional networks for 3D human pose estimation. In *BMVC*.