HopFIR: Hop-wise GraphFormer with Intragroup Joint Refinement for 3D Human Pose Estimation

Kai Zhai[1*] Qiang Nie[2*] Bo Ouyang[1†] Xiang Li[3] Shanlin Yang[1]
[1]Hefei University of Technology [2]Youtu Lab, Tencent [3]Tsinghua University

Abstract

2D-to-3D human pose lifting is fundamental for 3D human pose estimation (HPE), for which graph convolutional networks (GCNs) have proven inherently suitable for modeling the human skeletal topology. However, the current GCN-based 3D HPE methods update the node features by aggregating their neighbors' information without considering the interaction of joints in different joint synergies. Although some studies have proposed importing limb information to learn the movement patterns, the latent synergies among joints, such as maintaining balance are seldom investigated. We propose the Hop-wise GraphFormer with Intragroup Joint Refinement (HopFIR) architecture to tackle the 3D HPE problem. HopFIR mainly consists of a novel hop-wise GraphFormer (HGF) module and an intragroup joint refinement (IJR) module. The HGF module groups the joints by k-hop neighbors and applies a hop-wise transformer-like attention mechanism to these groups to discover latent joint synergies. The IJR module leverages the prior limb information for peripheral joint refinement. Extensive experimental results show that HopFIR outperforms the SOTA methods by a large margin, with a mean per-joint position error (MPJPE) on the Human3.6M dataset of 32.67 mm. We also demonstrate that the state-of-the-art GCN-based methods can benefit from the proposed hop-wise attention mechanism with a significant improvement in performance: SemGCN [42] and MGCN [49] are improved by 8.9% and 4.5%, respectively.

1. Introduction

Monocular 3D human pose estimation aims to accurately regress the 3D locations of human joints in the camera coordinate system from a single image. It plays an important role in many applications, such as action recognition and human–computer interaction. Compared with the monocular systems, multi-view capture systems are expensive and inconvenient to set up and operate, which prevents them from being widely used in practice. To tackle the monocular 3D HPE task, some approaches [2, 15, 31, 32, 38] estimates 3D joint coordinates or heat maps directly from an image via a convolutional neural network (CNN) [13, 14]. However, direct regression from the image space suffers from the problem of a large parameter searching space, which always leads to a sub-optimal solution. Recently, Martinez et al. [20] constructed a simple fully connected network using only 2D keypoints as input and achieved promising 3D HPE performance, showing that the 3D human pose can be efficiently and accurately estimated from 2D joint po-

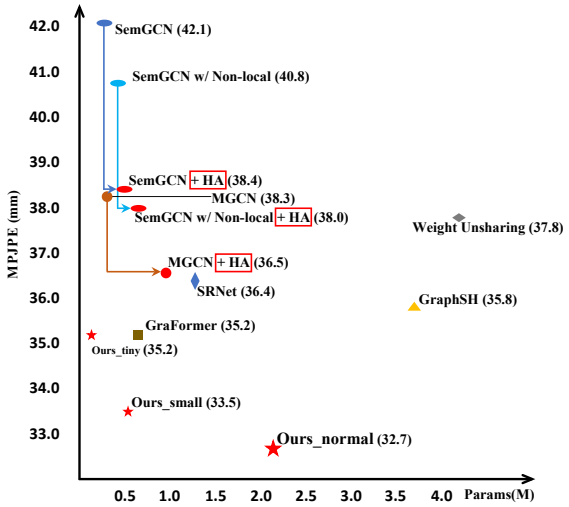

Figure 1. Comparison of performance and model size between the proposed HopFIR and SOTA methods, namely Modulated GCN (MGCN) [49], SemGCN [42], Weight Unsharing [18], SRNet [39], GraphSH [36], and GraFormer [43]. The methods are evaluated on the Human3.6M dataset [10] with ground truth 2D joints as input. The arrow shows the performance improvement obtained by inserting the HA layer into the other networks. Tiny, small, and normal denote the feature dimension of the HopFIR are 32, 64, and 128, respectively.

*Equal Contribution
†Corresponding Author

sitions. Inspired by them and considering the thoroughly investigated 2D HPE, many works decompose the problem into two subtasks, i.e., the 2D HPE and 2D-to-3D pose lifting [18, 40, 49, 36]. 2D-to-3D pose lifting, therefore, has emerged as a fundamental task in this area that our work devotes to.

A key consideration in 2D-to-3D pose lifting is that the human skeleton topology is inherently sparse and graph-structured. Fully connected neural networks are less effective in modeling graph-structured data due to their simple connections among all nodes and the probability of overfitting. To leverage the information of the human skeletal topology, some works [36, 40, 49] have proposed to model the human body with GCNs and have achieved SOTA results. For example, Ci et al. [4] introduced a locally connected network to enhance the representation capability of the GCN, and Liu et al. [18] explored the weight sharing and feature transformation that occurs before or after feature aggregation in the GCN. One limitation of these GCN-based 3D HPE methods, however, is that they update the node features by aggregating their neighbors' information without considering the different contributions of these nodes to different joint synergies.

Instead of considering all the joints of a skeleton as a whole, Xue et al. [37] demonstrated that the human skeleton exhibits obvious part-wise inconsistency in its motion patterns, as also reported in SRNet [39]. However, these works are limited to considering prior structural information of limb groupings and ignore investigating the latent groups underlying joint synergies. For example, the relative positions of the 1-hop neighbors of joint 0 are almost constant in "Discussion" subject, which can be a latent group, as shown in Fig. 2. Moreover, these works consider the joints in a limb as a whole to calculate the relationship with other limbs, which resulted in lower accuracy for peripheral joints, such as wrists and feet.

To address the abovementioned problems in monocular 3D HPE, we propose a novel architecture: the Hop-wise GraphFormer with Intragroup Joint Refinement (HopFIR). The first key component of HopFIR is a novel hop-wise GraphFormer (HGF) module that considers k-hop neighbors. In the HGF module, the information of every hop of every joint is aggregated into the hidden space, such that $N \times k$ groups of features are obtained for a skeleton model with N joints. Meanwhile, a hop-wise transformer-like attention mechanism is designed to extract the correlation among feature groups, which computes similarity by the dot product of the node feature and the group feature. The proposed HGF module enables the network to discover latent joint interactions considering human joint synergy. Because the HGF leverages little prior information about the human body and ignores the interaction among joints in a limb, especially the interactions of peripheral joints associated with a limb, we introduce an intragroup joint refinement (IJR) module to strengthen the intragroup correlation of joints grouped by limb prior information. Specifically, a residual block is built from two HGF modules followed by one IJR module. The proposed HopFIR architecture achieves optimal regression accuracy with a stack of three blocks.

To summarize, our work makes the following contributions:

- To the best of our knowledge, we design the first Hop-wise GraphFormer module to explore potential joint correlations underlying human joint synergy. We also prove that other GCN-based methods can benefit from the proposed HGF module efficiently, as shown in Fig. 1.
- We design an Intragroup Joint Refinement module, which attends to intragroup joints to refine joint features through the associated limb, especially the wrists and feet. The IJR module enables HGF modules to discover the latent synergies among joints.
- We propose the novel Hop-wise GraphFormer with Intragroup Joint Refinement (HopFIR) architecture for 3D HPE, which is built entirely from HGF and IJR modules. Specifically, two HGF modules and one IJR module are coupled into a block.
- Extensive experiments demonstrate the effectiveness and generalizability of the proposed modules and HopFIR architecture by providing new state-of-the-art results on two challenging datasets, i.e., Human3.6M [10] and MPI-INF-3DHP [21].

2. Related Work

3D Human Pose Estimation. Early works [26, 29] use handcrafted features, perspective relationships, and geometric constraints to estimate the 3D human pose. Recent pose estimation approaches can be generally divided into two categories. The first category of networks regresses 3D human joints directly from the image [24, 46]. Pavlakos et al. [24] adopted a CNN to predict the voxel-wise likelihoods for each joint, and Zhou et al. [46] directly embedded a kinematic object model into the networks to learn the general multi-articulate object pose. Approaches in the second category decouple the 3D HPE task into 2D pose estimation from an image and 3D pose estimation from the detected 2D joints (2D-to-3D). For example, Martinez et al. [20] proposed a simple yet effective baseline with fully-connect networks and proved that 3D human poses can be regressed simply and effectively from 2D keypoints. Our paper follows this pipeline and focuses on the 2D-to-3D pose lifting. For promoting 3D human pose regression accuracy, it is crucial to group joints with consideration of their interactions rather than treating all joints of a skeleton as a whole. Xue et al. [37] divided the human skeleton graph into five groups according to the limbs to explore part-wise motion inconsistency, and Zeng et al. [39] split the human joints into

local regions and recombined the global information from the rest of the joints. Our proposed HopFIR differs from these approaches by grouping joints by the k-hop neighbors of each joint and prior limb information, which enables the network to discover latent connections between groups in different human joint synergies.

Graph Convolutional Networks. GCNs [5, 7, 12, 27] generalize the capability of CNNs by performing convolution operations on graph-structured data. GCNs can be divided into two categories: the spectral-based approaches [5] and the spatial-based approaches [12]. Our approach falls into the second category, which applies message-passing operations on the graph nodes and their neighbors.

Due to the graph-structure topology of the human skeleton, many works [4, 18] have introduced GCN to tackle the 3D HPE task. Zhao et al. [42] proposed a SemGCN to learn the semantic relationships between human joints, and Zou et al. [49] proposed a weight modulation and an affinity modulation based on the SemGCN. These methods aggregate the first-order neighborhood messages to update the feature matrix by assigning different weights to different nodes. Some works [47, 48] have extended the first-order neighbors to high-order neighbors in the spatial domain directly. Zeng et al. [40] designed a hierarchical fusion block by dividing the fusion procedure into two stages, where all the high-order neighbors of a node are aggregated into a feature in the first stage and fuse it with the node feature and the first-order neighbor in the second stage. Zhao et al. [43] introduced Chebyshev graph convolution to fuse information among the k-hop neighbors of a joint directly. These works updated features by aggregating each node's own k-hop neighborhood information in a GCN layer. However, HopFIR considers the k-hop groups of all nodes to reconstruct the k-hop feature of a node through the proposed attention mechanism, which can enhance the representation capability of GCNs.

Graph Attention. Graph attention networks [34] is a pioneer work that pay attention to the data in a graph structure by assigning an attention weight to each node. The introduction of the transformer [33] for machine translation tasks has proven the capacity of attention for sequential input. ViT [6] introduces the transformer into computer vision and achieves excellent performance. Inspired by them, some researchers have adopted the transformer for 3D HPE. Zhao et al. [43] applied self-attention to capture global information by calculating the similarity of all nodes. PoseFormer [44] directly applies Transformer Encoder (TE) by viewing each joint and each frame as tokens in the spatial and temporal domains, respectively. MixSTE [41] is based on [44], where each joint feature is represented by the temporal TE to model the joint motion. P-stmo [28] replaces the spatial TE with MLP and applies Stride Transformer Encoder to map N frames to one frame. MHFormer [17] generates multi-hypothesis at different depths of stacked spatial TEs by viewing all frames of each joint feature as a token, then communicates these hypotheses with cross-attention and self-attention. Different from the existing graph attention for 3D HPE, we propose the intergroup multi-head attention mechanism among the k-hop groups of all nodes, which assigns the attention weights by computing the similarity between the node feature and k-hop group feature. Moreover, we introduce the intragroup multi-head self-attention in limb groups to refine the joint features and promote the HGF module to discover the latent synergies among joints.

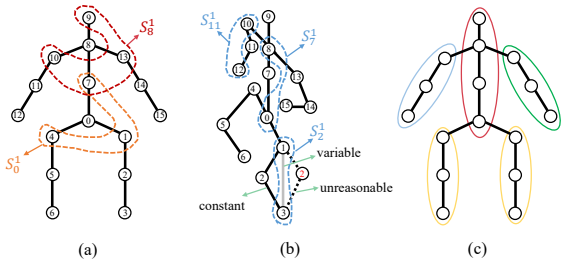

Figure 2. An illustration of the human skeleton graph and the groups in HGF and IJR modules. The k-hop neighbors of a joint are set as a group in HGF modules where (a) indicates the 1-hop groups of different joints and (b) indicates several latent groups because of physical limitations. (c) indicates the joints grouped by prior limb information.

3. The Proposed HopFIR

This paper proposes a novel architecture to regress the 3D human pose from N given 2D keypoints $X \in \mathbb{R}^{N\times 2}$. The proposed framework mainly consists of the HGF and IJR modules. In this section, we first review the vanilla Graph Convolutional Network and Transformer in Sec. 3.1. We then introduce the HGF and IJR modules in Sec. 3.2 and Sec. 3.3, respectively. Finally, we present the network architecture in Sec. 3.4.

3.1. Vanilla GCN and Transformer

GCN. A graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of N nodes and $\mathcal{E}$ is the adjacency matrix representing the edges between the nodes. Given a collection of input features $H^l \in \mathbb{R}^{N\times D}$, a generic GCN layer that aggregates neighborhood information can be formulated as follows:

$$H^{(l+1)} = \sigma(\tilde{A}H^l W) \tag{1}$$

where $W \in \mathbb{R}^{D\times D'}$ is the learnable weight matrix that transforms the feature dimension from D to D', and $\sigma(.)$ is the activation function, such as ReLU [22]. $H^{(l+1)}$ is the updated feature matrix, $\tilde{A} \in \mathbb{R}^{N\times N}$ is the symmetrically normalized affinity matrix [12] with added self-connections, and $A \in \{0,1\}^{N\times N}$ is the adjacency ma-

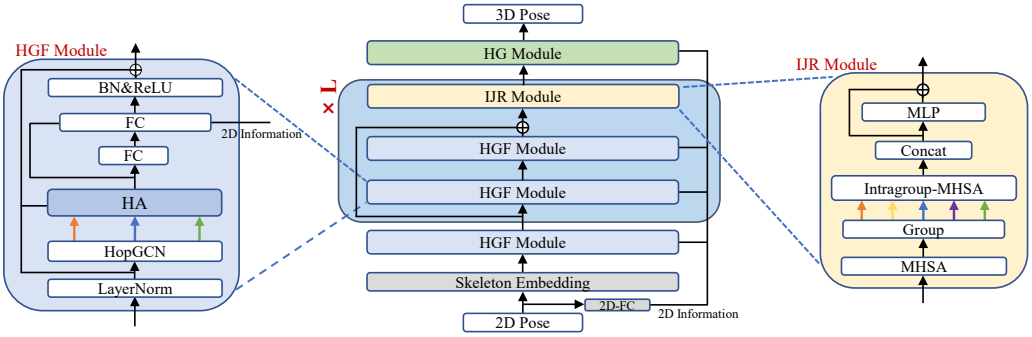

Figure 3. The HopFIR architecture, with details of the HGF and IJR modules. The designed block is a residual block [9] built by two HGF modules followed by one IJR module. The proposed architecture with three blocks achieves optimal performance. Arrows of different colors represent different hops and groups in HGF and IJR, respectively.

trix. The $(i,j)th$ entry $a_{ij} = 1$ representing node j is the neighbor of node i. Otherwise, they are not connected and $a_{ij} = 0$. Therefore, the none-neighbor nodes have a weak influence on each other in the vanilla GCN, which hinders the modeling of underlying joint synergy in 3D HPE.

Transformer. Transformer architecture relies entirely on self-attention to compute representations of its input and output. The self-attention function maps the inputs to the queries Q, keys K, and values V by weight matrices W_Q, W_K, and W_V, respectively, and the matrix of outputs is calculated as:

$$Attention(Q,K,V) = Softmax(QK^T/\sqrt{d})V \quad (2)$$

where d is the feature dimension of Q, and $\frac{1}{\sqrt{d}}$ is a scaling factor to prevent extremely small gradients. The multi-head self-attention (MHSA), which performs self-attention in parallel, projects the queries, keys, and values P times with different linear projections to the respective subspaces, as follows:

$$MultiHead(Q,K,V) = Concat(D_1,...,D_P)W_o \quad (3)$$

where W_O is the projection matrix of outputs, $D_p = Attention(QW_Q^p, KW_K^p, VW_V^p)$, and $p \in [1,..,P]$.

3.2. Hop-wise GraphFormer

Previous GCN studies for 3D HPE aggregate multi-hop neighborhood information [40, 48] or assign an attention weight to each first-order neighbor [42, 49]. To effectively capture the node's neighborhood message and increase the representational capacity of GCNs, we introduce the hop-wise GraphFormer (HGF) module, which treats each hop as a group (Fig.2 a,b) and computes the attention weights for each hop (more intuitive descriptions for k-hop can be found in the supplementary material). By considering the relationship within k hops, we can obtain $N \times k$ groups for a skeleton graph with N joints, which provides enough combinations of joints to discover the latent correlations among joints in different human joint synergy.

We first define the k-hop matrix A^k as

$$a_{ij}^k = \begin{cases} 1, & d(v_i,v_j) = k \\ 0, & otherwise \end{cases} \quad (4)$$

where $d(v_i,v_j)$ denotes the distance of the shortest path between v_i and v_j on the skeleton graph. The k-hop neighborhood information is aggregated with a weighted sum of the target node's k-hop neighbors, named HopGCN:

$$s_i^k = \sum_j a_{ij}^k h_j W^k \quad (5)$$

where $s_i^k \in \mathbb{R}^D$ is a hidden representation of the k-hop neighborhood information. Eq. 5 is similar to Eq. 1 but with the extended definition of A to k hops. The weight matrix W is assigned to the respective hops.

Before aggregating the hidden representation s_i^k to the target node, we propose a transformer-like attention mechanism computing the similarity between the node feature h_i and the k-hop hidden representation, as shown in Eq. 6:

$$z_i^k = \sum_j softmax_j(\frac{h_i {s_j^k}^T}{\sqrt{d}}) s_j^k. \quad (6)$$

Due to the structure of the human skeleton and the joint synergy, the k-hop neighborhood of joint i is related to that of other joints. In Fig. 2, for example, the synergy of hop S_5^1 is related to hop S_6^1 and S_3^1. By packing z_i^k, h_i, and s_i^k into matrices Z^k, H, and S^k, we can calculate Z^k in parallel:

$$Z^k = softmax(\frac{H{S^k}^T}{\sqrt{d}}) S^k \quad (7)$$

Z^k can also be reconstructed using the self-attention mechanism purely on hop features or swap the positions of H and

S^k. A performance comparison of these cases is discussed in the experimental results.

We further reduce the dimension of hidden representation by using a fully connected(FC) layer based on the order of the neighborhood, considering different amounts of information related to the target node, as formulated in Eq. 8.

$$r_i^k = F^k(z_i^k) \tag{8}$$

where r_i^k is the refined representation of the k-hop neighborhood information with respect to node i, and F^k is the mapping function. The refined representation is then concatenated to the updated node feature.

All the refined representations of hop-wise neighborhoods and the target node feature h_i are encoded to a D-dimensional vector:

$$h_i' = F(h_i, r_i^1, r_i^2, r_i^3, ..., r_i^k) \tag{9}$$

where h_i' is the final updated feature of node i in this layer, and F is the aggregation function. Packing together the updated features of all nodes, Eq. 9 can be rewritten as:

$$H' = F(H, R^1, R^2, R^3, ..., R^k). \tag{10}$$

We thereby obtain as the core layer of the HGF module the hop-wise attention (HA) layer, which extracts latent correlations between feature groups and aggregates k-hop neighborhood information. With the increment of k, we get more N groups on which to explore the underlying joint synergies. The value of k can be adjusted based on the tasks and pipelines. According to the experimental results, three hops achieve optimal performance in the HopFIR architecture.

To fuse the current global information and original 2D information, we concatenate all the joint features in a batch and feed them into a FC layer to extract the current global information. We then concatenate the extracted global information, 2D information, and the output of the HA layer on the feature dimension of the joint feature to fuse all the information in another FC layer.

3.3. Intragroup Joint Refinement

The proposed HGF module splits the skeleton graph into different groups based on the k-hop neighborhood of each joint, which attends to the key groups in joint synergies. However, HGF leverages little prior information about the human body and ignores the interaction among joints in a limb, especially the interaction of the peripheral joints associated with limbs, such as the wrists and feet. We introduce an intragroup joint refinement (IJR) module to strengthen the intragroup correlation of joints grouped by limb prior information, as shown in Fig. 2(c). The HGF features of the joints in each limb group are refined in the IJR module by using multi-head self-attention [33], as in Eq. 11.

$$H_g = MHSA(MHSA(H)_g) \tag{11}$$

where H is the feature matrix from the HGF module, $MHSA(H)_g$ is the feature matrix of the group g updated by global multi-head self-attention, and H_g is the final feature matrix of the group g updated by the IJR module. More details of the IJR module are provided in Fig. 3.

3.4. The HopFIR Architecture

The HopFIR architecture consists of the proposed HGF and IJR modules, as illustrated in Fig. 3. The residual block, which contains two HGF modules and one IJR module, is designed as the basic block in HopFIR. Moreover, we define a linear embedding layer to map the input to the latent space and a HG module to transform the output into 3D space. The HG module is a variant of HGF designed for the output layer, more details of which are provided in the supplementary material. HopFIR accepts 2D keypoints as input, which can be obtained via an off-the-shelf 2D detector. The graph is obtained by adding a normalized globally learnable k-hop graph to the skeleton graph and then symmetrically normalizing it, as in [49]. We use the L1-norm loss and L2-norm loss to compute the error between ground truth and prediction with the weighted sum as follows:

$$L = \alpha \sum_{n=1}^{N} \|Y_n - \hat{Y_n}\|_2 + \beta \sum_{n=1}^{N} \|Y_n - \hat{Y_n}\|_1 \tag{12}$$

where N is the joint number, $\hat{Y_n}$ is the predicted 3D position of joint n, Y_n is the ground truth, α = 1, and β = 0.1.

4. Experiments

In this section, we first introduce the experimental setup and implementation details of the HopFIR networks. We then present our experimental results and comparisons with state-of-the-art methods. Finally, we conduct several ablation studies of the proposed architecture.

4.1. Datasets and Evaluation Protocols

Datasets. Human3.6M [10] is currently the largest publicly available dataset for 3D human pose estimation, with 3.6 million video frames. It captures accurate 3D human joint positions from four camera viewpoints and records 11 subjects performing 15 assigned actions. Following previous works [49, 36, 42], we train our model on five subjects (S1, S5, S6, S7, S8) and test it on two subjects (S9, S11). In contrast to Human3.6M, MPI-INF-3DHP [21] includes complex outdoor scenes, which are commonly used to evaluate the generalizability of proposed methods. Accordingly, we use the test set of MPI-INF-3DHP to verify the generalizability of our model.

Evaluation Protocols. For Human3.6M [10], we evaluate our model on two standard evaluation protocols: the mean per-joint position error (MPJPE) and the mean per-joint position error after Procrustes alignment (P-MPJPE).

Protocol #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez et al. [20]	ICCV2017	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhao et al. [42](†)	CVPR2019	47.3	60.7	51.4	60.5	61.1	**49.9**	47.3	68.1	86.2	**55.0**	67.8	61.0	**42.1**	60.6	45.3	57.6
Ci et al. [4](†)	ICCV2019	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
Pavllo et al. [25]	CVPR2019	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
Cai et al. [2](†)	ICCV2019	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
Liu et al. [18](†)	ECCV2020	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
zeng et al. [39]	ECCV2020	44.5	48.2	47.1	**47.8**	51.2	56.8	50.1	45.6	59.9	66.4	52.1	**45.3**	54.2	39.1	40.3	49.9
Zou et al. [49](†)	ICCV2021	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
Xu et al. [36](†)	CVPR2021	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. [43] (Δ)(†)	CVPR2022	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
Ours(Δ)(†)		**43.9**	**47.6**	45.5	48.9	**50.1**	58.0	**46.2**	**44.5**	**55.7**	62.9	**49.0**	45.8	51.8	**38.0**	**39.9**	**48.5**

Table 1. Quantitative comparison on Human3.6M with detected 2D poses as input under Protocol #1, in millimeters. The best results are highlighted in bold and the second-best results are underlined. (†) indicates GCN-based methods and (Δ) indicates Transformer-based methods.

Protocol #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Zhou et al. [45](+)	ICCV2019	34.4	42.4	36.6	42.1	38.2	**39.8**	34.7	40.2	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9
Ci et al. [4](+)(∗)(†)	ICCV2019	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
Martinez et al. [20]	ICCV2017	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Zhao et al. [42](†)	CVPR2019	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
Cai et al. [2](†)	ICCV2019	33.4	39.0	33.8	37.0	38.1	47.3	39.5	37.3	43.2	46.2	37.7	38.0	38.6	30.4	32.1	38.1
Liu et al. [18](†)	ECCV2020	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
zeng et al. [39]	ECCV2020	35.9	36.7	29.3	34.5	36.0	42.8	37.7	31.7	40.1	44.3	35.8	37.2	36.2	33.7	34.0	36.4
Zou et al. [49](†)	ICCV2021	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.4
Xu et al. [36](†)	CVPR2021	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. [43] (Δ)(†)	CVPR2022	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(Δ)(†)		**31.3**	**34.0**	**28.0**	**32.0**	**33.1**	42.1	**34.1**	**28.1**	**33.6**	39.8	**31.7**	**32.9**	**33.8**	**26.7**	**28.9**	**32.7**

Table 2. Quantitative comparison on Human3.6M with ground truth 2D keypoints as input under Protocol #1, in millimeters. (+) uses additional data from MPII [1]. (∗) uses pose scales in both training and testing. The best results are highlighted in bold.

These are referred to Protocol #1 and Protocol #2, respectively. The MPJPE and P-MPJPE are given in millimeters. For MPI-INF-3DHP [21], we follow previous works [4, 42, 49] by reporting the percentage of correct keypoints (PCK) with a threshold of 150 mm and the area under the curve (AUC) for a range of PCK thresholds.

4.2. Implementation Details

Following previous work [25], we obtain the detected 2D poses using the cascaded pyramid network(CPN) [3]. We do not use data augmentation during training and testing with the 2D ground truth input to verify the efficacy of our model. We adopt Adam [11] optimizer and all experiments are conducted on a single NVIDIA RTX 3090 GPU. 3D pose regression from 2D detections is more challenging than that from 2D ground truth because the former needs to deal with some extra uncertainty in the 2D space. To manage this uncertainty, we set different configurations for them. In experiments with 2D ground truth as the input, we train the HopFIR networks with an initial learning rate of 0.001, a decay factor of 0.90 per 4 epochs, a batch size of 64, channels of 128, and PReLU activation [8]. When using detected 2D poses, we train the HopFIR networks with the initial learning rate of 0.006, a decay factor of 0.95 per 4 epochs (but 0.2 for the first 4 epochs), a batch size of 256, 256 channels, and LeaklyReLU activation [19]. To avoid overfitting, we apply Dropout[30] with a dropout rate of 0.5.

4.3. Comparison with State-of-the-art

We compare the performance of HopFIR with some SOTA methods on Human3.6M under Protocol #1 and Protocol #2, with the results shown in Table 1. Our method reaches an MPJPE of 48.50 mm and outperforms the best of the existing approaches [49] on all 15 actions. Given the uncertainty of 2D detections, we also investigate the capability of HopFIR networks using ground truth 2D key points as input. As shown in Table 2, HopFIR obtains surprisingly better performance when given precise 2D joint information and produces SOTA results, which verifies its effectiveness.

4.4. Ablation Study

We conduct a comprehensive ablation study on Human3.6M to validate the individual effectiveness of each component of the proposed HopFIR architecture under controlled settings. We follow previous works [18, 36, 49] conducting the ablation experiments using GT as inputs to avoid the influence of the 2D pose detector.

Effectiveness of Different Modules. We separately verify the effectiveness of the HGF module and IJR module and conduct experiments on first-order neighbors, removing all modules and the multi-hop mechanism. Note that GCN and HopGCN refer to applying only vanilla GCN (Eq. 1) and HopGCN (Eq. 5) to Fig. 3, respectively. The results in Table 3 show that each module improves the performance

Method	Channels	Params	MPJPE	P-MPJPE
GCN	128	0.36M	40.63	31.65
HopGCN	128	0.59M	39.15	31.40
HopGCN & IJR	128	2.15M	36.62	29.23
HopGCN & HGF	128	1.05M	35.19	28.81
HopGCN & IJR	64	0.57M	36.69	29.68
HopGCN & HGF	80	0.48M	36.01	29.58
HopGCN & HopFIR	128	2.15M	**32.67**	**26.20**
HopGCN & HopFIR	64	0.54M	33.52	27.37
HopGCN & HopFIR	32	0.14M	35.19	28.71

Table 3. Ablation experiments on the proposed modules.

over the GCN-only approach, and coupling the two layers to form a HopFIR block achieves further performance improvement. The HopFIR networks reduce the MPJPE to 32.67 mm, which represents a 7.2% improvement over GraFormer [43]. By reducing feature channels in HopGCN & IJR and HopGCN & HGF, we decrease parameters to 0.57M and 0.48M. However, the models still achieved errors of 36.69mm and 36.01mm, compared to the HopGCN with 39.15mm error and 0.59M parameters, which should be attributed to the structure design other than the model size. Moreover, we reduce the HopFIR network parameters by changing the channels to 64 and 32, respectively, which are also superior to SOTA.

Effectiveness of the HA Layer. In sec. 3.2, we introduce k-hop groups to discover latent joint interactions in human joint synergies. The attention matrices in Fig. 5 show the latent joint synergies captured by HopFIR, in which each weight of row i indicates a discovered latent group for the corresponding joint i. To verify the effectiveness of HA layer, we explore the correlation between all joints with a transformer encoder instead of HA layer, which explores the correlation between individual nodes but ignores the synergy between groups of joints in the human body and obtained 36.21 mm error. Moreover, we remove the human body prior by using random graph instead of skeleton graph without changing the HopFIR architecture and reached 34.68 mm error, suggesting that latent group correlations can be explored by k-hop groups, but group correlations underlying joint synergies can be better explored based on the human body prior.

In Table 4, we show the experimental results of three different ways to design the HA layer. HSS is the method selected in this paper, where H, S, and S are Q, K, and V, and H and S represent the node feature and the k-hop group feature, respectively. As we do not follow [33] in applying a linear transformation of Q, K, and V, we also show the result of such a linear transformation, which is denoted as HA+W. The experiment results show that HSS similarity achieves better performance in the HopFIR architecture, but one can choose the type of similarity according to the network property.

We further insert the HA layer into SOTA pose estimation methods, namely SemGCN [42] and Modulated GCN [49], to investigate its generalizability. No changes are made to their source code, with the HA layer inserted before the information aggregation stage of these methods. The experimental results in Table 5 show that the HA layer improves these previous SOTA networks to a large degree; especially, the MPJPE of SemGCN [42], is reduced from 42.14 mm to 38.41 mm, representing an 8.9% improvement. Moreover, HA layer with linear transformation (W) makes learning more stable, so We test HA and HA+W in SemGCN and MGCN, respectively, to show both of them are effective. For a fair comparison of parameters, we tested MLP and MHSA instead of HA, both of them with 1.03M parameters, and achieved 39.12mm and 39.22mm errors, respectively. More details can be found in the supplementary material. Experiments on the above methods of aggregating first-order neighbor information demonstrate the effectiveness of HA and also indicate that latent joint grouping can recognize the human joint synergies.

Error on Peripheral Joints. We report the regres-

Attention	HA			HA+W		
	HSS	SSS	SHH	HSS	SSS	SHH
Params	2.15M			2.50M		
MPJPE	**32.67**	34.18	34.28	**33.29**	33.53	33.90
P-MPJPE	**26.20**	27.70	27.71	**27.16**	27.40	27.41

Table 4. Quantitative comparison of HA layers with different similarity computing approaches.

Method	Channels	Params	MPJPE	P-MPJPE
SemGCN [42]	128	0.27M	42.14	33.53
SemGCN + HA(HSS)	128	0.49M	**38.41**	**30.56**
SemGCN + HA(SSS)	128	0.49M	41.30	33.07
SemGCN + HA(SHH)	128	0.49M	38.81	31.05
SemGCN [42] w/ Non-local [35]	128	0.43M	40.78	31.46
SemGCN w/ Non-local +HA(HSS)	128	0.66M	38.03	30.50
SemGCN w/ Non-local +HA(SSS)	128	0.66M	**37.75**	30.17
SemGCN w/ Non-local +HA(SHH)	128	0.66M	37.94	**29.71**
Modulated GCN [49]	128	0.29M	38.25	30.06
Modulated GCN +HA(HSS)+W	128	0.96M	36.54	29.09
Modulated GCN +HA(SSS)+W	128	0.96M	**36.14**	**29.02**
Modulated GCN +HA(SHH)+W	128	0.96M	37.38	30.02

Table 5. Comparison of the improved performance of proposed HA layer added on different methods. We test on two GCN-based methods: SemGCN [42] and MGCN [49].

Num-k	Channels	Params	MPJPE	P-MPJPE
1	128	1.88M	35.88	28.76
2	128	2.03M	34.96	27.74
3	128	2.15M	**32.67**	**26.20**
4	128	2.27M	35.58	28.35
Num-Block	Channels	Params	MPJPE	P-MPJPE
1	128	0.80M	37.33	30.71
2	128	1.47M	34.21	27.63
3	128	2.15M	**32.67**	**26.20**
4	128	2.82M	33.84	27.63

Table 6. Ablation study for number of k-hop and designed blocks. The units of MPJPE and P-MPJPE are millimeters (mm).

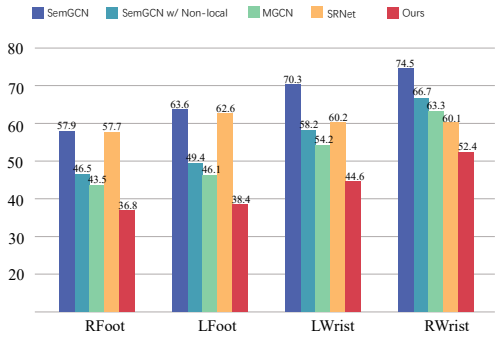

Figure 4. Comparison of MPJPE for peripheral joints on the test set of the Human3.6M. R and L denote right and left, respectively.

Block	Channels	Params	MPJPE	P-MPJPE
(H)(I)	128	1.78M	34.43	27.79
(I)(H)	128	1.78M	34.13	27.56
(H)(I)(H)	128	2.15M	35.08	27.98
(I)(H)(H)	128	2.15M	34.20	27.60
(H)(H)(I)	128	2.15M	**32.67**	**26.20**

Table 7. Ablation study for arrangements of the designed block. (I) and (H) denote IJR module and HGF module, respectively.

	MPJPE	P-MPJPE	MPJVE
PoseFormer [44] (T=81)	44.3	36.5	3.1
MixSTE [41] (T=243)	**40.9**	32.6	2.3
MHFormer [17] (T=351)	43.0	-	-
P-STMO [28] (T=243)	42.1	34.4	-
Ours (T=243)	41.1	**32.5**	**2.1**

Table 8. Quantitative comparision on Human3.6M with detected 2D pose (CPN) in video. T denotes the number of input frames.

Methods	PCK				AUC
	GS	no GS	Outdoor	All	
Martinez et al. [20]	49.8	42.5	31.2	42.5	17.0
Ci et al. [4]	74.8	70.8	77.3	74.0	36.7
zeng et al. [39]	-	-	80.3	77.6	43.8
Li et al. [16]	70.1	68.2	66.6	66.9	-
Zhao et al. [43]	80.1	77.9	74.1	79.0	43.8
Liu et al. [18] (weight unsharing)	77.6	80.5	80.1	79.3	47.6
Xu et al. [36]	81.5	81.7	75.2	80.1	45.8
Nie et al. [23]	-	-	-	83.5	45.9
Zou et al. [49]	86.4	**86.0**	85.7	86.1	53.7
Ours	**89.1**	85.9	**85.9**	**87.2**	**57.0**

Table 9. Quantitative comparisons on the MPI-INF-3DHP test set. GS denotes green screen.

sion accuracy for the peripheral joints (wrists and feet) in Fig. 4 in comparison with some previously proposed methods [49, 42, 39]. The HopFIR network with IJR modules outperforms SOTA methods on the right foot (RFoot) by 6.7 mm, left foot (LFoot) by 7.7 mm, left wrist (LWrist) by 9.6 mm, and right wrist (RWrist) by 7.7 mm. The experimental results verify that the intragroup joint attention within each limb group strengthens the capabilities of the HopFIR.

Different Numbers of k-Hops. As HopFIR is designed

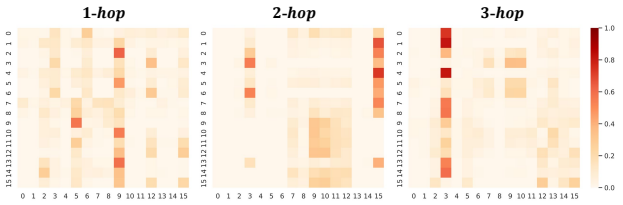

Figure 5. Attention weight of the j-th k-hop for the i-th joint, deeper color indicates higher correlation. i-th row and j-th col represent i-th joint and k-hop of j-th joint, respectively.

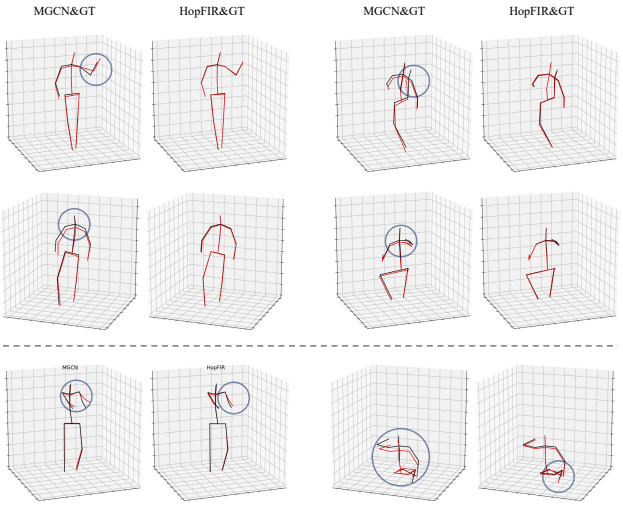

Figure 6. Qualitative visual results for HopFIR and MGCN [49] on the Human3.6M. The black lines are the ground truth (GT) and the red lines are the predictions of HopFIR and MGCN. Wrong predictions are circled. The bottom row shows our failure case.

to extract the correlation between feature groups, we set different k values to discover various latent connections underlying the human joint synergies, with the results shown at the top of Table 6. The MPJPE gradually decreases as the number of hops increases, and reaches the best performance at 3 hops. Therefore, the optimal number of hops for 3D HPE is 3, which entails that we obtain 16×3 groups from the skeleton graph. Each of the 16 groups corresponds to a potential correlation among coupled nodes at different distances, and three hops is sufficient to recognize the joint synergies.

Arrangement of the Designed Block. To investigate the optimal structure of the designed block, experiments are performed with various block numbers and various combinations of HGF and IJR modules. As shown at the bottom of Table 6, the error gradually decreases as the number of blocks increases, until the best performance is achieved at 3 blocks. As shown in Table 7, the (H)(H)(I) arrangement achieves the optimal results by reducing the error to 32.67 mm. HGF treats each hop as a group and applies a hop-

wise attention mechanism to these groups to discover latent joint synergy. IJR utilizes the limb prior for peripheral joint refinement. Thus, (H)(H)(I) first integrates the complete joint information and then refines it by IJR. While (I)(H)(H) reverses this procedure resulting in the insufficient utilization of joint information. That the results are superior to GraFormer [43] in all cases except for the single-block model indicates that the HopFIR has significant human pose representation capabilities.

Extend to Temporal Domain. Without a specific design to integrate temporal information, we extend to the temporal domain by adding two TEs after each block and replacing the HG module with a linear layer, and achieved competitive results as shown in table 8.

Cross-Dataset Results on MPI-INF-3DHP. Table 9 further compares HopFIR with previous methods on cross-dataset scenarios to validate its generalizability. For these experiments, we train our model on the Human3.6M dataset and test it on the test set of the MPI-INF-3DHP dataset. The results show that our approach obtains better results than other methods, which verifies the generalizability of our approach to unseen scenarios.

Qualitative Results. In Fig. 6, we show the visual results on Human3.6M in the world space. The bottom of the figure shows some failure cases of HopFIR, which predict some wrong joint positions. The figure shows that HopFIR is able to predict 3D joint positions more accurately, even for poses that cause difficulties for MGCN.

5. Conclusions

We present the Hop-wise GraphFormer with Intragroup Joint Refinement (HopFIR) as a novel architecture for 3D human pose estimation. The proposed architecture mainly comprises the HGF and IJR modules. The HGF module improves on the GCN-based pose estimation networks by grouping the joints by k-hop neighborhood and capturing the potential joint correlations in the different joint synergies. Because the peripheral joints strongly interact with intra-limb joints, the proposed IJR module applies intragroup attention to refine the peripheral joint features through the associated limb. The proposed method achieves new state-of-the-art results while maintaining a modest model size.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2021YFC0122602), in part by the Joint Funds Program of the National Natural Science Foundation of China (Grant No. U21A20517), in part by the Basic Science Centre Program of National Natural Science Foundation of China (Grant No. 72188101).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6

[2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2272–2281, 2019. 1, 6

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 6

[4] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019. 2, 3, 6, 8

[5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2:729–734 vol. 2, 2005. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 2, 5

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[15] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 1

[16] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 8

[17] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 3, 8

[18] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *European Conference on Computer Vision*, pages 318–334. Springer, 2020. 1, 2, 3, 6, 8

[19] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA, 2013. 6

[20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 1, 2, 6, 8

[21] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 2, 5, 6

[22] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 3

[23] Qiang Nie, Ziwei Liu, and Yunhui Liu. Lifting 2d human pose to 3d with domain adapted 3d body concept. *International Journal of Computer Vision*, 131(5):1250–1268, 2023. 8

[24] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2

[25] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 6

[26] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European conference on computer vision*, pages 573–586. Springer, 2012. 2

[27] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 3

[28] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 3, 8

[29] Cristian Sminchisescu. 3d human motion analysis in monocular video: techniques and challenges. In *Human motion*, pages 185–211. Springer, 2008. 2

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 6

[31] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 1

[32] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018. 1

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 5, 7

[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 3

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 7

[36] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16105–16114, 2021. 1, 2, 5, 6, 8

[37] Youze Xue, Jiansheng Chen, Xiangming Gu, Huimin Ma, and Hongbing Ma. Boosting monocular 3d human pose estimation with part aware attention. *IEEE Transactions on Image Processing*, 31:4278–4291, 2022. 2

[38] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5255–5264, 2018. 1

[39] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020. 1, 2, 6, 8

[40] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11436–11445, 2021. 2, 3, 4

[41] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022. 3, 8

[42] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019. 1, 3, 4, 5, 6, 7, 8

[43] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022. 1, 3, 6, 7, 8, 9

[44] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 3, 8

[45] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2344–2353, 2019. 6

[46] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 2

[47] Yiran Zhu, Xing Xu, Fumin Shen, Yanli Ji, Lianli Gao, and Heng Tao Shen. Posegtac: Graph transformer encoder-decoder with atrous convolution for 3d human pose estimation. In *IJCAI*, pages 1359–1365, 2021. 3

[48] Zhiming Zou, Kenkun Liu, Le Wang 0003, and Wei Tang. High-order graph convolutional networks for 3d human pose estimation. In *BMVC*, 2020. 3, 4

[49] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11477–11487, 2021. 1, 2, 3, 4, 5, 6, 7, 8