Pattern Recognition Letters 189 (2025) 175–181

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A cross-feature interaction network for 3D human pose estimation

Jihua Peng [a], Yanghong Zhou [a,c], P.Y. Mok [a,b,d,e,*]

[a] School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong
[b] Laboratory for Artificial Intelligence in Design, Hong Kong
[c] Research Centre of Textiles for Future Fashion, The Hong Kong Polytechnic University, Hong Kong
[d] Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong
[e] Division of Integrative Systems and Design, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Editor: Antonio Fernández-Caballero

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D human pose estimation
graph convolutional network (GCN)
self-attention
cross-attention

ABSTRACT

The task of estimating 3D human poses from single monocular images is challenging because, unlike video sequences, single images can hardly provide any temporal information for the prediction. Most existing methods attempt to predict 3D poses by modeling the spatial dependencies inherent in the anatomical structure of the human skeleton, yet these methods fail to capture the complex local and global relationships that exist among various joints. To solve this problem, we propose a novel Cross-Feature Interaction Network to effectively model spatial correlations between body joints. Specifically, we exploit graph convolutional networks (GCNs) to learn the local features between neighboring joints and the self-attention structure to learn the global features among all joints. We then design a cross-feature interaction (CFI) module to facilitate cross-feature communications among the three different features, namely the local features, global features, and initial 2D pose features, aggregating them to form enhanced spatial representations of human pose. Furthermore, a novel graph-enhanced module (GraMLP) with parallel GCN and multi-layer perceptron is introduced to inject the skeletal knowledge of the human body into the final representation of 3D pose. Extensive experiments on two datasets (Human3.6M (Ionescu et al., 2013) and MPI-INF-3DHP (Mehta et al., 2017)) show the superior performance of our method in comparison to existing state-of-the-art (SOTA) models. The code and data are shared at https://github.com/JihuaPeng/CFI-3DHPE

1. Introduction

The goal of 3D human pose estimation (HPE) is to predict the 3D coordinates of body joints from input human images. It is a prominent research field in computer vision with versatile applications across various domains, such as action recognition [1], human–robot interaction [2], and virtual reality [3]. Among various approaches for 3D HPE, the 2D-to-3D lifting methods [4–7] represent the mainstream approach, which infers 3D poses from the estimated 2D poses of the input images. This approach, benefiting from the remarkable performance of 2D pose detectors [8,9], has achieved state-of-the-art performance, outperforming other one-step methods [10–12]. However, the task of lifting an individual 2D pose into 3D space is an ill-posed problem, as multiple 3D poses can yield the same 2D projections. Furthermore, the absence of temporal information in single monocular images further amplifies the issue of depth ambiguity in this task. The focus of this research centers on the investigation for strategies to thoroughly explore and effectively capture spatial information.

One the other hand, the graph convolutional networks (GCNs) [13,14] have recently been used for single-frame 3D pose estimation with outstanding performance. Such GCNs-based methods [15–17] utilize the topological information of the human skeleton by aggregating features of the neighboring body joints. However, these methods [15–17] focus only on modeling the motion characteristics of adjacent or connecting joints, namely the *local information*. There are, in fact, additional implicit kinematic information between joints that are not physically connected. For example, in the action of 'walking a dog', the joints of two hands and two feet move in the same direction along the dog's motion. In order to better capture the *global information* of human skeleton representations, some transformer-based methods [18–20] were proposed, exploiting the self-attention mechanism to effectively model the spatial dependencies among all body joints. In addition, some other studies [21–23] combined GCNs and transformers to facilitate the learning of spatial correlations in human skeleton. Nevertheless, all of them utilize GCNs and transformer blocks in a

* Corresponding author.
E-mail addresses: ji-hua.peng@connect.polyu.hk (J. Peng), yanghong.zhou@connect.polyu.hk (Y. Zhou), tracy.mok@polyu.edu.hk, tracy.mok@ust.hk (P.Y. Mok).

https://doi.org/10.1016/j.patrec.2025.01.016
Received 30 May 2024; Received in revised form 4 December 2024; Accepted 18 January 2025
Available online 1 February 2025

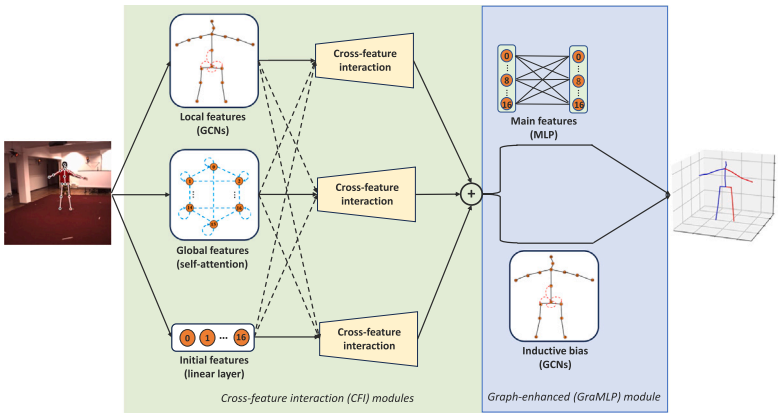

Fig. 1. Schematic architecture of the proposed method. We extract the initial, local and global features, respectively, by a linear layer, GCNs and self-attention mechanism. Next, cross-feature interaction (CFI) modules are introduced that facilitate cross-feature communications among these three features. A GraMLP module integrates human skeletal knowledge as an inductive bias into the final representation of 3D pose.

sequential manner, either using the output of GCNs as the input for a transformer block or vice versa. The resulting features from both GCNs and transformers lack direct interaction, which may hinter the model's performance, preventing it from fully leveraging the strengths of both components.

Tao et al. [24] tacked the problem of communication constraints within network by quantizing the input and output signals. Wang et al. [25] employed the Q-learning algorithm to continuously update the action–value function based on interactions between the agent and the environment. Song et al. [26] developed a neural adaptive quantized control strategy to alleviate the communication burden in interconnected nonlinear systems. Inspired by these interaction-based methods [24–26], for the task of single-frame 3D HPE, we propose a novel Cross-Feature Interaction (CFI) Network to effectively improve the learning of spatial representations of human skeleton. The primary motivation is to enable the network to effectively leverage the local features derived by GCNs, the global features captured by self-attention, and the initial features, and simultaneously facilitate cross-feature communication among them. Fig. 1 shows the schematic architecture of our method. As shown, we capture the local and global features by GCNs and self-attention mechanisms, respectively. We also obtain the initial 2D pose features by a linear embedding. The initial features, often neglected by other methods, can serve as an residual connection, to effectively compensate for the information loss that occurs during the layer-to-layer propagation of the other two types of features. Moreover, we design a specific multi-head cross-attention (MHCA) to facilitate cross-feature interaction among the three different features, namely the local features, global features, and the initial 2D pose features. This specially designed MHCA, named as cross-feature interaction (CFI) module, can effectively model dependencies between multiple features and enable the other two features to complement the features of the current branch. Next, these three types of features derived from individual CFI modules are aggregated to form the enhanced spatial features. Finally, we develop a graph-enhanced module (GraMLP) with parallel structure of GCN and multi-layer perceptron (MLP) to incorporate the human skeletal knowledge as an inductive bias into the final representation of 3D pose. The key contributions of this paper are summarized as follows:

- We develop a novel Cross-Feature Interaction Network for single-frame 3D pose estimation. A cross-feature interaction (CFI) module is designed to effectively model dependencies among local features, global features, and the initial features, which are further aggregated as the enhanced spatial features.
- A graph-enhanced module 'GraMLP' is introduced to integrate vanilla MLP with GCN, improving the accuracy of 3D pose estimation.
- Extensive experiments on two benchmarks show that our method outperforms other SOTA models.

2. Related work

2D-to-3D lifting methods. Different from one-step methods [27–29] that directly regress the 3D pose from input images, the two-step methods utilize the intermediate 2D pose detectors [8,9]. They first obtain 2D joint coordinates from input images using off-the-shelf 2D pose detectors, and then design a 2D-to-3D lifting network to lift these 2D poses into 3D space. Martinez et al. [4] proposed a simple yet effective baseline network to regress the single 3D pose, which demonstrates the outstanding performance of estimated 3D pose obtained by utilizing accurately predicted 2D pose locations as inputs. Fang et al. [30] integrated bi-directional RNNs with cascaded linear layers to encode the human body configurations into a knowledge set. Zhang et al. [31] proposed a human structure aware network to refine the coordinates of hard joints.

GCN-based methods. Given that the human skeletal structure can be represented as a graph, several methods [15–17] leverage GCNs to model spatial correlations among body joints. Zhao et al. [15] introduced a learnable mask to scale up the receptive field of convolution filters in GCNs, capturing semantic information among all nodes. Liu et al. [16] explored different weight sharing schemes in GCNs and proposed a pre-aggregation graph convolution to aggregate node information with varying weights. Zou and Tang [17] introduced a weight modulation vector and a matrix modulation vector to efficiently enhance the performance of GCN-based pose estimation.

Transformer-based methods. Transformer, a deep learning network, has revolutionized first in natural language processing (NLP) and later in computer vision since its introduction in 2017 [32]. Zheng et al. [18] first utilized the transformer to learn spatio-temporal representations for 3D pose estimation. Zhu et al. [21] inserted the graph convolution into transformer to model long-range correlations among multi-top neighboring nodes. Zhao et al. [22] replaced the MLP in the transformer with learnable graph convolution layers to form the GraAttention block, capturing global information among all nodes. Li et al. [20] encoded the relative distance between a pair of joints and used the distance information as the attention bias in the self-attention module. However, these methods fail to effectively leverage the individual strengths of GCNs and self-attention mechanisms in extracting local and global features, respectively, nor model the interactions between the features.

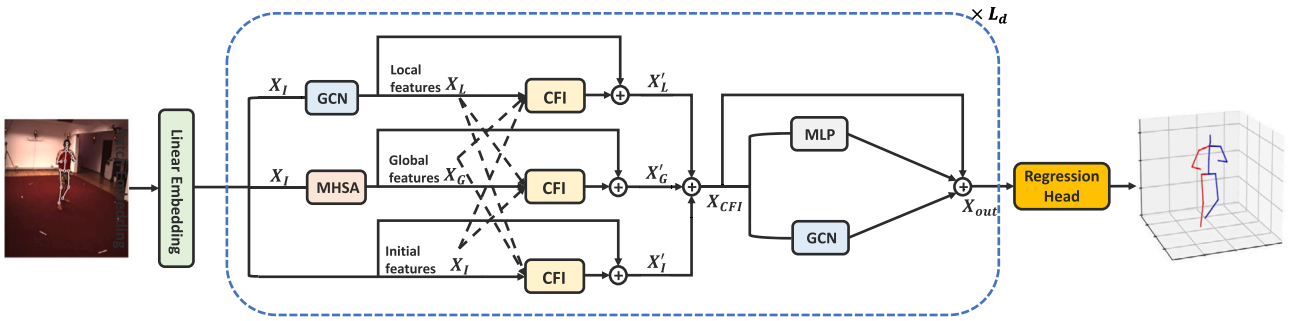

Fig. 2. An overview of Cross-Feature Interaction Network. The input 2D pose joints $X \in \mathbb{R}^{N\times 2}$ are projected to initial features $X_I \in \mathbb{R}^{N\times D}$ by a linear embedding. Then, the initial features are processed by GCNs and self-attention, respectively, generating local features $X_L \in \mathbb{R}^{N\times D}$ and global features $X_G \in \mathbb{R}^{N\times D}$. Three CFI modules output three enhanced features X'_G, X'_L, X'_I, which are then added together to form X_{CFI} to input to the GraMLP. Lastly, the GraMLP output the X_{out} to the regression head to generate the final 3D pose.

3. Proposed method

In this paper, we design a network to facilitate the communication between local and global features, simultaneously considering the modeling capabilities of GCNs [33] and multi-head self-attention (MHSA) [32], as shown in Fig. 2. We first present a brief overview of GCNs and MHSA below.

3.1. Preliminary

Graph convolutional networks (GCNs) [33]. A graph can be defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a collection of nodes and $\mathcal{E}$ indicates a set of edges. The representation of edges can be realized through an adjacency matrix $A \in \{0,1\}^{N\times N}$, while the set of features of all nodes in the lth layer can be expressed as a matrix $H_l \in \mathbb{R}^{D\times N}$, where N is the number of nodes, and D represents the feature dimension. The graph convolution operation aggregates features from neighboring nodes in the lth layer as follows:

$$H_l = \sigma(W_l H_{l-1} \tilde{A}) \tag{1}$$

where $W_l \in \mathbb{R}^{D\times D}$ is a learnable weight matrix $\tilde{A}$, referring to the adjacency matrix of the graph with the inclusion of self-connections, $\tilde{A} = A + I_N$ and I_N is an identity matrix.

Multi-head self-attention (MHSA) [32]. The MHSA computes multiple attention heads via self-attention in parallel. Each attention head $(i = 1, \ldots, h)$ is computed by $head_i = \textsc{Softmax}\left(\frac{(ZW_i^Q)(ZW_i^K)^T}{\sqrt{d_m}}\right)(ZW_i^V)$, where $Z \in \mathbb{R}^{N\times D}$ is the input token, W_i^Q, W_i^K and $W_i^V \in \mathbb{R}^{D\times D}$ are learnable parameters. The function $\textsc{Softmax}(\cdot)$ normalizes the dot-product scores into a probability distribution. All h attention heads are then concatenated together:

$$Z_{MHSA} = \textsc{Concat}(head_1, \ldots, head_i, \ldots, head_h) \tag{2}$$

where function $\textsc{Concat}(\cdot)$ combines the outputs of multiple attention heads, followed by a linear transformation, to form the final output Z_{MHSA}.

3.2. Cross-feature interaction

Fig. 2 illustrates the proposed Cross-Feature Interaction Network, which consists of two main components of *Cross-Feature Interaction* module (CFI) and graph-enhanced module (GraMLP). The input 2D pose joints $X \in \mathbb{R}^{N\times 2}$ are initially embedded into high-dimensional tokens by a linear embedding layer, resulting in the initial features, denoted as $X_I \in \mathbb{R}^{N\times D}$. N represents the number of joints, and $N = 17$ for 3D HPE task, while D represents the feature dimension, which can be set to 256, 512, 1024, or other values. The initial features X_I is then fed into the GCN, yielding the local features $X_L \in \mathbb{R}^{N\times D}$:

$$X_L = \sigma(W X_I \tilde{A}) \tag{3}$$

where $\tilde{A}$ denotes the adjacency matrix of anatomical relationships in the human body. We obtain the global features $X_G \in \mathbb{R}^{N\times D}$ by Eq. (2) and each head is resulted from feeding initial features X_I to the MHSA:

$$head_i^G = \textsc{Softmax}\left(\frac{\left(X_I W_i^Q\right)\left(X_I W_i^K\right)^T}{\sqrt{d_m}}\right)\left(X_I W_i^V\right) \tag{4}$$

To facilitate communication and achieve mutual complementarity among the three types of features, we introduce a **cross-feature interaction** module, as shown in Fig. 3. For example, the initial features X_I, local features X_L, and global features X_G are regarded as queries, keys, and values, respectively, for a specific multi-head cross attention of the CFI unit as follows:

$$head_i = \textsc{Softmax}\left(\frac{\left(X_I W_i^Q\right)\left(X_L W_i^K\right)^T}{\sqrt{d_m}}\right)\left(X_G W_i^V\right) \tag{5}$$

In Eq. (5), $\textsc{Softmax}(\cdot)$ is used to compute attention weights that determine the contribution of each feature to the attended output. The *enhanced global features* $X'_G \in \mathbb{R}^{N\times D}$ can be obtained by:

$$X'_G = \textsc{Concat}(head_1, \ldots, head_i, \ldots, head_h) + X_G \tag{6}$$

where $\textsc{Concat}(\cdot)$ aggregates the attended features across multiple attention heads. By Eq. (5), the three features engage in interactions and exchange information with each other. The global features can compensate for the limited receptive field of GCN, providing additional implicit kinematic knowledge to the local features. The initial features can offer valuable information that may be lost during the process of feature aggregation by GCN from neighboring joints. Moreover, the residual term in Eq. (6) ensures that the current branch focus on the global features.

Similarly, we employ the CFI module to obtain the enhanced local features $X'_L \in \mathbb{R}^{N\times D}$ and initial features $X'_I \in \mathbb{R}^{N\times D}$. Hereafter, the enhanced features X'_G, X'_L and X'_I are sum up to form as the output sequence from the CFI modules:

$$X_{CFI} = X'_G + X'_L + X'_I \tag{7}$$

which X_{CFI} is input to the GraMLP module.

3.3. GraMLP

The MLP structure in a vanilla transformer is densely connected, which has limited ability to model topological structure information of human skeleton. To inject the human skeleton information into the final 3D pose, we introduce a parallel design of MLP and GCN, namely GraMLP. Considering that the MLP can introduce non-linearities to the input features, adding GCN in parallel can retain anatomical knowledge of the human body, serving as an inductive bias to enhance the representation of 3D pose. GraMLP processes the features from the CFI

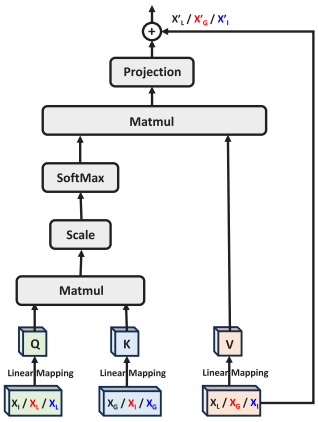

Fig. 3. Cross-feature interaction module (CFI) with three input features of X_I, X_L and X_G. One feature vector serves as the query while the others serve as key and value, enabling information exchange across different features. The output of this interaction module generates enhanced features X'_I, X'_L and X'_G, respectively.

module as follows:

$$X_{out} = X_{CFI} + MLP(X_{CFI}) + GCN(X_{CFI}) \tag{8}$$

where $MLP(\cdot)$ consists of multiple fully connected layers with GELU activation functions. $GCN(\cdot)$ refers to Eq. (3).

3.4. Regression head and loss function

In the regression head, a simple linear layer without an activation function is applied to predict, based on the output X_{out}, the 3D joint coordinates of the output pose $\tilde{J}$, as shown in Fig. 2. The loss function for our CFI network is thus given as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\|\tilde{J}_i - J_i\|_2^2) \tag{9}$$

where $\tilde{J}_i$ and $J_i \in \mathbb{R}^{N\times 3}$ denote the predicted and ground-truth 3D joint coordinates, respectively.

4. Experiments

4.1. Datasets and evaluation metrics

Two benchmark datasets were used in experimental evaluation in this study. **Human3.6M** [35] is an indoor scenes dataset with 3.6 million video frames. It has 11 professional actors, performing 15 actions under 4 synchronized camera views. Following previous work [20,22], we used subjects 1, 5, 6, 7 and 8 for training, and subjects 9 and 11 for testing. Two standard **evaluation metrics**, namely the mean per-joint position error (MPJPE) and the mean per-joint position error after procrustes alignment (P-MPJPE) were used in the evaluation.

MPI-INF-3DHP [36] is also a public large-scale dataset, including indoor and outdoor scenes. The test set comprises three distinct scenarios: a studio with green screen (GS), a studio without green screen (noGS), and outdoor scene (Outdoor). Following [20,22], the area under the curve (AUC) and the percentage of correct keypoints (PCK) were used as **evaluation metrics**. We employ the test set of this dataset to verify the generalization capability of our model trained on Human3.6M.

4.2. Implementation details

We implemented our method over the PyTorch™ [37] framework on one NVIDIA GeForce™ RTX 3090 GPU. We stacked the Cross-Feature Interaction Network for 3 loops, i.e., $L_d = 3$ in Fig. 2. We used the Adam optimizer to train our model for 20 epochs using mini-batch size of 512. The learning rate was initialized to 0.001 and decayed by 0.95 per epoch. We discuss the feature dimension and the number of heads in Section 4.4.

4.3. Comparison with state-of-the-art methods

Human3.6M. Table 1 compares the single-image estimation accuracy of our method with existing SOTA methods using 2D poses detected by CPN [9] as inputs. As shown, our method outperforms other SOTA models and achieve the same performance of 49.4 mm of MPJPE as Zou and Tang [17], which adopts a refinement module [6]. By applying the same refinement [6] to our model, the performance is improved from 49.4 mm to 48.6 mm, surpassing MGCN [17] by 0.8 mm error reduction. Moreover, our method obtains the best results of 38.8 mm and 38.7 mm in terms of P-MPJPE. In Table 2, we compare our results with those SOTA methods using 2D ground-truth poses as inputs. Our method attains SOTA performance, validating the effectiveness of our method for different types of input.

MPI-INF-3DHP. Table 3 reports the quantitative comparisons of our method with SOTA methods on cross-dataset scenarios. Our model was trained on the Human3.6M dataset and subsequently evaluated on the test set of the MPI-INF-3DHP dataset. The results show that our method achieves the best PCK and AUC performance in noGS, Outdoor and All scenarios. The PCK result in GS scenario is the second best. The possible reason is that the indoor GS data is relatively simple and limited in quantity, while the noGS and Outdoor data are more complex and abundant. Our model trained on large Human3.6M dataset may suffer from overfitting problem in GS scenarios.

4.4. Ablation study

To examine the effectiveness of each proposed module, we conducted ablation experiments on Human3.6M using 2D poses detected by CPN [9] as inputs. Table 4 shows the results of the ablation study. The vanilla transformer network, composed of the MHSA and MLP, is utilized as our baseline. For consistency, the transformer network is stacked for 3 loops, resulting in an overall accuracy of 51.9 mm MPJPE. The notation CFI(·) indicates the application of CFI module to feature representations of the said branch. For example, CFI(local) denotes the application of CFI module to the local features, i.e., Eq. (7) has only one component of X'_L. The results show that the application of three CFI modules, i.e., CFI(global), CFI(local) and CFI(initial), contribute 0.5 mm, 0.7 mm and 1.3 mm of error reduction, respectively. The incorporation of three CFI modules can result in 4% improvement of accuracy, reducing MPJPE from 51.9 mm to 49.8 mm. Table 4 also shows that the initial features play a crucial role in the interaction of local and global features, which brings the largest contribution of accuracy improvement. This is because the initial features processed by our CFI module can serve as an residual connection to effectively compensate for the information loss that occurs during the layer-to-layer propagation of the other two types of features. Lastly, by the introduction of the GraMLP module on top of three CFI modules, the estimation errors further drop 0.4 mm, achieving 49.4 mm of MPJPE. The ablation experiments demonstrate the effectiveness of each proposed module in our method.

Moreover, there are three hyper-parameters for our model (i.e., the depth of network L_d, the dimension of model D and the number of heads H in attention blocks). We tested different values of these parameters to verify which set of values yields the best results. As shown in Table 5, our model with $L_d = 3$, $D = 512$, $H = 8$ obtains the best results.

Table 1
Quantitative comparisons with SOTA methods of MPJPE and P-MPJPE based on Human3.6M with 2D poses detected by CPN [9] as inputs.

MPJPE (CPN)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez et al. [4] (ICCV'17)	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. [15] (CVPR'19)	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. [5] (ICCV'19)	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
Liu et al. [16] (ECCV'20)	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 [7] (CVPR'21)	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. [22] (CVPR'22)†	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
Cai et al. [6] (ICCV'19)∗	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
Li et al. [20] (AAAI'23)†	47.9	50.0	47.1	51.3	51.2	59.5	48.7	46.9	**56.0**	61.9	51.1	48.9	54.3	40.0	42.9	50.5
Zeng et al. [34] (ECCV'20)	**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 and Tang [17] (ICCV'21)∗	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†	45.4	49.5	46.1	49.3	51.7	56.7	**47.3**	**44.6**	58.6	63.0	50.4	47.2	51.8	38.2	41.3	49.4
Ours† ∗	45.0	50.3	45.8	48.4	**49.7**	**55.8**	**47.3**	45.4	56.4	59.4	49.9	46.5	50.9	**38.0**	**39.6**	**48.6**
P-MPJPE (CPN)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez et al. [4] (ICCV'17)	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Ci et al. [5] (ICCV'19)	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
Liu et al. [16] (ECCV'20)	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
Cai et al. [6] (ICCV'19)∗	36.8	38.7	38.2	41.7	40.7	46.8	37.9	35.6	47.6	51.7	41.3	36.8	42.7	31.0	34.7	40.2
Zeng et al. [34] (ECCV'20)	35.8	39.2	36.6	**36.9**	39.8	45.1	38.4	36.9	47.7	54.4	**38.6**	36.3	**39.4**	30.3	35.4	39.4
Zou and Tang [17] (ICCV'21)∗	35.7	38.6	36.3	40.5	**39.2**	44.5	37.0	35.4	46.4	51.2	40.5	**35.6**	41.7	30.7	33.9	39.1
Ours†	**35.3**	**37.8**	36.8	40.1	40.1	**43.6**	36.2	**34.3**	46.4	50.2	40.8	**35.6**	41.1	**30.0**	34.0	38.8
Ours† ∗	35.5	38.1	**35.9**	40.4	39.9	43.7	**36.0**	34.7	**46.1**	**48.4**	40.5	35.7	41.3	30.2	**33.7**	**38.7**

∗ denotes using the refinement module [6]. † indicates the transformer-based methods. Best results are showed in **bold**.

Table 2
Quantitative comparisons with SOTA methods of MPJPE based on Human3.6M with ground-truth (GT) 2D poses as inputs.

MPJPE (GT)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Martinez et al. [4] (ICCV'17)	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. [15] (CVPR'19)	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. [6] (ICCV'19)∗	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
Zhu et al. [21] (IJCAI'21)†	37.2	42.2	32.6	38.6	38.0	44.0	40.7	35.2	41.0	45.5	38.2	39.5	38.2	29.8	33.0	38.2
Liu et al. [16] (ECCV'20)	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 [17] (ICCV'21)∗	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.4
Zeng et al. [34] (ECCV'20)	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
Xu and Takano [7] (CVPR'21)	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. [22] (CVPR'22)†	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
Li et al. [20] (AAAI'23)†	32.9	38.3	**28.3**	33.8	34.9	38.7	37.2	30.7	**34.5**	39.7	33.9	**34.7**	**34.3**	**26.1**	28.9	33.8
Ours†	35.4	38.7	29.8	34.8	33.6	**36.8**	39.8	30.9	36.6	**36.3**	34.9	37.6	34.4	28.3	30.4	34.6
Ours† ∗	**29.1**	37.1	29.5	**31.8**	**33.2**	41.1	36.0	**29.8**	38.2	39.3	**33.3**	36.2	35.8	27.3	**28.6**	**33.7**

∗ denotes using the refinement module [6]. † indicates the transformer-based methods. Best results are showed in **bold**.

Table 3
Quantitative comparisons with SOTA methods of PCK and AUC performance based on MPI-INF-3DHP.

Methods	PCK ↑				AUC ↑
	GS	noGS	Outdoor	All	
Martinez et al. [4] (ICCV'17)	49.8	42.5	31.2	42.5	17.0
Ci et al. [5] (ICCV'19)	74.8	70.8	77.3	74.0	36.7
Zeng et al. [34] (ECCV'20)	–	–	80.3	77.6	43.8
Zhao et al. [22] (CVPR'22)	80.1	77.9	74.1	79.0	43.8
Liu et al. [16] (ECCV'20)	77.6	80.5	80.1	79.3	47.6
Xu and Takano [7] (CVPR'21)	81.5	81.7	75.2	80.1	45.8
Li et al. [20] (AAAI'23)	**86.2**	84.7	81.9	84.1	53.7
Ours	85.0	**86.1**	**85.7**	**85.6**	**54.0**

Table 4
Ablation study on the effectiveness of different modules.

CFI(initial) X'_I	CFI(local) X'_L	CFI(global) X'_G	GraMLP	MPJPE (mm)
				51.9
			✓	51.0
✓				50.6
	✓			51.2
		✓		51.4
✓	✓			50.2
✓		✓		50.5
	✓	✓		50.6
✓	✓	✓		49.8
✓	✓	✓	✓	**49.4**

✓indicates the corresponding module is being included.

Table 5
Ablation study of hyperparameter setting. L_d is the depth of the network, D is the dimension of model, and H is the number of heads in attention blocks.

L_d	D	H	MPJPE (mm)	P-MPJPE (mm)
2	512	8	49.7	39.1
3	512	8	**49.4**	**38.8**
4	512	8	50.0	39.2
3	256	8	49.8	**38.8**
3	512	8	**49.4**	**38.8**
3	1024	8	49.7	39.0
3	512	4	49.6	39.0
3	512	8	**49.4**	**38.8**
3	512	16	49.8	39.1

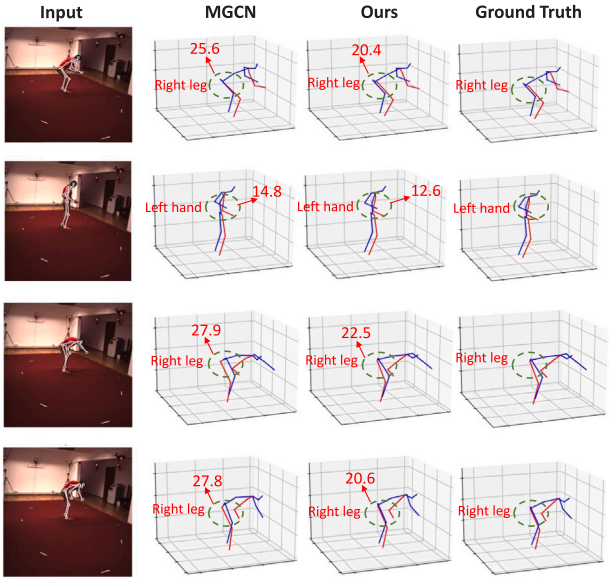

Fig. 4. Qualitative comparison with MGCN [17] on Human3.6M dataset. The red numbers represent the MPJPE errors.

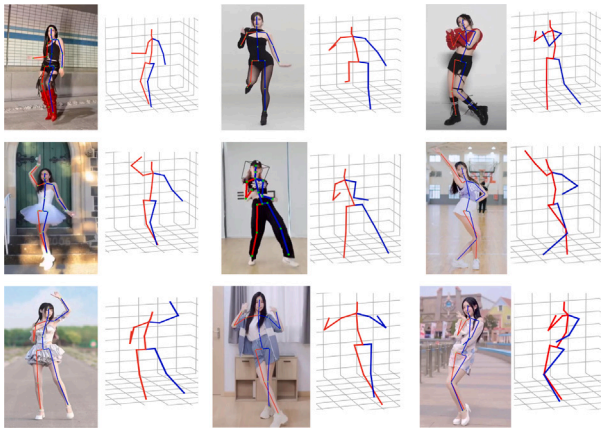

Fig. 5. Results of 3D pose estimation with in-the-wild input images.

4.5. Qualitative results

We visualize some 3D pose estimation results to validate the efficacy of our method in comparison to MGCN [17] in Fig. 4. The red numbers represent the MPJPE errors for the joints in the green circled areas (e.g., hands, legs). As shown, we can achieve more accurate 3D pose predictions compared to MGCN [17]. Moreover, we collected some in-the-wild images to test and validate the applicability of our method to real-world situation, as shown in Fig. 5.

5. Conclusions and future work

In this paper, we develop a Cross-Feature Interaction Network for 3D human pose estimation, which contains two core modules, namely the cross-feature interaction (CFI) module and the GraMLP module. In our CFI network, local features and global features are first extracted by GCN and MHSA, respectively. Next, the CFI module can facilitate communication among three types of features (initial, local, and global). The GraMLP module, a parallel structure with GCN and MLP, then aggregates the resulting enhanced features in a single layer, generating the final representation of 3D pose. Experimental results on two benchmarks have demonstrated the effectiveness of our method for single-frame 3D pose estimation.

Although our method has achieved promising results, the performance of our model trained on Human3.6M may not generalize well on diverse scenes and actions. This is because the Human3.6M dataset is somehow recorded in relatively homogeneous environments and with limited human actions. In the future, we will explore domain adaptation and use of synthetic poses to improve the generalization of our method.

CRediT authorship contribution statement

Jihua Peng: Writing – original draft, Visualization, Methodology, Investigation, Data curation. **Yanghong Zhou:** Writing – review & editing, Validation, Investigation. **P.Y. Mok:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work described in this paper was supported, in part, by The Hong Kong Polytechnic University (Project codes: P0043858/CD6P, P0051330/BDVH) and the Laboratory for Artificial Intelligence in Design (Project Code: RP1-1) under InnoHK Research Clusters, Hong Kong SAR.

Data availability

I have shared the link in the manuscript.

References

[1] M. Liu, J. Yuan, Recognizing human actions as the evolution of pose estimation maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1159–1168.
[2] M. Svenstrup, S. Tranberg, H.J. Andersen, T. Bak, Pose estimation and adaptive robot behaviour for human-robot interaction, in: 2009 IEEE International Conference on Robotics and Automation, IEEE, 2009, pp. 3571–3576.
[3] N. Hagbi, O. Bergig, J. El-Sana, M. Billinghurst, Shape recognition and pose estimation for mobile augmented reality, IEEE Trans. Vis. Comput. Graphics 17 (10) (2010) 1369–1379.
[4] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2640–2649.
[5] H. Ci, C. Wang, X. Ma, Y. Wang, Optimizing network structure for 3d human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2262–2271.
[6] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, N.M. Thalmann, Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272–2281.
[7] T. Xu, W. Takano, Graph stacked hourglass networks for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16105–16114.

[8] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 483–499.
[9] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.
[10] S. Li, A.B. Chan, 3D human pose estimation from monocular images with deep convolutional neural network, in: Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12, Springer, 2015, pp. 332–347.
[11] S. Park, J. Hwang, N. Kwak, 3D human pose estimation using convolutional neural networks with 2d pose information, in: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, Springer, 2016, pp. 156–169.
[12] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3d human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7307–7316.
[13] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005, vol. 2, IEEE, 2005, pp. 729–734.
[14] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2008) 61–80.
[15] L. Zhao, X. Peng, Y. Tian, M. Kapadia, D.N. Metaxas, Semantic graph convolutional networks for 3d human pose regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3425–3435.
[16] K. Liu, R. Ding, Z. Zou, L. Wang, W. Tang, A comprehensive study of weight sharing in graph networks for 3d human pose estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, Springer, 2020, pp. 318–334.
[17] Z. Zou, W. Tang, Modulated graph convolutional network for 3d human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11477–11487.
[18] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11656–11665.
[19] W. Li, H. Liu, H. Tang, P. Wang, L. Van Gool, Mhformer: Multi-hypothesis transformer for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13147–13156.
[20] H. Li, B. Shi, W. Dai, H. Zheng, B. Wang, Y. Sun, M. Guo, C. Li, J. Zou, H. Xiong, Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (1) 2023, pp. 1296–1304.
[21] Y. Zhu, X. Xu, F. Shen, Y. Ji, L. Gao, H.T. Shen, Posegtac: Graph transformer encoder-decoder with atrous convolution for 3D human pose estimation, in: IJCAI, 2021, pp. 1359–1365.
[22] W. Zhao, W. Wang, Y. Tian, Graformer: Graph-oriented transformer for 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20438–20447.
[23] J. Peng, Y. Zhou, P. Mok, Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1123–1132.
[24] Y. Tao, H. Tao, Z. Zhuang, V. Stojanovic, W. Paszke, Quantized iterative learning control of communication-constrained systems with encoding and decoding mechanism, Trans. Inst. Meas. Control 46 (10) (2024) 1943–1954.
[25] R. Wang, Z. Zhuang, H. Tao, W. Paszke, V. Stojanovic, Q-learning based fault estimation and fault tolerant iterative learning control for MIMO systems, ISA Trans. 142 (2023) 123–135.
[26] X. Song, P. Sun, S. Song, V. Stojanovic, Quantized neural adaptive finite-time preassigned performance control for interconnected nonlinear systems, Neural Comput. Appl. 35 (21) (2023) 15429–15446.
[27] A. Kanazawa, M.J. Black, D.W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7122–7131.
[28] N. Kolotouros, G. Pavlakos, M.J. Black, K. Daniilidis, Learning to reconstruct 3D human pose and shape via model-fitting in the loop, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2252–2261.
[29] B. Biggs, D. Novotny, S. Ehrhardt, H. Joo, B. Graham, A. Vedaldi, 3D multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data, Adv. Neural Inf. Process. Syst. 33 (2020) 20496–20507.
[30] H.-S. Fang, Y. Xu, W. Wang, X. Liu, S.-C. Zhu, Learning pose grammar to encode human body configuration for 3d pose estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, (1) 2018.
[31] X. Zhang, Z. Tang, J. Hou, Y. Hao, 3D human pose estimation via human structure-aware fully connected network, Pattern Recognit. Lett. 125 (2019) 404–410.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
[33] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[34] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, S. Lin, Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 507–523.
[35] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2013) 1325–1339.
[36] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, in: 2017 International Conference on 3D Vision, 3DV, IEEE, 2017, pp. 506–516.
[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, vol. 32, 2019.