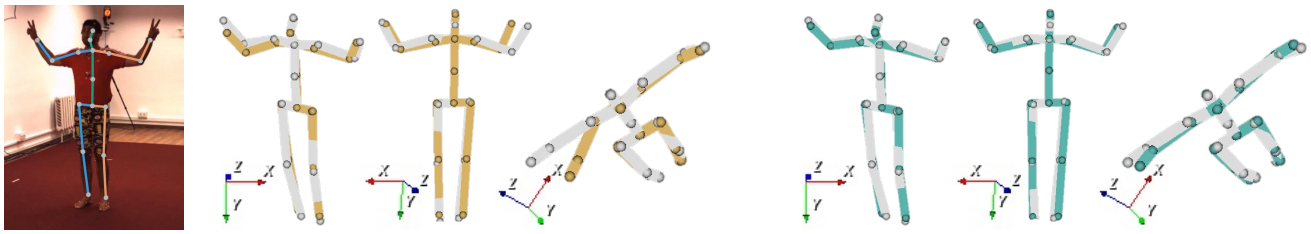
Conditional Directed Graph Convolution for 3D Human Pose Estimation

Wenbo Hu[1,2], Changgong Zhang[2], Fangneng Zhan[3], Lei Zhang[2,4], Tien-Tsin Wong[1,†]
[1] The Chinese University of Hong Kong [2] DAMO Academy, Alibaba Group
[3] Nanyang Technological University [4] The Hong Kong Polytechnic University
{wbhu, ttwong}@cse.cuhk.edu.hk, changgong.zcg@alibaba-inc.com, fnzhan@ntu.edu.sg, cslzhang@comp.polyu.edu.hk



Input 2D pose | Estimated 3D pose by UGCN | Estimated 3D pose by our method

Figure 1: Given a sequence of 2D human poses that are estimated by an off-the-shelf 2D pose estimator, *e.g.*, HR-Net [40], our method can produce more precise 3D poses compared with state-of-the-art approach, UGCN [48]. To better evaluate 3D poses' quality, we show them under three different viewpoints as indicated by the 3D orientation markers. And ground-truth 3D poses are shown in gray as a reference.

ABSTRACT

Graph convolutional networks have significantly improved 3D human pose estimation by representing the human skeleton as an undirected graph. However, this representation fails to reflect the articulated characteristic of human skeletons as the hierarchical orders among the joints are not explicitly presented. In this paper, we propose to represent the human skeleton as a *directed* graph with the joints as nodes and bones as edges that are directed from parent joints to child joints. By so doing, the directions of edges can explicitly reflect the hierarchical relationships among the nodes. Based on this representation, we adopt the spatial-temporal directed graph convolution (ST-DGConv) to extract features from 2D poses represented in a temporal sequence of directed graphs. We further propose a spatial-temporal conditional directed graph convolution (ST-CondDGConv) to leverage varying non-local dependence for different poses by conditioning the graph topology on input poses. Altogether, we form a U-shaped network with ST-DGConv and ST-CondDGConv layers, named *U-shaped Conditional Directed Graph Convolutional Network* (U-CondDGCN), for 3D human pose estimation from monocular videos. To evaluate the effectiveness of our U-CondDGCN, we conducted extensive experiments on two challenging large-scale benchmarks: Human3.6M and MPI-INF-3DHP. Both quantitative and qualitative results show that our method achieves top performance. Also, ablation studies show that directed graphs can better exploit the hierarchy of articulated human skeletons than undirected graphs, and the conditional connections can yield adaptive graph topologies for different kinds of poses.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**.

[†]Corresponding author

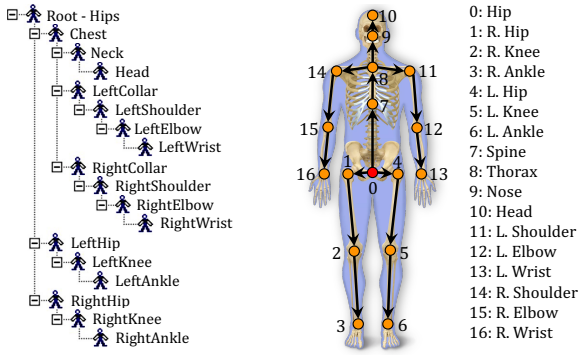
KEYWORDS

3D human pose, conditional directed graph convolution

1 INTRODUCTION

Human pose estimation from monocular videos plays a critical role in a wide spectrum of applications, *e.g.*, action recognition [25, 52], athletic training [47], data-driven computer animation, and gaming. Compared with the 2D pose in image space, the 3D pose in physical space is more informative. However, estimating 3D poses from monocular videos is much more challenging due to the depth ambiguity, metric inconsistency (*i.e.*, millimeters instead of pixels), and the high non-linearity of human dynamics. Given a monocular video of human motions acquired, for example, from consumer-level cameras, our ultimate goal is to estimate the human pose sequence in the 3D physical space. Following [27, 34, 48], we define the 3D pose as the 3D locations of joints, including *"head", "shoulders", "knees", "elbows",* and so no, of the human body.

Thanks to the development of deep learning, we have witnessed remarkable achievements in 3D pose reasoning [2, 6, 7, 17, 23, 24, 27, 34, 39, 45, 48, 50]. State-of-the-art approaches [2, 6, 7, 24, 34, 48] typically divide the task into two stages: 2D pose estimation to localize the keypoints in the image space, and predicting joint positions in the 3D space from 2D pixel coordinates. We follow this strategy and focus on the second stage, lifting 2D pixel coordinates to 3D positions, while the first stage, 2D pose estimation, is also a popular vision task that is quite well-studied in [3, 5, 40]. Recent efforts [2, 48] have explored to represent the human pose as an undirected spatial-temporal graph and thereafter employ graph convolution networks to estimate the 3D pose. Compared with representing human pose as a time sequence of independent joint location vectors [7, 24, 27, 34], the undirected graph representation is more relevant to the inherent nature of human skeletons.



(a) Biovision Hierarchy (BVH) file format

(b) Directed graph for the articulated human skeleton

Figure 2: The BVH file format (a) is a concrete representation of the hierarchical structure of human skeletons. We adopt the directed graph (b) to represent the articulated human skeleton, with the hierarchy is represented by the directions of edges. The "hip" joint (red dot in (b)) is set as the directed graph's root node.

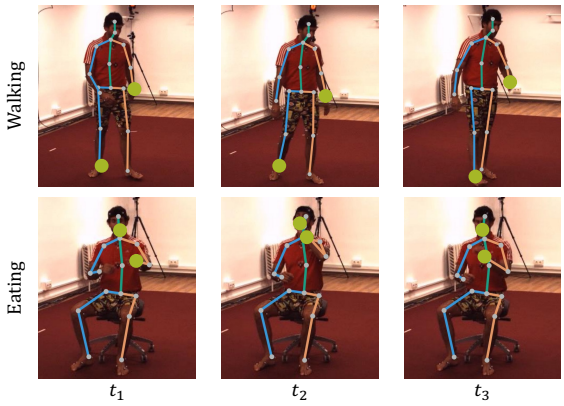


Figure 3: Varying non-local dependence among joints. Dependence between "left hand" and "right foot" (marked in green) is significant when people are walking (first row) since this pose can help to keep balance. While dependence between "left hand" and "head" is high when people are eating (second row). t_1, t_2, and t_3 are time indices.

However, the undirected graph representation does not take into account the hierarchical structure of bones, which is one of the most significant characteristics of human anatomy. For example, when a person moves the shoulder joints, the following joints (*i.e.* elbows and wrists) are moved as well according to the anatomical articulation while not the other way around. This hierarchy is widely utilized to represent human motion in computer animation as evidenced by the Biovision Hierarchy (BVH) file format [*], which presents a typical hierarchical structure as illustrated in Figure 2 (a). To this end, we propose to represent the human skeleton as a *directed* graph with the joints as nodes and bones as edges that are directed from parent joints to child joints, as shown in Figure 2 (b). By doing so, the hierarchical relationships among all the joints are explicitly presented by directions of the edges.

After representing 2D poses as a sequence of directed graphs, we can employ the spatial-temporal directed graph convolution (ST-DGConv) to extract features. However, ST-DGConv shares the same graph topology among all kinds of poses, which may not be optimal since there is non-local dependence among the nodes and the non-local dependence varies a lot for different poses. For example, as shown in Figure 3, the dependence between *"left hand"* and *"right foot"* joints is obviously significant when people are walking (since this pose can help keep balance). In contrast, the dependence between *"hands"* and *"head"* joints would be high when eating. Inspired by the conditional convolution [44, 53] that enables different data samples using different convolution kernels, we propose a *spatial-temporal conditional directed graph convolution* (ST-CondDGConv) to condition the connections of the directed graph on input poses, such that different kinds of poses can adopt appropriate connections to optimally leverage non-local dependence. To the best of our knowledge, this is the first attempt to introduce conditional connections to directed graph convolution for 3D human pose estimation. Overall, we composite a U-shaped network with the ST-DGConv and ST-CondDGConv layers, named *U-shaped Conditional Directed Graph Convolutional Network* (U-CondDGCN), to capture temporal relationship in both short temporal intervals and long temporal ranges.

We evaluated our model on two large-scale 3D human pose estimation benchmarks: Human3.6M [16] and MPI-INF-3DHP [28]. Both quantitative and qualitative experimental results show our method achieves top performance. Moreover, we conducted extensive ablation studies to demonstrate that the directed graph can better utilize the hierarchy of articulated human skeletons, and the conditional connections can yield adaptive graph topologies for different kinds of poses. Our contributions are summarized below.

- We argue that the hierarchy of articulated skeletons is beneficial for 3D pose reasoning, and directed graphs are more suitable to model the hierarchy than undirected graphs.
- We propose a novel conditional directed graph convolution to enable adaptive graph topologies for different pose samples at inference time, such that different poses can benefit from appropriate non-local dependence.
- We present a U-shaped Conditional Directed Graph Convolutional Network (U-CondDGCN) for 3D human pose estimation from detected 2D keypoints. Our method achieves top performance on the challenging Human3.6M and MPI-INF-3DHP benchmarks, thus demonstrating its effectiveness.

2 RELATED WORK

2.1 3D Human Pose Estimation

With the development of deep learning, 2D human pose estimation [3, 5, 30, 40] has shown remarkable progress, while 3D pose estimation remains more challenging due to the depth ambiguity. Several methods [13, 18, 21, 35, 45] propose to relieve this issue by adopting multi-view images/videos as input. However, multi-view observations are expensive to obtain in daily life scenarios. Thus, 3D human pose estimation from monocular images/videos is highly demanded. Some works explored to directly infer 3D human pose from monocular images/videos with end-to-end deep neural networks [33, 42]. This end-to-end idea is elegant and free of accumulated errors. However, we need paired data (images/videos

[*] https://en.wikipedia.org/wiki/Biovision_Hierarchy

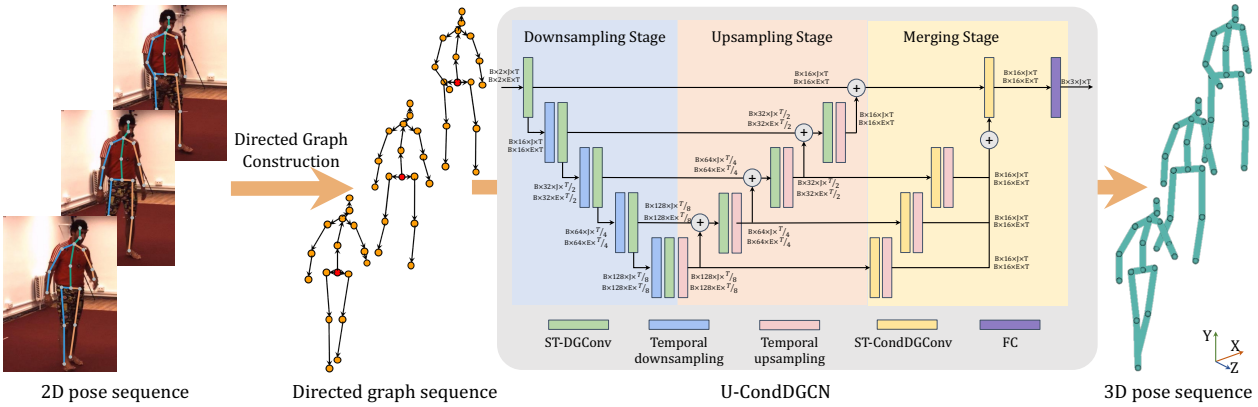


Figure 4: Overview of our framework. Given a sequence of 2D poses estimated by any off-the-shelf 2D pose estimators, we first construct a sequence of directed graphs and then estimate the 3D poses with our U-shaped conditional directed graph convolutional network (U-CondDGCN).

and corresponding 3D poses) to supervise the training, while manually labeling 3D human poses for images/videos is impractical. To this end, Martinez *et al.* [27] proposed to divide 3D human pose estimation into 2D keypoint detection followed by lifting 2D joint locations to 3D positions, such that the first stage can be trained with manually labeled 2D poses and the training data for the second stage can be obtained by projecting 3D poses, obtained from motion capture devices, to 2D space. This divide-and-conquer strategy benefits from intermediate supervision and outperforms the end-to-end counterparts. A family of approaches [4, 7, 15, 34, 48] followed this strategy and focused more on lifting 2D to 3D. Our method can also be categorized into this group.

Recently, Hossain *et al.* [15] proposed to employ LSTM [14] to leverage temporal information for lifting 2D joint locations to 3D positions. Dilated temporal convolutions [34] and temporal attention [24] are further explored for better temporal information aggregation. Besides, some works focus on specific issues in the 3D human pose estimation, *e.g.*, tackling occlusion problems [6, 7], relieving the unreliable input issue by kinematics analysis [50], or addressing the lack of 3D annotations in the wild and overfitting issues by weakly-supervised learning [17, 34, 46]. Also, Lin *et al.* [22] proposed to predict the 3D human pose in trajectory space by factoring the pose sequence into a trajectory base matrix and a coefficient matrix. The above methods represent human poses as a temporal sequence of independent joint location vectors. However, this representation cannot fully express the dependence among highly correlated human joints.

2.2 Graph Convolution Network (GCN)

GCNs [8, 12, 36] generalize conventional convolution operators to graphs, and can be roughly categorized into spectral [8, 19, 20] and spatial [1, 10, 12] perspective GCNs. Our proposed U-CondDGCN is more related to the latter one. Interested readers are referred to [49] for a complete survey of GCNs.

By representing human skeletons as graphs, GCNs have significantly improved a series of human-related reasoning tasks, *e.g.*, action recognition [37, 38, 52], action synthesis [51] and pose tracking [31]. Moreover, several works [2, 23, 48, 54, 55] adopt undirected graphs to represent human skeletons and apply GCN to utilize the prior knowledge of human skeleton. However, the undirected graph representation fails to reflect the articulated characteristic of human skeletons as the hierarchical orders among joints are not explicitly presented. In contrast, directed graph representation explicitly models hierarchical relationships among the nodes by the directions of edges. Recently, Shi *et al.* [37] also employed the directed graph representation and proposed a directed graph neural network for action recognition. Their method learns the graph topology from the training data rather than simply defining it based on the natural structure of human skeletons. However, the graph topology is fixed at inference time, which means different data samples still share the same topology. Differently, our conditional directed graph convolution allows different poses to benefit from appropriate non-local dependence at both training and inference time, by conditioning the graph topology on input poses. This mechanism is crucial for 3D human pose estimation since the optimal non-local dependence varies a lot for different poses, as shown in Figure 3.

3 APPROACH

Our full pipeline is illustrated in Figure 4. We first construct a temporal sequence of directed graphs from a sequence of human poses in the 2D image space $P_{2D} = \{\boldsymbol{X}_{t,j} \in \mathbb{R}^2 \mid t = 1, 2, 3, ..., T;\ j = 1, 2, 3, ..., J\}$, where T and J denote the number of frames in the sequence and joints on the human skeleton, respectively. The input 2D pose sequence P_{2D} can be estimated from monocular videos by any off-the-shelf 2D pose estimators, *e.g.*, CPN [5], HR-Net [40], or OpenPose [3]. The nodes in the directed graph represent major joints of the human skeleton, while the edges represent the bones among the joints, as shown in Figure 2 (b). We set the directions of edges following the convention definition in the BVH file format. And the "hip" joint, marked as the red dot in Figure 2 (b), is set as the root node since it is the gravity center of the human body. The features associated with nodes and edges are initialized as the joints' locations and their first-order derivatives (the difference between the child joint and parent joint), respectively. Formally, a temporal sequence of directed graphs can be formulated as $\mathcal{G}_{2D} = \{G_t = (\mathcal{N}, \mathcal{E}) \mid t = 1, 2, 3, ..., T\}$, where $\mathcal{N}$ is the set of nodes, and $\mathcal{E}$ is the set of directed edges. We then apply our *U-shaped Conditional Directed Graph Convolutional Network* (U-CondDGCN) to estimate the pose sequence in the 3D physical space $P_{3D} = \{\boldsymbol{X}_{t,j} \in \mathbb{R}^3 \mid t = 1, 2, 3, ..., T;\ j = 1, 2, 3, ..., J\}$.

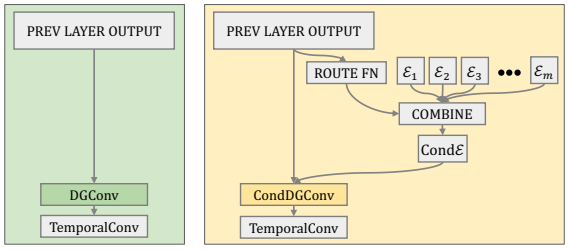
3.1 Network Blocks

After representing the 2D human pose as a sequence of directed graphs, the problem now lies in how to extract features from them to estimate the 3D pose. To aggregate features both spatially and temporally, we use five types of blocks in our network, *i.e.*, spatial-temporal directed graph convolution (ST-DGConv), spatial-temporal conditional directed graph convolution (ST-CondDGConv), temporal downsampling, temporal upsampling, and FC blocks, as shown in Figure 4.

ST-DGConv. The ST-DGConv block consists of a directed graph convolution (DGConv) followed by a temporal convolution, as shown in Figure 5 (a). DGConv exploits the spatial relationship by aggregating features from neighboring edges or nodes. Details of DGConv are to be presented in Section 3.2. On the other hand, to take advantage of the temporal relationship, we employ a temporal convolution, which is a conventional 1D convolution, since the temporal sequence of directed graphs has a regular grid structure along the temporal dimension.

ST-CondDGConv. The aforementioned ST-DGConv is based on a fixed directed graph connection $\mathcal{E}$, which is defined according to the natural structure of human skeletons. However, the predefined connections can only utilize the local dependence among hierarchically neighboring joints. More importantly, sharing the same connections among all kinds of poses may not be optimal since the non-local dependence varies a lot for different poses, as shown in Figure 3. Inspired by the conditional convolution [44, 53] that enables different data samples using different convolution kernels, we propose a spatial-temporal conditional directed graph convolution (ST-CondDGConv) to condition the connections of the directed graph on input poses, such that different poses can adopt appropriate connections to exploit varied non-local dependence. As shown in Figure 5 (b), the ST-CondDGConv predicts the conditional connections Cond$\mathcal{E}$ from previous layer's output. Specifically, there are a series of trainable connection matrix bases $\{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, ..., \mathcal{E}_m\}$, where $\mathcal{E}_i \in \mathbb{R}^{J\times J}$ and m is the number of bases (set to 16 in our experiments), for the network to learn. To encourage the sparsity of the connection matrix bases, we employ the sparse initialization [26] before the network training. We use a routing function to predict the blending weights for the connection matrix bases from the previous layer's output. The network structure of the routing function is the same as that used in conditional convolution [53], which is a global average pooling layer followed by a fully connected layer and a Sigmoid activation. After that, we linearly combine the bases by the predicted blending weights to produce the conditional connections Cond$\mathcal{E}$, which are then fed into the CondDGConv to aggregate the spatial information both locally and non-locally. Note that non-local connections here are conditioned on the previous layer's output. Thus, the routing function is able to differentiate input samples to allow different kinds of poses to leverage appropriate non-local dependence at both training and inference time. Finally, a temporal convolution is employed to aggregate the temporal information.

Besides the above two major blocks, the other three types of blocks can be easily derived. The temporal downsampling block is the ST-DGConv with the inside temporal convolution's stride setting to two. It is used to downsample the temporal resolution for the larger receptive field. The temporal upsampling block is the conventional bilinear interpolation along the temporal axis to recover higher temporal resolution. The FC block is the standard fully-connected layer to predict the final 3D pose from the extracted features on directed graphs.



(a) ST-DGConv (b) ST-CondDGConv

Figure 5: Spatial-temporal directed graph convolution (ST-DGConv) and spatial-temporal conditional directed graph convolution (ST-CondDGConv) blocks.

3.2 Conditional Directed Graph Convolution

In this section, we first formulate our conditional directed graph convolution (CondDGConv), which is the key operator in the ST-CondDGConv block. The directed graph convolution (DGConv), the key operator in ST-DGConv, can be easily derived from it. As shown in Figure 6 (a), the input directed graph for CondDGConv has the predefined connections $\mathcal{E} = \{e_i \mid i = 1, 2, ..., E\}$ and the predicted conditional connections Cond$\mathcal{E} = \{\text{Cond}e_i \mid i = 1, 2, ..., C\}$, where E and C are the number of predefined connections and conditional connections, respectively. The predefined connections and conditional connections are shown as arrows and blue dash-line arrows in Figure 6, respectively. The CondDGConv can be formulated with three steps, as shown in Figure 6 (b), (c), and (d).

Nodes updating. Connections of the directed graph used in this step are the predefined connections, $\mathcal{E} = \{e_i \mid i = 1, 2, ..., E\}$. For each node n_i, we have the set of incoming edges that are heading into it, $\mathcal{E}_i^-$, and the set of outgoing edges that are heading out from it, $\mathcal{E}_i^+$. Following the idea of conventional convolution, aggregating neighboring features, the nodes updating is defined as the aggregation of its incoming edge set, its outgoing edge set, and itself. However, the number of elements in the outgoing edge set is varying. For example, as shown in Figure 6 (b), the incoming edge set of node n_4 is $\mathcal{E}_4^- = \{e_3\}$ while the outgoing edge set is $\mathcal{E}_4^+ = \{e_4, e_5\}$. Therefore, we employ a pooling function to summarize features of the outgoing edge set. Mathematically, the nodes updating step can be formulated as:

$$f(n_i') = \sigma(\mathbf{w} \cdot [f(\mathcal{E}_i^-);\ f(n_i);\ \mathcal{P}(f(\mathcal{E}_i^+))]^{\mathrm{T}} + \mathbf{b}), \tag{1}$$

where $f(\cdot)$ is the mapping from nodes/edges to their associated features; $\mathcal{P}(\cdot)$ is the pooling function (average pooling is adopted in our implementation); $\mathbf{w}$ and $\mathbf{b}$ are the trainable parameters that are similar with the kernel and bias of conventional convolution, respectively; and σ is the activation function (ReLU is used in our implementation).

Nodes updating with conditional connections. This step is designed to address the varying non-local dependence for different poses. Thus, connections of the directed graph used in this step are the predicted conditional connections, Cond$\mathcal{E} = \{\text{Cond}e_i \mid i = 1, 2, ..., C\}$. For each node n_i', we have the parent nodes set $\mathcal{N}_i^{\mathrm{p}}$ that

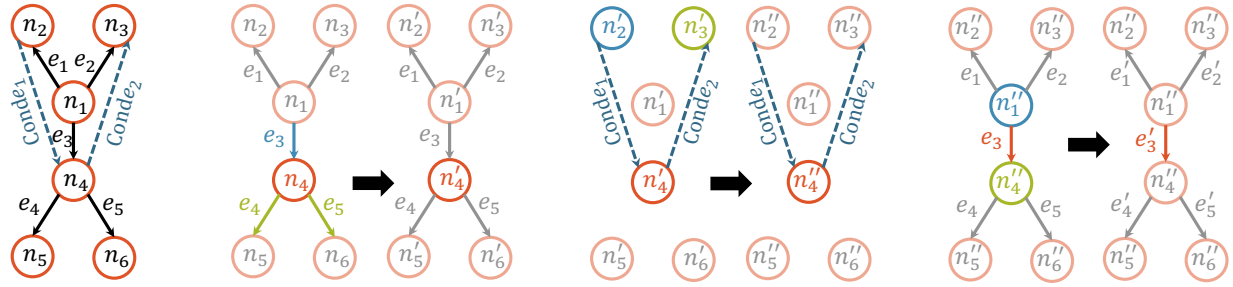


Figure 6: Conditional directed graph convolution (CondDGConv). Part (a) is the input directed graph with predefined connections (e_i) and predicted conditional connections (Conde_i). Part (b), (c) and (d) are three steps of the CondDGConv.

have edges directed at n'_i and the child nodes set $\mathcal{N}_i^{\mathrm{c}}$ that have edges directed from n'_i. For example, as shown in Figure 6 (c), node n'_4 has the parent nodes set $\mathcal{N}_4^{\mathrm{p}} = \{n'_2\}$ and the child nodes set $\mathcal{N}_4^{\mathrm{c}} = \{n'_3\}$. Similarly, the nodes updating with conditional connections step can be formulated as:

$$f(n''_i) = \sigma(\mathbf{w} \cdot [\mathcal{P}(f(\mathcal{N}_i^p));\ f(n'_i);\ \mathcal{P}(f(\mathcal{N}_i^c))]^{\mathrm{T}} + \mathbf{b}). \quad (2)$$

Edges updating. Connections of the directed graph used in this step are the predefined connections, $\mathcal{E} = \{e_i \mid i = 1, 2, ..., E\}$. For each edge e_i, we have the source node $\mathcal{N}_i^s$ and the target node $\mathcal{N}_i^t$, *e.g.*, as shown in Figure 6 (d), the source node and target node of edge e_3 are n''_1 and n''_4, respectively. Again, we can formulate the edges updating step as:

$$f(e'_i) = \sigma(\mathbf{w} \cdot [f(\mathcal{N}_i^s);\ f(e_i);\ f(\mathcal{N}_i^t)]^{\mathrm{T}} + \mathbf{b}). \quad (3)$$

After these three steps, features associated with both nodes and edges are updated. Note that even the edges updating step adopts the predefined connections; the non-local dependence can also be aggregated into the edges' features since they are updated by the source and target nodes' features that have aggregated the non-local information. Removing step (ii), nodes updating with conditional connections, can yield the directed graph convolution (DGConv), which is purely based on the predefined connections. Thus, it can utilize the prior knowledge of the natural structure of human skeletons but fails to leverage the varying non-local dependence for different poses.

3.3 Full Network

We construct the full network as a U-shaped conditional directed graph convolutional network (UCondDGCN) using the aforementioned blocks to capture temporal relationships in both short temporal intervals and long temporal ranges. As shown in the middle part of Figure 4, the UCondDGCN has three stages: i) downsampling stage to aggregate information at long-time ranges by temporal pooling; ii) upsampling stage to recover the temporal resolution, and there are skip connections between the downsampling and upsampling stages to integrate the low-level details; iii) merging stage to combine multi-scale feature maps to predict the final 3D poses. The numbers shown in the figure indicate the shape of features in the network, where B is the batch size, J is the number of nodes, E is the number of edges, and T is the sequence length. As discussed in Section 3.2, the ST-DGConv can better utilize the prior knowledge of the natural structure of human skeletons, whereas ST-CondDGConv allows different kinds of poses to adopt appropriate non-local dependence. To balance the stability and flexibility, we adopt ST-DGConv in the downsampling and upsampling stages while adopting ST-CondDGConv in the merging stage. This configuration is further justified with ablation study as shown in Table 4. Overall, the input of our UCondDGCN is the 2D human poses represented in directed graphs, and the output is the corresponding 3D poses, as shown in Figure 4.

4 EXPERIMENTS

We formulate the loss function similar to UGCN [48],

$$\mathcal{L} = \mathcal{L}_p + \lambda \mathcal{L}_m, \quad (4)$$

where $\mathcal{L}_p$ is 3D joint position loss that is known as the mean per-joint position error (MPJPE), $\mathcal{L}_m$ is motion loss [48], and λ is a weight to balance them (set to 0.1 in our implementation). The motion loss is originally proposed in UGCN [48] to supervise the temporal structure of the predicted pose sequence. It first encodes the positions of the same joint at two different temporal instants into pairwise motion encodings for the predicted pose sequence and the ground truth, respectively; and then computes the L1 loss between them. The number of layers, kernel size, and input sequence length are also followed UGCN [48].

We implemented UCondDGCN on the PyTorch platform [32] and conducted experiments on a single NVIDIA TITAN V GPU. We optimized the model by the AdaMod optimizer [9] for 110 epochs with a batch size of 256, in which the learning rate was initially set to 5×10^{-3} and decayed by 0.1 after the 80^{th}, 90^{th}, and 100^{th} epochs. To avoid over-fitting, we set the weight decay factor to 10^{-5} and the dropout rate to 0.3. We followed UGCN [48] to adopt the sliding window algorithm with a step length of five frames to estimate variable sequence length at inference time.

4.1 Dataset and Evaluation Metrics

Human3.6M. Human3.6M [16] is the most widely used evaluation benchmark, containing 3.6 million video frames captured from four synchronized cameras with different locations and poses at 50 Hz. There are 11 subjects performing 15 kinds of actions, *e.g.*, "walking", "sitting", and "eating". Following previous works [7, 27, 33, 34, 41, 43, 48], we adopted the 17-joint pose, trained a single model on five subjects (S1, S5, S6, S7, S8) for all kinds of actions, and tested it on the remaining two subjects (S9 and S11).

Table 1: Quantitative comparisons with state-of-the-art methods on Human3.6M under protocol #1 and protocol #2, where methods marked with † are video-based; T denotes the number of input frames; and CPN and HR-Net denote the input 2D poses are estimated by [5] and [40], respectively. The best and second-best results are marked in bold and underlined, respectively.

Protocol #1		Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [27]	(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
Fang *et al.* [11]	(AAAI'18)	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Zhao *et al.* [54]	(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
Liu *et al.* [23]	(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
Cai *et al.* [2]† (CPN, T=7)	(ICCV'19)	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Pavllo *et al.* [34]† (CPN, T=243)	(CVPR'19)	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Xu *et al.* [50]† (CPN, T=9)	(CVPR'20)	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6
Liu *et al.* [24]†(CPN, T=243)	(CVPR'20)	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
UGCN [48]† (CPN, T=96)	(ECCV'20)	41.3	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
UGCN [48]† (HR-Net, T=96)	(ECCV'20)	38.2	**41.0**	45.9	39.7	**41.4**	51.4	41.6	41.4	52.0	57.4	41.8	44.4	41.6	33.1	30.0	42.6
Ours† (CPN, T=96)		38.0	43.3	39.1	39.4	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Ours† (HR-Net, T=96)		**35.5**	41.3	**36.6**	**39.1**	42.4	**49.0**	**39.9**	**37.0**	**51.9**	63.3	**40.9**	**41.3**	**40.3**	**29.8**	**28.9**	**41.1**
Protocol #2		Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [27]	(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
Fang *et al.* [11]	(AAAI'18)	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Liu *et al.* [23]	(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.* [2]† (CPN, T=7)	(ICCV'19)	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Pavllo *et al.* [34]† (CPN, T=243)	(CVPR'19)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Xu *et al.* [50]† (CPN, T=9)	(CVPR'20)	31.0	34.8	34.7	34.4	36.2	43.9	31.6	33.5	42.3	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu *et al.* [24]†(CPN, T=243)	(CVPR'20)	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
UGCN [48]† (CPN, T=96)	(ECCV'20)	32.9	35.2	35.6	34.4	36.4	42.7	31.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
UGCN [48]† (HR-Net, T=96)	(ECCV'20)	28.4	**32.5**	34.4	32.3	**32.5**	40.9	30.4	29.3	42.6	**45.2**	33.0	32.0	33.2	24.2	22.9	32.7
Ours† (CPN, T=96)		29.8	34.4	31.9	31.5	35.1	40.0	30.3	30.8	42.6	49.0	35.9	31.8	35.0	25.7	23.6	33.8
Ours† (HR-Net, T=96)		**27.7**	32.7	**29.4**	**31.3**	**32.5**	**37.2**	**29.3**	**28.5**	**39.2**	50.9	**32.9**	**31.4**	**32.1**	**23.6**	**22.8**	**32.1**

Table 2: Results on Human3.6M with ground-truth 2D poses as input. Video-based methods are marked with †.

Method		MPJPE
Martinez *et al.* [27]	(ICCV'17)	45.5
Zhao *et al.* [54]	(CVPR'19)	43.8
Liu *et al.* [23]	(ECCV'20)	37.8
Cai *et al.* [2]†	(ICCV'19)	37.2
Pavllo *et al.* [34]†	(CVPR'19)	37.2
Liu *et al.* [24]†	(CVPR'20)	34.7
UGCN [48]†	(ECCV'20)	25.6
Ours†		**22.7**

Table 3: Results on MPI-INF-3DHP with three metrics, where PCK and AUC are the higher, the better, while MPJPE is the lower, the better.

Method	PCK[↑]	AUC[↑]	MPJPE[↓]
Mehta *et al.* [28]	75.7	39.3	-
VNect (ResNet50) [29]	77.8	41.0	-
VNect (ResNet101) [29]	79.4	41.6	-
TrajectoryPose3D [22]	83.6	51.4	79.8
UGCN [48]	86.9	62.1	68.1
Ours	**97.9**	**69.5**	**42.5**

MPI-INF-3DHP. MPI-INF-3DHP [28] is a relatively new dataset captured in both indoor and outdoor environments. Similar to Human3.6M, it contains various subjects, actions, and camera settings, and we followed [22, 29, 48] to split the training and testing set.

Evaluation Metrics. For Human3.6M, we adopt the most widely used two metrics: *Protocol 1* is the mean per-joint position error (MPJPE) that is the mean Euclidean distance between the estimated joint positions and ground truth in millimeters; and *Protocol 2* is the error after alignment with the ground truth in translation, rotation, and scale (P-MPJPE). For MPI-INF-3DHP, we also report the percentage of correct keypoints (PCK) [28] score with the threshold of 150*mm* and the area under the curve (AUC) [28] of the PCK scores with different error thresholds, following [22, 28, 29, 48].

4.2 Quantitative Evaluation

Results on Human3.6M. To evaluate the effectiveness of our U-CondDGCN, we first quantitatively compared our method with state-of-the-art methods on the Human3.6M benchmark in Table 1, including image-based methods [11, 23, 27, 54] and video-based methods [2, 24, 34, 48, 50] (marked with †). The input 2D poses are estimated from monocular images/videos by either CPN [5] or HR-Net [40] (a more powerful 2D pose estimator). We can see that video-based methods generally perform better than image-based methods, which justifies that the temporal information is beneficial to the 3D human pose estimation. More importantly, from both Protocol #1 and Protocol #2 results, we can see our method consistently outperforms all the others by a large margin no matter with CPN or HR-Net input 2D poses (*1.7mm* and *1.5mm* improvements in terms of Protocol #1, respectively). This demonstrates the effectiveness of our proposed conditional directed graph convolution.

To further explore the upper bound of our U-CondDGCN for lifting 2D poses to 3D poses, we compared our method with several state-of-the-art methods with ground-truth 2D poses as input since doing so can eliminate the influence of the 2D pose estimators applied. As shown in Table 2, we can see our method significantly outperforms all the others ($\geq 2.9mm$) in terms of MPJPE. It demonstrates if a more powerful 2D pose estimator is available, our U-CondDGCN is able to produce more accurate 3D poses.

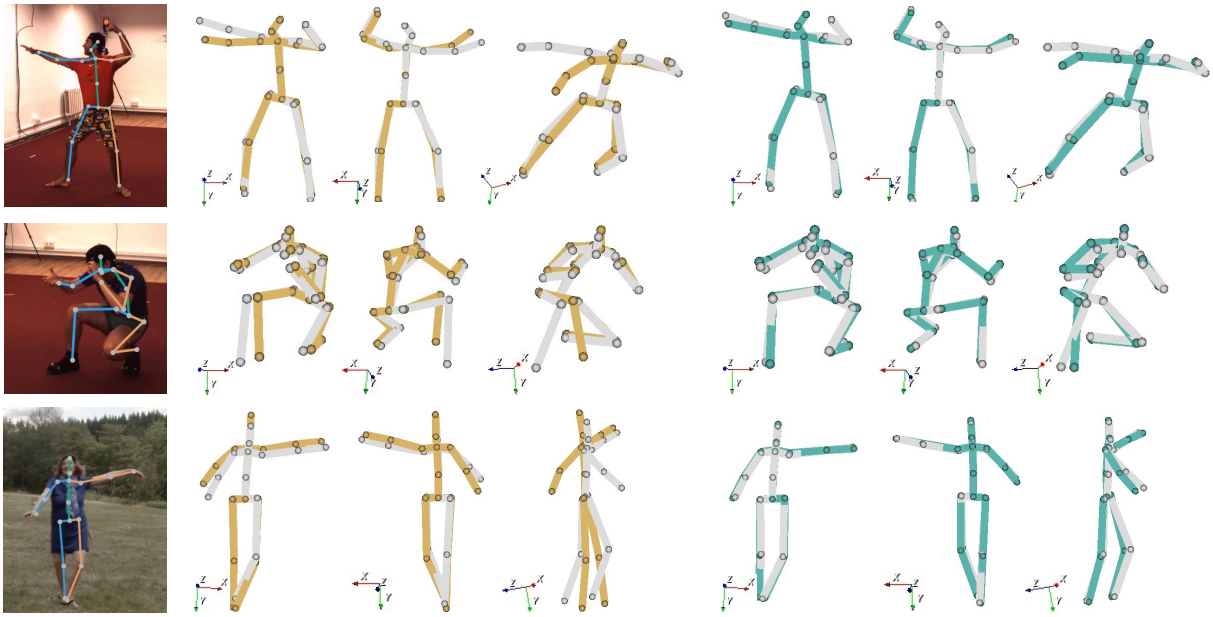


Figure 7: Qualitative comparison between and the baseline, UGCN [48] and our method on Human3.6M and MPI-INF-3DHP. To better evaluate 3D poses' quality, we show them under three different viewpoints as indicated by 3D orientation markers. And ground-truth 3D poses are shown in gray as a reference.

Results on MPI-INF-3DHP. To evaluate the generalization ability, we compared our method with state-of-the-art methods on the MPI-INF-3DHP benchmark, as shown in Table 3. We followed the experimental setting in [22, 48] to adopt ground-truth 2D poses as input. Our method achieves significant improvements no matter in terms of PCK (11.0% improvement), AUC (7.4% improvement), or MPJPE (25.6*mm* improvement), which demonstrates our method generalizes well on various datasets.

4.3 Qualitative Evaluation

To qualitatively evaluate our U-CondDGCN, we compared it against the baseline, UGCN [48]. As shown in Figure 7, we show several example results of UGCN and ours under various viewpoints in orange and teal, respectively, while the ground truth 3D poses are shown in neutral gray as a reference. The first two examples are from Human3.6M, and the last example is from MPI-INF-3DHP. The input 2D poses for Human3.6M are estimated from monocular videos by the 2D pose estimator, HR-Net [40], while that for MPI-INF-3DHP is the ground-truth 2D poses. We can see our results are more consistent with the ground-truth 3D poses, especially for the end-effectors, *e.g.*, the "wrist" and "ankle" joints. It qualitatively demonstrates the effectiveness of our U-CondDCCN for estimating 3D poses. Readers are highly recommended to watch the supplementary video to better explore the performance.

4.4 Analysis

Effectiveness of CondDGConv. To explore the effectiveness of directed graph representation and our proposed CondDGConv, we conducted ablation experiments on the Human3.6M dataset by considering the following methods:

- UGCN [48] is the baseline, which adopts the undirected graph representation and has the same U-shaped structure;
- UDGCN is a variant of our method that replaces all the CondDGConv with the DGConv; and
- U-CondDGCN is the full version of our method.

We adopted the ground-truth 2D poses as input to eliminate the influence of the 2D pose estimator. The results are shown in the top part of Table 4. Comparing the results of UGCN and UDGCN, we can see that UDGCN outperforms the UGCN by 1.7*mm* in terms of MPJPE. It shows the directed graph can better model the hierarchy of the articulated human skeleton, and the hierarchy is beneficial to the 3D pose reasoning. Moreover, comparing the results of UDGCN and U-CondDGCN, we can find CondDGConv can further improve the performance by 1.2*mm*. It demonstrates adaptive graph topologies for different pose samples can better leverage the varying non-local dependence.

Table 4: Ablation study. We compared results of the baseline (UGCN [48]), the variant of our method (UDGCN), our U-CondDGCN, and different configurations for our U-CondDGCN on Human3.6M. The Δ denotes the improvements compared with the baseline.

Method	Cond$\mathcal{E}$	MPJPE	Δ
UGCN	-	25.6	-
UDGCN	-	23.9	1.7
U-CondDGCN	Merging stage	**22.7**	**2.9**
U-CondDGCN	Upsampling stage	24.1	1.5
U-CondDGCN	Downsampling stage	25.3	0.3
U-CondDGCN	All stages	23.6	2.0

Configuration of U-CondDGCN. To explore the best stage to utilize the conditional connections (Cond$\mathcal{E}$), we conducted experiments by adopting Cond$\mathcal{E}$ at various stages, *i.e.*, downsampling

Table 5: Anatomy accuracy of the baseline (UGCN) and our method in terms of bone length error (in mm), bone direction error (in degree), and left-right symmetric error (in mm).

Method	Length error	Direction error	Symmetric error
UGCN	12.0	7.5	6.1
Ours	**11.1**	**7.2**	**4.9**

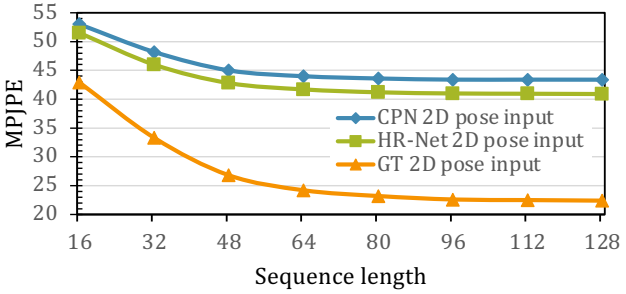
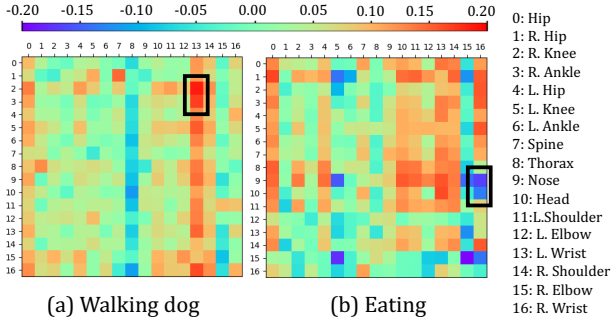


Figure 8: Influence of the input sequence length. Curves show the quality of estimated 3D poses against the input sequence length with the input 2D poses from CPN [5], HR-Net [40], and ground-truth, respectively.

stage, upsampling stage, merging stage, and all stages, and show the results in the bottom part of Table 4. We can see that adopting Cond$\mathcal{E}$ at the downsampling or upsampling stage would decrease the performance compared with the UDGCN, whereas adopting the Cond$\mathcal{E}$ at the merging stage can improve the performance. This is because introducing Cond$\mathcal{E}$ at too early stages would prevent the network from utilizing the prior knowledge of the natural structure of human skeletons while introducing Cond$\mathcal{E}$ at the late stage, *i.e.*, the merging stage, can both keep the prior knowledge and allow different kinds of poses to adopt optimal non-local dependence. Compared with adopting Cond$\mathcal{E}$ at all the stages (yields 2.0*mm* improvement), adopting it at the merging stage can better balance the stability and flexibility (yields 2.9*mm* improvement).

Influence of the input sequence length. To explore it, we measured the MPJPE of estimated 3D poses under various lengths of the input sequence with the input 2D poses from CPN [5], HR-Net [40], and ground-truth, and plotted curves of the MPJPE against the input sequence length in Figure 8. We can see that no matter with what kinds of input 2D poses, the performance always goes better when the sequence length increases. It shows longer sequence can provide more temporal information for the 3D pose reasoning.

Anatomy accuracy. The evaluation metrics above only consider the accuracy of joint positions. However, the anatomy accuracy is also important for human poses, like the lengths of bones, the directions of bones, and the symmetric accuracy between the left and right bone lengths. Therefore, we compared our method against the UGCN [48] on Human3.6M to analyze the anatomy accuracy. We fed the 2D poses estimated by HR-Net [40] to our method and UGCN and measured the bone length error, bone direction error, and the left-right symmetric error from the estimated 3D poses. As shown in Table 5, we can see our method achieves lower errors on all three aspects than UGCN, even though no loss terms are added to explicitly constrain them.



(a) Walking dog (b) Eating

Figure 9: Visualization of the predicted conditional connection matrices from two kinds of actions. Two significant dependency examples are marked in black rectangles.

Visualization of predicted conditional connections. To verify the predicted conditional connections for different input poses, we visualized the connection matrices (from the last layer of U-CondDGCN) predicted for two pose sequences in the "walking dog" and "eating" action classes, respectively. The connection matrix Cond$\mathcal{E}$ has a shape of $J \times J$, where J is the number of nodes. The absolute value of entry Cond$\mathcal{E}_{i,j}$ means the extent of dependence between node i and j, while the positive/negative sign stands for node i belongs to the child/parent set of node j, where i and j are the row and column indices, respectively. As shown in Figure 9, we can see these two connection matrices are very different. For example, the dependency between the left wrist and the right knee/ankle nodes is significant for the "walking dog" sequence, whereas the dependency between the right wrist and the nose/head nodes is significant for the "eating" sequence, as marked with black rectangles. It demonstrates our U-CondDGCN can differentiate input poses to predict appropriate non-local connections at inference time.

Inference speed. The number of parameters of our method is 3.42M, while that of UGCN [48] is 1.69M. The difference mainly comes from that conventional GCN only considers nodes' features while our CondDGConv considers both nodes' and edges' features. However, such a number of parameters is still far less than that of temporal-convolution-based methods such as VideoPose3D [34], which has 16.95M parameters. Under the experimental setting stated in Sec. 4, our method takes around 3.6ms to estimate a 3D pose from the 2D poses.

5 CONCLUSION

We present the conditional directed graph convolution for 3D human pose estimation from monocular videos. The employed directed graph representation can better model the articulated hierarchy of human skeletons than the undirected graph. Moreover, we present a novel conditional connection mechanism for the directed graph convolution, such that different kinds of poses can adopt appropriate non-local dependence to facilitate the 3D pose reasoning. Extensive quantitative and qualitative evaluations demonstrate that our method achieves top performance on two large-scale challenging benchmarks, Human3.6M and MPI-INF-3DHP. Also, we conducted a wide spectrum of analyses to verify our method. We believe the insight behind the conditional directed graph convolution can also benefit other tasks where the articulated structure is involved, like the 3D hand gesture estimation and recognition.

REFERENCES

[1] James Atwood and Don Towsley. 2016. Diffusion-convolutional neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 43, 1 (2019), 172–186.

[4] Ching-Hang Chen and Deva Ramanan. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 2020. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *AAAI Conference on Artificial Intelligence*.

[7] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[8] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[9] Jianbang Ding, Xuancheng Ren, Ruixuan Luo, and Xu Sun. 2019. An Adaptive and Momental Bound Method for Stochastic Learning. *arXiv preprint arXiv:1910.12249* (2019).

[10] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[11] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI Conference on Artificial Intelligence*.

[12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*.

[13] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. 2020. Epipolar transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Mir Rayat Imtiaz Hossain and James J Little. 2018. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*.

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 36, 7 (2013), 1325–1339.

[17] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. 2020. Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable triangulation of human pose. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[19] Ron Levie, Federico Monti, Xavier Bresson, and Michael M Bronstein. 2018. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing* 67, 1 (2018), 97–109.

[20] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018. Adaptive graph convolutional neural networks. In *AAAI Conference on Artificial Intelligence*.

[21] Junbang Liang and Ming C Lin. 2019. Shape-aware human pose and shape reconstruction using multi-view images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[22] Jiahao Lin and Gim Hee Lee. 2019. Trajectory Space Factorization for Deep Video-Based 3D Human Pose Estimation. In *British Machine Vision Conference (BMVC)*.

[23] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. 2020. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*.

[24] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. 2020. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Diogo C Luvizon, David Picard, and Hedi Tabia. 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] James Martens. 2010. Deep learning via hessian-free optimization.. In *International Conference on Machine Learning (ICML)*.

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[28] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*.

[29] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (SIGGRAPH)* 36, 4 (2017), 1–14.

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*.

[31] Guanghan Ning, Jian Pei, and Heng Huang. 2020. Lighttrack: A generic framework for online top-down human pose tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[33] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[34] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. 2019. Cross view fusion for 3d human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-based action recognition with directed graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. MotioNet: 3D Human Motion Reconstruction from Monocular Video with Skeleton Consistency. *ACM Transactions on Graphics (SIGGRAPH Asia)* 40, 1 (2020), 1–15.

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[41] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[42] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. 2016. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference (BMVC)*.

[43] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. 2016. Direct prediction of 3d body poses from motion compensated sequences. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[44] Zhi Tian, Chunhua Shen, and Hao Chen. 2020. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision (ECCV)*.

[45] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. 2020. VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment. In *European Conference on Computer Vision (ECCV)*.

[46] Bastian Wandt and Bodo Rosenhahn. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019. AI coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *ACM International Conference on Multimedia (MM)*.

[48] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2020. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision (ECCV)*.

[49] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[50] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. 2020. Deep kinematics analysis for monocular 3d human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[51] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

[52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*.

[53] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. 2019. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[54] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. 2019. Semantic graph convolutional networks for 3d human pose regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[55] Zhiming Zou, Kenkun Liu, Le Wang, and Wei Tang. 2020. High-order Graph Convolutional Networks for 3D Human Pose Estimation. In *British Machine Vision Conference (BMVC)*.