IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 9, 2021, accepted February 14, 2021, date of publication February 26, 2021, date of current version March 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062426

3D Human Pose Estimation With Spatial Structure Information

XIAOSHAN HUANG, JUN HUANG, AND ZENGMING TANG
Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China

Corresponding author: Jun Huang (huangj@sari.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0192800, and in part by the NSFC under Grant 61972007.

ABSTRACT Estimating 3D human poses from 2D poses is a challenging problem due to joints self-occlusion, weak generalization, and inherent ambiguity of recovering depth. Actually, there exists spatial structure dependence on human body key points which can be used to alleviate the problem of joints self-occlusion. Therefore, we represent human pose as a directed graph and propose a network implemented with graph convolution to predict 3D poses from the given 2D poses. In the digraph, we determine the connection weight of each edge according to the error distribution of joints estimation. This makes our model robust to noise. By optimizing coarse 3D estimation and adversarial learning, our algorithm can successfully improve the accuracy of estimation and relieve the ambiguity of mapping. Through testing on Human 3.6M and MPI-INF-3DHP datasets, we achieve excellent quantitative performance. More importantly, our algorithm also has a superior generalization to outdoor dataset MPII by the pre-training process.

INDEX TERMS 3D human poses, graph convolutional networks, adversarial learning, geometric priors, gradient vanish, in-the-wild scenes.

I. INTRODUCTION

Locating the 3D joints of human poses in images or videos is a fundamental problem in computer vision [1]–[4]. There has been great success in 2D human pose estimation due to the availability of large annotated datasets. However, advances in 3D human pose estimation remain limited by joints self-occlusion, weak generalization, and ambiguous mapping.

Note that current methods rarely utilize the spatial structure information among joints to estimate 3D human poses. As shown in Fig. 1, the network fails to reconstruct the 3D pose correctly when those self-occluded joints tend to be indistinguishable, because it cannot get effective information from such a 2D pose. Especially, the location of joints is mainly influenced by their adjacent joints according to the human skeletal structure. And such structural information can be used to relieve the problem of joints self-occlusion. So we model the human pose as a digraph and propose a network to predict 3D human poses from 2D poses which uses the graph convolution to capture structural dependence among joints.

Currently, almost all the algorithms for monocular 3D human pose estimation can be divided into two approaches: (i) estimating 3D joints from 2D images directly [5], [6]; (ii) estimating 3D human poses from 2D observations, such as joint heatmaps or coordinate vectors [7], [8]. Our paper belongs to the second method. Actually, approach (ii) is more effective than approach (i) since using 2D human poses as input makes the process of 2D-to-3D invariant to environment, light, background, etc. And the whole process of 3D human pose estimation is shown in Fig. 2.

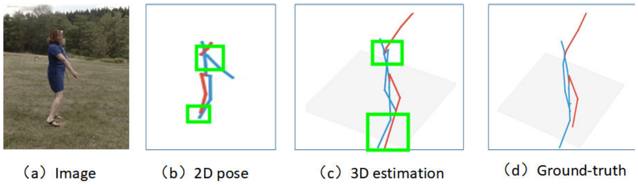
(a) Image (b) 2D pose (c) 3D estimation (d) Ground-truth

FIGURE 1. 3D human pose estimation with joints self-occlusion. (a) is the image from MPI-INF-3DHP dataset. (b) is the 2D ground truth with some joints overlapped in the green box area. And (c) is the 3D estimation predicted from the given 2D pose. Compared to the 3D ground truth (d), those self-occluded joints are inaccurate in the depth axis. Actually, It is a challenging problem for the network to predict 3D pose from such a 2D pose because the network cannot acquire useful structure knowledge.

In standard graph convolutional networks (GCNs), the construction of adjacency matrix only depends on the graph structure, which makes models more sensitive to noise. Some

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

studies show that the estimation error mainly comes from these joints with a large range of activities. Assuming the estimation error satisfies a certain distribution, we determine the adjacent matrix according to this error distribution. Meanwhile, the most commonly large-scale 3D datasets are captured in a restricted laboratory environment. And the 3D models trained on such datasets do not generalize well to other datasets in the wild. Hence, it is desirable to exploit multiple datasets. Specifically, we first train our network on an outdoor 2D dataset and then fine-tune it on a 3D dataset.

There is no 3D information available, which results in ambiguous mapping. And different 3D poses can be projected to a same 2D pose. So we utilize our network to generate coarse 3D poses and extract pose features from them. Then we concatenate low-dimensional 2D and 3D pose features to further refine these coarse 3D estimations. Also, we train our network in an adversarial manner. Our discriminator provides feedback to the generator allowing it to learn priors of 3D human poses. By using an optimized loss function on the discriminator, our method successfully avoids the problems of gradient vanishing and model collapse.

We evaluate our approach on three datasets Human 3.6M, MPI-INF-3DHP, and MPII. On all datasets, our network gets remarkable performance. To sum up, our main contributions are as follows:

- We model the human pose as a digraph and propose a network to predict 3D human poses from 2D poses which can capture spatial structure information among joints to relieve the problem of joints self-occlusion.
- We determine the connection weights of edges in graphs according to the error distribution of joints estimation, which makes our model robust to noise. Also, we strengthen the generalization by pre-training our network on an outdoor dataset.
- We also introduce a method to refine coarse 3D human poses in the current task. This method fuses 2D and 3D pose information to alleviate ambiguity of recovering depth.
- Our method avoids the problems of gradient vanish and model collapse in adversarial learning. Through quantitative and qualitative analysis, our approach achieves excellent performance.

II. RELATED WORK

A. 2D POSE TO 3D POSE

Recent work has decoupled the 3D human pose estimation into two stages. These methods first use a state-of-the-art 2D pose detector [9]–[12] to get a 2D estimation and then lift the 2D pose to 3D by regression modules. The state-of-the-art work gets a remarkable improvement in accuracy using a simple feedforward network [13]. Their results further suggest the feasibility of predicting 3D human poses from the 2D poses. Li and Lee [14], [15] argue that 3D human pose estimation is an inverse problem where multiple solutions can exist. So they propose several types of networks to generate multiple feasible hypotheses from 2D poses. Meanwhile, it is easy to collect the relative depth information of joints which is sufficient for training and evaluating the 3D pose estimation algorithms. Therefore, Ronchi *et al.* [16] and Pavlakos *et al.* [17] use such information as a supervision signal to avoid the inversion of joints. Fang *et al.* [18] and Park and Kwak [19] divide the human body joints into several groups to model the relationship among different groups.

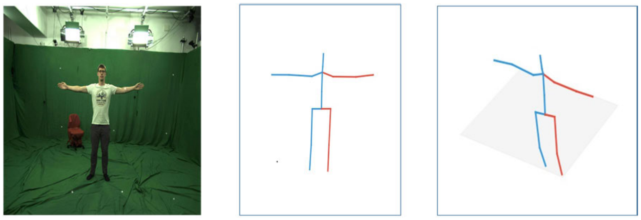
FIGURE 2. The process of 3D human pose estimation with two stages. Left is an input RGB image from MPI-INF-3DHP dataset. Middle is the 2D human pose predicted from the given image. And right is the final 3D human pose by lifting the 2D human pose to 3D.

However, the above approaches fail to reconstruct 3D poses accurately when the 2D poses are considerably different from training examples or contain self-occluded human poses. Because they rarely consider the spatial structural information among human body key points. Specifically, the location of each joint mainly depends on its adjacent joints. Therefore, we further utilize such structure information to address the problem of effectively learning 3D poses from 2D poses.

B. GRAPH CONVOLUTIONAL NETWORKS

GCNs [20], [21], [22] as the special networks of CNNs are often used to process data represented in the graph domain. Recent researches have achieved state-of-the-art performance in modeling the relations of visual temporal sequences [23], [24] and visual objects [25], [26] with GCNs. Ci *et al.* [27] present a locally connected network to improve the representation capability of GCNs. Liu *et al.* [28] first systematically analyze the mechanism of weight sharing in graphs. Inspirited by them, we apply GCNs to 3D human pose prediction. Advances in GCNs are often categorized as spectral and spatial approaches. For spectral approaches, graph convolutional operation is defined in the Fourier domain by computing the eigenvalue decomposition of Laplacian matrix. Yet spatial graph convolution aggregates the joint features directly among adjacency points. Our work belongs to the second approach.

In general, the format of adjacency matrix only depends on the graph structure. The 3D estimations will produce a large deviation when 2D poses are inaccurate. To make our network robust to noise, we first assume that the estimation error satisfies a certain distribution. And then we determine the weights among joints according to this distribution. Our experimental results prove that it is a feasible solution for the noise issue.

C. OUTDOOR 3D POSE ESTIMATION

Many researchers have attempted to estimate the 3D human pose from an image captured in the wild. However, there are

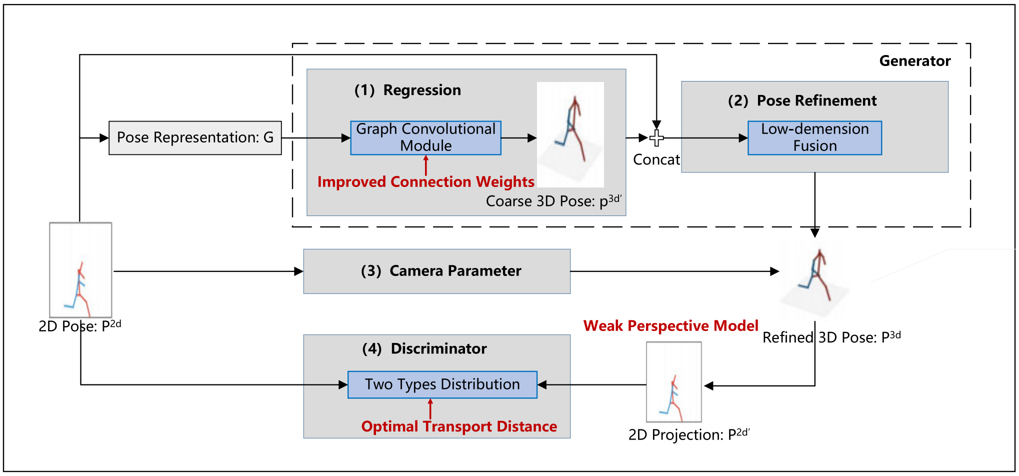

FIGURE 3. **The proposed network contains four parts of modules: a regression module used to estimate coarse 3D pose, a discriminator module used to distinguish between ground-truth and projection, a camera parameter module used to project the 3D pose into 2D, and a pose refinement module used to optimize coarse 3D estimation. There are also some detailed implementations (red) for regression, discriminator, and projection.**

no large-scale outdoor datasets. Therefore, Wang *et al.* [29] propose a novel stereo inspired neural network to generate high quality 3D pose labels for in-the-wild images. Li *et al.* [30] and Chen *et al.* [31] address this problem by training networks in self-supervision without annotated 3D data. Ramakrishna *et al.* [32] propose an algorithm based on Projected Matching Pursuit. Meanwhile, there is an ambiguous mapping from 2D to 3D due to a lack of 3D information. So some methods apply generative adversarial networks to learn 3D pose priors [8], [33], [34]. Others propose several geometric priors to constrain the 3D human poses, such as angle limits [35], physical plausibility [36], and anthropometric regularization [32], [37], [38].

Our method is similar to theirs, but we combine the advantages of these algorithms in a framework. We acquire robust detectors by pre-training the network on a small outdoor dataset. In addition, our approach successfully avoids the problems of gradient vanish and model collapse by using an optimized loss in the discriminator during adversarial learning.

III. METHOD

The main idea behind our method is to utilize the spatial structure information to predict 3D human pose $P^{3d} \in R^{N\times3}$ from the given 2D pose $P^{2d} \in R^{N\times2}$. As shown in Fig. 3, the total network consists of four components. To establish the relations among joints, we represent the human pose as a digraph G. First, the pose regression module (1) is implemented with stacked graph convolutional layers [39], which captures the structural dependence to relieve the problem of joints self-occlusion and generates coarse 3D estimation $P^{3d'}$ from G. Later, 2D input results in the ambiguity of recovering depth. So the pose refinement module (2) concatenates 2D and 3D pose features to learn a refined pose P^{3d} from $P^{3d'}$. Then, to estimate the accurate 3D pose consistent with 2D input, the refined 3D estimation P^{3d} is projected into 2D through a weak perspective model. And the camera parameter module (3) is used to predict the projection matrix in the absence of ground truth. Finally, the discriminator module (4) learns to distinguish between the 2D ground truth P^{2d} and projection $P^{2d'}$. Our regression module can acquire 3D pose priors and further relieve the ambiguity from 2D to 3D by adversarial learning.

A. REPRESENTATION OF HUMAN POSE

Following the basic kinematic constraints of joints, the representation of human pose is shown in the right of Fig. 4. To consist with the definition in Hourglass network [9], we choose the most common applied 16 joints and define the human pose as a triplet $G = (V, \xi, A)$, where V denotes a set of pose joints $[V_1, V_2, \ldots, V_N]$, ξ is the edges indicating the connection relationship among adjacent joints, and A is the adjacent matrix with $a_{ij} = 0$ if (i, j) not in ξ, $a_{ij} = 1$ if (i, j) in ξ. For each joint V_i, its feature is denoted as $H_i = [H_i^1, H_i^2, \ldots H_i^n]$ where n is the dimension of feature.

B. REGRESSION MODULE

1) GRAPH CONVOLUTION MODULE

The graph convolution in our work is used to aggregate joint features among adjacent nodes, which can explicitly capture the structural dependence of joints. When joints are occluded, their location can be inferred by such structural information. The architecture of our graph convolutional module is illustrated in Fig. 5. Inspired by the stacked neural network in [11], this module consists of a residual block that contains two graph convolutional layers with batch normalization, dropout, and non-linear activation after each layer. A residual connection is added from the first layer to the final layer with 512 channels. The main principle of constructing graph convolution is to maximize the receptive field and avoid the indistinguishability between graph nodes due to excessive

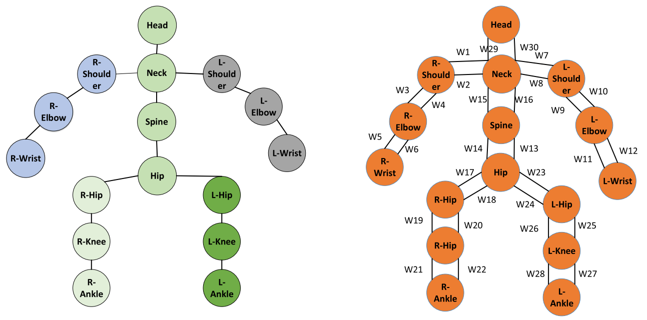

FIGURE 4. **Spatial graph representation of human poses. The left is the structure used in many previous approaches which divide the human pose into five groups according to semantic information. Different color represents different semantic groups. The right is the pose representation in our paper where the pose is modeled as a spatial digraph. W_i denotes the weights, where paired joints have different influence factors.**

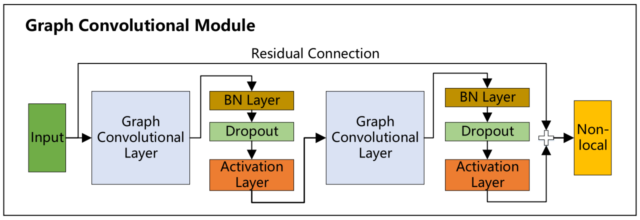

FIGURE 5. **The graph convolutional module.**

network layers. Actually, capturing long-range relationships can address the problem of the limited receptive field as well as obtain complete pose information. Therefore, we add a non-local layer [40] after each residual block which is favorable for information interaction.

Feature updating mechanism among joints is defined as (1) where A is the adjacent matrix and δ (.) is the non-linear activation function. First, joint feature H is gathered to joint V_i according to the adjacent matrix. And then these aggregated features are transformed to objective dimension by learnable weight W. If 3D ground truth is available, the loss of distance L_{ori} between estimation and ground truth is calculated as (2). The p_g and p_i are ground truth and estimation.

$$H^{l+1} = \delta(AH^lW) \tag{1}$$

$$L_{ori} = \sum_{i=1}^{n} ||p_i - p_g||_2 \tag{2}$$

2) CONNECTION WEIGHTS OF EDGES

In standard graph convolution, the adjacency matrix $A \in R^{N\times N}$ is computed as (3), where D is a degree matrix and I denotes the identity matrix. By multiplying by the degree matrix, the adjacency matrix is normalized.

$$A = D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}} \tag{3}$$

However, the fixed adjacent matrix limits the robustness of the model to noise. The final 3D estimation will produce a grand deviation when the 2D pose is inaccurate. Our main idea is to construct an adjacent matrix according to the distribution of the estimation error. Thus, we adopt a method based on iterative error feedback. The graph convolution operation is redefined as (4).

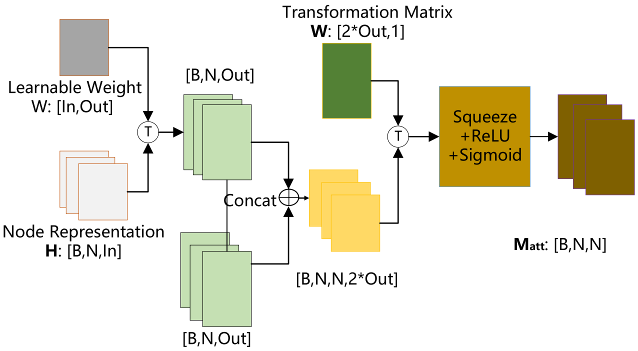

FIGURE 6. **The process of calculating the weight matrix M_{att}. T denotes the linear transformation.**

$$H^{l+1} = \delta(WH^lA') \tag{4}$$

A' is the adjacent matrix fitted by networks with initial A. First, node representations are transformed into objective dimension by the learnable weight W. Then these feature representations are aggregated to current joints from its adjacent nodes. In each convolutional layer, our network calculates a refined adjacent matrix which is used as the input of next layer to propagate the characteristics. This process is defined in following (5) and (6).

$$A' = A \cdot {M_{att}}^T \tag{5}$$

$$M_{att} = Sigmoid(\delta((concat H_i H_j)W)), \quad i, j \in N \tag{6}$$

$M_{att} \in R^{N\times N}$ is the weight matrix with formats of $[[m_{11}, m_{12}, \ldots, m_{1N}], \ldots, [m_{N1}, m_{N2}, \ldots, m_{NN}]]$. Each element m_{ij} denotes the influence factor between joint i and j. Shown as (6), we first concatenate each joint with others to compute the similarity among them. The weight coefficient is obtained by learnable transformation matrix W. Through ReLU and Sigmoid functions, our network gains the weight matrix M_{att}. Finally, we update the adjacent matrix by dot multiplication. Whole network is implemented in a differentiable manner. The estimation error in each round of iteration back propagates to front layers. Therefore, those joints with a large range of activities will get higher weights than others. And the process of calculating weight matrix M_{att} is shown in Fig. 6.

C. POSE REFINEMENT MODULE

To relieve the ambiguity of recovering depth, our network generates coarse 3D estimations and then the pose refinement module concatenates 2D and 3D pose features to learn refined 3D poses. Meanwhile, the high-dimensional pose features generated by the intermediate layer contains redundant information, which may result in inaccurate estimation. So we move this process to front layers. In current 2D coordinate space, data is collected from different perspectives. And all joints seem to be occluded except head and neck joints. But these two joints are also affected by other joints. Therefore,

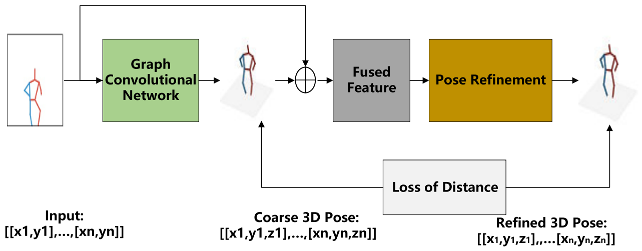

FIGURE 7. The process of human pose refinement.

our pose refinement module is applied to whole human poses. The method of optimizing coarse 3D estimation is defined as (7) and (8), where P^{3d} denotes the refined pose and W denotes the learnable parameters of pose refinement module.

$$P^{3d} = F(P', W) \tag{7}$$

$$P' = P^{3d'} concat P^{2d} \tag{8}$$

Once obtaining the coarse 3D human pose, we concatenate 2D and 3D poses. The fused vector is denoted as $P' = [N, D]$ where D is the dimension of joint features. By concatenating the features of different dimension, the input vector contains original location and orientation information. Finally, we use the pose refinement module F to learn the accurate location from the fused representation of joints. The above process of pose refinement is shown as Fig. 7. And the structure of pose refinement module is similar to Wei *et al.* [11], which achieve excellent performance for 3D human pose estimation with their neural network.

Our approach also imposes a penalty on those invalid 3D poses with inversion and swap. Especially, we use loss functions derived from joints angle limits and skeletons symmetries. Most body joints are constrained within an angular limit. Shown as (9), the illegal angle loss L_{ang} encapsulates this constraint for four pose joints the left/right elbow $L/R(e)$ and left/right knee $L/R(k)$ with their bending beyond 180 degrees [41]. Supposing we are given the human skeletons as $B = ((p_1, p_2, t_{12}), (p_2, p_3, t_{23}), \ldots)$ where each tuple (p_i, p_j, t_{ij}) specifies the pair keypoints i, j and its length t. The bones symmetry loss L_{ske} is defined as (10) where $len(p_i, p_j)$ denotes the length of bones predicted by our network and $||.||_F$ stands for F-norm.

$$L_{ang} = L(e) + L(k) - R(e) - R(k) \tag{9}$$

$$L_{ske} = \sum_{(i,j)\in B} ||len(p_i, p_j) - t||_F \tag{10}$$

D. CAMERA PARAMETER MODULE

In outdoor scenes, the ground truth of camera parameters and 3D poses are not available. Therefore, our camera parameter module is used to predict the projection matrix. We do not set any assumptions about this matrix. The 3D poses are valid only if their projections are consistent with 2D input. This module has a similar architecture to the refinement module but with the output of a 6-dimension vector $R \in R^{3\times 2}$. Each module in our camera parameter network consists of two consecutive residual blocks. For activation functions, we produce the best result with Leaky ReLUs. As shown in (11), the refined 3D human pose P^{3d} is projected into 2D by a weak perspective camera model.

$$P^{2d'} = P^{3d} R \tag{11}$$

E. DISCRIMINATOR MODULE

To further relieve ambiguity of mapping and constrain inaccurate 3D human poses, the discriminator module is utilized in our network to discriminate between 2D ground truth P^{2d} and projection $P^{2d'}$. Total network shown in Fig. 3 is trained alternatively between the generator and discriminator. Our generator can acquire prior information of 3D human poses when the model converges.

We set the true labels as 1 and 0 for 2D ground-truth and projection, respectively. And the most used loss function on the last layer of discriminator module is Cross-Entropy. Shown as (12), P and Q represent two different distributions.

$$H(P, Q) = -\frac{1}{N}\sum_{i}^{N}[p_i \log q_i + (1 - q_i)\log(1 - q_i)] \tag{12}$$

Especially, minimizing this loss function may result in the problems of model collapse and gradient vanish when the model manifold and the true distribution's support have not a non-negligible intersection with random initialization. Therefore, we use the optimal transport distance [45] as a loss function on the discriminator module. Shown as (13), $\prod(P, Q)$ denotes the set of all feasible distribution whose marginals are P and Q. We observe that there often exhibits sharp gradients around some regions. Thus, a gradient penalty scheme is also utilized on the discriminator as (14). We calculate the distance from gradient to pre-defined parameter K. And α denotes the punishment intensity.

$$L(P, Q) = min\, E_{(x,y)\sim\gamma}[||x - y||],\ \gamma \in \prod(P, Q) \tag{13}$$

$$\begin{aligned} L_{adv}(P, Q) &= max E_{x\sim P}[f(x)] - E_{x\sim Q}[f(x)] \\ &\quad + \alpha||\nabla D - K||_2,\ ||f||_L < k \end{aligned} \tag{14}$$

Finally, the total loss function is shown as (15) where λ denotes the weight coefficient of pose prior terms.

$$L_{total} = L_{ori} + L_{adv} + \lambda(L_{ske} + L_{ang}) \tag{15}$$

IV. EXPERIMENTS

We perform our experiments on three datasets: Human 3.6M, MPI-INF-3DHP, and MPII. The first two datasets are the benchmark datasets for 3D human pose estimation with 2D and 3D labels. We use Human 3.6M training set to train our network. Consistent with reference methods, we also utilize the 2D estimation of stacked hourglass networks as input in quantitative analysis. For denoting the generalization of our model on unseen data, we evaluate the trained network on MPI-INF-3DHP dataset. Finally, the qualitative results for unusual poses are shown on MPII.

A. DATASET DESCRIPTION

1) HUMAN 3.6M

Human 3.6M [46] is the largest available dataset for 3D human pose estimation with the image under the motion capture system. This dataset contains 3.6 million RGB images with four camera perspectives. A total of 15 different actions are classified with 11 test subjects in this dataset such as walking, eating, sitting, and smoking. Meanwhile the camera projection parameters and body proportions are also available.

2) MPI-INF-3DHP

MPI-INF-3DHP [5] is a newly released dataset with different scenes including both indoor (background with a green screen (GS) and no green screen (NoGS)) and outdoor settings. This is a more challenging dataset than Human3.6M, whose data is collected only in an indoor setting. In our experiments, we only directly use the test split of this dataset to demonstrate the generalization ability of our trained model quantitatively.

3) MPII

MPII [47] is the widely used dataset for 2D pose estimation which is collected from short YouTube videos covering daily human activities. It contains 25k training images and 2,957 validation images. The human pose is annotated with sixteen 2D joints. Since this dataset consists of a large variety of poses, we use it for qualitative analysis and pre-training the network in a self-supervised manner.

B. IMPLEMENTATION DETAILS

In the stage of data pre-processing, we apply the standard normalization procedure to 2D and 3D poses by subtracting the mean and dividing by the standard deviation. Firstly, our network is pre-trained on the MPII dataset with 100 epochs in a self-supervised manner where 3D ground truth is not available. The generated 3D human poses are projected into 2D using camera projection matrix which is predicted by our camera parameter module. Finally, the pre-trained model is fine-tuned on Human 3.6M dataset for another 100 epochs with ground truth.

Total experiments are carried out in our laboratory platform with GeForce RTX 3080 Graphic Card. The training rate between generator and discriminator is set as 1:3. The weight coefficient of geometric pose priors and the number of stacked convolutional layers are analyzed in ablation experiment. We use an initial learning rate of 0.001 with exponential decay every 10 epochs.

C. QUANTITATIVE EVALUATION

Our main contribution is to propose a network to regress the 3D poses from 2D poses. The network can maintain a meaningful pose even if the 2D poses have deviation. Therefore, in quantitative evaluation part, we compare our model with those approaches which mainly focus on the process from 2D human poses to 3D poses.

1) FOR HUMAN 3.6M

There are two different evaluation protocols: Protocol-I and Protocol-II. Protocol-II employs a rigid alignment between the prediction and ground truth by Procrustes Analysis whereas protocol-I does not. We compute Mean Per Joint Positioning Error (MPJPE) in millimeters (mm) as evaluation metric. Meanwhile, we use the data of subjects S1, S5, S6, S7, S8 for training and S9, S11 for testing. Most of the contrastive methods give the results under both ground truth (GT) and 2D estimation of hourglass networks (HG). Therefore, we differentiate the test results with these two different input.

In Table 1, we evaluate our method with the input of GT. Our approach outperforms all the methods based on original regression network. In these methods, the spatial structure information among joints is rarely considered. This is our motivation of using graph convolutional networks to regress 3D human poses. On average our approach reduces estimation error by 6.8 millimeters under protocol-I, compared to the method [13]. We are also aware of the fact that our method will not outperform these methods based on the weight non-sharing mechanism in graphs [27]. Because weight non-sharing mechanism enhances the model's complexity and representation capability. As shown in this table, they achieve state-of-the-art performance in the current task.

Table 2 shows the results of 3D pose reconstruction with the input of HG. All the numbers are taken from reference papers. In this table, we also achieve a significant improvement over Martinez *et al.* [13] on all the actions. Our result is 8.5 millimeters and 3.6 millimeters smaller than its, under Protocol -I and Protocol-II. Meanwhile, we outperform about 7.1% on protocol-I over the state-of-the-art method based on joints spatial structure information [19]. The estimation error in Pavlakos *et al.* [17] is 41.8 millimeters, which indicates the advantage of using extra ordinal annotations in predicting depth.

2) FOR MPI-INF-3DHP

To verify the generalization of our model to unseen human poses. We compare our method with several approaches that use adapted Percentage of Correct Keypoints (PCK) and Area Under Curve (AUC) as metrics in MPI-INF-3DHP dataset. Especially, we choose a threshold of 150mm in the calculation of PCK. Table 3 shows the quantitative results of different scenes from which our results are closest to the method of [27]. In outdoor scenes, our result is 5.7% higher than Chang *et al.* [44]. They model the estimation error and synthesis the input to make the network robust to error. But it cannot generalize well to outdoor scenes. Our results on MPI-INF-3DHP dataset further suggest the fact that utilizing graph convolutional networks to capture the spatial structure information of human poses can relieve the problem of joints self-occlusion, improve the accuracy of estimation, and also enhance the generalization.

TABLE 1. The quantitative results on Human 3.6M dataset following protocol-I (no rigid alignment) and protocol-II (rigid alignment). The input of our network is 2D ground truth (GT). The contrastive methods also focus on the process from 2D pose to 3D pose. All the numbers are taken from reference papers. – denotes the values not given. * denotes the method based on weight non-sharing mechanism. The lower MPJPE (mm) denotes the better results.

Protocol-I	Direct.	Disc.	Eat	Great	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
FangH et al. [18]	50.1	54.3	57.0	57.1	66.6	66.6	53.4	55.7	72.8	88.6	60.3	60.3	62.7	47.5	50.6	60.4
Li Y et al. [30]	48.7	53.6	54.7	55.1	61.3	76.1	51.5	50.3	68.0	75.9	56.7	53.8	58.8	42.6	47.9	57.0
Li c et al. [14]	43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
Chen et al. [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51.0
wandt et al. [33]	50.0	53.5	44.7	51.6	49.0	58.7	48.8	51.3	51.1	66.0	46.6	50.6	42.5	38.8	60.4	50.9
Ronchi et al. [16]	40.2	44.4	46.6	48.3	53.3	41.7	56.7	42.4	55.9	68.1	50.8	46.2	51.8	40.4	45.5	48.8
Martinez et al. [13]	37.7	44.4	40.3	42.1	54.6	58.0	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Biswas et al. [42]	35.7	42.3	39.0	40.6	44.4	52.5	42.9	38.8	53.1	53.9	42.1	43.3	43.9	33.3	36.5	42.8
Sharma S et al. [43]	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	29.4	34.3	40.9
Ci et al. [27]	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	36.2	39.5	34.4	38.4	38.2	31.3	34.2	***36.3**
Ours	35.1	39.1	33.5	38.9	37.3	45.7	41.2	37.4	39.6	42.1	36.8	41.5	40.2	35.5	36.6	**38.7**
Protocol-II	**Direct.**	**Disc.**	**Eat**	**Great**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD**	**Smoke**	**Wait**	**Walk**	**WalkD**	**WalkT**	**Avg.**
FangH et al. [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
Li Y et al. [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Park et al. [19]	38.3	42.5	41.5	43.3	47.5	53.0	39.3	37.1	54.1	64.3	46.0	42.0	44.8	34.7	38.7	45.0
li c et al. [14]	35.5	39.8	41.3	42.3	46.0	48.9	36.9	37.3	51.0	60.6	44.9	40.2	44.1	33.1	36.9	42.6
Chang et al. [44]	32.7	36.0	45.1	38.3	37.2	36.3	32.4	36.7	58.4	41.2	46.2	36.2	27.2	47.3	32.5	39.1
Wandt et al. [33]	33.6	38.8	32.6	37.5	36.0	44.1	37.8	34.9	39.2	52.0	37.5	39.8	34.1	40.3	34.9	38.2
Ronchi et al. [16]	31.1	37.0	34.3	36.3	37.2	42.5	36.6	33.8	39.9	49.3	37.0	37.7	38.8	33.1	37.5	37.5
Chen et al. [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.0
Ci et al. [27]	24.6	28.6	24.0	27.9	27.1	31.0	28.0	25.0	31.2	35.1	27.6	28.0	29.1	24.3	26.9	***27.9**
Ours	26.6	30.5	26.3	29.8	30.1	34.2	30.3	28.4	34.7	37.7	29.5	31.0	31.4	27.4	29.6	**30.3**

TABLE 2. Quantitative results for the reconstruction of 3D pose on Human 3.6M. The input data is 2D estimation by Hourglass network (HG). Other methods are under the same data conditions as we did. – denotes the values not given. * is the method based on weight non-sharing mechanism. + denotes the method using extra ordinal annotations.

Protocol-I	Direct.	Disc.	Eat	Great	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg.
Wandt et al. [33]	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	79.0	89.9
Chen et al. [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ronchi et al. [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Martinez et al. [13]	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
Li Y et al. [30]	49.7	54.5	58.0	56.8	63.4	80.0	52.4	52.7	71.4	78.3	58.9	55.2	60.0	43.8	49.6	59.0
Park et al. [19]	49.4	54.3	51.6	55.0	61.0	73.3	53.7	50.0	68.5	88.7	58.6	56.8	57.8	46.2	48.6	58.6
Ci et al. [27]	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**
Ours	46.3	54.2	46.5	52.8	53.1	71.6	51.7	47.9	64.1	81.4	51.7	55.3	56.9	44.6	44.2	**54.4**
Protocol-II	**Direct.**	**Disc.**	**Eat**	**Great**	**Phone**	**Photo**	**Pose**	**Purch.**	**Sit**	**SitD**	**Smoke**	**Wait**	**Walk**	**WalkD**	**WalkT**	**Avg.**
Wandt et al. [33]	53.0	58.3	59.6	66.5	72.8	71.0	56.7	69.6	78.3	95.2	66.6	58.5	63.2	57.5	49.9	65.1
Chen at al. [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.0
Ronchi et al. [16]	40.2	44.4	46.6	48.3	53.3	41.7	56.7	42.4	55.9	68.1	50.8	46.2	51.8	40.4	45.5	48.8
Martinez et al. [13]	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
Li Y et al [30]	39.6	42.6	45.7	46.0	47.6	57.1	41.0	39.2	55.4	59.9	46.4	42.5	47.1	34.4	41.0	45.7
Ci et al. [27]	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**
Pavlakos et al. [17]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	**+41.8**
Ours	36.8	42.8	38.5	43.7	44.2	53.8	39.9	36.3	51.3	63.5	45.4	41.3	44.5	34.1	40.3	**44.1**

D. QUALITATIVE RESULT

Firstly, we show some qualitative results on Human 3.6M dataset in Fig. 8. The input to our model is 2D ground truths from the test set. Estimation results are shown in the right column. We can see our method successfully reconstructing 3D human poses from the given input. Our network can still recover the correct depth, even if the 2D joints are occluded. Then, Fig. 9 shows the qualitative results on MPII dataset. These results also prove that our method can get an excellent performance across different datasets.

E. ABLATION STUDY

In this section, we show the ablative analysis of different network components and designed parameters used during training. All of the ablation experiments are performed on Human 3.6M dataset with the test metrics of MPJPE. The *Model 1* is a fully connected network similar to Martinez *et al.* [13].

Table 4 shows the ablative results of different network components. From this table, we can see that the basic graph convolutional network with fixed adjacent matrix *Model 2* and fitted adjacent matrix *Model 3* can decrease the error by 1.7mm and 3.4mm compared to *Model 1*. In *Model 4*, we use the optimal transmission distance based on gradient penalty as a loss function and train the whole network in an adversarial manner. This model further outperforms about 3.1% compared to *Model 2*. By adding a refinement network, the *Model 5* reduces the estimation error by 0.9mm compared

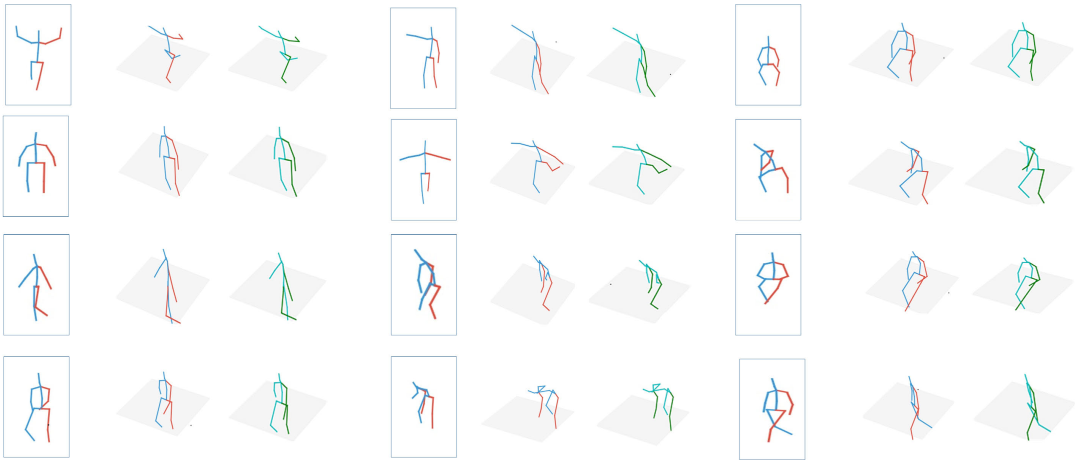

FIGURE 8. The qualitative results on the test set of Human 3.6M. The **left**: 2D inputs. The **middel**: 3D ground truths. The **right** (green): 3D predictions.

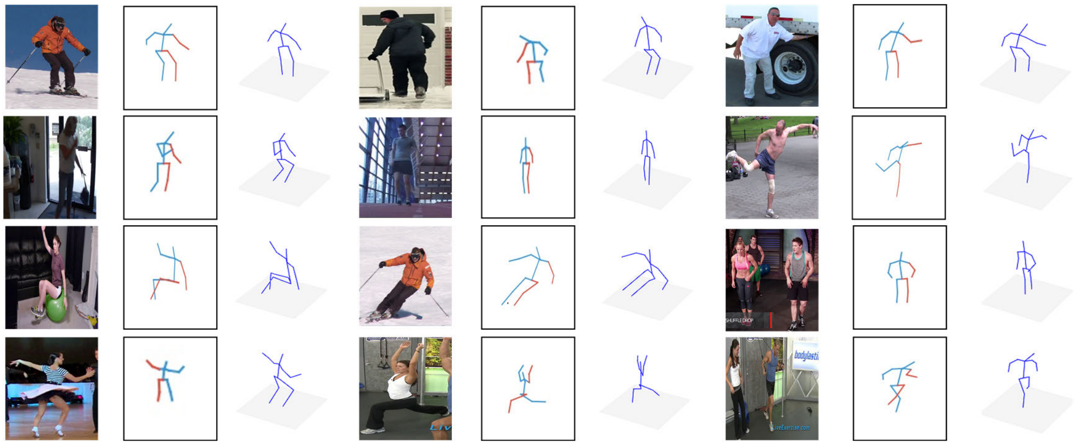

FIGURE 9. The qualitative results on the MPII test set. In this section, we use Stacked Hourglass [9] as the 2D detector to get 2D pose estimations from images. And then the 2D poses are lifted to 3D by our network. The **left**: input images. The **middle**: 2D estimations. The **right** (purple): 3D predictions.

TABLE 3. Quantitative results on MPI-INF-3DHP test set with different scenes. All the results of other methods are taken from the reference papers. – means the values are not given. The higher PCK(%) and AUC denote the better results. Reference [44] is a method with data augmentation. Algorithm [27] is based on weighted non-sharing mechanism. * denotes the methods without capturing spatial structure information.

Method	Training Data	PCK				AUC
		GS	NoGS	Outdoor	ALL	ALL
Chang et al. [44]	H36m	**81.6**	**73.6**	72.5	**76.5**	**40.2**
Ci et al. [27]	H36m	74.8	70.8	**77.3**	74.0	36.7
*Biswas et al. [42]	H36m+MPII	74.2	66.9	71.4	70.8	34.5
*Wang et al. [29]	H36m	-	-	-	71.2	33.8
*Li et al. [14]	H36m	70.1	68.2	66.6	67.9	-
*Chen et al. [31]	H36m	-	-	-	64.3	31.6
*Biswas et al. [42]	H36m+MPII	66.9	63.0	67.4	65.8	31.2
Ours	H36m+MPII	72.3	68.7	76.6	72.5	35.3

to the *Model 3*. Finally, we combine all methods together and the final error is decreased to 38.7mm on protocol-I.

Table 5 shows the ablation results of differently designed hyperparameters. The first row denotes different weight coefficients of geometric priors in self-supervised learning and the third row denotes the number of stacked graph convolutional modules during fine-tuning training. We note that our results are getting worse with the increase of weight coefficients and the excessive number of convolutional layers. As shown in Table 5, the best appropriate parameter for weight coefficients is 0.1 and for the amount of stacked convolutional layers is 4. Actually, the longest skeleton contains 4 joints according to the representation of human poses.

TABLE 4. Ablation results for different algorithm module. A denotes the fixed adjacent matrix. A′ denotes the fitted adjacent matrix by network. And WE denotes the Wassertein loss function.

Model	Method	MPJPE (mm)
Model 1	Full connection	45.5
Model 2	GCN(A)	43.8
Model 3	GCN(A$'$)	42.1
Model 4	GCN(A$'$)+WE+Gradient-penalty	40.8
Model 5	GCN(A$'$)+Refinement	41.2
Total network	*Model 3*+*Model 4*	**38.7**

TABLE 5. Ablation studies for different hyperparameters during network learning. Weight denotes the weight coefficient λ in (15). Amount denotes the number of stacked graph module.

Weight	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
MPJPE (mm)	**119**	121	122	125	130
Amount	**2**	**3**	**4**	**5**	**6**
MPJPE (mm)	49.1	44.1	**42.9**	43.7	45.5

Through the superposition of convolution layers, the receptive field just covers a complete skeleton. Therefore, our results are consistent with this prior.

V. CONCLUSION

In this paper, we model the human pose as a digraph and propose a neural network for estimating 3D human poses from 2D poses which can explicitly capture the spatial structure information among joints to alleviate the problem of joints self-occlusion. To make our model robust to noise, we determine the connection weights of edges in graphs according to the error distribution of joints estimation. The pre-training and geometry prior term are also proved necessary for strong generalization and accurate 3D human pose. Through adversarial learning and refining coarse estimation, our method relieves the ambiguity of recovering depth. Using an optimized loss function in discriminator module, our regression module successfully avoids the gradient vanish and model collapse. By validating on three datasets, our method achieves excellent performance. More importantly, our experimental results further prove that graph convolution and adversarial learning based on geometric constraint are feasible solutions for the problems of joints self-occlusion and ambiguous mapping.

REFERENCES

[1] M. Zolfaghari, A. Jourabloo, S. G. Gozlou, B. Pedrood, and M. T. Manzuri-Shalmani, "3D human pose estimation from image using couple sparse coding," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1489–1499, Aug. 2014.

[2] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2014, pp. 332–347.

[3] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.

[4] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.

[5] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (DV)*, Oct. 2017, pp. 506–516.

[6] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1263–1272.

[7] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1077–1086.

[8] D. Drover, M. V. Rohith, C.-H. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh, "Can 3D pose be learned from 2D projections alone," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 78–94.

[9] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 483–499.

[10] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[11] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[12] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[13] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2640–2649.

[14] C. Li and G. H. Lee, "Generating multiple hypotheses for 3D human pose estimation with mixture density network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9887–9895.

[15] C. Li and G. H. Lee, "Weakly supervised generative network for multiple 3D human pose hypotheses," 2020, *arXiv:2008.05770*. [Online]. Available: http://arxiv.org/abs/2008.05770

[16] M. R. Ronchi, O. M. Aodha, R. Eng, and P. Perona, "It's all relative: Monocular 3D human pose estimation from weakly supervised data," 2018, *arXiv:1805.06880*. [Online]. Available: http://arxiv.org/abs/1805.06880

[17] G. Pavlakos, X. Zhou, and K. Daniilidis, "Ordinal depth supervision for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7307–7316.

[18] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," 2017, *arXiv:1710.06513*. [Online]. Available: http://arxiv.org/abs/1710.06513

[19] S. Park and N. Kwak, "3D human pose estimation with relational networks," 2018, *arXiv:1805.08961*. [Online]. Available: http://arxiv.org/abs/1805.08961

[20] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, Jul. 2005, pp. 729–734.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[22] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[23] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 399–417.

[24] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," 2018, *arXiv:1801.07455*. [Online]. Available: http://arxiv.org/abs/1801.07455

[25] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 670–685.

[26] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.

[27] H. Ci, C. Wang, X. Ma, and Y. Wang, "Optimizing network structure for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2262–2271.

[28] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang, "A comprehensive study of weight sharing in graph networks for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 318–334.

[29] L. Wang, Y. Chen, Z. Guo, K. Qian, M. Lin, H. Li, and J. S. Ren, "Generalizing monocular 3D human pose estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[30] Y. Li, K. Li, S. Jiang, Z. Zhang, C. Huang, and R. Y. D. Xu, "Geometry-driven self-supervised method for 3D human pose estimation," in *Proc. AAAI*, 2020, pp. 11442–11449.

[31] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, R. Mv, S. Stojanov, and J. M. Rehg, "Unsupervised 3D pose estimation with geometric self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5714–5724.

[32] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2012, pp. 573–586.

[33] B. Wandt and B. Rosenhahn, "RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7782–7791.

[34] Y. Kudo, K. Ogaki, Y. Matsui, and Y. Odagiri, "Unsupervised adversarial learning of 3D human pose from 2D joint locations," 2018, *arXiv:1803.08244*. [Online]. Available: http://arxiv.org/abs/1803.08244

[35] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1446–1455.

[36] P. Zell, B. Wandt, and B. Rosenhahn, "Joint 3D human motion capture and physical analysis from monocular videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 17–26.

[37] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. Moreno-Noguer, "Single image 3D human pose estimation from noisy observations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2673–2680.

[38] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2361–2368.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018, pp. 7444–7452.

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[41] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3D human pose from structure and motion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 679–696.

[42] S. Biswas, S. Sinha, K. Gupta, and B. Bhowmick, "Lifting 2D human pose to 3D: A weakly supervised approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–9.

[43] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3D human pose estimation by generation and ordinal ranking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2325–2334.

[44] J. Y. Chang, G. Moon, and K. M. Lee, "PoseLifter: Absolute 3D human pose lifting network from a single noisy 2D human pose," 2019, *arXiv:1910.12029*. [Online]. Available: http://arxiv.org/abs/1910.12029

[45] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[46] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[47] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

XIAOSHAN HUANG received the B.S. degree in computer science and technology from Northeastern University, China, in 2018. He is currently pursuing the M.S. degree in electronics and communication engineering with the Chinese Academy of Sciences. His research interests include computer vision and pattern recognition.

JUN HUANG received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2011. He is currently an Associate Professor with the Shanghai Advanced Research Institute, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, and media analysis.

ZENGMING TANG received the B.S. degree in computer science and technology from Nanchang University, in 2019. He is currently pursuing the M.S. degree in electronics and communication engineering with the Shanghai Advanced Research Institute, Chinese Academy of Sciences. His research interests include computer vision and pattern recognition.