Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images

Size Wu[1,3] Sheng Jin[2,3] Wentao Liu[3*] Lei Bai[4] Chen Qian[3] Dong Liu[1] Wanli Ouyang[4]
[1] University of Science and Technology of China [2] The University of Hong Kong
[3] SenseTime Research and Tetras.AI [4] The University of Sydney
wsz327471010@mail.ustc.edu.cn {jinsheng, liuwentao, qianchen}@sensetime.com
baisanshi@gmail.com dongeliu@ustc.edu.cn wanli.ouyang@sydney.edu.au

Abstract

This paper studies the task of estimating the 3D human poses of multiple persons from multiple calibrated camera views. Following the top-down paradigm, we decompose the task into two stages, i.e. *person localization and pose estimation. Both stages are processed in coarse-to-fine manners. And we propose three task-specific graph neural networks for effective message passing. For 3D person localization, we first use Multi-view Matching Graph Module (MMG) to learn the cross-view association and recover coarse human proposals. The Center Refinement Graph Module (CRG) further refines the results via flexible point-based prediction. For 3D pose estimation, the Pose Regression Graph Module (PRG) learns both the multi-view geometry and structural relations between human joints. Our approach achieves state-of-the-art performance on CMU Panoptic and Shelf datasets with significantly lower computation complexity.*

1. Introduction

The task of estimating 3D human poses of multiple persons from multiple views is a long-standing problem. It has attracted increasing attention for its wide range of applications, *e.g.* sports broadcasting [6] and retail analysis [35].

Recent research on 3D multi-person pose estimation using multi-view images generally follows two streams: 2D-to-3D lifting-based approaches and direct 3D estimation approaches. As shown in Figrue 1(a), 2D-to-3D lifting approaches [3, 4] first estimate 2D joints in each view through monocular pose estimator, then associate 2D poses across views, and finally lift the matched 2D single-view poses to 3D via triangulation [2] or Pictorial Structure Models (PSM) [11]. Such approaches are generally efficient and are the de-facto standard when seeking real-time performance [31]. However, the 3D reconstruction accuracy is limited by the 2D pose estimation, which is not robust to occlusion. As shown in Figure 1(b), direct 3D approaches [35] construct the discretized 3D volumetric representations [28, 29] by gathering multi-view features and directly operate in the 3D space. Such approaches avoid making incorrect decisions in 2D camera views. However, their computation cost increases cubically with the size of the space. They also suffer the quantization errors caused by space discretization [35].

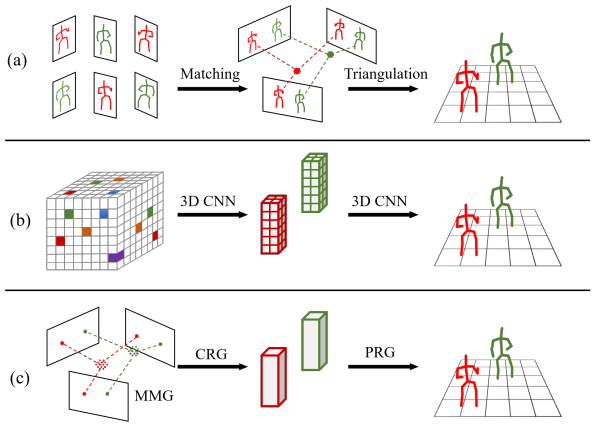

Figure 1. Overview of mainstream multi-view 3D pose estimation frameworks. (a) 2D-to-3D lifting-based approaches (b) Direct 3D pose estimation approaches. (c) Our approach applies graph-based matching algorithm to detect human centers, and applies a graph-based pose refinement model to effectively utilize both geometric cues and human structural prior to achieve better performance.

As shown in Figure 1(c), we combine the virtues of both approaches by adopting 2D-to-3D lifting for efficient *3D human center detection* in the first stage, and direct 3D estimation approach for accurate *single-person 3D pose estimation* in the second stage. To strike a balance between accuracy and efficiency, both stages are processed in coarse-to-fine manners with task-specific graph neural networks.

For coarse-level 3D human center detection in the first stage, we generate coarse human center predictions via multi-view matching. Previous methods perform associa-

*Corresponding author.

tion across views by multi-view geometric constraints [18] and appearance similarity [11]. However, their matching criteria are hand-crafted and not learnable, which may suffer from tedious hyper-parameter tuning and inaccurate matching results. To solve this problem, we propose the Multi-view Matching Graph Module (MMG) to *learn from data* to match people across views by considering both the visual and geometric cues. It also captures the relationship among multiple views to make more reliable predictions.

For fine-level 3D human center detection in the first stage, we propose a graph-based point predictor, *i.e*. Center Refinement Graph Module (CRG), to refine the coarse human center locations. Previous works [1, 6, 29, 28, 35] mostly discretize the space into voxels and operate on a regular grid. CRG instead adopts implicit field representations [21, 32, 33] and directly operates on the continuous 3D space to predict whether a point is a human center or not. It gives us the flexibility to balance between accuracy and speed, by sampling with arbitrary step sizes. Additionally, we propose to use graph models to learn to fuse multi-view features, which are not well-exploited in literature.

For coarse-level single-person pose estimation, we simply use an off-the-shelf pose estimator to generate initial 3D poses based on the detected human proposals. For fine-level single-person pose estimation, we propose the Pose Regression Graph Module (PRG) to refine the initial 3D poses, by exploiting both the spatial relations between body joints and the geometric relations across multiple views.

The three graph modules can alleviate the aforementioned weakness caused by inaccurate 2D detection or space discretization and improve the pose estimation accuracy.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first attempt of using task-specific graph neural networks for multi-view 3D pose estimation. We propose a novel coarse-to-fine framework that significantly outperforms the previous approaches both in accuracy and efficiency.
- We propose Multi-view Matching Graph Module (MMG) to significantly improve the performance of multi-view human association via learnable matching.
- We propose Center Refinement Graph Module (CRG) for point-based human center refinement, which effectively aggregates multi-view features via graph neural networks, and adaptively samples points to achieve more efficient and accurate localization.
- We propose a powerful graph-based model, termed Pose Regression Graph (PRG) for 3D human pose refinement. It accounts for both the human body structural information and the multi-view geometry to generate more accurate 3D human poses.

2. Related Work

2.1. Single-view 3D pose estimation

For *single-person 3D pose estimation* from a monocular camera, we briefly classify the existing works into three categories: (1) from 2D poses to 3D poses [8, 23, 41] (2) jointly learning 2D and 3D poses [27, 28], and (3) directly regressing 3D poses [29, 43] from images. They have shown remarkable results in reconstructing 3D poses, which motivates more research efforts on the more challenging multi-person tasks. *Multi-person 3D pose estimation* from a single RGB image generally follows two streams: top-down and bottom-up. Top-down approaches [10, 26, 39] first use a human detector to produce human locations and then apply single-person pose estimation for each detected person. Bottom-up approaches [24, 40] directly localize keypoints of all people and perform keypoint-to-person association.

Single-view 3D pose estimation has achieved significant progress in recent years. However, inferring 3D poses from a single view is an ill-posed problem. And its reconstruction accuracy is not comparable with that of the multi-view approaches.

2.2. Multi-view 3D pose estimation

We mainly focus on the multi-person 3D pose estimation from multiple views. Existing approaches can be mainly categorized into 2D-to-3D pose lifting approaches [1, 3, 4, 6, 11, 13, 15, 22, 44] and direct 3D pose estimation approaches [35].

2D-to-3D lifting approaches [1, 3, 4, 6, 11, 13] first estimate 2D joints of the same person in each view through monocular pose estimator, then lift the matched 2D single-view poses to 3D locations. Belagiannis *et al*. [3, 4] first extends 2D PSM to 3D Pictorial Structure Model (3DPS) to encode body joint locations and pairwise relations in between. Other works [6, 15] first solve multi-person 2d pose detection and associate poses in multiple camera views. The 3D poses are recovered using triangulation [6] or single-person 3D PSM [11]. Concurrently Lin *et al*. [22] propose to use 1D convolution to jointly address the cross-view fusion and 3D pose reconstruction based on plane sweep stereo. However, such approaches heavily rely on 2D detection results, and the gross errors in 2D may largely degrade 3D reconstruction. In comparison, our approach makes predictions in a coarse-to-fine manner. It models the interaction between multiple camera views using graph neural networks, which are much more efficient and accurate.

Direct 3D pose estimation approaches [35] discretize the 3D space with volumetric representation and gather features from all camera views via multi-view geometry. Tu *et al*. proposes to solve multi-person multi-view 3D pose estimation following the top-down paradigm. Specifically, it first discretizes 3D space with voxels and intensively oper-

ates on 3D space via 3DCNN to give human proposals. For each human proposal, another 3DCNN is applied to recover 3D human poses. Such approaches reliably recover 3D poses but are computationally demanding. In comparison, our approach introduces MMG to significantly reduce the searching space using the multi-view geometric cues. Combined with point-based predictor CRG, we achieve higher accuracy with less computation complexity.

Aggregating features from arbitrary views is important but not well-exploited in literature. Traditional methods aggregate multi-view features by concatenation or average-pooling [35]. Feature concatenation can hardly generalize to different camera settings by design. Average-pooling is permutation invariant but ignores the relations between views. In this paper, we propose a novel graph neural network model to learn to combine geometric knowledge with the corresponding 2D visual features from different views.

2.3. Graph Neural Networks

Graph Convolutional Networks (GCN) generalizes convolutional neural networks to handle graphic data. GCNs have shown effectiveness in message passing, and global relations modeling in various tasks, *e.g*. action recognition [38] and tracking [14]. Recent GCNs can be categorized into spectral approaches [7, 20] and spatial approaches [12, 36]. In this paper, we use spatial approaches for better efficiency and generalizability.

Recently, GCN have shown effectiveness in modeling human body structure for *single-view* 2D human pose estimation. Zhang *et al*. [42] proposes to use PGNN to learn the structured representation of keypoints for 2D single-person pose estimation. Qiu *et al*. [30] proposes OPEC-Net to handle occlusions for 2D top-down pose estimation. Jin *et al*. [16] proposes the hierarchical graph grouping module to learn to associate joints for 2D bottom-up pose estimation. There are also works for *single-view* single-person 3D pose estimation. Zhao *et al*. [45] proposes SemGCN to capture both local and global semantic relationships between joints. Zou *et al*. [46] proposes to capture the long-range dependencies via high-order graph convolution.

We propose to use graph-based models to learn to aggregate features from multiple camera views via multi-view geometry, which was not investigated in existing GCN works. In Pose Refinement Graph Module (PRG), both the body structure priori and the geometric correspondence of multiple views are encoded for more robust and accurate human pose estimation. Moreover, we propose EdgeConv-E, a variant of EdgeConv [36], to explicitly incorporate geometric correspondence as the edge attributes in GCN.

2.4. Implicit Field Representations

Most 3D multi-view pose estimators [1, 6, 28, 29, 35] use 3D volumetric representations, where 3D space is discretized into regular grids. However, constructing a 3D volume suffers from the cubic scaling problem. This limits the resolution of the volumetric representations, leading to large quantization errors. Using finer grids can improve the performance, but it incurs prohibitive memory costs and computation complexity.

Recently, implicit neural representation or implicit field [9, 25, 32, 33] have become popular. Such approaches learn 3D reconstruction in *continuous* function space. Kirillov *et al*. proposes PointRend [21] to select a set of points at which to make predictions for instance segmentation. Inspired by PointRend [21], we propose Center Refinement Graph (CRG), a point-based predictor, to operate on continuous 3D space in a coarse-to-fine manner. We are able to achieve higher accuracy with significantly lower computation complexity.

3. Method

3.1. Overview

We directly use the same pre-trained 2D bottom-up pose estimator from Tu *et al*. [35] to localize 2D human centers in each camera view and to provide feature maps for our task-specific GCNs.

To predict the 3D human centers from 2D locations, we propose Multi-view Matching Graph Module (MMG) to match the centers from different camera views corresponding to the same person. Then we obtain coarse 3D human center locations from the matching results via simple triangulation [2]. The coarse center candidates are further refined by the Center Refinement Graph Module (CRG).

After 3D human centers are predicted, we follow Tu *et al*. [35] to generate 3D bounding boxes with the fixed orientation and size, and apply the 3D pose estimator [35] to generate initial 3D poses. To improve the pose estimation accuracy, the predicted initial 3D poses are further refined by our proposed Pose Regression Graph Module (PRG).

3.2. Multi-view Matching Graph Module (MMG)

Given the 2D human centers generated by the 2D pose estimator, the proposed Multi-view Matching Graph Module (MMG) aims to match them across different camera views, and lift the 2D human centers to coarse 3D human centers via triangulation [2]. We construct a multi-view matching graph, where a vertex represents a human center candidate in a view and an edge represents the connectivity between a pair of human centers in two camera views. The edge connectivity is a binary value in $\{0, 1\}$ representing whether the two corresponding vertices belong to the same person or not. Therefore, the problem of multi-view matching is formulated as the edge connectivity prediction problem. Our MMG applies a graph-based model to solve this problem.

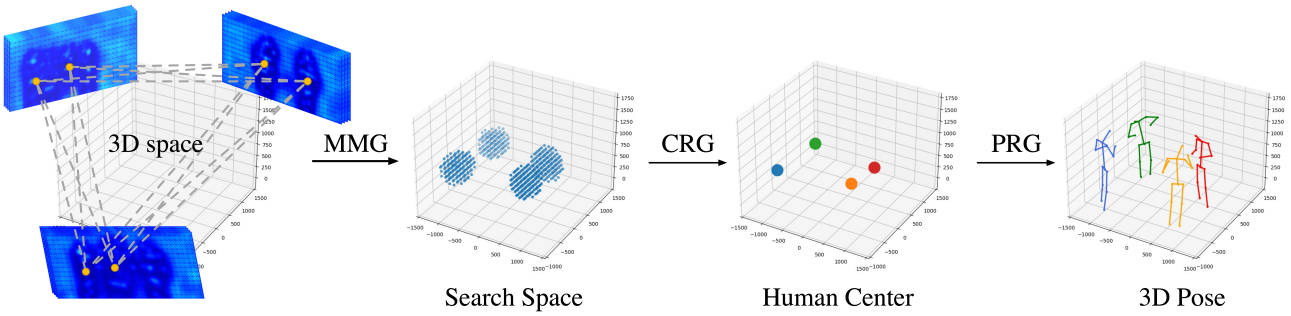

Figure 2. Overview of our approach. The whole pipeline follows the top-down paradigm. It first applies Multi-view Matching Graph Module (MMG) to obtain coarse human center candidates, which are used to limit the search space. Center Refinement Graph Module (CRG) adaptively performs point-based prediction in the search space for more accurate human detection. Finally, Pose Regression Graph Module (PRG) is applied to each detected human proposal to predict the 3D poses in a coarse-to-fine manner.

The graph model consists of two layers of EdgeConv-E (see Sec. 3.2.1) followed by two fully-connected layers. It takes both the vertex features and edge features as input, extracts representative features via message passing, and learns to predict the edge connectivity scores.

The vertex feature encodes the 2D visual cues which are obtained from the feature maps of the 2D backbone networks. Specifically, the vertex feature vector $\mathbb{R}^{512}$ is extracted at each human center location. The edge feature encodes the pair-wise geometric correspondences of two 2D human centers from two distinct views via epipolar geometry [2]. Specifically, we first compute the symmetric epipolar distance [2] d between the two centers. Then the correspondence score s_{corr} can be calculated by $s_{corr} = e^{-m \cdot d}$, where m is a constant and is empirically set to 10.0 in our implementation. In this way, We explicitly use the geometric correspondence score s_{corr} as the edge feature in MMG.

3.2.1 Incorporating edge attributes with EdgeConv-E

EdgeConv [36] is a popular graph convolution prediction to capture local structure and learn the embeddings for the edges. Mathematically, EdgeConv can be represented as:

$$x_v \doteq \max_{v' \in \mathcal{N}(v)} h_\theta \left(\text{Concat}(x_v, x_{v'} - x_v) \right), \tag{1}$$

where x_v and $x_{v'}$ represent the node features at v and v'. 'Concat' denotes the feature concatenation operation. $\mathcal{N}(v)$ is the neighbor vertices of v. h_θ is a neural network, *i.e.* a multi-layer perceptron (MLP).

In standard EdgeConv (Eq.1), the feature aggregation procedure only takes into account the node features x_v and the relative relation of two neighboring nodes $(x_{v'} - x_v)$. It does not explicitly utilize edge attributes for message passing. Based on EdgeConv [36], we propose EdgeConv-E to explicitly incorporate edge attributes $e_{(v,v')}$ into the aggregation procedure. The propagation rule of EdgeConv-E is illustrated in Eq.2.

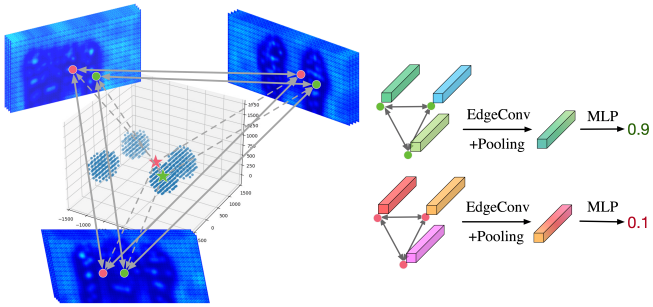

Figure 3. Center Refinement Graph Module (CRG) iteratively applies point-based prediction on selected query points to detect human centers. The graph is constructed by linking the 2D projections of the 3D query in all camera views. Through a few graph convolutions, graph pooling, and MLP, we obtain the confidence score for each proposal.

$$x_v \doteq \max_{v' \in \mathcal{N}(v)} h_\theta \left(\text{Concat}(x_v, x_{v'} - x_v, e_{(v,v')}) \right). \tag{2}$$

3.2.2 Training

We first construct a multi-view graph, where the vertices are generated using the 2D human centers, and the edges connect each pair of 2D human centers in distinct camera views. The target edge connectivity is assigned "1"s for edges connecting the same persons, and "0"s otherwise. To avoid overfitting, we augment by adding uniform noises ranging from 0 to 25 pixels to the ground-truth 2D human center coordinates. Binary cross-entropy loss between the predicted and the target edge connectivity is used for training. We adopt Adam optimizer [19] with a learning rate of 10^{-4} to train the model for 2 epochs.

3.3. Center Refinement Graph Module (CRG)

Center Refinement Graph Module (CRG) is built on top of MMG to refine the 3D human center detection results. CRG adaptively samples query points in the 3D search

space, and predicts the possibility of the query point being a human center. It replaces the commonly used volumetric representations with the implicit field representations, which enables querying at any *real-value* point for more flexible search and accurate localization in the 3D space.

Search space. Instead of operating on the whole 3D space, we propose to restrict the search space based on the matching results from MMG. For each pair of matched 2D human centers, we recover a coarse 3D human center proposal via triangulation [2]. We generate a 3D ball surrounding each 3D human center proposal within a radius of $r_0 = 300$mm. The search space (denoted as Ω_0) is thus the union of these 3D balls.

Feature extraction. Each query 3D point is first projected to all 2D camera views to get its corresponding 2D locations. Then the point-wise feature representations of the corresponding 2D point locations are obtained from the 2D feature maps. Features for a real-value 2D location are obtained via bilinear interpolation, using the surrounding four nearest neighbors located on the regular grid.

We first introduce a baseline model, which concatenates the point-wise features from different views and processes with a learnable multi-layer perceptron (MLP). For each candidate point, the MLP outputs a confidence score of being a human center. We refer to this approach as *MLP-Baseline*. Although intuitive, we argue that this approach is limited for two reasons: (1) it assigns the same weights to all views, and cannot handle occlusion in some viewpoints. (2) it cannot generalize to other camera settings (different number of cameras) by design.

To alleviate these limitations, we propose to use graph neural networks for efficient message passing across views. Our Center Refinement Graph Module (CRG) learns to fuse information from multiple views and verify the proposals from the previous stage. As shown in Figure 3, for each 3D query point, we construct a multi-view graph. The vertices represent the 2D projections in each camera view. The vertex features include (1) visual features $\mathbb{R}^{512}$ extracted in the image plane (2) normalized 3D coordinates $\mathbb{R}^3$ of the query point. (3) 2D center confidence score from the 2D backbone. The edges densely connect these 2D projections to each other, enabling cross-view feature aggregation.

Our CRG uses three layers of EdgeConv for cross-view feature message passing, followed by a max-pooling layer for feature fusion and one fully-connected (FC) layer to predict the center confidence score. We use the standard EdgeConv instead of EdgeConv-E, because CRG does not have explicit edge features for aggregation.

3.3.1 Point Selection

Inference. Given a search region from MMG, we iteratively search for the human centers in a coarse-to-fine manner. CRG starts with the search space Ω_0 described in Sec. 3.3. In the iteration t, it uniformly samples query points in the search space, with the step size τ_t. The graph model processes the sampled queries and predicts their possibility of being a human center. The point with the highest confidence score is selected as the refined human center x_t. We update the search space for the next iteration, Ω_{t+1}, as the 3D ball subspace surrounding the human center x_t with a radius of $r_{t+1} = r_t \cdot \gamma$. We shrink the sampling step size by *i.e.* $\tau_{t+1} = \gamma' \cdot \tau_t$. The iteration continues until the step size reaches the desired precision (ϵ).

Complexity analysis. In Tu *et al*. [35], the search space of human center proposals, as well as the time complexity, is $O(L \times W \times H)$, where L, W, and H are the size of the 3D space. Applying our proposed MMG and CRG, the size of the search space is significantly reduced to $O(N)$, where N is the number of people. Here we omit the size of the search region of an instance, which is a constant. It is noticeable that the complexity is independent of the size of the space, making it applicable to large space applications, *e.g*. the football field. In the experiments, we set the initial step size $\tau_0 = 200$mm, the shrinking factor $\gamma = 0.6$ and $\gamma' = 0.25$, the desired precision $\epsilon = 50$mm. On CMU Panoptic [17] dataset, we record an average of 1,830 queries per frame compared with 128,000 queries taken by Tu *et al*. [35].

3.3.2 Training

The model learns to predict the confidence score for each query point. We develop an effective sampling strategy for selecting training samples to train CRG. Two types of samples are considered for training: positive samples that are located around the ground-truth human centers and negative ones that are far away from human locations. We take positive samples around ground-truth human centers following the Gaussian distributions with the standard deviation $\sigma_{pos} = 400$mm. For negative ones, we take samples uniformly in the entire 3D space. Empirically, the ratio of the number of positive and negative samples is $4:1$.

For a sample located at $\mathbf{X}$, the target confidence score is calculated by

$$s^*_{conf} = \max_{j=1:N} exp\left\{-\frac{\|\mathbf{X} - \mathbf{X}^*_j\|_2^2}{2\sigma^2}\right\}, \qquad (3)$$

where N is the number of human instances and $\mathbf{X}^*_j$ is the 3D coordinate of the center point of person j. And σ is the standard deviation of the Gaussian distribution, which is set as $\sigma = 200$mm. The training loss of CRG is the ℓ_2 loss between the predicted and the target confidence score. We adopt Adam optimizer [19] with a learning rate of 10^{-4}. It takes 4 epochs to reach the best performance.

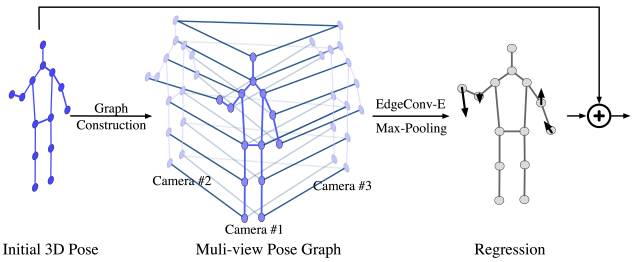

Figure 4. Overview of the 3D pose estimation stage. The initial 3D pose is projected to all camera views to construct the multi-view pose graph. With effective message passing and feature fusion, PRG predicts the regression offsets for 3D pose refinement.

3.4. Pose Regression Graph Module (PRG)

Existing 3D pose regression models produce reliable results on joints that are visible in most views, but will generate inaccurate localization results for occluded joints. Human beings can easily recognize occluded poses, mainly because of their prior knowledge of bio-mechanical body structure constraints and multi-view geometry. The knowledge helps remove ambiguity in localization caused by self-occlusion or inter-occlusion. In light of this, we design the Pose Regression Graph Module (PRG) to learn to refine joint locations considering both the multi-view geometry and structural relations between human joints.

An overview of the 3D pose estimation stage is illustrated in Figure 4. We applied PRG to each individual to further improve the accuracy. The PRG module takes an initial 3D pose as the input. In our implementation, we simply use the pose regressor of [35] to generate the initial 3D pose. The initial 3D pose is projected to all camera views to obtain multiple 2D poses. We construct a multi-view pose graph based on the projected 2D poses in different camera views. The graph predicts the offsets of each keypoint in 3D space, which are added to the initial 3D pose for refinement.

For the multi-view pose graph, the vertices represent the 2D keypoints in a certain camera view. We concatenate the following features the initialize all the nodes in the graph: (1) visual features $\mathbb{R}^{512}$ obtained from the feature maps of the 2D backbone networks at the projected 2D location. (2) one-hot representation of the joint type $\mathbb{R}^{K}$ (3) normalized initial 3D coordinates $\mathbb{R}^{3}$.

The multi-view pose graph consists of two types of edges: (1) single-view edges that connect two keypoints of different types in the canonical skeleton structure in a certain camera view. (2) cross-view edges that connect two keypoints of the same type in different views. We use the one-hot feature vector $\mathbb{R}^{2}$ to distinguish these two types of edges. The one-hot edge features are passed to the EdgeConv-E defined by Eq. 2.

Our graph model of PRG first uses two consecutive EdgeConv-E layers for message passing between neighboring body joints and multiple camera views. Then a max-pooling layer is applied to aggregate the cross-view features and coarsen the graph. The max pooled features are updated by the following three EdgeConv-E layers via effective information flow between the body joints. Finally, the extracted features are passed to one MLP with two fully-connected (FC) layers to regress a refinement vector for each joint.

Training. The target offset is the difference between the ground-truth 3D pose and the initial 3D pose. We use ℓ_1 regression loss between the predicted offset and the target offset to train PRG. Note that the loss gradients of PRG can be back-propagated to the 2D backbone network, which will further improve its feature representation ability. We train PRG using the Adam optimizer [19] with a learning rate of 5×10^{-5}. We train it for 4 epochs to obtain the best model.

4. Experiments

4.1. Datasets

CMU Panoptic [17]: The CMU Panoptic dataset is currently the largest real-world dataset for multi-person 3D pose estimation. It is captured in a studio laboratory, with multiple people doing social activities. In total, it contains 65 sequences (5.5 hours) and 1.5 million of 3D skeletons with 30+ HD camera views. We follow [35, 37] to split the dataset into training and testing subsets. However, we lack the '160906band3' training subset due to broken images. Mean Average Precision (mAP) and mean Average Recall (mAR) are popular metrics for comprehensive evaluation. We calculate mAP and mAR by taking the mean of AP and AR over all the Mean Per Joint Position Error (MPJPE) thresholds (from 25mm to 150mm with a step size of 25mm). We report mAP and mAR along with MPJPE for evaluating the performance of both 3D human center detection and 3D human pose estimation.

Shelf [3]: The Shelf dataset consists of four people disassembling a shelf captured by five cameras. It is challenging due to the complex environment and heavy occlusion. We follow [3, 11, 35] to prepare the training and testing datasets. Following [35], we use the same 2D pose estimator trained on the COCO dataset. We follow [3, 4, 5, 11, 13] to use the percentage of correctly estimated parts (PCP3D) to evaluate the estimated 3D poses.

4.2. Comparisons to the state-of-the-arts

In this section, we compare with the state-of-the-art approaches on CMU Panoptic [17] and Shelf [3] datasets.

On CMU Panoptic dataset, we follow [35] to experiment with the five camera setups. To make fair comparisons, we use the same HD camera views (id: 3, 6, 12, 13, 23). As the AP_{75}, AP_{125} and mAR are not reported in the original paper of Tu *et al*. [35], we reproduce the re-

Table 1. Comparisons to the state-of-the-art approaches on CMU Panoptic dataset [17]. The symbol ↑ means that the higher score the better, while ↓ means that the lower the better. '*' indicates the mean value of four AP_K metrics reported in [35, 22]. '‡' indicates that better 2D pose estimator [34] is used.

	mAP ↑	mAR ↑	MPJPE ↓
Tu *et al*. [35]	95.40*	-	17.68mm
Tu *et al*. [35] (reproduce)	96.73	97.56	17.56mm
‡ Lin *et al*. [22]	97.68*	-	16.75mm
Ours	**98.10**	**98.70**	**15.84mm**

sults by running the publicly available official codes[1] with the recommended hyper-parameters. We find that our re-implementation achieves a slightly better result (17.56mm vs 17.68mm). We show that our approach significantly improves upon Tu *et al*. [35] on mAP, mAR, and MPJPE. Compared with Tu *et al*. [35], our approach has higher accuracy (98.10 mAP vs 96.73 mAP) and also higher recall (98.70 mAR vs 97.56 mAR). Especially, the MPJPE remarkably decreases from 17.56mm to 15.84mm, demonstrating the effectiveness of our proposed method in reducing the quantization error caused by space discretization.

The quantitative evaluation results on Shelf [3] dataset are presented in Table 2. In the experiments, we follow the evaluation protocol of Tu *et al*. [35]. We show that our approach achieves the state-of-the-art performance.

Table 2. Quantitative comparisons to the state-of-the-art approaches on Shelf [3] datasets. The metric is the percentage of correctly estimated parts (PCP3D). '†' means method with temporal information.

Shelf	Actor1	Actor2	Actor3	Average
Belagiannis *et al*. [3]	66.1	65.0	83.2	71.4
†Belagiannis *et al*. [5]	75.0	67.0	86.0	76.0
Belagiannis *et al*. [4]	75.3	69.7	87.6	77.5
Ershadi *et al*. [13]	93.3	75.9	94.8	88.0
Dong *et al*. [11]	98.8	94.1	97.8	96.9
Tu *et al*. [35]	99.3	94.1	97.6	97.0
Huang *et al*. [15]	98.8	96.2	97.2	97.4
†Zhang *et al*. [44]	99.0	96.2	97.6	97.6
Ours	99.3	96.5	97.3	**97.7**

4.3. Ablation study

In this section, we conduct ablative experiments to analyze each component in our proposed framework in detail.

Effect of MMG. In Table 3, we evaluate the performance of Multi-view Matching Graph Module (MMG) on 3D human center detection and 3D human pose estimation. All results use the same 2D detections and 3D human centers are recovered using multi-view triangulation [2]. Traditional methods perform association across views using epipolar constraints [18]. However, they do not generate reliable matching results in occluded scenes. MMG *learns* from data to match people across views. We observe significant improvement in the matching performance (75.91 mAP vs 61.65 mAP). We also notice that replacing MMG with the ground-truth matching results does not notably improve the human center detection results (78.70 mAP vs 75.91 mAP). This implies that the human association results generated by MMG are already very accurate.

Effect of CRG. The Center Refinement Graph Module (CRG) aims at refining the coarse human center predictions. To show the effectiveness of the graph reasoning for human center prediction, we compare CRG with the *MLP-Baseline* introduced in Sec. 3.3 on CMU Panoptic dataset. For fair comparisons, we make both models share the same input features, and have roughly the same number of parameters. As shown in Table 3, CRG outperforms the *MLP-Baseline* in terms of both human detection accuracy (82.10 mAP vs 81.38 mAP) and 3D human pose estimation accuracy (98.10 mAP vs 97.82 mAP). This indicates the importance of learning the multi-view relationship via graph-based message passing.

Table 3. Effect of MMG and CRG on human ceter detection and 3D human pose estimation. Pose results in this table are all obtained by PRG. 'Epi' means epipolar matching. 'GT' means using ground-truth matching results.

Method	Center mAP ↑	Pose mAP ↑	Pose mAR ↑	Pose MPJPE ↓
Epi+Triangulation	61.65	86.02	91.08	24.46mm
MMG+Triangulation	75.91	95.11	97.60	16.99mm
GT+Triangulation	78.70	96.77	98.44	16.08mm
MMG+*MLP-Baseline*	81.38	97.82	97.89	16.06mm
Epi+**CRG**	79.80	95.68	95.68	16.03mm
MMG+CRG (final)	**82.10**	**98.10**	**98.70**	**15.84mm**

Effect of PRG. To analyze the effect of the Pose Regression Graph (PRG), we conduct experiments on CMU Panoptic dataset with multiple initial 3D pose regressors of different accuracy. These models are obtained by varying the granularity of the voxels, *i.e*. 32^3, 48^3, and 64^3. We report the accuracy of the poses before and after the PRG refinement in Table 4. Our PRG is a general pose refiner, which can be applied to various pose estimators to consistently improve the 3D pose estimation accuracy. Note that the 3D pose estimator of (c), is from Tu *et al*. [35].

Table 4. Improvement of 3D pose estimation (MPJPE ↓) when PRG is applied to different initial 3D pose regressors.

	Before PRG	After PRG	Improvement
(a)	18.12mm	16.63mm	1.49mm (8.2%)
(b)	17.78mm	16.44mm	1.34mm (7.5%)
(c)	17.09mm	15.84mm	1.25mm (7.3%)

[1] https://github.com/microsoft/voxelpose-pytorch

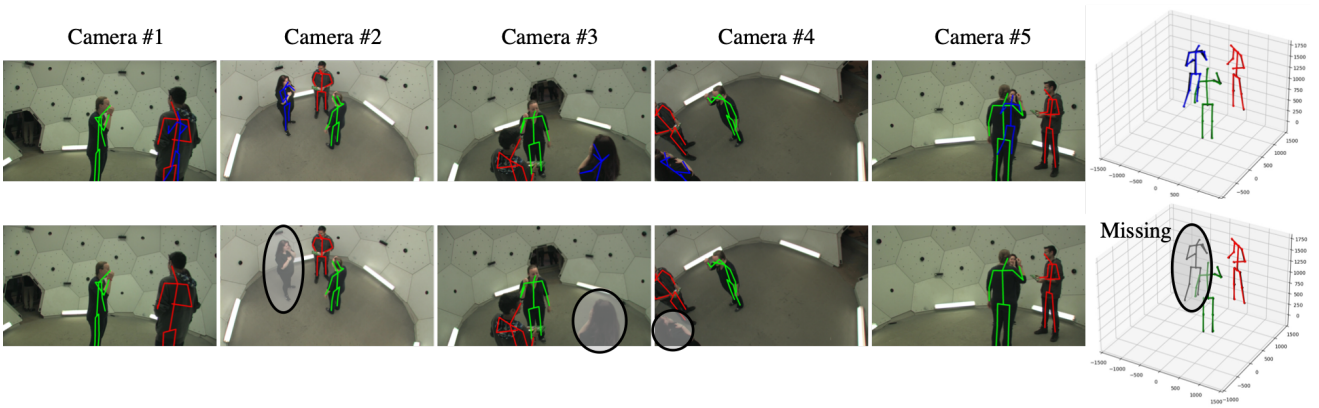

Figure 5. **Qualitative analysis.** Estimated 3D poses and their 2D projections of ours (the 1st row), and Tu *et al*. [35] (the 2nd row). The last column illustrates the ground-truth (black) and the predicted 3D poses (red, green, and blue). Missing poses are highlighted with circles.

4.4. Qualitative Study

We qualitatively compare our results with those of Tu *et al*. [35] in Figure 5. In this example, the body of the woman (blue) is only clearly captured by one camera (view #2), while it is either truncated or occluded in other views. Tu *et al*. [35] simply averages features from all the views with the same weights. This will make the features unreliable, leading to false negatives (FN). In comparison, our approach learns the multi-view feature fusion via GCN. We obtain more comprehensive features which allows us to make more robust estimation. Our approach also gets fewer false positives (FP) and predicts human poses with higher precision. Please see the supplementary for more examples.

4.5. Memory and Runtime Analysis

Table 5. Memory and runtime analysis on CMU Panoptic dataset. Runtime is tested with one Titan X GPU. * denotes the cost of processing one person proposal.

	CPN [35]	PRN* [35]	MMG	CRG	PRG*
Memory	1.10GB	2.38GB	7.10MB	1.08MB	20.3MB
Runtime	26ms	52ms	2.4ms	5.6ms	6.8ms

Table 5 reports the memory and runtime on the sequences with 5 camera views on CMU Panoptic dataset. The results are tested on a desktop with one Titan X GPU. Tu *et al*. [35] proposes CPN to localize people, and PRN to regress 3D poses. Both of them use volumetric representations, which suffer from large amount of memory. In comparison, the memory cost of our proposed graph neural networks is negligible. Our presented modules are also very efficient. On average, our unoptimized implementation takes only 2.4ms for multi-view matching (MMG) and 5.6ms for finer multi-person human center prediction (CRG). Compared with the CPN in [35], CRG requires tens of fewer sampling queries (1.8K vs 128K) due to smaller searching space. And the time cost of PRG is 6.8ms for each person.

When using the PRN as the initial pose estimator, our method facilitates the use of fewer bins of the voxel representation. Comparing #1 and #4 in Table 6, our method using 32^3 bins has about 1/4 computational cost and higher accuracy (1.84mm improvement) than Tu *et al*. [35]. Reducing the bins leads to smaller error increase for ours (0.11mm comparing #2 and #4), but large error increase for Tu *et al*. [35] (1.51mm comparing #1 and #3).

Table 6. Runtime comparison. N is the number of persons. 'avg is the average runtime (ms) when $N = 4$. '#bins' is the number of bins (voxel granularity) for PRN.

#	Method	#bins	Computational cost	avg	MPJPE ↓
1	Tu *et al*. [35]	64^3	$26 + 52 \times N$	234	17.68mm
2	Ours	64^3	$8 + (52 + 6.8) \times N$	243	15.84mm
3	Tu *et al*. [35]	32^3	$26 + 7.3 \times N$	55	19.19mm
4	Ours	32^3	$8 + (7.3 + 6.8) \times N$	64	15.95mm

5. Conclusion

In this paper, we propose a novel framework for multi-view multi-person 3D pose estimation. We elaborately design three task-specific graph neural network models to exploit multi-view features. We propose Multi-view Matching Graph Module (MMG) and Center Refinement Graph Module (CRG) to detect human centers by proposal-and-refinement, and Pose Regression Graph Module (PRG) to produce accurate pose estimation results. Comprehensive experiments demonstrate that the proposed approach significantly outperforms the previous approaches.

Acknowledgement. We would like to thank Lumin Xu and Wang Zeng for their valuable feedback to the paper. This work is supported by the Australian Research Council Grant DP200103223 and FT210100228, Australian Medical Research Future Fund MRFAI000085, the Natural Science Foundation of China under Grants 62036005 and 62021001, and the Fundamental Research Funds for the Central Universities under contract WK3490000005.

References

[1] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. In *Brit. Mach. Vis. Conf.*, volume 1, 2013. 2, 3

[2] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001. 1, 3, 4, 5, 7

[3] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1669–1676, 2014. 1, 2, 6, 7

[4] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):1929–1942, 2015. 1, 2, 6, 7

[5] Vasileios Belagiannis, Xinchao Wang, Bernt Schiele, Pascal Fua, Slobodan Ilic, and Nassir Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In *Eur. Conf. Comput. Vis.*, pages 742–754. Springer, 2014. 6, 7

[6] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 0–0, 2019. 1, 2, 3

[7] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *Int. Conf. Learn. Represent.*, 2014. 3

[8] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7035–7043, 2017. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5939–5948, 2019. 3

[10] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Eur. Conf. Comput. Vis.*, pages 668–683, 2018. 2

[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7792–7801, 2019. 1, 2, 6, 7

[12] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Adv. Neural Inform. Process. Syst.*, 2015. 3

[13] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018. 2, 6, 7

[14] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4649–4659, 2019. 3

[15] Congzhentao Huang, Shuai Jiang, Yang Li, Ziyue Zhang, Jason Traish, Chen Deng, Sam Ferguson, and Richard Yi Da Xu. End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In *Eur. Conf. Comput. Vis.*, pages 477–493. Springer, 2020. 2, 7

[16] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *Eur. Conf. Comput. Vis.*, pages 718–734. Springer, 2020. 3

[17] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1):190–204, 2017. 5, 6, 7

[18] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3d human pose regression. *Machine Vision and Applications*, 32(1):1–14, 2021. 2, 7

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5, 6

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Int. Conf. Learn. Represent.*, 2017. 3

[21] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9799–9808, 2020. 2, 3

[22] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11886–11895, 2021. 2, 7

[23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2640–2649, 2017. 2

[24] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4460–4470, 2019. 3

[26] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Int. Conf. Comput. Vis.*, pages 10133–10142, 2019. 2

[27] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *Int. Conf. Comput. Vis.*, pages 3467–3475. IEEE, 2017. 2

[28] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7307–7316, 2018. 1, 2, 3

[29] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric predic-

tion for single-image 3d human pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7025–7034, 2017. 1, 2, 3

[30] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *Eur. Conf. Comput. Vis.*, pages 488–504. Springer, 2020. 3

[31] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6040–6049, 2020. 1

[32] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Int. Conf. Comput. Vis.*, pages 2304–2314, 2019. 2, 3

[33] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 84–93, 2020. 2, 3

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 7

[35] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3, 5, 6, 7, 8

[36] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 2019. 3, 4

[37] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10965–10974, 2019. 6

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 3

[39] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2148–2157, 2018. 2

[40] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. *Adv. Neural Inform. Process. Syst.*, 31:8410–8419, 2018. 2

[41] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 2

[42] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. *arXiv preprint arXiv:1901.01760*, 2019. 3

[43] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Int. Conf. Comput. Vis.*, 2021. 2

[44] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1324–1333, 2020. 2, 7

[45] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3425–3435, 2019. 3

[46] Zhiming Zou, Kenkun Liu, Le Wang, and Wei Tang. High-order graph convolutional networks for 3d human pose estimation. In *Brit. Mach. Vis. Conf.*, 2020. 3