

# Reconstructing Groups of People with Hypergraph Relational Reasoning

Buzhen Huang<sup>1</sup> Jingyi Ju<sup>1</sup> Zhihao Li<sup>2</sup> Yangang Wang<sup>1\*</sup>

<sup>1</sup>Southeast University, China

<sup>2</sup>Huawei Noah's Ark Lab



Figure 1: We exploit human collectiveness and correlation in crowds to improve human mesh recovery in large-scale crowded scenes (more than 50 people).

## Abstract

Due to the mutual occlusion, severe scale variation, and complex spatial distribution, the current multi-person mesh recovery methods cannot produce accurate absolute body poses and shapes in large-scale crowded scenes. To address the obstacles, we fully exploit crowd features for reconstructing groups of people from a monocular image. A novel hypergraph relational reasoning network is proposed to formulate the complex and high-order relation correlations among individuals and groups in the crowd. We first extract compact human features and location information from the original high-resolution image. By conducting the relational reasoning on the extracted individual features, the underlying crowd collectiveness and interaction relationship can provide additional group information for the reconstruction. Finally, the updated individual features and the localization information are used to regress human meshes in camera coordinates. To facilitate the network training, we further build pseudo ground-

truth on two crowd datasets, which may also promote future research on pose estimation and human behavior understanding in crowded scenes. The experimental results show that our approach outperforms other baseline methods both in crowded and common scenarios. The code and datasets are publicly available at <https://github.com/boycehbz/GroupRec>.

## 1. Introduction

Although immense progress has been made in monocular multi-person human mesh recovery in recent years, the existing methods still cannot accurately reconstruct groups of people from large-scale crowded scenes. The top-down formulation [16, 38, 29, 34] iteratively predicts each individual from tightly cropped image patches, which discards the interaction relationships and location information in the original camera coordinates. Alternatively, the bottom-up formulation [64, 75, 63, 74] parses inter-person interactions with global pixel-level cues and enables its impressive performance on occluded cases. However, bottom-up methods always fail in large-scale scenes like Fig. 1 since they require downsampling images to low-resolution (e.g.,  $512 \times 512$ ) to satisfy computational constraints.

Recently, a few works have attempted to estimate hu-

\*Corresponding author. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), the Natural Science Foundation of Jiangsu Province (No. BK20220127).

man poses in large-scale crowded scenes. Several techniques like composite fields [41] and occlusion augmentation [23, 28] in 2D pose estimation are proposed for addressing the low-resolution inputs. PandaNet [7] further lifts the root-relative 3D poses from the 2D detections with an anchor-based representation. Nevertheless, these works cannot be used to reconstruct absolute body meshes in camera coordinates due to the inherent coupling between depth and body shape. In addition, the challenges of a huge number of people, severe mutual occlusions, and complex spatial distribution make the problem far from being solved.

Different from a few people or single-person cases, it is very common that the crowd in large-scale scenes show significant interactive and collective motions [79, 52]. As shown in Fig. 2, the individuals in the same group even have similar poses. Based on this observation, **our key-idea is to fully exploit the collectiveness and social interaction in crowds, and promote human mesh recovery in large-scale scenes with a relational reasoning**.

However, the idea faces two technical obstacles. First, without a compact representation, the limited hardware memory cannot afford the relational reasoning for a large number of people. Second, the existing networks can hardly formulate the complex and high-order correlation among different individuals and groups in the crowd. To address the obstacles, we propose a multiscale hypergraph to represent the individuals and groups in different scales, and discard the redundant image features in the relational reasoning. Specifically, we first detect bounding-boxes [4, 22] and extract valid human features in the original image. Different from previous top-down methods, we also record the bounding-box information, which preserves the vital global location cues [44] to regress humans in absolute camera coordinates. The compact features and corresponding bounding-boxes depict expressive and high-resolution human information in the crowd image. Then, a multiscale hypergraph network is constructed for the relational reasoning. Based on the hypergraph structure [21], we represent the individuals with hypergraph nodes, and the nodes on the same hyperedge are regarded as a group. Since human groups in a crowd image have unordered structure, the connection relationships of hyperedges cannot be defined with hand-crafted adjacency matrix like previous graph-based pose estimation methods [9, 15]. We thus introduce a differentiable optimization to infer the graph topology, and then assign the individuals with high human feature similarity to the same group. Subsequently, we initialize the nodes with individual human features, and the features for different individuals and groups can pass through the hypergraph via node-to-hyperedge and hyperedge-to-node phases. After the relational reasoning, the updated individual features with group information in the nodes can be utilized to regress the groups of people with absolute positions.



Figure 2: Collective motions are common in human crowd. In addition, since no existing 3D human dataset is captured in real large-scale scenes, we further build pseudo ground-truth on Panda [67] and CrowdPose [42] to relieve the domain gap for synthetic data. The datasets may promote future research on pose estimation and human behavior understanding in large-scale scenes. The main contributions of this work are summarized as follows.

- We reconstruct crowds from single color images and verify that crowds can provide essential knowledge for multi-person mesh recovery.
- We propose a hypergraph relational reasoning network to formulate correlations among individuals and groups, which exploits crowd collectiveness and social interaction to improve human mesh recovery in crowded scenes.
- We build pseudo ground-truth on 2 crowd datasets to promote the research on pose estimation and human behavior understanding in large-scale crowded scenes.

## 2. Related Work

**2D multi-person pose estimation.** 2D multi-person pose estimation explicitly considers person-person interactions and occlusions [82], which can be generally divided into two categories: top-down and bottom-up methods. The top-down strategy [11, 57, 26, 25] iteratively performs pose estimation on each individual in the image. The method achieves high accuracy, but the detection errors in crowded scenes may result in poor performance [20]. The bottom-up strategy [59, 31, 33, 40] distinguishes the body parts of different people simultaneously and produces more robust results in interactive cases. Some representative grouping approaches like Part Affinity Field [10], Associative Embedding [55], and mid-range offset fields [56] are introduced to assemble limbs. However, directly applying these methods in large-scale images (*e.g.*, gigapixel-level [67] and surveillance [17] video) may fail to obtain satisfactory results. The top-down models discard the interactive cues in the original image from the very beginning, while the bottom-up models confront severe scale variations. Only a few works attempt to address the challenges of low resolution and mutual occlusions in larger-scale crowd images with synthesis data [19, 23], composite fields [41] and association mechanism [42]. Nevertheless, all of them do not utilize relationships among individuals like pose similarity and crowd collectiveness [79] in the pose estimation.

**3D multi-person pose and shape reconstruction.** 3D multi-person pose estimation [53, 61, 51, 18, 60] directly regresses joint positions from images, which faces inherent depth ambiguity. To obtain correct depth order in camera-centric coordinate, compressed volumetric heatmap [18], ordinal relation [66], camera prior knowledge [53, 50, 27], and root depth map [77, 35, 46, 12, 68] are proposed for absolute pose prediction. Due to the inherent shape-depth coupling, the multi-person shape reconstruction is more ambiguous than pure pose estimation. The absolute position may not be available [63, 16], or the estimation requires additional ground plane constraints [73, 65]. Some works [74, 29] can regress translations with 2D poses and focal length, but the strategy is strongly affected by the accuracy of 2D poses and predicted body shapes. Other works utilize 6D pose estimation [54], point-based representation [75], bird’s-eye-view-based representation [64], and depth ordering-aware loss [34, 38] to address the obstacles. However, these solutions in multi-person mesh recovery cannot easily be applied to large-scale scenes due to the low resolution and computational constraints. Recently, Crowd3D [69] estimates SMPL maps for cropped patches, and relies on a calibrated ground plane to combine the results in global coordinates. In this work, we incorporate group features and location information [44] in the network inference, and supervise 3D humans in the original camera coordinate system. Unlike the recent relation-aware work [39], our method explicitly considers the group-wise relations with crowd collectiveness, producing more accurate results for the occluded people in crowded scenes.

**Crowd analysis.** Crowd analysis has broad applications in visual surveillance, social behavior understanding, density measurement, and abnormal activity detection. The earliest work in crowd analysis is found in crowd counting [24], which counts individuals [78] or approximates the density of the crowd [47, 62]. Other works further exploit the interactions within crowds for behavior understanding. The underlying relationships among people are utilized in activity recognition [30], dominant motion extraction [14], and trajectory prediction [71, 70]. Nevertheless, previous works in crowd analysis always adopt simplified models (*e.g.*, particle system [5]) to represent the crowd, which discards a lot of human details. In this work, we recover multi-person meshes with absolute positions from a monocular crowd image. To address the occlusions and interactions in crowds, we exploit the crowd collectiveness [79, 52], which indicates the degree of individuals acting as a union in collective motion, and formulate the complex and high-order relation correlation with a hypergraph relational reasoning network. The reconstructed high-fidelity 3D human provides more information to describe the crowd, which may promote future research on human behavior understanding.

### 3. Method

Our method reconstructs groups of people from a large-scale crowd image. The compact human features are first extracted for each individual (Sec. 3.2). Then, a hypergraph relational reasoning network is constructed to fully exploit the collectiveness and interaction relationship among individuals and groups in the crowd (Sec. 3.3). Finally, the group features can compensate for insufficient individual information to regress the 3D crowd with accurate body poses and shapes (Sec. 3.4).

#### 3.1. Preliminaries

**Representation.** We adopt SMPL model [48] with 6D rotation representation [81] to represent the 3D human. The model consists of pose  $\theta \in \mathbb{R}^{144}$ , shape  $\beta \in \mathbb{R}^{10}$  and translation  $t \in \mathbb{R}^3$  parameters. Finally, the output of our network for  $N$  people are  $\{\theta_1, \beta_1, t_1, \dots, \theta_N, \beta_N, t_N\} \in \mathbb{R}^{N \times 157}$ .

**Hypergraph neural networks (HGNN).** HGNN can formulate complex and high-order data correlation with high efficiency through its hypergraph structure [21]. It can be defined as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  and  $\mathcal{E}$  are the set of nodes and hyperedges. Different from the simple graph, a hyperedge connects two or more nodes, where the connection relationships are defined by an adjacency matrix  $\mathcal{H}$ . In this work, we adopt the nodes and hyperedges to represent individuals and groups, and exploit the crowd collectiveness and interactions with a hypergraph relational reasoning.

#### 3.2. Individual feature extraction

We first extract human features from the crowd image for the relational reasoning. Since each person occupies only a small proportion of pixels in a large-scale image [67], it is nontrivial to extract valid and high-quality human features from such inputs. The previous bottom-up methods directly rescale the original image for network input, which results in extremely low resolution and then leads to poor reconstruction performance. Consequently, we predict bounding-boxes for each human and then extract valid image features from the original image as input.

Specifically, we first predict all bounding-boxes [22] from the large image. To preserve the localization information, we transform the box coordinates to  $b_n = \frac{1}{f} [c_x, c_y, d]$ , where  $n \in [1, \dots, N]$ .  $N$  is the number of people in the image.  $(c_x, c_y)$  is the box location relative to the original image center, and  $d$  is its size.  $f$  is the focal length of the original image. With the predicted bounding-boxes, the image patches for all people  $\mathcal{I}^{2D} = \{I_n\}$  on the original image can be cropped. We then extract the high-resolution human image features  $q_n \in \mathbb{R}^m$  from the image patch with a backbone network.

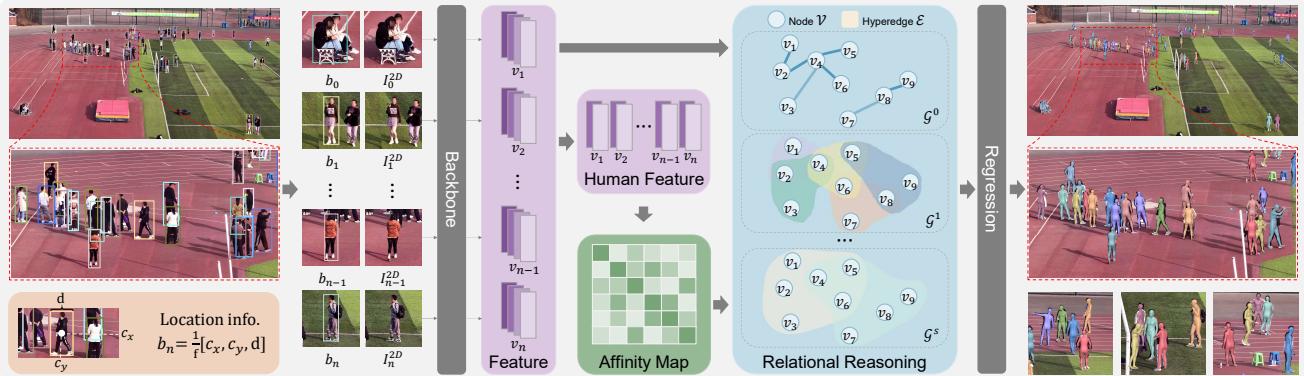


Figure 3: Overview of our method. We first extract location information  $b$  and high-resolution human features  $q$  from the original crowd image. Then, we infer the graph topology according to crowd collectiveness, and then represent the individuals and groups in the crowd with nodes  $\mathcal{V}$  and hyperedges  $\mathcal{E}$  of multiscale hypergraphs  $\mathcal{G}$ . By conducting the hypergraph relational reasoning, we exploit the group features to provide additional cues to regress a crowd of people in camera coordinates.

### 3.3. Hypergraph relational reasoning

Due to the occlusions and depth ambiguity, the individual features are insufficient to regress an accurate 3D human in crowded scenes. However, crowds always show significant collective and interactive motions. The group information can provide additional knowledge for the reconstruction. The core of our work is to exploit the collectiveness and interaction relationship in crowds for multi-person mesh recovery. We propose a novel multiscale hypergraph network to formulate complex correlations among individuals and groups. Mathematically, the individuals are denoted as nodes  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ , and the groups at scale  $s$  are represented with hyperedges  $\mathcal{E}^{(s)} = \{e_1^{(s)}, e_2^{(s)}, \dots, e_{M_s}^{(s)}\}$ . The nodes on the same hyperedge belong to the same group, and a larger  $s$  indicates a larger group size. Hence, the multiscale hypergraph are defined as  $\mathcal{G} = \{\mathcal{G}^{(0)}, \mathcal{G}^{(1)}, \dots, \mathcal{G}^{(S)}\}$ , where  $\mathcal{G}^{(s)} = (\mathcal{V}, \mathcal{E}^{(s)})$ .

Previous graph-based relational reasoning methods [3, 43] only focus on modeling the pair-wise interaction, which ignores the group-wise correlations. In contrast, our hypergraph explicitly forms group structures to exploit human collectiveness and considers the group’s influence on individuals. Besides, the multiscale design also alleviates the over-smoothing in conventional graph networks [80].

**Collectiveness based group inference.** To define the connection relationship of hyperedges, adjacency matrices  $\mathcal{H}^{(s)} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}^{(s)}|}$  are also required. They show the topology of hypergraph, where  $\mathcal{H}_{i,j}^{(s)} = 1$  if the  $i$ th node is included in the  $j$ th hyperedge, otherwise  $\mathcal{H}_{i,j}^{(s)} = 0$ . Existing graph-based pose estimation [9, 15] builds hand-crafted adjacency matrices with known human skeletal and symmetrical relationships. However, human groups are unordered structures without intuitive interpretations. Thus, we infer the topology with an implicit human pose similarity.

The people with similar human features  $v_n = [q_n, b_n] \in$

$\mathbb{R}^{m+3}$  are assigned to the same group. We first compute an affinity matrix  $\mathcal{A} \in \mathbb{R}^{N \times N}$  based on the human feature correlation:

$$\mathcal{A}_{i,j} = v_i^\top v_j / (\|v_i\|_2 \|v_j\|_2). \quad (1)$$

The element of  $\mathcal{A}_{i,j}$  measures the pose similarity and spatial proximity between  $i$ th and  $j$ th individuals. For  $\mathcal{G}^{(0)}$ , we consider the common pair-wise relationship. The two nodes with the largest affinity scores will be connected, leading to adjacency matrix  $\mathcal{H}^{(0)}$  and hyperedge  $\mathcal{E}^{(0)}$ . The other hypergraphs consider group-wise relationships. Assuming the group size at  $s$ th scale is  $K^{(s)}$ , we then find the  $K^{(s)} \times K^{(s)}$  high-density submatrices in  $\mathcal{A}$ . The  $K^{(s)}$  nodes in the group have the highest correlation, and we use a hyperedge  $e_i^{(s)}$  at scale  $s$  to represent the group. The hyperedge can be obtained with:

$$e_i^{(s)} = \arg \max_{\Omega \subseteq \mathcal{V}} \|\mathcal{A}_{\Omega, \Omega}\|_{1,1} \quad (2)$$

$$\text{s.t. } |\Omega| = K^{(s)}, v_i \in \Omega, i = 1, \dots, N,$$

where  $\|\cdot\|_{1,1}$  is the sum of the absolute values of all elements. The optimization can be efficiently solved with a greedy algorithm. For each node  $v_i$ , we find other  $K^{(s)} - 1$  nodes to form a group. Therefore, the hypergraph at scale  $s$  has  $N$  hyperedges. That is, the hypergraphs at different scales have the same number of nodes. Finally, we obtain all  $\mathcal{H} = \{\mathcal{H}^{(0)}, \mathcal{H}^{(1)}, \dots, \mathcal{H}^{(S)}\}$  to construct the hypergraphs.

**Group message passing** Once the graph topologies are constructed, we initialize the node with the individual features  $v_n$ . Different from simple graphs, we can directly exploit group-wise correlations of all group members with the hypergraph structure, and then use the group features to compensate for each individual. To achieve the group message passing, we design node-to-hyperedge and hyperedge-to-node phases. In the node-to-hyperedge phase, the individual features in the nodes are first aggregated to hy-

peredge to obtain group features. Then, the group features are used to update the corresponding individual in the hyperedge-to-node phase. We iteratively execute the two phases at different scales. Finally, the individual features at all scales are concatenated to decode the human pose parameters.

Specifically, the group features are obtained with the following function in the node-to-hyperedge phase:

$$\mathbf{e}_i = c_i \mathcal{F}_e \left( \sum_{v_j \in e_i} \lambda_j v_j \right), \quad (3)$$

where  $\lambda_j = \mathcal{F}_\lambda(v_j, \sum_{v_m \in e_i} v_m)$ , which denotes the contribution of the  $j$ th node to the  $i$ th group.  $c_i$  is the group collectiveness factor:

$$c_i = \sigma \left( \mathcal{F}_c \left( \sum_{v_j \in e_i} (v_j - \bar{v}_i) \right) \right), \quad (4)$$

where  $\sigma(\cdot)$  is a sigmoid function, and  $\bar{v}_i$  is the average features of the  $i$ th group.  $\mathcal{F}_e$ ,  $\mathcal{F}_\lambda$ , and  $\mathcal{F}_c$  are learnable functions implemented by MLPs.

In the hyperedge-to-node phase, the aggregated group features on all associated hyperedges are used to update the individual features.

$$v_i = \mathcal{F}_v \left( v_i, \sum_{e_j \in \mathcal{E}_i} \mathbf{e}_j \right), \quad (5)$$

where  $\mathcal{E}_i = \{e_j \mid v_i \in e_j\}$  denotes the all hyperedges that associate with  $v_i$ .  $\mathcal{F}_v$  is also implemented with MLPs.

The node-to-hyperedge and hyperedge-to-node phases are simultaneously repeated for several times in all scales of hypergraphs. The final individual features on the nodes contain group features and interaction information, which promote more reasonable spatial distribution. Furthermore, the group with high pose similarity can provide additional gesture knowledge for the occluded person to infer a plausible 3D mesh.

#### 3.4. Human parameter regression

After the relational reasoning, the node features on different scales are concatenated with the bounding-box information to obtain the final individual representation  $v'_i = [v_i^{(0)}, v_i^{(1)}, \dots, v_i^{(S)}, b_i]$ , which is used to regress the pose  $\theta$ , shape  $\beta$  and camera  $[f_c, t_x, t_y]$  parameters. The predicted camera can be further transformed into absolute translation:

$$t_X = t_x + \frac{2c_x}{df_c}, t_Y = t_y + \frac{2c_y}{df_c}, t_Z = \frac{2f}{df_c}, \quad (6)$$

where  $t = [t_X, t_Y, t_Z]$  is the translation. More details on the transformation can refer to [44]. Finally, the network output all SMPL parameters  $\{\theta_1, \beta_1, t_1, \dots, \theta_N, \beta_N, t_N\} \in \mathbb{R}^{N \times 157}$  for  $N$  people.

#### 3.5. Network training

The network is trained in an end-to-end manner with the following loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{reproj}} + \lambda_2 \mathcal{L}_{\text{smp}} + \lambda_3 \mathcal{L}_{\text{joint}} + \lambda_4 \mathcal{L}_{\text{crowd}}, \quad (7)$$

where  $\lambda_1 = 5.0$ ,  $\lambda_2 = 5.0$ ,  $\lambda_3 = 5.0$ , and  $\lambda_4 = 0.1$  are loss weights. With the transformation in Equ. (6), we can supervise the reprojection error in the original image, which enforces the network to regress reasonable absolute translations. Specifically, we add the translation  $t$  to the SMPL 3D joint positions  $J_{3D}$  and calculate the loss with following function:

$$\mathcal{L}_{\text{reproj}} = \frac{1}{N} \sum_{n=1}^N \|\Pi(J_{3D}^n + t^n) - J_{2D}^n\|_2^2, \quad (8)$$

where  $\Pi$  projects the 3D joints to 2D image with camera parameters, and  $J_{2D}^n$  is ground-truth 2D pose for  $n$ th person. The SMPL parameters and 3D joint positions are also used for supervision:

$$\mathcal{L}_{\text{smp}} = \frac{1}{N} \sum_{n=1}^N \|[\beta^n, \theta^n] - [\hat{\beta}^n, \hat{\theta}^n]\|_2^2. \quad (9)$$

$$\mathcal{L}_{\text{joint}} = \frac{1}{N} \sum_{n=1}^N \|J_{3D}^n - J_{3D}^n\|_2^2. \quad (10)$$

The  $\hat{\beta}$ ,  $\hat{\theta}$ , and  $J_{3D}^n$  are ground-truth annotations. Although the network can produce accurate body poses with the above constraints, the absolute positions may be totally unreasonable due to the depth-shape coupling. For example, a short person close to the camera can get a similar reprojection error as a tall person in the distance. Previous works [73, 50] rely on a known ground plane to decouple the ambiguities. However, the ground plane is not always available in a single in-the-wild crowd image. Thus, we further exploit the crowd cues and propose a crowd constraint to promote more accurate absolute position prediction.

$$\mathcal{L}_{\text{crowd}} = \text{std}(J_{\text{root}} \cdot l), \quad (11)$$

where  $\text{std}(\cdot)$  denotes standard deviation, and  $J_{\text{root}} \in \mathbb{R}^{N \times 3}$  is the root positions of all people in the image.  $(\cdot)$  means dot product.

$$l = \frac{1}{N} \sum_{n=1}^N \frac{J_{\text{top}}^n - J_{\text{bottom}}^n}{\|J_{\text{top}}^n - J_{\text{bottom}}^n\|}. \quad (12)$$

$J_{\text{top}}$  is 3D keypoint on the head, and  $J_{\text{bottom}}$  is the middle point of two ankle keypoints. The constraint penalizes unreasonable absolute positions and enforces more accurate body shapes. We found the constraint can be pretty robust in common crowded scenes with an appropriate loss weight.



Figure 4: Our method produces accurate body poses and reasonable spatial distribution on **Internet images** [2].

#### 4. Pseudo ground-truth for crowd data

Although 3D human data have seen prosperous developments in recent years, the crowd data in large-scale scenes is still scarce due to the requirement of complex hardware [67] and expensive annotation [32]. To promote the research on crowd analysis, recent works [19, 58] produce photorealistic crowd data using game engines, rendering techniques or generative models [76]. However, a large domain gap exists between synthetic and real data since the illumination conditions and human textures are more complex in the real world. In addition, the natural human behavior in crowds can hardly be simulated in virtual environments.

Therefore, we follow the previous pseudo annotator [37] to build 3D pseudo ground-truth (GT) for Panda [67] and CrowdPose [42]. CrowdPose is a crowd dataset for 2D pose estimation, and Panda is the first gigapixel-level human dataset, which captures real-world crowds ranging from 80 to 900 people. Since Panda does not contain 2D poses, we first predict the 2D keypoints with ViTPose [72]. To generate 3D annotations, we first train our hypergraph relational reasoning network on common crowd data. Once the network is trained, we then estimate initial SMPL parameters from the 2D poses. Due to the domain gap, the initial values may not be accurate enough. We thus finetune the network parameters to adapt to 2D poses via reprojection error in Equ. (8). Like [37], we optimize the network parameters for several iterations and finally output the estimated results as the pseudo GT. Additional constraints are also used in the finetuning. The detailed procedure can be found in the Sup. Mat.

We also manually filter the incorrect estimations in the

camera view. Different from previous pseudo annotators [37, 44], the adaption explicitly considers the crowd interactions and constraints in multi-person scenarios. The 3D models in the final dataset have plausible ordinal relationships and are consistent with image observations. The experiment in Tab. 1 shows that crowd reconstruction methods can gain significant improvement with the proposed datasets.

### 5. Experiments

#### 5.1. Datasets

We use 3 benchmarks, Panoptic [36], GigaCrowd [1], and JTA [19], to evaluate our method. Panoptic is a multi-person dataset captured in an indoor environment, and we use it for evaluation only. We follow previous work [64] to train the network on Human3.6M [32], MuCo-3DHP [51], MSCOCO [45], MPII [6] and CrowdPose [42]. To further evaluate our method on more complex crowded scenes, we use GigaCrowd [1], a large-scale 3D crowd reconstruction dataset containing 3D root positions and 2D poses, as a benchmark. For the evaluation on GigaCrowd, besides the mentioned training data, the proposed Panda dataset is also used for training. On JTA dataset, we use its standard train and test split protocols to conduct the experiments. More details about each dataset can be found in Sup. Mat.

#### 5.2. Metrics

We follow [69] to use the Procrustes-aligned pair-wise percentual distance similarity (PA-PPDS) [69] and object keypoint similarity (OKS) [45] to evaluate the absolute po-



Figure 5: Qualitative comparison with BEV [64] on **GigaCrowd**. Our method is more robust to scale variations and occlusions. In addition, the proposed approach can also reconstruct crowds with more reasonable ordinal relationships.

Method	PA-PPDS↑	OKS↑	PCOD↑	RP↓
CRMH [34]	63.29	64.52	75.28	<b>0.17</b>
BEV [64]	71.37	71.96	83.27	0.23
<b>Ours</b>	<b>82.21</b>	<b>77.31</b>	<b>88.21</b>	<b>0.17</b>
CRMH w/o Panda [34]	52.16	56.31	60.48	<b>0.17</b>
BEV w/o Panda [64]	55.41	62.47	62.38	0.22
<b>Ours w/o Panda</b>	<b>67.22</b>	<b>70.80</b>	<b>71.42</b>	<b>0.17</b>

Table 1: **Comparisons on GigaCrowd.** “w/o Panda” means the model is trained without our Panda dataset.

sitions and pose accuracy on GigaCrowd. In addition, the percentage of correct ordinal depth (PCOD) [77] is adopted to measure the correctness of ordinal depth relations. The redundant punishment (RP) [69] is also used to penalize redundant detections. On other datasets, we adopt the 3D extension of the Percentage of Correct Keypoints (3DPCK) and the Mean per Joint Position Error (MPJPE) to measure the joint accuracy. To consider the missing detections, we follow [12, 18] to use F1-score with thresholds 0.4m, 0.8m, and 1.2m for evaluating absolute positions. The detailed definition of metrics can be found in Sup. Mat.

### 5.3. Comparison to state-of-the-art methods

We conduct several experiments to demonstrate the effectiveness of our method on large-scale crowded scenes. Tab. 1 shows a quantitative comparison on GigaCrowd with CRMH and BEV. CRMH and BEV are the current SOTA methods that can obtain absolute body meshes in large-scale scenes. For a fair comparison, we train the baseline methods with the same data. Since rescaling the full image for BEV input causes extremely low resolution, we use BEV’s released code to crop the original image and combine the predicted results from each patch. Due to the crowd constraints, our method can obtain more reasonable absolute positions and thus results in better performance in terms of PA-PPDS and PCOD. Since some crowds in GigaCrowd dataset show significant collectiveness, our method benefits from the group features and can outperform previous top-down and bottom-up approaches by a large margin on OKS. Besides, we found that the models trained on common multi-person data do not generalize well on large im-

Method	3DPCK <sub>all</sub> ↑	F1(0.4)↑	F1(0.8)↑	F1(1.2)↑
PandaNet [7]	83.2	—	—	—
Benzine <i>et al.</i> [8]	43.9	—	—	—
LoCO [18]	—	50.82	64.76	70.44
Cheng <i>et al.</i> [13]	—	57.22	68.51	72.86
Cheng <i>et al.</i> [12]	—	58.15	69.32	74.19
<b>Ours</b>	<b>86.7</b>	<b>59.59</b>	<b>70.81</b>	<b>76.67</b>

Table 2: **Comparisons on JTA.** Due to the lack of SMPL annotations, we regress joint positions on this dataset for fair comparisons. Our method outperforms previous joint regression baseline methods. “—” means the results are not available.

Method	Haggling↓	Mafia↓	Ultim↓	Pizza↓	Mean↓
Zanfir <i>et al.</i> [73]	140.0	165.9	150.7	156.0	153.4
MubyNet [74]	141.4	152.3	145.0	162.5	150.3
CRMH [34]	129.6	133.5	153.0	156.7	143.2
BMP [75]	120.4	132.7	140.9	147.5	135.4
Pose2UV [29]	104.2	136.0	123.2	151.0	128.6
ROMP [63]	110.8	122.8	141.6	137.6	128.2
3DCrowdNet [16]	109.6	135.9	129.8	135.6	127.3
Luvizon <i>et al.</i> [49]	93.6	—	133.8	145.9	—
BEV [64]	90.7	<b>103.7</b>	113.1	125.2	109.5
<b>Ours</b>	<b>86.8</b>	107.8	<b>110.7</b>	<b>121.1</b>	<b>106.6</b>

Table 3: Comparison with multi-person mesh recovery methods on **Panoptic** dataset. **All methods do not use the data from Panoptic for training.** The results for baseline methods are directly obtained from the original papers. The numbers are MPJPE.

ages due to severe scale variations and complex spatial distributions. The comparisons between rows 1-3 and rows 4-6 in Tab. 1 reveal that the proposed Panda dataset can close the gap between common and large-scale scenarios. In Fig. 5, although BEV estimates the crowd from the cropped images, it still misses some people in the distance. Besides, BEV fails to estimate correct absolute positions in large-scale scenes, while our method produces a reasonable spatial distribution with the relational reasoning.

With the group information and crowd constraints, our method also produces accurate body meshes and reasonable absolute positions on Internet images in Fig. 4. For the images that show significant collectiveness, the occluded people can obtain additional knowledge from others in the same group and result in appropriate poses.

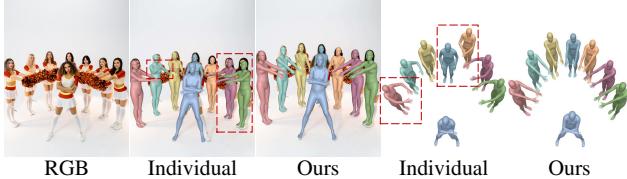


Figure 6: "Individual" removes the relational reasoning. Our method can produce reasonable absolute positions and accurate body poses with the relational reasoning.

Method	CRMH [34]	BEV [64]	Individual	Ours
FPS	21.9	24.1	26.4	23.2

Table 4: Running time on Panoptic with an RTX 3090 GPU. All top-down methods use YOLOX [22] for detection.

We further conduct a comparison with previous pose estimation methods on JTA dataset in Tab. 2. PandaNet [7] is the first 3D pose estimation framework designed for large-scale scenes. Due to our top-down strategy, our method outperforms PandaNet in terms of 3DPCK. We also follow recent works [12, 18] to use F1-score to validate the effectiveness of our method for absolute multi-person position prediction. For a fair comparison, we also regress 3D joint positions on this dataset. Our method constrains individuals with group features and still outperforms these works on all metrics.

To demonstrate the performance of our work on common multi-person scenarios, we compare existing multi-person mesh recovery works on Panoptic. Panoptic is captured in an indoor studio with designed activities, and the people in the image have a high pose similarity. Although the dataset has different camera views and severe mutual occlusions, our method still works well under this challenging setting. In Tab. 3, the Pizza sequence contains truncations, object and human occlusions, and our method also outperforms other baseline methods. The results show that the group features can compensate for insufficient individual information to address the occlusion, which reveals the importance of group features in multi-person mesh recovery.

#### 5.4. Ablation study

**Relational reasoning.** We investigate the importance of group features for multi-person mesh recovery. In Tab. 5 (Individual), we adopt the network without relational reasoning and directly regress human meshes from individual features via the head network in Sec. 3.4, which shows a significant decrease. In Fig. 6, all individuals are improved from the collectiveness with the relational reasoning. Then, we replace the proposed hypergraph network with a transformer for the relational reasoning (Transformer). The transformer-based network is similar to [39], which receives individual features for  $N$  people and outputs corresponding SMPL parameters. It implicitly learns human correlations with attention mechanisms and ignores group-wise relations. Conversely, our hypergraph relational

Method	PA-PPDS↑	OKS↑	Params↓
Individual	70.20	69.44	26.47M
Transformer	76.47	70.28	29.24M
hypergraph-(1)	75.71	71.06	27.15M
hypergraph-(1,3)	78.11	73.30	29.17M
hypergraph-(1,3,5)	82.21	77.31	29.18M
hypergraph-(1,2,3,5)	82.61	76.78	30.19M
hypergraph-(1,3,5,11)	81.84	75.41	30.19M
hypergraph-(1,3,5) w/o $\mathcal{L}_{\text{crowd}}$	77.07	73.41	29.18M

Table 5: **Ablation studies on GigaCrowd.** "Transformer" uses a transformer-based network for relational reasoning. "(1,3,5)" means 3 scales with group sizes of 1, 3, and 5.

reasoning explicitly forms groups with crowd collectiveness and considers the group behavior's influence, which leads to superior performance.

**Group size and scales.** We analyze the impact of group size and scales in Tab. 5. The performance increases with more scales at first and then becomes stable. In addition, the people in large groups (e.g., 11) in most cases have unobvious crowd collectiveness, and the group information may introduce noises in the reasoning.

**Crowd constraints.** Due to the depth ambiguity, regressing reasonable absolute positions from monocular image is an ill-posed problem. The PA-PPDS in Tab. 5 shows that the ambiguity can be greatly alleviated by incorporating the crowd constraints in the loss function.

**Computational complexity.** We compare the running time and network parameters in Tab. 4 and Tab. 5. The results show that the relational reasoning is compact, and our method has competitive running efficiency.

## 6. Limitation and future work

Although our method can reconstruct human groups in large-scale crowd images, there still exist some limitations. First, when the number of people in an image exceeds the maximum, the relational reasoning can only afford a limited number of individuals at a time. Although we can still simultaneously estimate all people in the image by assigning them to different samples of a batch, an interactive pair may be assigned to different samples and can not provide additional cues for each other. In the future, the network can be improved to aggregate similar body poses in the same node to get better compatibility. Second, we may require to decrease the crowd constraint loss weight for some special cases where people are in different planes (e.g., audience in a theater). A too-large crowd loss weight may drag the people to the same plane. To address this problem, to incorporate the scene semantics for future crowd reconstruction might be a feasible solution.

## 7. Conclusion

In this work, we propose a novel hypergraph relational reasoning network to exploit crowd features for reconstruct-

ing groups of people from a large-scale monocular image. To construct the graph topology, crowd collectiveness is used to infer the connection relationships. The proposed network explicitly considers both pair-wise and group-wise relations with a compact individual representation, and promotes accurate body pose and absolute position prediction. In addition, we also build pseudo ground-truth for two crowd datasets. The proposed datasets may promote future research on pose estimation and human behavior understanding in crowded scenes.

**Acknowledgments.** The authors would like to thank the anonymous reviewers for their valuable comments. They also thank Yizhu Li for helpful discussions.

## References

- [1] Gigacrowd challenge. <https://www.gigavision.cn/track/track/?nav=GigaCrowd>. 6
- [2] Pexels. <https://www.pexels.com>. 6
- [3] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *ICCV*, pages 13390–13400, 2021. 4
- [4] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. *ICIP*, pages 966–970, 2022. 2
- [5] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, pages 1–6, 2007. 3
- [6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 6
- [7] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *CVPR*, pages 6856–6865, 2020. 2, 7, 8
- [8] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Single-shot 3d multi-person pose estimation in complex images. *Pattern Recognition*, 112:107534, 2021. 7
- [9] Yujun Cai, Liuqiao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, pages 2272–2281, 2019. 2, 4
- [10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 43(1):172–186, 2019. 2
- [11] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 2
- [12] Yu Cheng, Bo Wang, and Robby Tan. Dual networks based 3d multi-person pose estimation from monocular video. *TPAMI*, 2022. 3, 7, 8
- [13] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In *CVPR*, pages 7649–7659, 2021. 7
- [14] Anil M Cheriyadat and Richard J Radke. Detecting dominant motions in dense crowds. *IEEE Journal of Selected Topics in Signal Processing*, 2(4):568–581, 2008. 3
- [15] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787, 2020. 2, 4
- [16] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, pages 1475–1484, 2022. 1, 3, 7
- [17] Mickael Cormier, Fabian Röpke, Thomas Golda, and Jürgen Beyerer. Interactive labeling for human pose estimation in surveillance videos. In *ICCV*, pages 1649–1658, 2021. 2
- [18] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, pages 7204–7213, 2020. 3, 7, 8
- [19] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, pages 430–446, 2018. 2, 6
- [20] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, pages 2334–2343, 2017. 2
- [21] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI*, pages 3558–3565, 2019. 2, 3
- [22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 3, 8
- [23] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. In *AVSS*, pages 1–8, 2019. 2
- [24] Jason M Grant and Patrick J Flynn. Crowd scene understanding from video: a survey. *TOMM*, 13(2):1–23, 2017. 3
- [25] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 2
- [26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2
- [27] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*, pages 710–720, 2021. 3
- [28] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *TPAMI*, pages 5010–5026, 2022. 2
- [29] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior. *TIP*, 31:4679–4692, 2022. 1, 3, 7

- [30] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, pages 721–736, 2018. 3
- [31] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, pages 34–50, 2016. 2
- [32] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 6
- [33] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCV*, pages 627–642, 2016. 2
- [34] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 1, 3, 7, 8
- [35] Lei Jin, Chenyang Xu, Xiaojuan Wang, Yabo Xiao, Yandong Guo, Xuecheng Nie, and Jian Zhao. Single-stage is enough: Multi-person absolute 3d pose estimation. In *CVPR*, pages 13086–13095, 2022. 3
- [36] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 6
- [37] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, pages 42–52, 2021. 6
- [38] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, pages 1715–1725, 2022. 1, 3
- [39] GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, 2022. 3, 8
- [40] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, pages 417–433, 2018. 2
- [41] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, pages 11977–11986, 2019. 2
- [42] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 2, 6
- [43] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *NeurIPS*, pages 19783–19794, 2020. 4
- [44] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2, 3, 5, 6
- [45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6
- [46] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3d human pose estimation. In *ECCV*, 2022. 3
- [47] Weizhe Liu, Nikita Durasov, and Pascal Fua. Leveraging self-supervision for cross-domain crowd counting. In *CVPR*, pages 5341–5352, 2022. 3
- [48] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3
- [49] Diogo C. Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3d multi-human motion capture from a single camera, 2023. 7
- [50] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *TOG*, 39(4):82–1, 2020. 3, 5
- [51] 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 *3DV*, pages 120–130, 2018. 3, 6
- [52] Ling Mei, Jianghuang Lai, Zeyu Chen, and Xiaohua Xie. Measuring crowd collectiveness via global motion correlation. In *ICCV Workshops*, 2019. 2, 3
- [53] 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 *ICCV*, pages 10133–10142, 2019. 3
- [54] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *CVPR*, pages 14474–14483, 2021. 3
- [55] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *NeurIPS*, 30, 2017. 2
- [56] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, pages 269–286, 2018. 2
- [57] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, pages 4903–4911, 2017. 2
- [58] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 6
- [59] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, 2016. 2
- [60] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tesseltrack: End-to-end learnable multi-person articulated 3d pose tracking. In *CVPR*, pages 15190–15200, 2021. 3

- [61] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 42(5):1146–1161, 2019. 3
- [62] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *CVPR*, pages 19618–19627, 2022. 3
- [63] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 1, 3, 7
- [64] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 1, 3, 6, 7, 8
- [65] Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *3DV*, pages 53–63, 2021. 3
- [66] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259, 2020. 3
- [67] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *CVPR*, pages 3268–3278, 2020. 2, 3, 6
- [68] Zitian Wang, Xuecheng Nie, Xiaochao Qu, Yunpeng Chen, and Si Liu. Distribution-aware single-stage models for multi-person 3d pose estimation. In *CVPR*, pages 13096–13105, 2022. 3
- [69] Hao Wen, Jing Huang, Huili Cui, Haozhe Lin, Yu-Kun Lai, Lu Fang, and Kun Li. Crowd3d: Towards hundreds of people reconstruction from a single image. In *CVPR*, pages 8937–8946, 2023. 3, 6, 7
- [70] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *CVPR*, pages 6498–6507, 2022. 3
- [71] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *CVPR*, pages 5275–5284, 2018. 3
- [72] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 6
- [73] 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 *CVPR*, pages 2148–2157, 2018. 3, 5, 7
- [74] 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. *NeurIPS*, 2018. 1, 3, 7
- [75] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556, 2021. 1, 3, 7
- [76] Zimeng Zhao, Binghui Zuo, Zhiyu Long, and Yangang Wang. Semi-supervised hand appearance recovery via structure disentanglement and dual adversarial discrimination. In *CVPR*, pages 12125–12136, 2023. 6
- [77] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, pages 550–566, 2020. 3, 7
- [78] Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, and Jian Sun. Progressive end-to-end object detection in crowded scenes. In *CVPR*, pages 857–866, 2022. 3
- [79] Bolei Zhou, Xiaou Tang, and Xiaogang Wang. Measuring crowd collectiveness. In *CVPR*, pages 3049–3056, 2013. 2, 3
- [80] Kaixiong Zhou, Xiao Huang, Yuening Li, Daochen Zha, Rui Chen, and Xia Hu. Towards deeper graph neural networks with differentiable group normalization. *NeurIPS*, pages 4917–4928, 2020. 4
- [81] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 3
- [82] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, 2023. 2