Crowd3D: Towards Hundreds of People Reconstruction from a Single Image

Hao Wen[1,†], Jing Huang[1,†], Huili Cui[1], Haozhe Lin[2], Yu-Kun Lai[3], Lu Fang[2], Kun Li[1,*]
[1]Tianjin University, China [2]Tsinghua University, China [3]Cardiff University, United Kingdom
{wenhao, hj00, huilicui_1, lik}@tju.edu.cn, {linhz, fanglu}@tsinghua.edu.cn,
LaiY4@cardiff.ac.uk

Figure 1. Given a single large-scene image with hundreds of people, our method can reconstruct 3D poses, shapes and locations of these people in a global camera space with coherency with the scene. Please zoom in for more details.

Abstract

Image-based multi-person reconstruction in wide-field large scenes is critical for crowd analysis and security alert. However, existing methods cannot deal with large scenes containing hundreds of people, which encounter the challenges of large number of people, large variations in human scale, and complex spatial distribution. In this paper, we propose Crowd3D, the first framework to reconstruct the 3D poses, shapes and locations of hundreds of people with global consistency from a single large-scene image. The core of our approach is to convert the problem of complex crowd localization into pixel localization with the help of our newly defined concept, Human-scene Virtual Interaction Point (HVIP). To reconstruct the crowd with global consistency, we propose a progressive reconstruction network based on HVIP by pre-estimating a scene-level camera and a ground plane. To deal with a large number of persons and various human sizes, we also design an adaptive human-centric cropping scheme. Besides, we contribute a benchmark dataset, LargeCrowd, for crowd reconstruction in a large scene. Experimental results demonstrate the effectiveness of the proposed method. The code and the dataset are available at http://cic.tju.edu.cn/faculty/likun/projects/Crowd3D.

1. Introduction

3D pose, shape and location reconstruction for hundreds of people in a large scene will help with modeling crowd behavior for simulation and security monitoring. However, no existing methods can achieve this with global consistency. In this paper, we aim to reconstruct the 3D poses, shapes and locations of hundreds of people in the global camera space from a single large-scene image, as shown in Fig. 1.

Although monocular human pose and shape estimation [15, 37, 43] has been extensively explored over the past years, estimating global space locations together with human poses and shapes for multiple people from a single image is still a difficult problem due to the depth ambiguity. Existing methods [13, 35] reconstruct 3D poses, shapes and relative positions of the reconstructed human meshes by assuming a constant focal length. But the methods are limited to small scenes with a common FoV (Field of View). These methods cannot regress the people from a whole large-scene image [41] due to the relatively small and varying human scales in comparison to the image size. Even with an image cropping strategy, these methods cannot obtain consistent

† Equal contribution.
* Corresponding author.

reconstructions in the global camera space due to independent inference from the cropped images. Besides, existing methods hardly consider the coherence of the reconstructed people with the outdoor scene, especially with the ground, since the ground is a common and significant element of outdoor scenes. Taking a usual urban scene as an example, these methods may include wrong positions and rotations so that the reconstructed people do not appear to be standing or walking on the ground.

In general, there are three challenges in reconstructing hundreds of people with global consistency from a single large-scene image: 1) there are a large number of people with relatively small and highly varying 2D scales; 2) due to the depth ambiguity from a single view, it is difficult to directly estimate absolute 3D positions and 3D poses of people in the large scene; 3) there is no large-scene image datasets with hundreds of people for supervising crowd reconstruction in large scenes.

In this paper, to address these challenges, we propose *Crowd3D*, the first framework for crowd reconstruction from a single large-scene image. To deal with the large number of people and various human scales, we propose an adaptive human-centric cropping scheme for a consistent scale proportion of people among different cropped images by leveraging the observation of pyramid-like changes in the scales of people in large-scene images. To ensure the globally consistent spatial locations and coherence with the scene, we propose a progressive ground-guided reconstruction network Crowd3DNet to reconstruct globally consistent human body meshes from the cropped images by pre-estimating a global scene-level camera and a ground plane. To alleviate the ambiguity brought in by directly estimating absolute 3D locations from a single image, we present a novel concept called *Human-scene Virtual Interaction Point (HVIP)* for effectively converting the 3D crowd spatial localization problem into a progressive 2D pixel localization problem with intermediate supervisions. Benefiting from HVIP, our model can reconstruct the people with various poses including non-standing.

We also construct *LargeCrowd*, a benchmark dataset with over 100K labeled humans (2D bounding boxes, 2D keypoints, 3D ground plane and HVIPs) in 733 gigapixel images (19200×6480) of 9 different scenes. To our best knowledge, this is the first large-scene crowd dataset, which enables the training and evaluation on large-scene images with hundreds of people. Experimental results demonstrate that our method achieves globally consistent crowd reconstruction in a large scene. Fig. 1 gives an example.

To summarize, our main contributions include:

1) We propose Crowd3D, a multi-person 3D pose, shape and location estimation framework for large-scale scenes with hundreds of people. We design an adaptive human-centric cropping scheme and a joint local and global strategy to achieve the globally consistent reconstruction.
2) We propose a progressive reconstruction network with the newly defined HVIP, to alleviate the depth ambiguity and obtain global reconstructions in harmony with the scene.
3) We contribute *LargeCrowd*, a benchmark dataset with over 100K labeled crowded people in 733 gigapixel large-scene images (19200×6480), which are valuable for the training and evaluation of crowd reconstruction and spatial reasoning in large scenes.

2. Related Work

Multi-person 3D Pose Estimation. These methods can be divided into top-down [2, 29, 33, 40] or bottom-up [4, 7, 20, 27, 28, 45] paradigms. The top-down methods first detect the people and then estimate the 3D pose of each person separately. Moon *et al*. [29] estimate root location and root-relative pose separately after detecting the persons. They regard the area of 2D bounding box as a prior and adopt a neural network to learn a correction factor. HMOR [40] divides human relations into three levels and formulates pair-wise ordinal relations in each level. Different from the top-down paradigm, the bottom-up methods directly detect all the joints and group them. However, most methods either optimize the translation in a post-processing way [27] or ignore the root localization. Inspired by monocular depth estimation methods, SMAP [45] utilizes a deep convolutional neural network (CNN) to estimate a normalized root depth map and part relative-depth maps. The final root map is recovered with the given focal length and hence the camera parameters need to be known to obtain the absolute positions.

All the above methods only estimate 3D poses in the form of skeletons while missing shape information that is important for many applications, such as interpenetration reasoning to avoid impossible poses, person re-identification and crowd analysis.

Multi-person 3D Pose and Shape Estimation. Parametric human body models, *e.g*., SMPL [25], have been widely adopted to represent the 3D pose and shape of a person. Single-person 3D pose and shape estimation has been achieved with tremendous progress [3, 15, 17–19, 30, 32, 36, 42, 43], while multi-person 3D pose and shape estimation still faces many challenges.

Some methods adopt a two-stage framework by utilizing a single-person reconstruction method for each detected person. 3DCrowdNet [6] leverages 2D poses to distinguish different people and uses a joint-based regressor to estimate human model parameters. This kind of approaches focuses more on the accuracy of pose and shape but ignores 3D spatial locations of the people which are important for holistic understanding of the scene. To get coherent reconstruc-

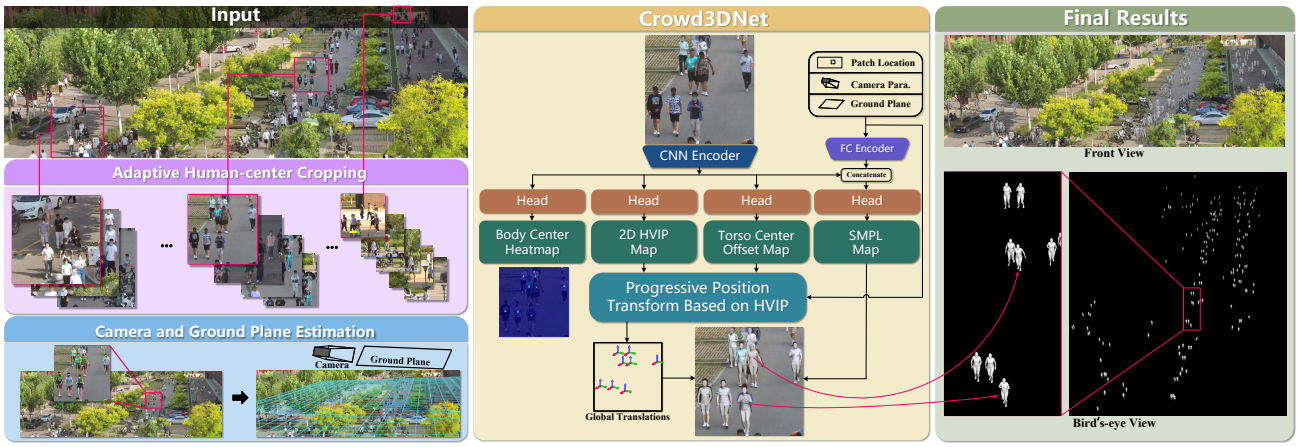

Figure 2. Overview of Crowd3D framework. First, Crowd3D adopts an adaptive human-centric cropping scheme to crop the large-scene image into patches with hierarchical sizes for more appropriate scales of people. Then, Crowd3D estimates the scene-camera intrinsics and ground plane equation with human pose priors. Finally, Crowd3DNet takes the cropped image, the patch location, the estimated camera and ground parameters as inputs and outputs the crowd reconstruction with consistent spatial locations in the global camera space.

tion results, Jiang *et al*. [13] propose CRMH, an R-CNN-based architecture, to detect all the people in an image and estimate their SMPL parameters by using an interpenetration loss and a depth ordering-aware loss in training. This method calculates human depths based on the assumption that people are consistent in height, which will estimate excessive depths for short individuals like kids. To solve the inherent body size and depth ambiguity problem, Ugrinovic *et al*. [38] propose a multi-stage optimization-based method to optimize the 3D translations and scales of body meshes estimated by CRMH [13]. Different from multi-stage methods with computation redundancy, BMP [44] is a single-stage solution for multi-person mesh regression, which correlates the depth of a person with the features of different scales. In ROMP [34], the mesh and location information can be obtained in combination with the camera map and SMPL map according to the center map. However, this method is based on the assumption of weak perspective projection and can only reason about the 2D locations of people in the image plane. It uses an approximation method to obtain depth ordering. To address this, BEV [35] uses Bird's-Eye-View representation to simultaneously reason about body centers in image and in depth. As mentioned by itself [35], BEV is not trained or designed to deal with large "crowds" (*e.g*., 100s of people) with a constant focal length assumption. In general, all the above methods can only get relative depths rather than absolute 3D positions, and they cannot be applied directly to large scenes.

Multi-person Datasets. Multi-person datasets can be collected in indoor controlled environments or in outdoor scenes. Datasets such as *Panoptic* [14] build multi-view capture systems to obtain relatively high accurate ground-truths. For outdoor scenes, some datasets enable in-the-wild 3D capture from videos with IMUs [39] or from videos in which humans have to stay still [21], while other datasets [26] annotate in-the-wild images in 2D only. Besides, datasets such as *Agora* [31] and *MuCo-3DHP* [28] generate synthetic images including 3D people and background images or 3D background scenes. However, all the above datasets only contain a few people in small scenes.

In this paper, We propose the first work to reconstruct hundreds of people in a large scene with global consistency from a single RGB image. We also contribute a benchmark dataset, *LargeCrowd*, for the training and evaluation of crowd reconstruction in large scenes.

3. Method

Our work aims to recover a globally coherent reconstruction of crowd from a single large-scene image with hundreds of persons. Fig. 2 shows the framework of our method. The highlight of our method is that we design progressive position transform with our newly defined concept HVIP to establish a mapping between local image points and global spatial positions. Our method consists of three main steps: 1) we adopt an adaptive human-centric cropping scheme (Sec. 3.1) to crop the large-scene image into patches with hierarchical sizes which ensures that people in different cropped images have appropriate scales; 2) we estimate the camera intrinsics and ground plane equation (Sec. 3.2) of the scene with human pose priors for subsequent inference; 3) taking the cropped images, ground plane and camera parameters as inputs, we design the Crowd3DNet (Sec. 3.3) with the progressive position transform based on HVIP (Sec. 3.3.1) to directly estimate the human meshes in the large-scene camera coordinate system.

3.1. Adaptive Human-centric Cropping

Instead of using uniform cropping [10, 22] that cannot deal with people of various image sizes, we propose an

adaptive human-centric cropping strategy to ensure that the height ratio between people and the corresponding cropped image is as consistent as possible among different cropped images. It is crucial for accurate and reasonable estimation. Inspired by the observation that human heights hierarchically vary like a pyramid in the vertical direction of large-scene image, the sizes of the cropped images should also conform to a similar hierarchical change. Heuristically, we use a geometric sequence to simulate the hierarchical change, which is simple but effective. Define the heights of the persons at the top and the bottom of the large-scene image as h_{t} and h_{b}, respectively. The upper and lower bounds of the image area to be processed are defined as b_u and b_l. Considering non-overlapping square blocks in the vertical direction of image, we represent the sizes of blocks from top to bottom as $\{c_i\}_{i=1}^{n}$. When we set the height of people in a block to be half of the block size and make $\{c_i\}_{i=1}^{n}$ comply with the rule of geometric sequence, we have $c_1 = 2 \times h_{\text{t}}$, $c_i = c_1 \times q^{i-1}$ and $\sum_{i=1}^{n} c_i = b_l - b_u$, where q is the proportionality coefficient. This cropping problem is formulated as:

$$\arg\min_{n,q} |c_n - 2 \times h_{\text{b}}|. \tag{1}$$

To ensure each person can appear completely in some blocks, we further add overlapping blocks between adjacent rows of cropped images, with the size set to the average of cropped images in these rows. For the horizontal direction, we also add overlapping blocks with the same size as those in the row. The cropping parameters h_{t}, h_{b}, b_u, b_l can be set manually or automatically. Details are given in the supplementary document.

3.2. Camera and Ground Plane Estimation

We use the ground plane as a guidance for three reasons: 1) it is a common element in large scenes, especially surveillance scenarios; 2) it is the main object interacting with people in the large scenes, reflecting the harmony between people and the scene; 3) it provides the important global information to the local cropped images.

To estimate the ground plane equation and the scene-level camera parameters, the pose prior of people can be used for calibration. Note that the estimation of ground plane does not need too many people: more than ten people are enough as shown in the experiment (Sec. 4.5). Besides, our method can reconstruct people with various poses, but at the current stage, we only consider the standing or walking people who can be regarded as vertical lines on the ground plane. These people are automatically selected from the 2D keypoints detection obtained from RMPE [8]. We use a pinhole camera model with a focal length f ($f = f_x = f_y$) where the principal point (c_x, c_y) of the camera is the image center. We represent the ground equation as $N^T P_g + D = 0$, where $N = (x_n, y_n, z_n)$ is the ground normal with $\|N\|_2 = 1$, $P_g \in \mathbb{R}^3$ is the point on the ground plane and D is a constant term. For these standing people, we define the midpoints of their two ankle keypoints as $P_a \in \mathbb{R}^3$ and the midpoints of two shoulder keypoints as $P_s \in \mathbb{R}^3$. The projections of P_a and P_s are $p_a = (u_a, v_a)$ and $p_s = (u_s, v_s)$, respectively. Following perspective projection, we have $z_a \times \bar{p}_a = KP_a$, where $\bar{p} = (u, v, 1)^T$ represents the homogeneous coordinates of $p = (u, v)$, K is the intrinsic matrix of the scene-level camera and z_a is the depth of P_a. Similar to [9], we assume that P_a is on the ground plane, and the line from P_a to P_s is parallel to the ground normal. We also set a fixed height prior h. Therefore, we have $N^T P_a + D = 0$ resulting in

$$z_a = -\frac{D}{N^T K^{-1} \bar{p}_a}, \tag{2}$$

and P_s can be approximated by $P'_s = P_a + h \times N$. Then, the projection $\bar{p}'_s$ is computed by

$$z'_s \times \bar{p}'_s = z'_s \times \begin{bmatrix} u'_s \\ v'_s \\ 1 \end{bmatrix} = K(z_a \times K^{-1} \bar{p}_a + h \times N). \tag{3}$$

To solve the camera and ground plane parameters K, N, D, we adopt the following optimization loss:

$$\begin{aligned} L_{\text{param}} =& \lambda_{\text{angle}} L_{\cos}(p'_s - p_a, p_s - p_a) \\ &+ \lambda_{\text{mod}} \frac{|\|p'_s - p_a\|_2 - \|p_s - p_a\|_2|}{\|p_s - p_a\|_2}, \end{aligned} \tag{4}$$

where $L_{\cos}$ is the cosine distance, and λ_{angle} and λ_{mod} are the weights of the corresponding loss terms. Finally, we translate 0.1 meters along the ground normal direction to get the ground plane where people stand on.

3.3. Crowd3DNet

As shown in Fig. 2, Crowd3DNet is a one-stage multi-head network based on the body-center-guided representation [34]. Different from previous methods [34, 35, 44], we define a new concept, *Human-scene Virtual Interaction Point (HVIP)*, and a progressive position transform (Sec. 3.3.1) to better infer the global 3D positions of people. Crowd3DNet outputs four maps including a body center heatmap, a torso center offset map, a 2D HVIP map and a SMPL parameters map. The body center heatmap predicts the probability that each location is the center of a human body. If the body center heatmap gives positive responses, the network samples relevant parameters from other maps at the corresponding center locations to obtain 2D torso center offsets, 2D HVIPs and SMPL parameters of people. With progressive position transform based on HVIP, Crowd3DNet combines the sampled parameters, the input of ground plane equation and scene-level camera parameters to infer accurate 3D positions of people, achieving multi-person reconstruction in the large-scene camera system from the cropped images.

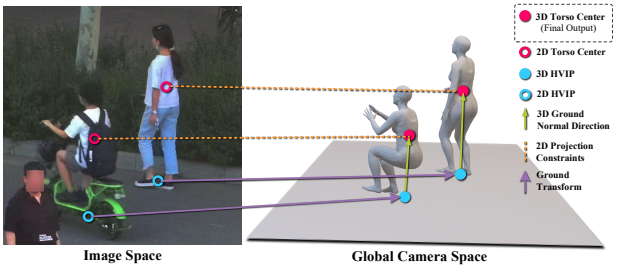

Figure 3. Progressive position transform based on HVIP.

3.3.1 Progressive Position Transform Based on HVIP

We design the progressive position transform based on Human-scene Virtual Interaction Point (HVIP) to help infer the accurate 3D locations of persons in the large-scene camera system. The core idea is to infer the global 3D position from 2D image pixel points by HVIP and ground transform to avoid the depth ambiguity of estimating from a single view directly. We define the HVIP which represents the projection point of a person's 3D torso center on the ground plane in the global camera space, marked as $P_v = (x_v, y_v, z_v)$. The torso center is a semantic point on human body, *i.e.*, the center of two shoulder and two hip joints, represented as $P_t = (x_t, y_t, z_t)$. As show in Fig. 3, HVIP is a point on the ground plane, which can participate in ground transform directly to establish the mapping from image pixels to 3D points on the ground plane. HVIP binds a person body's semantic point but it is not on the human body itself. Therefore, different from previous method [38] that forces people's ankle joints to be on the ground, which limits the posture of people, HVIP is determined by the 3D space position of people and can deal with people in various postures. Because the line from P_t to P_v is perpendicular to the ground, we have $P_t = P_v + d \times N$, where d represents the distance from P_t to the ground plane. We represent the projection points of P_v and P_t with $p_v = (u_v, v_v)$ and $p_t = (u_t, v_t)$, respectively. Refer to Eq. (2) and Eq. (3), we deduce

$$P_v = -\frac{D}{N^T K^{-1} \bar{p}_v} \times K^{-1} \bar{p}_v, \tag{5}$$

$$d = \frac{f \times y_v - (v_t - c_y) \times z_v}{(v_t - c_y) \times z_n - f \times y_n}. \tag{6}$$

Therefore, when the network predicts p_v and p_t, P_t can be uniquely determined. Then, the predicted body mesh M_{cam} in the global camera space follows $M_{cam} = M - P_{t-smpl} + P_t$, where $M \in \mathbb{R}^{6890\times3}$ and $P_{t-smpl} \in \mathbb{R}^3$ are the predicted vertices and torso center in SMPL [25] space, respectively. Finally, considering the cropping, our network takes the cropped image as input and predicts the local $p_{v-local}$ and $p_{t-local}$ on the cropped image. We have $p_v = p_{v-local} + t_{crop}$ and $p_t = p_{t-local} + t_{crop}$, where t_{crop} represents the pixel coordinates of the upper left corner of the cropped image. Our progressive position transform with HVIP builds a mapping from some pixel points on the cropped image to the global 3D location, which can simply and effectively predict the precise crowd positions.

3.3.2 Representations

Input Parameters. The network cannot perceive the whole scene information only from the cropped image, hence we take the estimated ground and camera parameters as extra inputs. We define the camera input as $(\frac{f}{W_s}, \frac{\hat{c}_x - c_x}{c}, \frac{\hat{c}_y - c_y}{c})$ which includes the information of FOV of scene and the principal point shift, where W_s, $(\hat{c}_x, \hat{c}_y)$ and c are the width of large-scene image, the image center of the cropped image and the size of the cropped image, respectively.

Body Center Heatmap. The body center heatmap $\boldsymbol{C_m}$ represents the body center likelihood by a Gaussian kernel combining body scales, where $C_m \in \mathbb{R}^{1\times H\times W}$ and $H = W = 64$. We define the body center the same as [34].

Torso Center Offset Map. Although we can directly define the body center as the 2D torso center, in practice, the body center heatmap tends to find a person's body salient point, especially when the person is occluded. Therefore, it is necessary to predict the human torso center separately. The torso center offset map $T_m \in \mathbb{R}^{2\times H\times W}$ contains the offset between 2D torso center $p_{t-local}$ and body center.

2D HVIP Map. The goal of 2D HVIP map $H_m \in \mathbb{R}^{1\times H\times W}$ is to obtain the 2D HVIP projection $p_{v-local}$ on the cropped image. The line from P_v to P_t is parallel to the ground normal, following the perspective theory, we have the projection points p_v, p_t and the vanishing point of the ground normal p_{vp} are collinear on image, where $p_{vp} = KN$. Therefore, we only need to estimate the 1D length from p_t to p_v to obtain 2D HVIPs.

SMPL Map. The SMPL map $S_m \in \mathbb{R}^{145\times H\times W}$ includes the parameters of SMPL [25] of people and a small 3D offset δt. The SMPL parametric model can represent various shape and pose with a small number of parameters. It takes the pose parameters θ and the shape parameters β as inputs and outputs a body mesh $M \in \mathbb{R}^{6890\times3}$. We adopt the 6D rotation representation [46] and drop the last two hand joints. Considering the error of the dataset annotations, we predict an offset δt to further refine the position of people by $P_t = P_v + d \times N + \delta t$.

3.3.3 Loss Function

Crowd3DNet is supervised by the weighted sum of multiple loss terms as follows:

$$\begin{aligned} L =& \lambda_{\text{center}} L_{\text{center}} + \lambda_{\text{mesh}} L_{\text{mesh}} + \lambda_{\text{hvip}} L_{\text{hvip}} \\ &+ \lambda_{\text{tc}} L_{\text{tc}} + \lambda_{\text{root}} L_{\text{root}} + \lambda_{\text{gn}} L_{\text{gn}} + \lambda_{\text{out}} L_{\text{out}}, \end{aligned} \tag{7}$$

$$\begin{aligned} L_{\text{mesh}} =& \lambda_{\text{pose}} L_{\text{pose}} + \lambda_{\text{shape}} L_{\text{shape}} + \lambda_{\text{j2D}} L_{\text{j2D}} \\ &+ \lambda_{\text{j3D}} L_{\text{j3D}} + \lambda_{\text{paj3D}} L_{\text{paj3D}} + \lambda_{\text{gm}} L_{\text{gm}}, \end{aligned} \tag{8}$$

where L_{center} is the 2D focal loss [23], and L_{mesh} is the common SMPL related L_2 loss including pose parameter

loss L_{pose}, shape parameter loss L_{shape}, 2D joint projection loss L_{j2D}, 3D joint loss L_{j3D} and 3D joint loss after Procrustes alignment L_{paj3D}. L_{hvip}, L_{tc} and L_{root} are all L_2 losses, which are used to supervise 2D HVIP projection, 2D torso center and absolute root position, respectively. $L_{\text{gn}} = L_{\cos}(P_s - P_a, N)$ is a ground normal regularization term to enhance the interaction consistency between people and ground plane, where $P_s - P_a$ is the approximated craniocaudal direction of human. We also use an out-of-bound loss to prevent people from penetrating the ground. More concretely, we use L_1 loss to punish the point with the most serious penetration into the ground plane, and the out-of-bound loss is defined as

$$L_{\text{out}} = \left|\min(\{\bar{v}_i \cdot G \mid \bar{v}_i \cdot G < 0\})\right|, \tag{9}$$

where $v_i \in M_{\text{cam}}$ and $G = [N^T, D]^T$.

3.4. Scene-specific Optimization and Merging

To improve the harmony between reconstructed crowd and scene, and the generalization to various camera and ground plane parameters, we add a scene-specific optimization for a new scene at test time. Please note that the scene-specific optimization is performed only once for a camera-fixed scene, *i.e.*, only one image of the scene is needed. Specifically, given a new scene at test time, we optimize a small set of weights in the head layer of Crowd3DNet with the ground normal and 2D poses estimated in the camera and ground plane module. The optimization loss L_{opt} is

$$L_{\text{opt}} = \lambda_{\text{j2D}} L_{\text{j2D}} + \lambda_{\text{gm}} L_{\text{gm}} + \lambda_{\text{gn}} L_{\text{gn}} + \lambda_{\text{out}} L_{\text{out}}. \tag{10}$$

We finally remove duplicated persons in the overlapped adjacent patches by merging. The merging operation retains the people farther away from the boundary of the overlapped region, which tends to keep more complete people to avoid truncation.

4. Experiments

4.1. Large-scene Crowd Dataset

To train and evaluate crowd reconstruction in a large scene, we contribute *LargeCrowd*, which is a benchmark dataset with over 100K labeled humans in 733 gigapixel images (19200×6480) of 9 different scenes (5 scenes for training and 4 scenes for testing). The images are extracted at a minimum interval of $3s$ from gigapixel streams which are captured by a ZoheTec JMC315 array camera. We annotate the bounding boxes, 2D poses and 2D HVIPs of all the visible people in the images, with the maximum error less than 5 pixels for 95% labels. We measure 3D landmarks in a world coordinate system and label the corresponding 2D points to solve the camera extrinsic matrix for each scene. Then, we compute the homography matrix for each ground plane. The homography matrices with labeled ground segmentations and HVIPs provide the true physical positions of the persons.

Table 1. Comparison on *LargeCrowd* dataset.

Method	PPDS↑	PA-PPDS↑	PCOD↑	OKS↑
SMAP [45]-Large	58.60	60.07	70.14	61.25
CRMH [13]-Large	59.16	64.79	80.25	67.24
BEV [35]-Large	74.21	75.05	87.31	66.15
Crowd3D w/o HVIP	80.45	88.95	92.42	64.17
Crowd3D	**81.53**	**89.36**	**92.63**	**71.72**

4.2. Implementation Details

We use HRNet-32 [5] as backbone, each head of which is composed of two ResNet [11] blocks with batch normalization. We resize input images to 512×512 with zero padding to keep the same aspect ratio. We also use the collision-aware representation of ROMP [34] to push apart close body centers. Our training process has two stages: 1) start by training the body center heatmap, torso center offset map and 2D HVIP map for 15 epochs to make sure that the subsequent learning about body mesh has a suitable initial position; 2) train the full model with all losses for 70 epochs. We implement Crowd3DNet with PyTorch and adopt the Adam [16] as optimizer with 5e-5 learning rate. We train our model on *LargeCrowd*, *Agora* [31], *MuCo-3DHP* [28] and a single person dataset *Human3.6M* [12].

4.3. Evaluation Metrics

We use the torso center as the location of people and evaluate the location distribution of crowd by the distances between people. We define a metric called pair-wise percentual distance similarity (PPDS) as

$$PPDS = \frac{\sum_{k=1}^{n-1}\sum_{i=k+1}^{n} 1 - \min\left(d_{ik}, 1\right)}{C_n^2}, \tag{11}$$

$$d_{ik} = \left| \frac{\|E_k - E_i\| - \|G_k - G_i\|}{\|G_k - G_i\|} \right|, \tag{12}$$

where n is the number of people in the image, and E_i and G_i represent the estimated and ground-truth locations of the i-th person, respectively. To evaluate the relative crowd distribution, we also define the procrustes-aligned pair-wise percentual distance similarity (PA-PPDS) which aligns the reconstructed crowd and the ground truth by Procrustes alignment to exclude the influence of scale and rotation. Due to the lack of 3D pose annotations, we use the object keypoint similarity (OKS) [24] to evaluate the 2D poses. The percentage of correct ordinal depth (PCOD) [45] is used to evaluate the ordinal depth relations between all pairs of people in the image.

4.4. Comparison

Because no existing methods can directly handle large-scene images with hundreds of people, we compare our method with three baselines that are modified from the state-of-the-art methods: SMAP [45], CRMH [13], and

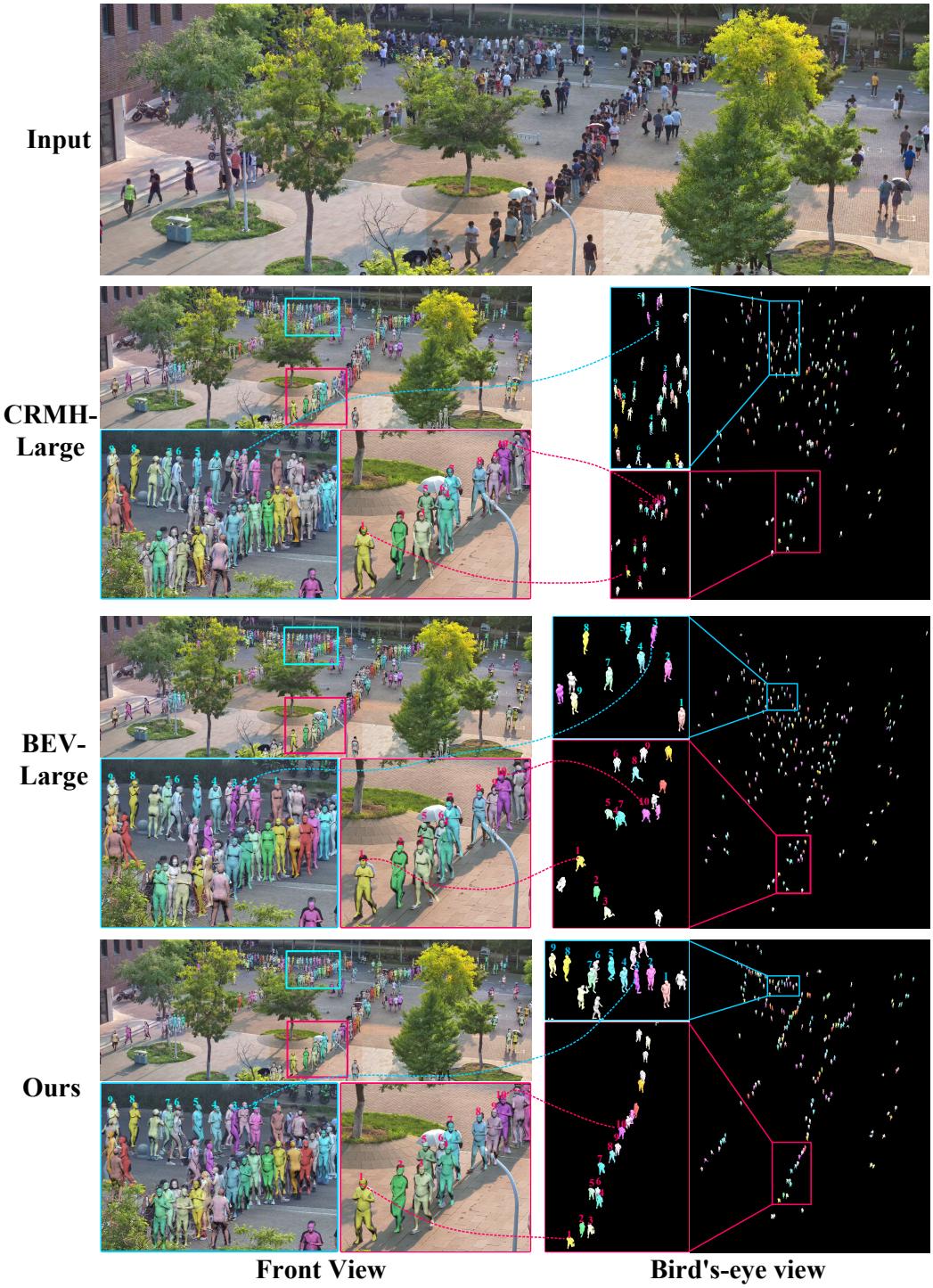

Figure 4. Qualitative results on *LargeCrowd*. The same color or number corresponds to the same person, and gray indicates that the person is not matched.

BEV [35]. We denote these baselines as SMAP-Large, CRMH-Large and BEV-Large. Specifically, we first use our adaptive human-centric cropping to obtain the hierarchical cropped images as their inputs and infer the respective reconstructed results on the cropped images. To obtain the global reconstruction results for these methods, we provide the scene-camera intrinsics estimated by our method to them. For CRMH [13] which predicts human bodies in bounding boxes by a weak perspective camera model, we use its transform from bounding box position to depth of

full image to infer the predicted locations in the global camera space. Both SMAP [45] and BEV [35] infer the locations through perspective camera models. Following previous method [1], we scale the depths of their results according to the focal length of the scene. Please refer to supplementary material for more details. For fair comparison, we fine-tune all the compared methods on *LargeCrowd*. Table 1 gives the quantitative results. Our method outperforms other approaches in terms of all the metrics. Especially, the obvious advantage in PPDS, PA-PPDS and PCOD shows that our method can predict accurate crowd location distribution, including physical distances and relative arrangements. Fig. 4 shows qualitative comparison results. The complete bird's-eye view on the right shows that our predicted crowd distribution is consistent with the input image, while the compared methods are not consistent. Taking the persons labeled with numbers for example, only our method recovers correct relative positions. The reconstructed people by the existing methods independently inferred from the cropped images are inconsistent in the global large-scene camera space. Besides, although these methods show reasonable projection results, the wrong global positions mean that their predicted 3D human bodies have wrong scales. We also provide comparison results on public small-scene datasets in supplementary material.

4.5. Ablation Study

Impact of the Number of People on the Estimated Ground and Camera. We explore the impact of the number of people on estimating ground and camera parameters by controlling the number of people used in optimization, and the newly added people are randomly selected. The metrics include a cosine distance for ground normal and a root mean square error for focal length. We calibrate the camera and obtain the ground-truth focal length (about 27000). As shown in Fig. 5, more than ten people are enough for estimating ground normal, which is common in real-world large-scale scenes, especially surveillance scenarios with hundreds of people. The focal length is not sensitive to the number of people.

Progressive Position Transform Based on HVIP. Our progressive position transform based on HVIP effectively helps the network to predict accurate global 3D positions of people. To verify this, we compare our full model with a variant of Crowd3DNet, Crowd3D w/o HVIP, which predicts 2D ankle joints and adopts the midpoint of ankle joints to participate in ground transform without using HVIP. The result is shown in Table 1. Benefiting from HVIP, which makes use of the ground plane without restricting human posture, Crowd3DNet has obvious advantages on OKS.

Adaptive Human-centric Cropping. To verify the adaptive human-centric cropping scheme, we denote a metric called cropping score, which counts the ratio of people with the appropriate scale after cropping. For a person with appropriate scale, we set the ratio of his height to the corresponding cropped image within [0.3, 0.8], and he is not truncated. The comparison result on *LargeCrowd* between adaptive human-centric cropping and uniform cropping is 0.923 *vs.* 0.806, which demonstrates the effectiveness of our adaptive human-centric cropping scheme.

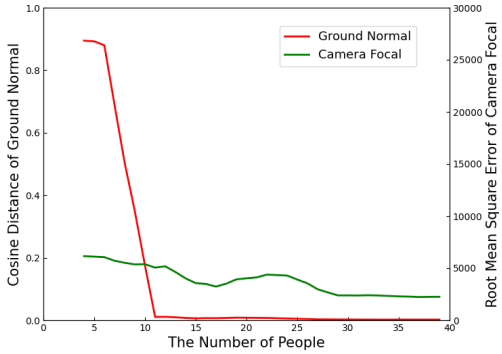

Figure 5. The impact of the number of people on the camera and ground plane estimation.

5. Conclusion and Discussion

Conclusion. We propose Crowd3D to reconstruct hundreds of people with global consistency from a single RGB large-scene image. Our method is a joint local and global inference framework which converts the complex crowd localization into pixel localization by our defined HVIP concept and the parameters of pre-estimated scene-level camera and ground plane. Our adaptive human-centric cropping scheme and progressive position transform based on HVIP solve the challenges of large number of people, large variations in human scale and complex spatial distribution in large scenes. We also contribute a large-scene dataset called *LargeCrowd* to help train and evaluate crowd reconstruction in large scenes with hundreds of people. Experimental results demonstrate that our method can achieve globally consistent crowd reconstruction in large scenes.

Limitations and Future Work. We focus on outdoor real-world large-scale scenes which contain one or several ground planes. Our method may be easily extended to multi-ground scenes by using the existing image-based ground plane segmentation methods or manual segmentation, which is taken as our future work. Although our Crowd3D shows effective crowd reconstruction in a global camera space, there are still some cases that we cannot solve well, *e.g.*, the people in complex ground conditions and the persons with complicated postures or severe occlusions. In future work, we will focus on a wider range of large-scale scenes with complex ground and crowd environments.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (62122058, 62125106, 61860206003 and 62171317).

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6DoF, face pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 8

[2] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. PandaNet: Anchor-based single-shot multi-person 3D pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6856–6865, 2020. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. European Conference on Computer Vision*, pages 561–578, 2016. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 2

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 6

[6] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[7] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3D pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2020. 2

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proc. IEEE/CVF International Conference on Computer Vision*, 2017. 4

[9] Xiaohan Fei, Henry Wang, Lin Lee Cheong, Xiangyu Zeng, Meng Wang, and Joseph Tighe. Single view physical distance estimation using human pose. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 12406–12416, 2021. 4

[10] Cristiane BR Ferreira, Helio Pedrini, Wanderley de Souza Alencar, William D Ferreira, Thyago Peres Carvalho, Naiane Sousa, and Fabrizzio Soares. Where's Wally: A gigapixel image study for face recognition in crowds. In *International Symposium on Visual Computing*, pages 386–397. Springer, 2020. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[12] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 6

[13] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 1, 3, 6, 7

[14] 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 *Proc. IEEE/CVF International Conference on Computer Vision*, pages 3334–3342, 2015. 3

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2

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

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 2

[18] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 2

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[20] Jogendra Nath Kundu, Ambareesh Revanur, Govind Vitthal Waghmare, Rahul Mysore Venkatesh, and R Venkatesh Babu. Unsupervised cross-modal alignment for multi-person 3D pose estimation. In *Proc. European Conference on Computer Vision*, pages 35–52, 2020. 2

[21] Vincent Leroy, Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, and Grégory Rogez. SMPLy benchmarking 3D human pose estimation in the wild. In *Proc. IEEE International Comference on 3D vision*, 2020. 3

[22] Lingling Li, Xiaohui Guo, Yan Wang, Jingjing Ma, Licheng Jiao, Fang Liu, and Xu Liu. Region NMS-based deep network for gigapixel level pedestrian detection with two-step cropping. *Neurocomputing*, 468:482–491, 2022. 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 2980–2988, 2017. 5

[24] 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 *Proc. European Conference on Computer Vision*, pages 740–755, 2014. 6

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned

multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 2, 5

[26] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proc. IEEE International Comference on 3D vision*, pages 506–516, 2017. 3

[27] 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. *ACM Transactions on Graphics*, 39(4):82–1, 2020. 2

[28] 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 *Proc. IEEE International Comference on 3D vision*, 2018. 2, 3, 6

[29] 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 *Proc. IEEE/CVF International Conference on Computer Vision*, pages 10133–10142, 2019. 2

[30] Gyeongsik Moon and Kyoung Mu Lee. Pose2Pose: 3D positional pose-guided 3D rotational pose prediction for expressive 3D human pose and mesh estimation. *arXiv preprint arXiv:2011.11534*, 2020. 2

[31] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 3, 6

[32] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019. 2

[33] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1146–1161, 2019. 2

[34] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 3, 4, 5, 6

[35] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 4, 6, 7, 8

[36] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019. 2

[37] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 1

[38] 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 *Proc. IEEE International Comference on 3D vision*, pages 53–63, 2021. 3, 5

[39] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proc. European Conference on Computer Vision*, 2018. 3

[40] 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 *Proc. European Conference on Computer Vision*, page 242–259, 2020. 2

[41] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3268–3278, 2020. 1

[42] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, 2020. 2

[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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11446–11456, 2021. 1, 2

[44] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 546–556, 2021. 3, 4

[45] 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 *Proc. European Conference on Computer Vision*, pages 550–566, 2020. 2, 6, 8

[46] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 5