arXiv:2411.06232v2 [cs.CV] 23 May 2025

RCR: Robust Crowd Reconstruction with Upright Space from a Single Large-scene Image

Jing Huang[1], Hao Wen[1], Tianyi Zhou[1], Haozhe Lin[2], Yu-Kun Lai[3], Kun Li[1*]

[1*]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China.
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 10084, China.
[3]School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, Unitied Kingdom.

*Corresponding author(s). E-mail(s): lik@tju.edu.com;
Contributing authors: hj00@tju.edu.com; wenhao@tju.edu.com; tianyi@tju.edu.com; linhz@tsinghua.edu.com; Yukun.Lai@cs.cardiff.ac.uk;

Abstract

This paper focuses on spatially consistent hundreds of human pose and shape reconstruction from a single large-scene image with various human scales under arbitrary camera FoVs (Fields of View). Due to the small and highly varying 2D human scales, depth ambiguity, and perspective distortion, no existing methods can achieve globally consistent reconstruction with correct reprojection. To address these challenges, we first propose a new concept, Human-scene Virtual Interaction Point (HVIP), to convert complex 3D human localization into 2D-pixel localization. We then extend it to RCR (Robust Crowd Reconstruction), which achieves globally consistent reconstruction and stable generalization on different camera FoVs without test-time optimization. To perceive humans in varying pixel sizes, we propose an Iterative Ground-aware Cropping to automatically crop the image and then merge the results. To eliminate the influence of the camera and cropping process during the reconstruction, we introduce a canonical Upright 3D Space and the corresponding Upright 2D Space. To link the canonical space and the camera space, we propose the Upright Normalization, which transforms the local crop input into the Upright 2D Space, and transforms the output from the Upright 3D Space into the unified camera space. Besides, we contribute two benchmark datasets, *LargeCrowd* and *SynCrowd*, for training and evaluating crowd reconstruction in large scenes. Experimental results demonstrate the effectiveness of the proposed method. *The source code and data will be publicly available for research purposes.*

Keywords: multi-human pose and shape, single image, camera space, various FoVs.

1 Introduction

3D pose, shape, and location reconstruction for hundreds of people from a single image will help model crowd behavior for simulation and security monitoring. In such surveillance scenes, human pixel sizes are small and variable, and the number of humans can also be large; therefore, these images can be named *large-scene images*. In this paper, we aim to reconstruct 3D poses, shapes,

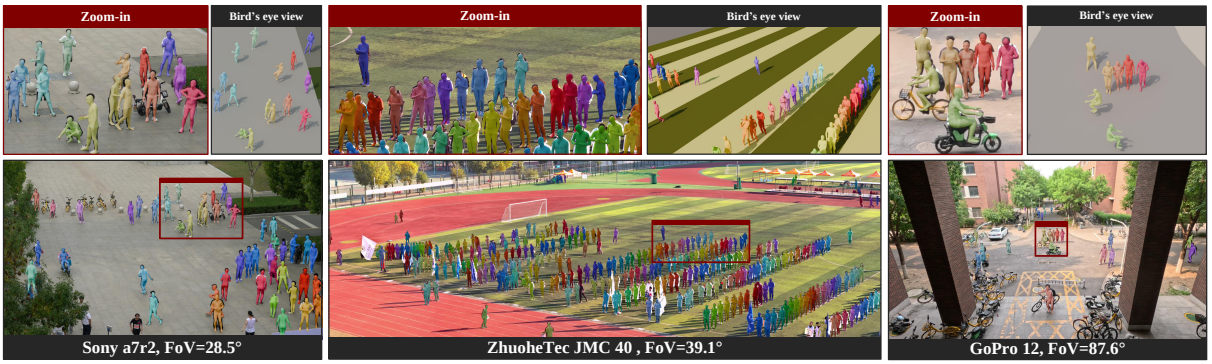


Fig. 1 Given a single image under arbitrary FoVs, our method can reconstruct human poses and shapes in a unified camera space with both spatial consistency and reprojection accuracy.

and locations of hundreds of people in the camera space from a single large-scene image with unknown and arbitrary camera parameters, as shown in Fig. 1.

Existing multi-person poses and shapes reconstruction methods [1–3] can estimate multiple human pose and shape parameters from a single image directly, but they can only handle small scenes where people occupy a significant portion of the image, such as a group photo. For a large-scene image, like *PANDA* [4], with small and varying human scales, these methods will result in a reconstruction that misses a large portion of people. On the other hand, a combination of human detection and single human reconstruction [5, 6] typically offers better adaptability to people of various scales in images. However, this combination has a limitation in terms of reconstruction outputs being in separate 3D spaces. Besides, adding a depth estimation branch to translate the single human reconstruction results to the camera space will lead to incorrect reprojection and relative positions of the humans. Recently, GroupRec [7] has made significant progress in large-scene crowd reconstruction. GroupRec adopts the pipeline of single reconstruction after detection, and proposes a hypergraph relational reasoning network to encode the inter-human and inter-group relations. It alleviates the reprojection issue in a regular FoV case by a test-time optimization strategy with a depth initialization [8] and additional correlations provided by the hypergraph network.

However, though GroupRec is enhanced by a test-time optimization with a single-plane prior and reprojection losses, it still suffers from the depth ambiguity and the perspective distortion in both regular and wide FoV cases, which leads to incorrect reprojection and relative positions of the humans, as shown in Fig. 2.

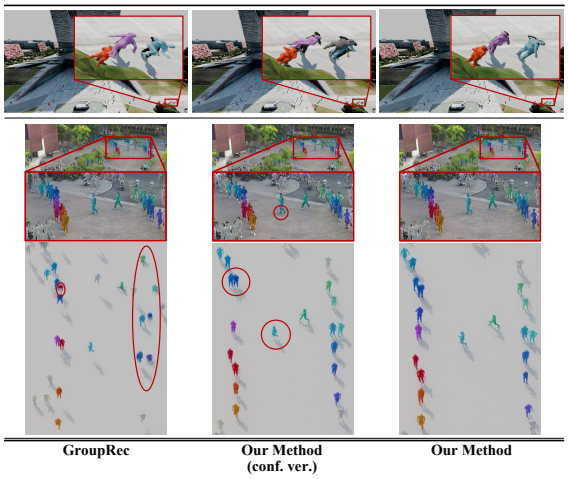


Fig. 2 Our method achieves accurate reprojection (top line), correct relative positions, and plausibility of human-ground interaction (bottom line), while the state-of-the-art method [7] and our conference version [9] do not perform well in these aspects.

In general, the main challenges for generalizing to arbitrary camera parameters in large-scene crowd reconstruction are: (1) Due to the depth ambiguity from a single view, it is difficult to directly estimate absolute 3D positions and 3D poses of people in a large scene; (2) there are a large number of people with relatively small and highly varying 2D scales; (3) the camera parameters and the cropping operations affect the 2D features of an actual 3D human, but learning how they affect the reconstruction results is difficult

due to the lack of ground-truth 3D annotations in various camera parameters. (4) There are no large-scene image datasets with hundreds of people to train and evaluate crowd reconstruction in large scenes.

In response to these challenges, we propose Robust Crowd Reconstruction (RCR), which achieves globally consistent crowd reconstruction in a unified camera space from a single large-scene image. To solve the depth ambiguity of 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 2D HVIP estimation problem with a robust human keypoint-based camera and ground estimation approach. To deal with a large number of people and various human scales, a simple idea is to crop the image into multiple local crops and then merge the results, which requires the local human pixel heights and positions as the cropping reference. However, the accurate local height information cannot be easily obtained by a coarse detection result, *e.g.*, detection from uniformed cropped images, because the coarse detection results do not contain enough information for how to better crop the image, *i.e.*, the maximum and minimum human pixel heights in a local area. Therefore, we propose an Iterative Ground-aware Cropping strategy that determines the cropping sizes by the actual 3D-2D mapping from the estimated ground plane and solves the circular dependency problem among the three processing steps: cropping, keypoints detection, and human keypoint-based camera and ground estimation.

To eliminate the influence of camera parameters and cropping operations during the single human reconstruction we define a canonical Upright 3D Space (U3 Space) and a corresponding Upright 2D Space (U2 Space) where the ground normal is the y-axis of U3 Space, and the projection from U3 Space to U2 Space is a fixed orthographic projection. To link the defined canonical space, we propose a Upright Normalization to transform the local crop input into the U2 Space and transform the output from the U3 Space into the unified camera space with the estimated 2D HVIP. We also propose an HVIPNet to estimate the 2D HVIP in the U2 Space where the 2D semantic scale and the 2D orientation of a human are normalized, For local human pose and shape, a single human reconstruction method with weak perspective modeling estimates these parameters in the U3 Space.

We also collect and annotate *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 scenes. To the best of our knowledge, this is the first large-scene crowd dataset that enables training and evaluation on large-scene images with hundreds of people. Besides, we also synthesize a virtual dataset *SynCrowd* at the 9600×5400 image resolution with ground-truth SMPL [10] label by the publicly released scenes and scanned human models for quantitative testing under various camera parameters.

To summarize, our main contributions include:

- We propose iterative ground-aware cropping to ensure stable human detection in various scales by the depth cues from the ground plane with a keypoints-based ground plane estimation and an iterative strategy.
- We propose the Human-scene Virtual Interaction Point (HVIP) concept and a keypoint-based camera and ground estimation approach to solve the depth ambiguity problem.
- We propose the Upright 3D Space and Upright 2D Space with the Upright Normalization to eliminate the influence of camera parameters and cropping operations during the single-person reconstruction.
- We contribute *LargeCrowd*, a benchmark dataset with over 100K labeled crowded people, and synthesize a virtual dataset *SynCrowd* for quantitative evaluation under various camera parameters.

This work is a significant extension to our earlier conference work Crowd3D [9] with the following enhancements:

- **Canonical Regression Space.** We propose the Upright 3D/2D Space and the corresponding Upright Normalization (Sec. 3.4) to eliminate the influence of camera parameters and cropping operations, while Crowd3D reconstructs humans directly in the camera space.
- **Automatic Cropping without Manual Settings.** We propose Iterative Ground-aware Cropping to automatically determine cropping

sizes by the actual 3D-2D mapping (Sec. 3.3), while the cropping of our conference version [9] determines the cropping sizes with a geometric sequence assumption and requires the maximum and minimum human pixel heights and positions as manual inputs for each image.
- **New Pipeline.** We propose a top-down pipeline with a single-person estimation network and a new HVIPNet (Sec. 3.4.2), while Crowd3D uses a multi-person estimation network for cropped square images.
- **New Datasets.** We contribute a synthesized dataset *SynCrowd* with SMPL labels for quantitative evaluation under various camera parameters, since previous large-scene datasets only provide 2D keypoints annotations and images with the same camera parameters.

These technical enhancements lead to the following improvements in effectiveness: (1) the reprojection performance is about 15% better than the conference version, and the bottom row of Fig. 2 illustrates the improvement of the reprojection and plausibility of human-ground interaction; (2) RCR can generalize to arbitrary FoV cases without any test-time optimization, while the conference version cannot, even with a test-time optimization, as shown in the top row of Fig. 2; (3) RCR can automatically process every new scene, while the conference version must manually set the cropping parameters.

2 Related Work

2.1 Monocular Single Person Pose and Shape Estimation

In most cases, human pose and shape estimation is formulated as the estimation of parameters of a parametric human model [10, 11]. For single-image inputs, there are optimization-based methods and regression-based methods. Optimization-based methods are performed by fitting the SMPL model to the 2D keypoints and silhouettes [11, 12]. In contrast, regression-based methods can be much faster [13, 14], and are not limited to the SMPL parameters representation. Some researchers propose the methods to use vertex representation with transformers [15], SMPL UV representation with an image-to-image network [16], and graph representation with graph neural networks [17]. The current state-of-the-art single-image methods remain the most popular approach, which is to regress the SMPL parameters using MLPs [6] or transformers [5] after the image feature extraction. PyMAF-X [6] focuses on image alignment performance with a coarse-to-fine feedback loop, and 4D-Humans [5] exploits the general transformer structure. For video input, recent methods are capable of utilizing temporal information to or output the results in camera space from videos captured by a static camera [18], or decouple the world space 3D pose and shape sequence and camera pose sequences from videos captured by a moving camera [19–21] with SLAM methodology or initialization, But this 3D localization ability of video-based methods can not be used for a single image. In general, all these single-image methods can not output results in camera space because they use a fixed weak perspective camera model to be compatible with the various 3D and 2D datasets.

2.2 Monocular Multi-person Pose and Shape Estimation

A closely related task is multi-person 3D pose estimation methods [22–28], which can be achieved by lifting 2D joint detections to 3D space in the depth directions to ensure reprojection performance. However, these pose-only methods sacrifice the plausibility of human pose and shape in the process of 3D lifting, so that they can not ensure important human priors, such as a natural body shape or equal-length arms and legs. For both human pose and shape, multi-person methods [1–3, 29, 30] typically estimate the SMPL [10, 11] pose and shape parameters to enforce human prior constraints similar to the single-person methods. Some methods ignore the 3D locations [2, 30]. ROMP [2] directly regresses multiple human SMPL parameters in a single stage. 3DCrowdNet [30] first detects 2D keypoints of all the individuals and leverages 2D poses to guide a joint-based regressor in estimating the human model parameters to address occlusions. Some methods can achieve camera space reconstruction, but with a fixed camera focal length [1, 3, 29]. CRMH [3] proposes a top-down architecture consisting of an R-CNN head and a SMPL parameter branch with additional 3D interpenetration loss and a depth ordering-aware loss to ensure the

coherence of the human meshes. BMP [29] is a single-stage method with a body mesh map, which also proposes an inter-instance ordinal depth loss to constrain the depth order. BEV [1] improves ROMP [2] by an additional bird's-eye view representation to simultaneously reason the 2D image positions and depths of the body center. Therefore, all of the above multi-person methods can not estimate results in camera space from an entire large-scene image due to the relatively small and varying human scales compared to the image size.

2.3 Pose and Shape Estimation from a Large-scene Image

The above multi-person methods can not reconstruct the 3D pose and shape in a unified camera space from a single large-scene image, which includes small and varying human pixel scales and is common in surveillance scenarios. Recently, GroupRec [7] has made significant progress in this topic. GroupRec proposes a hypergraph relational reasoning network to exploit the collectiveness and social interaction among crowds. It initializes each person's depth [8] and refines the results by a test-time optimization strategy with additional group features provided by the hypergraph network to alleviate the reprojection issue. Our conference version Crowd3D [9] proposes the concept of Human-scene Virtual Interaction Point (HVIP) with a human-centric ground estimate to solve the depth ambiguity. Specifically, it is a multi-stage method that crops the image, regresses the poses and shapes of multiple humans in the cropped image, and finally merges the results of the cropped images. Additionally, Crowd3D is also equipped with a self-supervised scene-specific optimization to refine the reprojection performance. However, training a network to perceive the influence of different camera FoVs and cropping positions is difficult, especially given the scarcity of datasets with diverse FoVs. Therefore, as shown in Fig. 2, the reprojection performance of Crowd3D and GroupRec is poor under arbitrary FoVs, with Crowd3D failing to ensure reasonable human-ground interactions, and GroupRec even producing incorrect relative positions or depths.

In this paper, we propose RCR (Robust Crowd Reconstruction) with Upright Space to reconstruct multiple humans from a single large-scene image under arbitrary camera FoVs without test-time optimization. To enable training and testing in large scenes, we contribute the *LargeCrowd* dataset with human keypoints level annotations and the *SynCrowd* dataset with 3D SMPL ground-truths and different camera FoVs.

3 Method

Our goal is to reconstruct 3D poses and shapes of all people from a single large-scene image as input. Our previous conference version [9] proposes a multi-person estimation network for cropped square images with an HVIP concept to locate each person and a simple cropping strategy that requires manually inputting the maximum and minimum human pixel heights.

In this paper, we keep the HVIP concept to deal with the depth ambiguity (Sec. 3.2) and propose a new pipeline, RCR, as illustrated in Fig. 3, which first detects each individual from the cropped images and then estimates the multi-person's pose and shape parameters and the 2D HVIP of each individual. To handle small and varying human scales in the image, we design an iterative ground-aware cropping (Sec. 3.3) to detect the keypoints of all individuals and estimate the camera and ground parameters in a loop without any manual inputs. To eliminate the influence of the camera parameters during the reconstruction, we reconstruct each detected human with the proposed *Upright 3D Space* (U3 Space) and *Upright 2D Space* (U2 Space) where the ground normal vector is parallel to the vertical axis and weak perspective projection is applicable (Sec. 3.4). We also propose an Upright Normalization to convert the cropped image to the U2 Space with the estimated camera and ground parameters. To locate the absolute 3D position of each individual in the camera space, we propose an HVIPNet (Sec. 3.4.2) to estimate 2D HVIP (Sec. 3.4.2) and reconstruct the SMPL parameters (Sec. 3.4.3) in the U3 Space. Finally, we convert all the results back to the camera space using the proposed Upright Normalization transform so that we achieve globally consistent reconstruction in the camera space.



Fig. 3 Overview of our method. Our method first detects all the humans and then reconstructs each person via the upright space (details in Fig. 4). With the help of the canonical Upright Space, though the SMPL estimation is separately conducted in the Upright 3D Space, the results can achieve both spatial consistency and reprojection accuracy in the camera space.

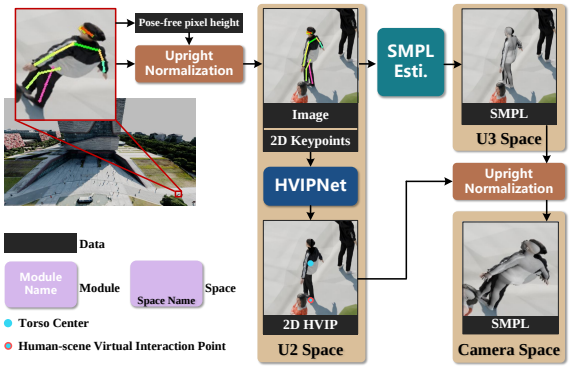


Fig. 4 Reconstructing each person via Upright 2D/3D Space, Upright Normalization, and HVIPNet.

3.1 Formulation

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 and assign K to be the camera intrinsic matrix. We formulate a single and arbitrary ground plane in the camera space 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 a point on the ground plane and $D \in \mathbb{R}$ constrains the distance from the camera center to the ground plane. The tuple $G = \{N, D\}$ is called the ground parameter. The projection from the camera space to the image space is defined as: $p = F_{\text{proj}}(K, P)$, where $P \in \mathbb{R}^3$ is a 3D point and $p = (u, v)$ is the 2D point in the image space. The reverse projection of a 2D point p from the image space to the camera space with the ground constrain is defined as: $P_g = F_{\text{rev-proj}}(K, N, D, p_g)$, where $p_g \in \mathbb{R}^2$ is the 2D point in the image, and $P_g \in \mathbb{R}^3$ is the 3D point in the camera space and on the ground plane (satisfying forward projection and $N^T P_g + D = 0$). Our method eventually output camera space SMPL [10] parameters $\theta \in \mathbb{R}^{72}$, shape $\beta \in \mathbb{R}^{10}$, scale $S_{\text{toCam}} \in \mathbb{R}^1$ and $T_{\text{toCam}} \in \mathbb{R}^1$, where θ and β are the pose and shape parameters of SMPL, S_{toCam} and T_{toCam} are the scale and translation from the SMPL space to the camera space which should be applied after the SMPL forward function $\mathcal{M} : (\mathbb{R}^{72}, \mathbb{R}^{10}) \to \mathbb{R}^{6890\times 3}$.

3.2 Human-scene Virtual Interaction Point (HVIP)

To help infer the accurate 3D locations of persons in the large-scene camera system, we inherit the concept of *Human-scene Virtual Interaction Point (HVIP)* from our conference version [9], which is the projection point of a person's 3D torso center on the ground plane in the camera space where the 3D torso center $P_t \in \mathbb{R}^3$ is defined as the center (arithmetic mean in all dimensions) of 4 primary body joints (2 shoulder joints and 2 hip joints).

Progressive Position Transform. We represent the 2D corresponding points in the image space of P_v and P_t as p_v and p_t, respectively. Since the direction of $P_t - P_v$ 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. Given a 2D torso center $p_t = (u_t, v_t)$, a 2D HVIP $p_v = (u_v, v_v)$, the camera and ground parameters K and G, we first compute the 3D HVIP $P_v = (x_v, y_v, z_v)$ by $F_{\text{rev-proj}}(K, N, D, p_v)$ since the HVIP is on the ground plane. With the equations (1) $P_v = F_{\text{rev-proj}}(K, N, D, p_v)$, (2) $P_t = P_v + d \times N$ and (3) $p_t = F_{\text{proj}}(K, P_t)$, we reduce the unknowns to only one variable d:

$$d = \frac{f \times y_v - (v_t - c_y) \times z_v}{(v_t - c_y) \times z_n - f \times y_n}. \tag{1}$$

where f, c_y are the defined camera parameters in K, and $(x_n, y_n, z_n) = N$ is the ground normal vector. Therefore, the progressive position transform

is a mapping:

$$P_t = F_{\text{prog-pos}}(K, G, p_t, p_v). \quad (2)$$

which means we can the 3D torso center P_t in the camera space by the 2D torso center p_t and 2D HVIP p_v when the camera K and ground parameters G are known. Since the 2D torso center can be easily extracted from the 2D keypoints and an SMPL mesh has a clear 3D torso center via the definition, we are able to locate any SMPL mesh by estimating the 2D HVIP of this person.

3.3 Iterative Ground-aware Cropping

The off-the-shelf human detection models, *e.g.*, MMPose [31], AlphaPose [32], and Openpose [33], only perform well when a person occupies a specific range of proportions in the image. To deal with the large-scene image with small human scales, the image cropping approach needs to ensure the detection performance. Different from the cropping approach in our conference version [9], which requires manual setting of the maximum and minimum human pixel heights, we propose an iterative ground-aware cropping strategy to automatically crop the image and merge the results. Therefore, there are two key problems to be solved in this section: (1) automatically deciding the size of the local cropping area, and (2) avoiding missing and duplication after merging the results.

To address the first problem, we note that the camera and ground parameters can indicate the possible human pixel height of a local area in the image. Camera and ground are unknown when we need to crop the original image, and estimating ground and camera may need the human 2D keypoints detection results. This shows a circular dependency problem among the three processing steps: cropping, keypoint detection, and camera and ground estimation. To overcome the circular dependency, we propose an iterative ground-aware cropping strategy as follows: (1) The iterative ground-aware cropping is initialized by several attempts of uniform cropping with different scales (cropped sizes are set to 512, 1024, ..., until reaching 1/4 of the image size, thus usually only one attempt); (2) The detection part takes the overlapped cropped images with cropping bounding boxes as input and outputs the keypoints of all people in the whole image without duplication (Sec. 3.3.1); (3) The camera and ground estimation part takes the keypoints of all individuals as input and outputs the camera and ground parameters (Sec. 3.3.2); (4) The ground-aware cropping part takes the camera and ground parameters as input and outputs the cropping bounding boxes.

3.3.1 Detecting and Filtering duplication

This subsection takes the cropped images as input and outputs the keypoints of all individuals while avoiding duplication of the same person in the original image, since overlapping exists among the neighboring cropped images. We run an off-the-shelf human detection toolkit [31] on the cropped images. To avoid low-quality detection, we first remove detections that are truncated by the cropping boundary. This operation is reasonable because the cropping strategy guarantees that each person will appear fully in at least one of the cropped images, and RCR does not process people truncated by the original image boundary. Particularly, we filter out the truncated keypoints that meet both of the following conditions: (1) with at least one keypoint with a low confidence score (< 0.2), (2) its corresponding bounding box area is very close to an image boundary which has a neighboring cropped image. After this filtering step, the detection part transforms all the keypoints to the original image space. To remove duplicated detections, we first find several groups of keypoints that are close to each other in terms of a defined 2D Pose Similarity in a KD-tree structure, where the 2D Pose Similarity is defined as the average Euclidean distance of all high-confidence (> 0.35) keypoints normalized by the pixel height. For each group, we keep the detection with the highest average confidence score to avoid duplication. Finally, this subsection outputs the keypoints of all individuals in the original image.

3.3.2 Keypoint-based Camera and Ground Estimation

This subsection takes the keypoints of all individuals as input and outputs the camera and ground parameters following our conference version [9]. For each detected individual, we compute the midpoints of two shoulder keypoints as the

top point $p_{\text{top}} \in \mathbb{R}^2$ and the midpoints of two ankle keypoints as the bottom point $p_{\text{bot}} \in \mathbb{R}^2$ from the detected keypoints. We consider that (1) p_{bot} is on the ground plane, (2) the direction $p_{\text{top}} - p_{\text{bot}}$ is close to the ground normal vector N, (3) and deviation in direction follows an almost normal distribution to solve the camera K and ground parameters G. Specifically, we define the reprojected top points as $p'_{\text{top}} = F_{\text{proj}}(K, hN + F_{\text{rev-proj}}(K, N, D, p_{\text{bot}}))$, where $h \in \mathbb{R}$ is the fixed height prior. We then compute the loss function of one individual as:

$$\begin{aligned} L_{\text{g}} =& \lambda_{\text{angle}} L_{\cos}(p'_{\text{top}} - p_{\text{bot}}, p_{\text{top}} - p_{\text{bot}}) \\ &+ \lambda_{\text{mod}} \frac{\left| \|p'_{\text{top}} - p_{\text{bot}}\|_2 - \|p_{\text{top}} - p_{\text{bot}}\|_2 \right|}{\|p_{\text{top}} - p_{\text{bot}}\|_2}, \end{aligned} \tag{3}$$

where $L_{\cos}$ is the cosine distance, and λ_{angle} and λ_{mod} are the weights of the corresponding loss terms. The two terms penalize inconsistencies between the original 2D keypoints and reprojection from angular and modular perspectives. Finally, we minimize the average loss of all individuals to solve the camera K and ground parameters G. Note that this optimization can also take camera intrinsics K as an optional input and only estimate the ground plane G for better accuracy.

3.3.3 Ground-aware Cropping

This subsection takes the camera K and ground parameters G as input and outputs the cropped image to ensure any person in the original image will be completely included in at least one cropped image in a proper ratio. The key idea is to utilize the depth prior provided by the estimated camera and ground. We set the expected human scale s_{ratio} for a cropped image and define the minimum human scale in pixels as s_{minPixel}. Unlike the cropping hyperparameters in our conference version [9], which need to be reset for different inputs, these values are only related to the capability of the detection toolkit. For example, when using a detection toolkit with higher precision in detecting human keypoints at very low resolutions, the value of s_{ratio} and s_{minPixel} may be set to a smaller value. For MMPose [31], they are set to 0.5 and 60, respectively. The algorithm is described in Algorithm 1 where the maximum human height in 3D space $H_{\text{personMax}}$ can be set to 2.1 meters, function F_{getMaxH} accepts a horizontal line referred to as d_{bot}, ground plane, camera intrinsic matrix, and the 3D length of a vector, and outputs the maximum 2D length of this vector in pixels if this vector starts from the ground plane, parallel to the ground normal and starts anywhere on this 2D horizontal line. This length can be negative which means the vector is upside down. This algorithm returns the cropping positions C_{bbox} of the cropped images.

Algorithm 1 Ground-aware cropping

Input: Camera intrinsic matrix K, Ground plane $G = \{N, D\}$, s_{minPixel}, maximum human height in 3D space $H_{\text{personMax}}$
Output: Cropping positions C_{bbox}

$C_{\text{bbox}} = \emptyset$;
$G_{\text{head}} = F_{\text{translate}}(G, H_{\text{personMax}})$; ▷ $F_{\text{translate}}$: translate the ground following the ground normal vector and get a new parallel ground plane.
$d_{\text{bot}} = 0$; ▷ Bottom position of the current line to crop.
while True **do**
  $h_{\text{maxPixel}} = F_{\text{getMaxH}}(d_{\text{bot}}, K, G, H_{\text{personMax}})$;
  if $h_{\text{maxPixel}} < s_{\text{minPixel}}$ **then**
    break; ▷ Too close to the vanishing line.
  end if
  $h_{\text{patch}} = h_{\text{maxPixel}} / s_{\text{ratio}}$;
  $w_{\text{patch}} = h_{\text{patch}}$;
  $C_{\text{bboxOneLine}} = F_{\text{CutRow}}(d_{\text{bot}}, h_{\text{patch}}, w_{\text{patch}})$;
  $C_{\text{bbox}} = C_{\text{bbox}} \cup C_{\text{bboxOneLine}}$;
  $h_{\text{maxPixel}} = F_{\text{getMaxH}}(d_{\text{bot}}, K, G_{\text{head}}, -H_{\text{personMax}})$;
  $d_{\text{bot}} = d_{\text{bot}} - h_{\text{maxPixel}}$;
  if $d_{\text{bot}} > H_{\text{wholeImage}}$ **then**
    break; ▷ Reach the top of the image.
  end if
end while

3.3.4 Iterations

Since the 3 steps (Sec. 3.3.1, Sec. 3.3.2, and Sec. 3.3.3) are circularly dependent, we need to iterate them until the ground parameters converge (usually 1 loop is enough) or the maximum number of iterations reaches 3. Finally, the iterative

ground-aware cropping outputs the detected keypoints of all individuals, the estimated camera K, and ground parameters $G = \{N, D\}$.

3.4 Reconstruction via Upright 2D/3D Space

This section aims to reconstruct each individual after their keypoints are detected. Unlike our conference version [9], which directly reconstructs people in the camera space with a test-time scene-specific optimization, this paper defines a pair of canonical 2D and 3D spaces to eliminate the local and individual influence of the camera parameters during the reconstruction and achieves higher accuracy. Specifically, we define a *Upright 2D Space* (U2 Space) and the *Upright 3D Space* (U3 Space) where the mapping from U3 Space to U2 Space is a fixed orthographic projection. The U2 Space of a person is defined to satisfy the following properties: the ground normal vector is parallel to the vertical axis of the image, and the body torso center is centered horizontally in the image. The corresponding U3 Space is defined by expanding the U2 Space in the depth direction so that the corresponding orthogonal projection is also defined. An example of an input in the U2 Space and corresponding mesh output in the U3 Space can be seen in the Fig. 4. To link the defined U2/U3 Space with the camera space of the original image, we propose the Upright Normalization, which consists of two directions so that the steps to reconstruct each individual can be divided into three parts: (1) converting the input image from input image space to U2 Space by the upright normalization, (2) estimating the 2D HVIP and SMPL parameters in the U3 Space, and (3) converting the results from to camera space by the upright normalization.

3.4.1 Input Image to Upright 2D Space

We formulate the 2D transform from input image to the U2 Space as a homography matrix $H_{\text{upright}} \in \mathbb{R}^{3\times3}$. The homography matrix H_{upright} can be determined by the ground plane equation G, the camera intrinsic matrix K and a 3D point P_l to represent where the person is located in the camera space since the depth differences between different parts of a person are tiny compared to their average depth in large scenes. In general, we solve H_{upright} with the least squares approach after sampling specific points in the input image and computing the corresponding points in the U2 Space. Specifically, we first extract the 2D torso center p_t from the keypoints and get a rough 3D torso center P_t by assuming the height of this P_t from the ground to be 1.15 meters, where the torso center is defined in Sec. 3.2. Although the accuracy of the P_t is rough, the following steps are enough to measure the local perspective distortion because the absolute depth of a person is much larger than the depth differences among different parts of this person.

We first define the basis vector in the 3D camera space as: $V_{\text{cam}_x} = F_{\text{normalize}}(N \times P_t)$, $V_{\text{cam}_z} = F_{\text{normalize}}(P_t)$ and $V_{\text{cam}_y} = V_{\text{cam}_z} \times V_{\text{cam}_x}$, where $F_{\text{normalize}}$ is a function that normalizes the input vector to have a unit length. We then define the basis vector in the Upright 2D Space as: $v_{x\_\text{upright}} = [1, 0]$ and $v_{y\_\text{upright}} = [0, 1]$. We sample 4 collinear points in the 3D camera space P_{cam_i} and assign the corresponding 2D points p_{up_i} in the U2 space:

$$
\begin{aligned}
P_{\text{cam}_1} &= P_t + V_{\text{cam}_x} + V_{\text{cam}_y}, \quad p_{\text{up}_1} = [1, 1], \\
P_{\text{cam}_2} &= P_t + V_{\text{cam}_x} - V_{\text{cam}_y}, \quad p_{\text{up}_2} = [1, -1], \\
P_{\text{cam}_3} &= P_t - V_{\text{cam}_x} + V_{\text{cam}_y}, \quad p_{\text{up}_3} = [-1, 1], \\
P_{\text{cam}_4} &= P_t - V_{\text{cam}_x} - V_{\text{cam}_y}, \quad p_{\text{up}_4} = [-1, -1].
\end{aligned}
$$

In a next step, we compute the 2D projected points of $P_{\text{cam}_i}, i = 1, 2, 3, 4$ and get p_{cam_i}, *i.e.*, $p_{\text{cam}_i} = F_{\text{proj}}(K, P_{\text{cam}_i})$. Thus we can solve a homography matrix H_{warp} with source points p_{cam_i} and target points p_{up_i} by the least squares method. We set the resolution of the warped image in the U2 Space to be 512×512. For the next step, we need to scale the human subject to a specific size in the U2 Space. To keep a semantic scale with various human poses, we define the *Pose-free Pixel Height* h_{pixel} of a human on the image to be the sum of the trunk part, the max of the two arms and the max of the two legs of the 2D keypoints. We then apply a scale transformation matrix to ensure that h_{pixel} occupies 75% of the image height and this scale transformation matrix is recorded as $H_{\text{scale}} \in \mathbb{R}^{3\times3}$. For the last transform step, we compute the translation $H_{\text{trans}} \in \mathbb{R}^{3\times3}$ by constraining the 2D torso center p_t to be horizontally centered at a specified distance from the top of the image. Finally, we

achieve the transform from the input image space to the Upright 2D Space via warping, scaling, and translating: $H_{\text{upright}} = H_{\text{trans}} \cdot H_{\text{scale}} \cdot H_{\text{warp}}$.

3.4.2 HVIP Esitmation

According to the previous Sec. 3.2, the 3D localization problem can be converted to 2D HVIP estimation as a 2D torso center can be obtained from 2D keypoint detection by Eq. (2). In this subsection, we estimate the 2D HVIP in the U2 Space with a proposed HVIPNet. Different from our conference version, which has to handle various human scales in one cropped image when estimating multiple 2D HVIPs, HVIPNet only needs to estimate the 2D HVIP of a single person in the U2 space, where the 2D ground normal is the vertical direction and a relatively constant semantic human scale ratio is kept by the proposed pose-free pixel height h_{pixel}. Our HVIPNet encodes an image in a fixed resolution of 512×512 with a CNN backbone [34] and flattens the feature map to a 1-D feature vector. After feature encoding, HVIPNet applies a multi-layer perceptron (MLP) to regress a normalized scalar value d_{hvip} from this encoded 1-D feature vector representing the relative height from the 2D torso center to the 2D HVIP, since the ground normal is the vertical direction in the U2 space. The number of neurons in the hidden layers of the MLP is 512, 256, 128, 64, and 32, respectively, and the activation function is LeakyReLU. The HVIPNet is trained with the ground-truth 2D HVIP, and 2D torso center with the L1 loss on the training set of *LargeCrowd* where the Upright Normalization (Sec. 3.4.1) is applied by the ground truth. The final estimated 2D HVIP in U2 Space is: $p_{v_{\text{up}}} = p_t + 512 \cdot [0, d_{\text{hvip}}]$, which can be used to compute the 3D torso center with the estimated ground and camera parameters in Sec. 3.3.

3.4.3 HVIP and SMPL Estimation in Upright 3D Space

For the image transformed into the upright space for each individual, we estimate SMPL pose parameters $\theta_{v_{\text{up}}}$ and shape parameters β with the state-of-the-art methods like [5, 6]. These methods will also output orthographic camera parameters $O = \{O_{s_x}, O_{s_y}, O_{t_x}, O_{t_y}\}$ which describe the orthogonal projection $u = P_x * O_{s_x} + O_{t_x}$ and $v = P_y * O_{s_y} + O_{t_y}$, where $O_{s_x}, O_{s_y}, O_{t_x}, O_{t_y}$ are scalar values, $P = [P_x, P_y, P_z]$ is the 3D point and $p = [u, v]$ are the 2D projected point. Hence, the estimated SMPL parameters $\theta_{v_{\text{up}}}$ and β are in the U3 Space and can be projected to the U2 Space by O.

3.4.4 Upright 3D Space to Camera Space

Since SMPL parameters $\theta_{v_{\text{up}}}, \beta$ in U3 Space are obtained, we convert the estimated SMPL parameters back to 3D camera space by computing the additional scaling factor S_{toCam}, rotation matrix R_{toCam} and translation vector T_{toCam} from the previous image-to-U2 Space transforms, the local orthographic camera parameters O, and the estimated 2D HVIP $p_{v_{\text{up}}}$ in the 2D Upright Space. Specifically, we first transform the upright space 2D HVIP $p_{v_{\text{up}}}$ to image space p_v by the inverse matrix of H_{upright}. Secondly, according to Eq. (2), the 3D torso center P_t in camera space can be computed by the 2D HVIP p_v, 2D torso center p_t, the camera intrinsic matrix K and the ground plane G. Consequently, the translation vector T_{toCam} is the difference between P_t and the 3D torso center of the SMPL model in U3 Space. For the rotation, R_{toCam} can be solved by the source basis vector $[V_{x\_\text{up}}, V_{y\_\text{up}}, V_{z\_\text{up}}] = I_3$ and the target basis vector $[V_{\text{cam}_x}, V_{\text{cam}_y}, V_{\text{cam}_z}]$. The scaling factor can be determined by equating the local orthogonal projection with the global perspective projection of a planar object. Since $O_{s_x} \equiv O_{s_y}$ in state-of-the-art methods, the scaling factor S_{toCam} can be 1-D and solved by the following equation:

$$\begin{aligned} O_{s_y} = \| & F_{\text{project}}(K, P_t + 0.5 \cdot N) \\ & - F_{\text{project}}(K, P_t - 0.5 \cdot N) \|_2, \end{aligned} \tag{4}$$

where $F_{\text{project}} \in (\mathbb{R}^{3\times3} \times \mathbb{R}^3) \to \mathbb{R}^2$ is a function that projects a 3D point to the 2D image space by camera K. We incorporate the additional rotation into the SMPL pose parameters and update the camera translation accordingly to simplify the outputs. Therefore, the final outputs for each individual are the 3D pose θ_{cam}, shape β, scale S_{toCam} and translation T_{toCam}.

4 Experiments

4.1 Large-scene Datasets

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 3 seconds 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.

4.1.1 SynCrowd

To evaluate the accuracy of 3D joints and generalization on different camera FOVs, we create a synthetic dataset named *SynCrowd*. We collect 26 scenes from GigaMVS [35] and BlendedMVS [36] datasets and use 2447 scanned human models from Thuman [37] and 2K2K [38] datasets. For each scene, we set several camera poses and a planar area for random human placement. The human scans are placed in the area with random rotations and positions where the lowest points contact the scene and minimum distances between any two humans are constrained. We render 68 images at 9600×5400 image resolution with camera FoV varying from 30 to 120 degrees with Blender [39] and its physically based rendering engine Cycles. *SynCrowd* is for testing only.

4.2 Evaluation Metrics

4.2.1 Metrics for Mutually Exclusive and Collectively Exhaustive.

The first and indispensable step of evaluating a multi-human reconstruction method is to match the predicted 3D human poses and shapes to the ground truth. However, it is not appropriate to count the matched pairs only because it would encourage the method to predict fewer people with higher confidence. The best results should be mutually exclusive and collectively exhaustive (MECE), *i.e.*, there is no missing or redundant prediction. Therefore, we follow BEV [1], which adopts the F1 score to show the MECE performance, punishes other metrics by the F1 score, and reports both the matched value and the value punished by the F1 score. The F1 score is the harmonic mean of precision and recall. In our evaluation, we exclude the cases occluded by the scene and truncated by the image edge when computing precision and recall. The details can be found in the supplementary material.

4.2.2 Metrics for Sptail Consistency.

To evaluate the spatial consistency of the reconstructed crowd, we use the percentage of correct ordinal depth (PCOD) [25] to evaluate the ordinal depth relations between all pairs of individuals. We further propose pair-wise percentage distance similarity (PPDS) and the Procrustes-aligned version (PA-PPDS) to evaluate the relative position distribution of the crowd. The pair-wise percentage distance similarity (PPDS) is defined as

$$\text{PPDS} = \frac{\sum_{k=1}^{n-1}\sum_{i=k+1}^{n} 1 - \min\left(d_{ik}, 1\right)}{C_n^2}, \tag{5}$$

$$d_{ik} = \left| \frac{\|E_k - E_i\| - \|G_k - G_i\|}{\|G_k - G_i\|} \right|, \tag{6}$$

where n is the number of people in the image, and E_i and G_i represent the estimated and ground-truth body torso centers of the i-th person respectively. To eliminate the influence of the scale, we also define the Procrustes-aligned pair-wise percentage distance similarity (PA-PPDS), which aligns the reconstructed crowd and the ground truth by Procrustes alignment. In the quantitative tables of this paper, the metrics are presented in both "match" and "norm" versions. "Match" means the average metrics of each individual in all the images, and "norm" refers to normalized, meaning the weighted average metrics that are punished by the F1 score of each image. Since metrics PCOD, PPDS, and PA-PPDS are scores (a higher score is better), the way of punishment is $\text{Score}_{\text{punished}} = \text{Score} \times \text{F1}$.

Table 1 Quantitative comparison on the *LargeCrowd* dataset. The first 4 metrics report in both "norm" and "match" values.

Metric	PyMAF-large [6]	4DHumans-large [5]	BEV-large [1]	GroupRec [7]	Our Method (conf. ver.)	Our method
PPDS↑	70.35 / 73.62	73.11 / 75.90	66.34 / 73.35	71.58 / 76.27	83.81 / 87.27	**88.92** / **92.01**
PA-PPDS↑	72.73 / 76.12	74.47 / 77.31	68.46 / 75.73	73.34 / 78.15	91.80 / 95.60	**92.76** / **95.98**
PCOD↑	83.79 / 87.72	84.63 / 87.87	79.67 / 88.21	83.41 / 88.90	94.46 / 98.38	**94.92** / **98.22**
OKS↑	70.53 / 73.81	80.54 / 83.63	67.08 / 74.28	78.40 / 83.55	73.55 / 76.59	**82.49** / **85.36**
Recall↑	96.06%	96.1%	**97.5**%	94.6%	94.6%	96.5%
Precision↑	95.1%	96.6%	84.3%	93.2%	**97.5**%	96.9%
F1↑	0.955	0.963	0.903	0.938	0.960	**0.966**

Table 2 Quantitative comparison on *SytheticCrowd* dataset. The first 6 metrics report in both "norm" and "match" values.

Metric	PyMAF-large [6]	4DHumans-large [5]	BEV-large [1]	GroupRec [7]	Our Method (conf. ver.)	Our Method
PPDS↑	84.25 / 85.93	85.26 / 86.61	71.04 / 75.04	81.60 / 85.58	84.03 / 87.14	**89.92** / **91.29**
PA-PPDS↑	86.91 / 88.65	87.76 / 89.13	72.71 / 76.80	83.84 / 87.92	90.60 / 93.94	**93.58** / **95.00**
PCOD↑	91.48 / 93.30	92.10 / 93.54	81.99 / 86.64	88.65 / 92.99	94.04 / 97.49	**96.28** / **97.74**
OKS↑	79.57 / 81.16	83.47 / 84.78	70.37 / 74.27	84.79 / 88.89	72.06 / 74.56	**85.30** / **86.61**
T-MPJPE↓	0.150 / 0.147	0.131 / 0.129	0.154 / 0.145	0.121 / 0.115	0.142 / 0.135	**0.112** / **0.110**
PA-MPJPE↓	0.0942 / 0.0923	0.0850 / 0.0836	0.0798 / **0.0753**	0.0858 / 0.0814	0.0951 / 0.0910	**0.0794** / 0.0782
Recall↑	98.6%	98.5%	**99.4**%	97.9%	96.7%	98.4%
Precision↑	97.6%	98.4%	90.6%	93.1%	96.7%	**98.6**%
F1↑	0.980	**0.985**	0.947	0.953	0.965	**0.985**

4.2.3 Metrics for Human Poses and Shapes.

Due to the unavoidable depth-ambiguity and the unknown camera intrinsic parameters, the widely-used mean per joint position error (MPJPE) exceeds the meaningful range ($> 2m$) for all the methods. Therefore, we adopt T-MPJPE and PA-MPJPE to evaluate the accuracy of the 3D joints, where the T-MPJPE is the MPJPE after aligning the root joints only by translation, and the PA-MPJPE is the MPJPE after Procrustes' Analysis. We also use the object keypoint similarity (OKS) [40] to evaluate the reprojected 2D joint accuracy. These metrics are also given in both "match" and "norm" versions, the same as the spatial consistency metrics. Since OKS, T-MPJPE, and PA-MPJPE are errors (a lower error is better), the way of punishment is $\text{Error}_{\text{punished}} = \text{Error}/\text{F1}$.

4.3 Comparison

4.3.1 Compared Method

Currently, only our methods and GroupRec [7] can handle large-scene images with small human sizes compared to the image size. To enable other state-of-the-art methods to process large-scene images, we equip the state-of-the-art multi-human reconstruction method [1] with our iterative ground-aware cropping in Sec. 3.3 and the state-of-the-art single human reconstruction method 4DHumans [5] and PyMAF [6] with the same detection results as our method. We use the solve-PnP algorithm to compute the position for each individual, which is also used in the comparison experiments of BEV [1]. At last, we employ the same strategy as our method to remove redundant people for BEV's outputs. The names of enhanced methods are marked with "-large" suffix, and detailed modifications can be found in the supplementary material.

4.3.2 Evaluation

For a fair comparison, our method and GroupRec [7] are trained on *LargeCrowd* and BEV[1] is finetuned on the cropped dataset of *LargeCrowd* (cropped by Sec. 3.3). Note that we finally chose the original weights of BEV for the best performance. A possible reason for the performance drop is that only the 2D keypoints are available during

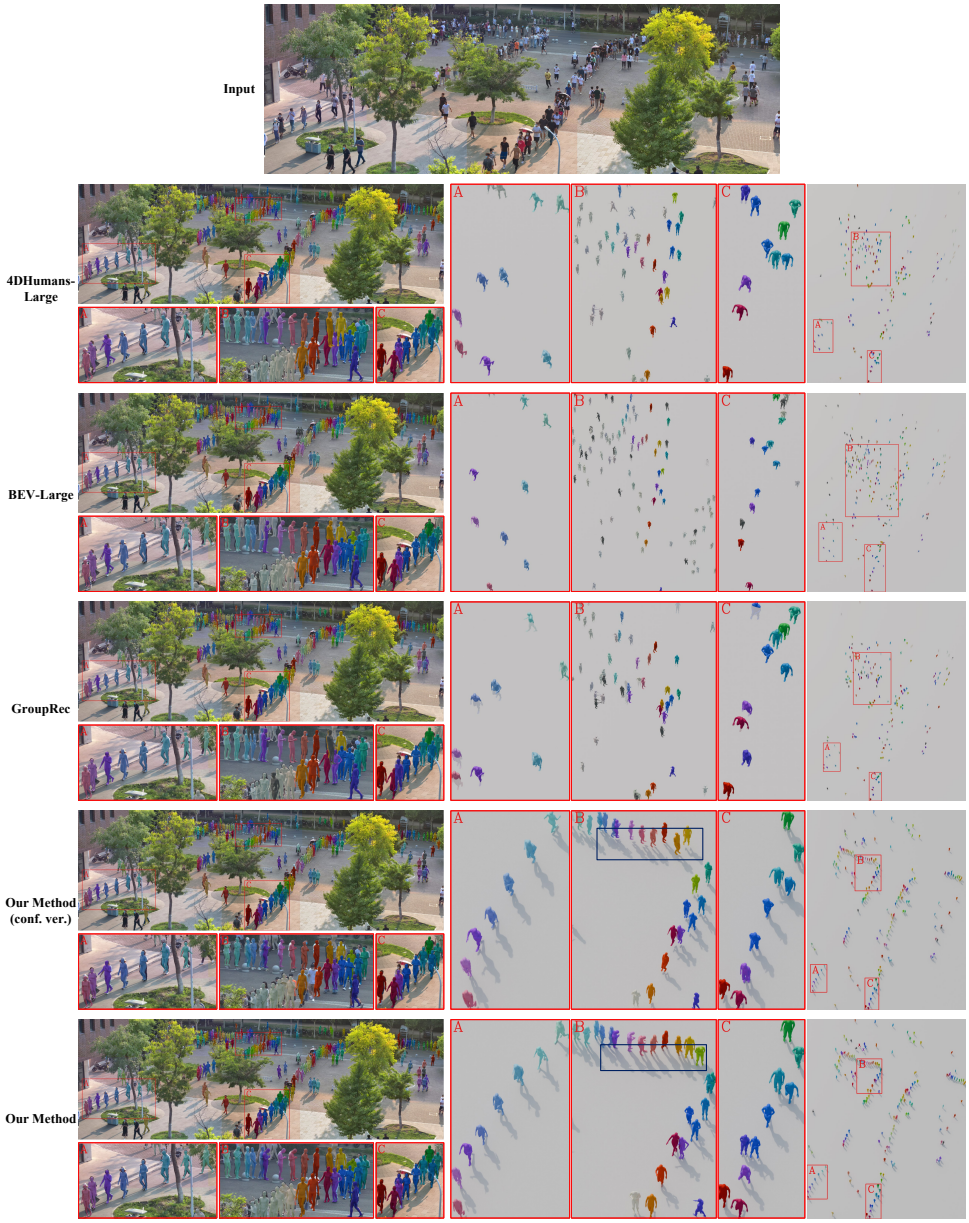


Fig. 5 Qualitative comparison on the *LargeCrowd* dataset. The color of a reconstructed human corresponds to the matched ground truth, while unmatched individuals are shown in gray. In the zoomed-in images A, B, and C, we use color saturation to differentiate whether the reconstructed results are fully within the cropped sub-images, *i.e.*, the lower saturation indicates that the result is not within the area. Shadows are ignored while rendering some compared methods because the results exhibit significant offsets from a unified ground plane.

the fine-tuning. Our method uses the pre-trained 4DHumans [5] as the single human reconstruction method in the upright space. The compared 4DHumans-large uses the same pre-trained model. None of the methods are trained on *SynCrowd*, and we only use it for testing. Because the compared methods must take the focal length as the input, these methods take the estimated camera intrinsics by our method as the input to compare fairly. Table 2 and Table 1 present the quantitative results on the *SynCrowd* dataset and the *LargeCrowd* dataset, respectively. The tables show that our method is optimal in terms of both unpenalized metrics ("match") and normalized metrics

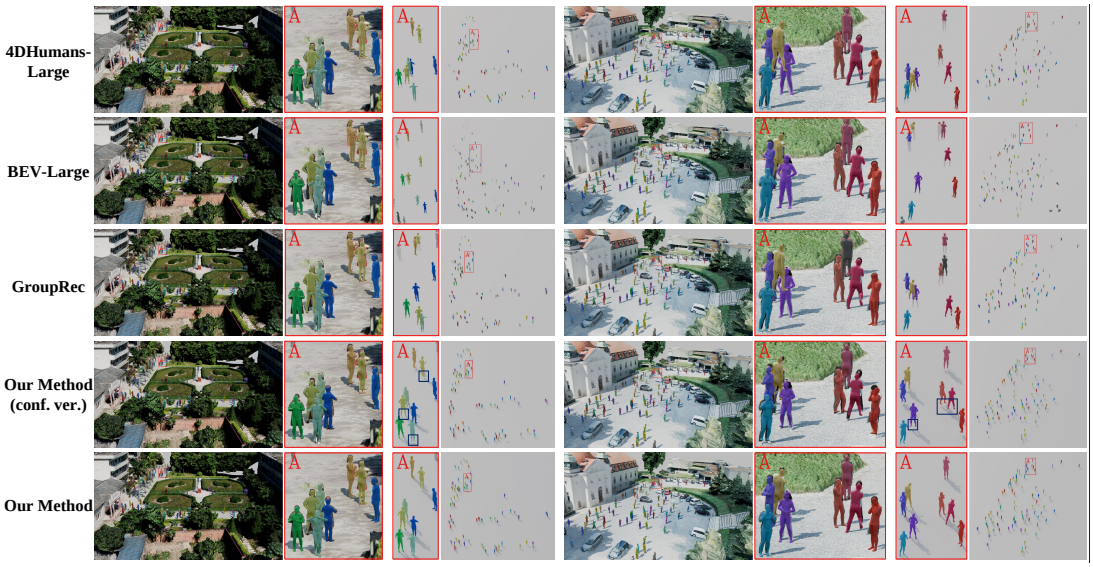


Fig. 6 Qualitative comparison on the *SynCrowd* dataset.

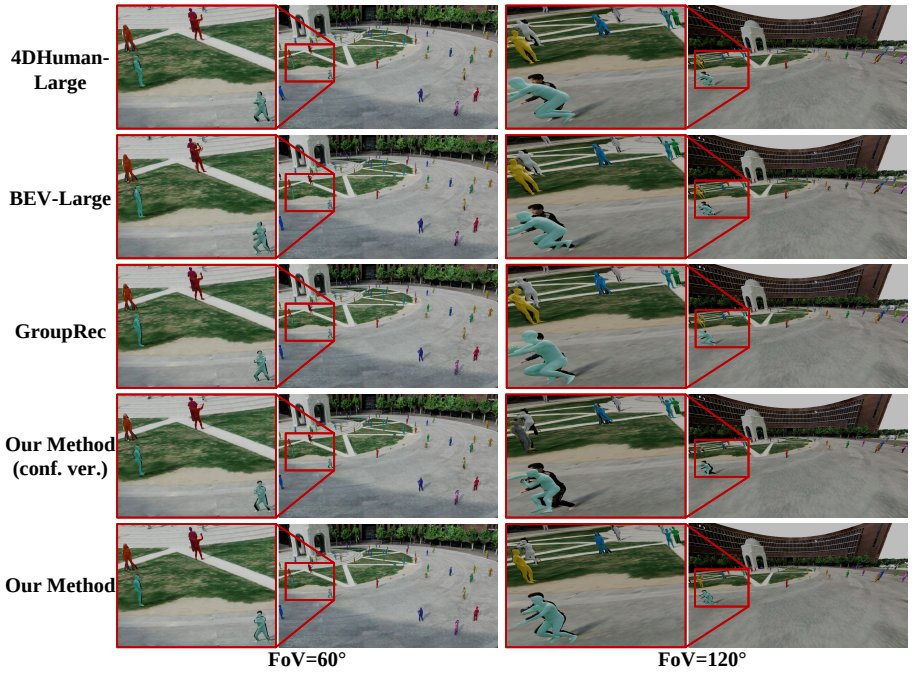


Fig. 7 Qualitative comparisons on the *SynCrowd* dataset under different FoVs.

("norm"). The advantage in PPDS, PA-PPDS, and PCOD shows that RCR achieves better spatial consistency, including more accurate physical distances and better relative arrangements. Fig. 6 and Fig. 5 show the qualitative comparison on the these datasets. The visualization of PyMAF [6] is ignored because the space is limited and it has already been compared by 4DHumans [5]. It can be seen that only RCR and our conference version can provide accurate relative positioning among individuals (bird's-eye view). Besides, RCR is much better in terms of reprojection, 3D poses, and interaction with the ground (highlighted by the dark blue bound boxes) against the conference

Table 3 Ablation studies on the *LargeCrowd*. Note that "Our Method with PyMAF" refers to our RCR pipeline with PyMAF [6] while "PyMAF-large" in the previous tables is a modified large-scene pipeline, *i.e.*, Pymaf enhanced by a human detection stage. The first 4 metrics report in both "norm" and "match" values.

Metric	Our Method w/o Upright Normalization	Our Method with PyMAF [6]	Our Method with 4DHumans [5]
PPDS↑	87.68 / 91.07	88.01 / 91.72	**88.92** / **92.01**
PA-PPDS↑	91.37 / 94.89	91.76 / 95.63	**92.76** / **95.98**
PCOD↑	94.16 / 97.79	94.09 / 98.07	**94.92** / **98.22**
OKS↑	77.08 / 80.07	74.72 / 77.87	**82.49** / **85.36**
Recall↑	**96.7**%	96.4%	96.5%
Precision↑	95.9%	95.6%	**96.9**%
F1↑	0.963	0.959	**0.966**

Table 4 Ablation studies on the *SynCrowd* dataset. The first 6 metrics report in both "norm" and "match" values.

Metric	Our Method w/o Upright Normalization	Our Method with PyMAF [6]	Our Method with 4DHumans [5]
PPDS↑	88.86 / 90.66	89.45 / 91.08	**89.92** / **91.29**
PA-PPDS↑	92.18 / 94.03	93.08 / 94.78	**93.58** / **95.00**
PCOD↑	95.56 / 97.47	95.90 / 97.66	**96.28** / **97.74**
OKS↑	77.15 / 78.71	80.35 / 81.81	**85.30** / **86.61**
T-MPJPE↓	0.132 / 0.129	0.138 / 0.135	**0.112** / **0.110**
PA-MPJPE↓	0.0832 / 0.0815	0.0932 / 0.0915	**0.0794** / **0.0782**
Recall↑	**98.7**%	98.5%	98.4%
Precision↑	97.5%	98.0%	**98.6**%
F1↑	0.980	0.982	**0.985**

version. Besides, benefiting from the canonical upright space, our method can infer more reasonable human poses and shapes with better reprojection under different FoVs, as demonstrated in Fig. 7.

4.4 Ablation Study

4.4.1 Efficacy of the Upright Spaces

In RCR, the defined Upright 3D/2D Spaces and the proposed Upright Normalization are designed to eliminate the effect of the camera parameters and the cropping operation on the global reconstruction. Although Fig. 7 has demonstrated the effectiveness of the upright normalization. We also design a variant, "RCR w/o Upright Normalization", which directly regresses SMPL from the cropped images with the same cropping strategy. To minimize the difference, "RCR w/o Upright Normalization" is provided with the same HVIPs as the full version of RCR for 3D localization. Table 4 reports that the full version of RCR outperforms the ablation version regarding all metrics. Because the same 2D HVIPs are used, the ablation version has a close performance in terms of PPDS, PA-PPDS, and PCOD. The full version has considerable advantages in terms of OKS, T-MPJPE, and PA-MPJPE, which indicates that the defined Upright 3D/2D Space is essential for human pose and shape reconstruction in large scenes.

4.4.2 Efficacy of the Iterative Ground-aware Cropping

In RCR, we design the iterative ground-aware cropping to detect the keypoints of all individuals to handle small and varying human scales in a large-scene image. To verify its efficacy, we make a uniform cropping baseline, which selects cropping windows roughly twice the pixel height of the largest person in each dataset, with a 50% overlap. For *LargeCrowd* and *SynCrowd*, the maximum pixel heights are about 1200 and 1500 so that the cropping window sizes are 2400 and 3000, respectively. We evaluate the detection performance of the iterative ground-aware cropping

Table 5 Ablation Studies for the efficacy of the Iterative Ground-aware Cropping.

Method	On *LargeCrowd*			On *SynCrowd*		
	Recall ↑	Precision ↑	F1 ↑	Recall ↑	Precision ↑	F1 ↑
Uniform Cropping	0.953	0.961	0.957	0.869	0.965	0.906
Iterative Ground-aware Cropping	**0.965**	**0.969**	**0.966**	**0.965**	**0.969**	**0.966**

and the uniform cropping baseline with the same detection strategy in Table 5. The results demonstrate that the iterative ground-aware cropping consistently outperforms the baseline across all metrics, without requiring any manual settings.

4.4.3 Compatibility with SOTA Reconstruction Methods.

RCR can be compatible with any state-of-the-art (SOTA) single-human reconstruction methods trained in weak perspective modeling, such as 4DHumans [5] and PyMAF [6]. They are noted as RCR with 4DHumans and RCR with PyMAF respectively, and their results are illustrated in the Table 3 and Table 4. The tables show that our method maintains stable positioning performance (PPDS, PA-PPDS, and PCOD) with different single-human reconstruction methods. The original performance of PyMAF can be referred to as PyMAF-large in Table 1 and Table 2. In terms of joint accuracy (OKS, T-MPJPE, and PA-MPJPE), the accuracy of our method is positively correlated with the accuracy of the single human reconstruction method. Additionally, as the performance of the chosen method improves (the OKS of 4DHumans-large is better than that of PyMAF-large), the performance of our method is also enhanced (RCR with 4DHumans is better than RCR with PyMAF according to Table 4, 3).

4.5 Discussions

We focus on the single-image crowd reconstruction from surveillance-like scenes. Since we adopt the Human-scene Virtual Interaction Point to solve the depth ambiguity, our method relies on the existence of a ground plane, which is common in a surveillance scene. For complex ground structures, the localization process of HVIP requires further designs to handle complex scenes like multiple floors and stairs by fusing the information of additional provided or estimated scene models and camera poses. Moreover, our method is based on a standard perspective camera model, which may not be directly applicable to certain types of cameras, such as fisheye or catadioptric cameras. However, once the camera model is identified and calibrated, upright normalization could potentially be extended to accommodate these camera types.

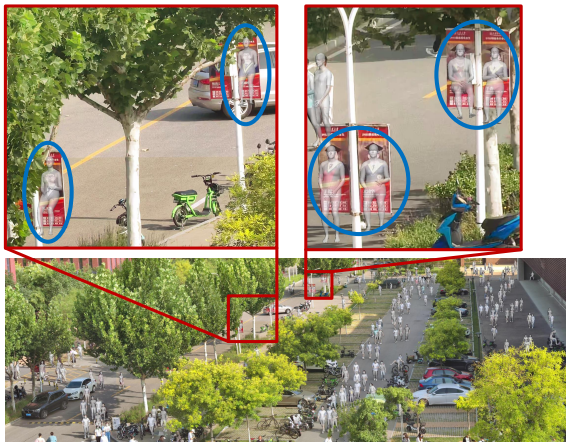


Fig. 8 Failure cases of redundant detection. It is challenging to distinguish whether these human figures are fake or real.

Failure Cases. In rare cases, the Iterative Ground-aware Cropping may detect wrong objects as humans, *e.g.*, Although the proposed Iterative Ground-aware Cropping can ensure a proper human scale for the human detection method while avoiding additional missing and redundant detection caused by the cropping and merging, our method can not remove some detection errors in the reconstruction stage, as shown in Fig. 8. Our method can remove extremely big or small detections by the depth cue, but it still relies on the detection results to some extent.

5 Conclusion

To reconstruct multiple human poses and shapes in the camera space from a single large-scene

image, we first propose the HVIP concept to effectively solve the depth ambiguity with a human keypoint-based camera and ground estimation approach. We then propose RCR (Robust Crowd Reconstruction) to reconstruct multiple human poses and shapes in a unified camera space from a single large-scene image. The proposed Upright 3D/2D Space and Upright Normalization eliminate the influence of the camera parameters and the cropping process during the reconstruction and ensure the reprojection performance under various FoVs without any test-time optimization or manually setting cropping parameters. We also contribute the *LargeCrowd* and *SynCrowd* datasets to help train and evaluate crowd reconstruction in large scenes under various camera parameters. Experimental results demonstrate that our method can achieve globally consistent crowd reconstruction in large scenes under various camera FoVs.

References

[1] Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3D people in depth. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13233–13242 (2022)

[2] Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3D people. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 11179–11188 (2021)

[3] Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2020)

[4] Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., *et al.*: PANDA: A gigapixel-level human-centric video dataset. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3268–3278 (2020)

[5] Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 14737–14748 (2023)

[6] Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12287–12303 (2023)

[7] Huang, B., Ju, J., Li, Z., Wang, Y.: Reconstructing groups of people with hypergraph relational reasoning. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 14827–14837 (2023)

[8] Kocabas, M., Huang, C.-H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 11035–11045 (2021)

[9] Hao, W., Jing, H., Huili, C., Haozhe, L., YuKun, L., Lu, F., Kun, L.: Crowd3D: Towards hundreds of people reconstruction from a single image. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8937–8946 (2023)

[10] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015)

[11] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10967–10977 (2019)

[12] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proc. European Conference on Computer Vision, pp.

561–578 (2016)

[13] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7122–7131 (2018)

[14] Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., Torre, F.: 3d human shape and pose from a single low-resolution image with self-supervised learning. In: Proc. European Conference on Computer Vision, pp. 284–300 (2020)

[15] Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1954–1963 (2021)

[16] Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3D human mesh regression with dense correspondence. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7054–7063 (2020)

[17] Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4496–4505 (2019)

[18] Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: HuMoR: 3D human motion model for robust pose estimation. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 11488–11499 (2021)

[19] Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: Proc. ACM SIGGRAPH Asia (2024)

[20] Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2070–2080 (2024)

[21] Wang, Y., Wang, Z., Liu, L., Daniilidis, K.: Tram: Global trajectory and motion of 3d humans from in-the-wild videos, pp. 467–487. Springer, ??? (2025)

[22] Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.-c., Asari, V.K.: Enhanced 3D Human Pose Estimation from Videos by Using Attention-Based Neural Network with Dilated Convolutions. International Journal of Computer Vision **129**(5), 1596–1615 (2021)

[23] Luvizon, D.C., Picard, D., Tabia, H.: Consensus-Based Optimization for 3D Human Pose Estimation in Camera Coordinates. International Journal of Computer Vision **130**(3), 869–882 (2022)

[24] Chen, Z., Huang, Y., Yu, H., Wang, L.: Learning a Robust Part-Aware Monocular 3D Human Pose Estimator via Neural Architecture Search. International Journal of Computer Vision **130**(1), 56–75 (2022)

[25] Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., Zhou, X.: SMAP: Single-shot multi-person absolute 3D pose estimation. In: Proc. European Conference on Computer Vision, pp. 550–566 (2020)

[26] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular RGB. In: Proc. IEEE International Comference on 3D Vision, pp. 120–130 (2018)

[27] Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)

[28] Benzine, A., Chabot, F., Luvison, B., Pham, Q.C., Achard, C.: PandaNet: Anchor-based single-shot multi-person 3D pose estimation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6856–6865 (2020)

[29] Zhang, J., Yu, D., Liew, J.H., Nie, X., Feng, J.: Body meshes as points. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 546–556 (2021)

[30] Choi, H., Moon, G., Park, J., Lee, K.M.: Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1465–1474 (2022)

[31] Contributors, M.: OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose (2020)

[32] Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[33] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)

[34] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. In: Proc. IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)

[35] Zhang, J., Zhang, J., Mao, S., Ji, M., Wang, G., Chen, Z., Zhang, T., Yuan, X., Dai, Q., Fang, L.: GigaMVS: A benchmark for ultra-large-scale gigapixel-level 3D reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7534–7550 (2022)

[36] Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1787–1796 (2020)

[37] Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5742–5752 (2021)

[38] Han, S.-H., Park, M.-G., Yoon, J.H., Kang, J.-M., Park, Y.-J., Jeon, H.-G.: High-fidelity 3D human digitization from single 2K resolution images. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12869–12879 (2023)

[39] Community, B.O.: Blender - a 3D Modelling and Rendering Package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018). Blender Foundation. http://www.blender.org

[40] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proc. European Conference on Computer Vision, pp. 740–755 (2014)