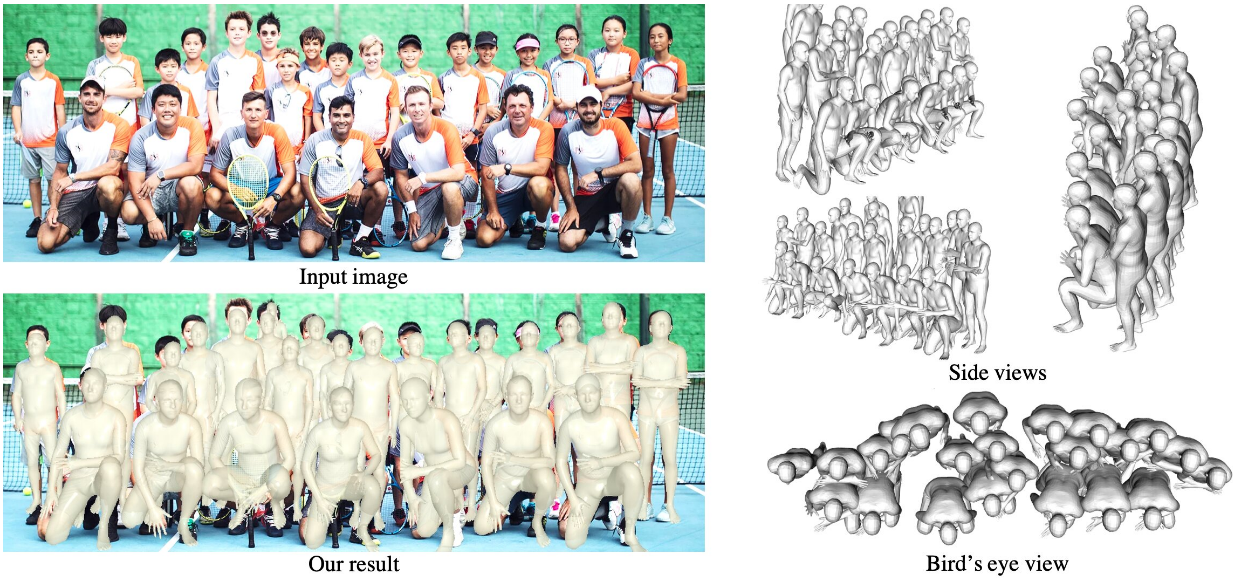
Putting People in their Place: Monocular Regression of 3D People in Depth

Yu Sun[1*] Wu Liu[2†] Qian Bao[2] Yili Fu[1†] Tao Mei[2] Michael J. Black[3]
[1]Harbin Institute of Technology, Harbin, China [2] Explore Academy of JD.com, Beijing, China
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany
yusun@stu.hit.edu.cn, liuwu1@jd.com, baoqian@jd.com, meylfu@hit.edu.cn
tmei@jd.com, black@tuebingen.mpg.de



Input image

Our result

Side views

Bird's eye view

Figure 1. **Monocular reconstruction of multiple 3D people with coherent depth reasoning.** We introduce BEV, a monocular one-stage method with an efficient new "bird's-eye-view" representation that enables the network to explicitly reason about people in 3D.

Abstract

Given an image with multiple people, our goal is to directly regress the pose and shape of all the people as well as their relative depth. Inferring the depth of a person in an image, however, is fundamentally ambiguous without knowing their height. This is particularly problematic when the scene contains people of very different sizes, e.g. from infants to adults. To solve this, we need several things. First, we develop a novel method to infer the poses and depth of multiple people in a single image. While previous work that estimates multiple people does so by reasoning in the image plane, our method, called BEV, adds an additional imaginary *Bird's-Eye-View representation to explicitly reason about depth. BEV reasons simultaneously about body centers in the image and in depth and, by combing these, estimates 3D body position. Unlike prior work, BEV is a single-shot method that is end-to-end differentiable. Second, height varies with age, making it impossible to resolve depth without also estimating the age of people in the image. To do so, we exploit a 3D body model space that lets BEV infer shapes from infants to adults. Third, to train BEV, we need a new dataset. Specifically, we create a "Relative Human" (RH) dataset that includes age labels and relative depth relationships between the people in the images. Extensive experiments on RH and AGORA demonstrate the effectiveness of the model and training scheme. BEV outperforms existing methods on depth reasoning, child shape estimation, and robustness to occlusion. The code*[1] *and dataset*[2] *are released for research purposes.*

[*] This work was done when Yu Sun was an intern at Explore Academy of JD.com.

[†] Corresponding author.

[1] https://github.com/Arthur151/ROMP

[2] https://github.com/Arthur151/Relative_Human

1. Introduction

In this article, we focus on simultaneously estimating the 3D pose and shape of all people in an RGB image along with their relative depth. There has been rapid progress [22] on regressing the 3D pose and shape of individual (cropped) people [4, 15, 16, 18, 19, 26, 29, 35, 44, 45, 47, 49] as well as the direct regression of groups [11, 34]. Neither class of methods explicitly reasons about the depth of people in the scene. Such depth reasoning is critical to enable a deeper understanding of the scene and the multi-person interactions within it. To address this, we propose a unified method that jointly regresses multiple people and their relative depth relations in one shot from an RGB image.

While previous multi-person methods perform well in constrained experimental settings, they struggle with severe occlusion, diverse body size and appearance, the ambiguity of monocular depth, and in-the-wild cases [11, 25, 38, 48]. These challenges lead to unsatisfactory performance in crowded scenes, including detection misses, similar predictions for overlapping people, and all predictions having a similar height. We observe two inter-related limitations that result in these failures. First, the architecture of the regression networks is closely tied to the 2D image, while the people actually inhabit 3D space. We address this with a new architecture that reasons in 3D. Second, depth estimation is fundamentally ambiguous due to the unknown height of the people in the image and it is difficult to obtain training data of images with ground-truth height and depth. To address this, we present a new dataset and novel losses that allow training without having metric depth.

We observe that crowded scenes contain rich information about the relative relationships between people, which can be exploited for both training and validation of depth reasoning. However, we still lack a powerful representations to learn from these cases. A few learning-based methods have been proposed for reasoning about the depth of predicted body meshes [11] or 3D poses [25, 38, 48]. Unfortunately, they all reason about depth via 2D representations, such as RoI-aligned features [11, 25] or a 2D depth map [38, 48]. These regression-based 2D representations have inherent drawbacks for representing the 3D world. The lack of an explicit 3D representation in the networks makes it challenging for these methods to deal with crowded scenes in which people overlap at different depths. Therefore, we argue that an explicit 3D representation is needed.

To achieve this, we develop BEV (for Bird's Eye View), a unified one-stage method for monocular reconstruction and depth reasoning of multiple 3D people. We take inspiration from ROMP [34], a one-stage, multi-person, regression method that directly estimates multiple 2D front-view maps for 2D human detection, positioning, and mesh parameter regression without depth reasoning. With ROMP, the network can only reason about the 2D location of people in the image plane. To go beyond this, we need to enable the network to efficiently reason about depth as well. To that end, we introduce a new *imaginary* 2D "bird's-eye-view" map that represents the likely centers of bodies in depth. To be clear, BEV takes only a single 2D image; the overhead view is inferred, not observed. BEV uses a powerful and efficient localization pipeline, performing bird's-eye-view-based coarse detection and fine localization in parallel. We employ the 2D heatmaps for coarse detection from both the front (image) and bird's eye views. BEV combines these heatmaps to obtain a 3D heatmap, as illustrated in Fig. 2. By learning the front and the bird's-eye view together, BEV explicitly models how people appear in images and in depth. This enables BEV to learn from available 2D and 3D annotations. BEV also uses a novel 3D Offset map to refine the initial coarse detections. From these coarse and fine maps, we obtain the 3D translation of all people in the scene. BEV transforms these predictions from the latent 3D Center-map space to an explicit camera-centric 3D space. Given these 3D translation predictions, BEV samples the features of all the people from a predicted mesh feature map and regresses the final SMPL [23] parameters. Distinguishing people at different depths enables BEV to estimate multiple people even with severe occlusion as illustrated in Fig. 1.

Even with a powerful 3D representation, we need an appropriate training scheme to ensure generalization. The main reason is that without knowing subject height, we lack effective constraints to alleviate the depth/height ambiguity under perspective projection. In particular, height varies with age, making it impossible to resolve depth without also estimating the age of people in the image. The ambiguity causes incorrect depth estimates for children and infants, limiting the generalization of existing methods. Unfortunately, existing 3D datasets with multiple people have limited diversity in height and age, so they cannot be used to improve or evaluate generalization.

Since collecting ground-truth 3D data in the wild is difficult, we instead train BEV using cost-effective weak labels of in-the-wild images. Specifically, we collect a dataset, named "Relative Human" (RH), that contains weak annotations of *depth layers* and human ages categorized into the groups adult, teenager, child, and infant. Moreover, we propose a weakly supervised training scheme (WST) to effectively learn from these weak supervision signals. For instance, we use a piece-wise loss function that exploits the depth layers to penalize incorrect relative depth orders. Exploiting age information to constrain height is tricky. While age and height are correlated, heights can vary significantly within the same age group. Consequently, we develop an ambiguity-compatible mixed loss function that encourages body shapes with heights that lie within an appropriate range for each age group.

We evaluate BEV on three multi-person datasets: in-the-wild using the 2D RH dataset and in 3D using the real CMU Panoptic [13] and the synthetic AGORA [28] datasets. On RH, compared with previous methods [11, 25, 38, 48], BEV is more accurate in relative depth reasoning and pose estimation. On CMU Panoptic, BEV outperforms previous methods [6, 11, 34, 42, 43] in 3D pose estimation. On AGORA, BEV significantly improves detection and achieves state-of-the-art results on "AGORA kids" in terms of the mesh reconstruction error. Also, fine-tunning on RH in a weakly supervised manner significantly improves the results for all age groups, especially for young people.

In summary, the main contributions are: (1) We construct a 3D representation to alleviate the monocular depth ambiguity via combining a front-view representation with an imaginary bird's eye view. (2) We collect the Relative Human dataset with weak annotations of in-the-wild images, which facilitates the training and evaluation on monocular depth reasoning in multi-person scenes. (3) We develop a weakly supervised training scheme to learn from weak depth annotations and to exploit age information.

2. Related Work

Monocular 3D mesh regression from natural scenes. Here, we focus on regressing a 3D body mesh using a parametric model like SMPL from a single RGB image. Most methods can be divided into multi-stage or single-stage approaches. For general multi-person cases, most existing methods [4, 15, 19, 26, 29] are based on a typical two-stage framework, which first detects people and then estimates the parameters of each person separately. Recent methods focus on exploring various supervision [33] signals, such as temporal coherence [16], contour alignment [7,31,39], self-contact [27], ground constraints [32, 40], or global human trajectory [41] to enhance the geometric/dynamic consistency. However, for depth reasoning about all people in the scene, these multi-stage methods are not ideal. The processing of individual cropped people cannot exploit the scene context or reason about depth ordering.

A few one-stage methods [24, 34] estimate multiple 3D people simultaneously. Given a single image, ROMP [34] outputs a 2D Body Center Heatmap, Camera Map, and Parameter Map for 2D human detection, positioning, and mesh parameter regression, respectively. At the position parsed from the 2D Body Center heatmap, ROMP samples the final mesh parameters from the Camera and Parameter maps. These one-stage methods enjoy a holistic view of the image, which is more suitable for depth reasoning. However, they are based on 2D representations that do not represent depth. Like most methods, they model adults (with SMPL), train on images of adults, and therefore only predict adults. To tackle the limitations of their 2D representation and age bias, we propose BEV and its training scheme of learning age priors that constrain body height.

Monocular depth reasoning. Most previous methods place bodies in depth via post-processing. Due to their 2D-based pipeline and lack of height prior for different age groups, their results are unsatisfying. A few learning-based methods, like 3DMPPE [25] and CRMH [11], address multi-stage depth reasoning. 3DMPPE uses image features to refine the bounding-box-based depth predictions. CRMH learns from instance segmentation to distinguish the relative depth between overlapping people. However, instance segmentation is expensive and unable to promote the learning of depth relations in cases without overlapping. SMAP [48] and HMOR [38] employ a 2D depth map to represent the root depth of 3D pose at each pixel. However, in crowded scenes, these 2D representations are ambiguous. In contrast, BEV adopts a novel bird's-eye-view-based 3D representation to distinguish people at different depths, therefore, it is more robust to the overlapping cases. Most recently, Ugrinovic et al. [36] propose an optimization-based method to refine the 3D translation of estimated body meshes. They fit the 3D body mesh to the detected 2D poses and force the feet to touch the ground. In contrast, our learning-based, one-stage, framework is more efficient and flexible, and can adapt to more scenarios, such as jumping. Albiero et al. [2] estimate the depth of all faces in a crowd in one shot by regressing their 6DoF pose; they do not deal with shape variation or articulation.

3. Method

3.1. Overview

The overall framework is illustrated in Fig. 2. BEV adopts a multi-head architecture. Given a single RGB image as input, BEV outputs 5 maps. For coarse-to-fine localization, we use the first 4 maps, which are the Body Center heatmaps and the Localization Offset maps in the front view and bird's-eye view. We first expand the front-/bird's-eye-view maps in depth/height and then combine them to generate the 3D Center/Offset maps. For coarse detection, we extract the rough 3D position of people from the 3D Center map. For fine localization, we sample the offset vectors from the 3D Offset map at the corresponding 3D center position. Adding these gives the 3D translation prediction. For 3D mesh parameter regression, we use the estimated 3D translation (x_i, y_i, d_i) and the Mesh Feature map. The depth value d_i of 3D translation is mapped to a depth encoding. At (x_i, y_i), we sample a feature vector from the Mesh Feature map and add it to the depth encoding for final parameter regression. Finally, we convert the estimated parameters to body meshes using the SMPL+A model.

3.2. SMPL+A: Mesh Representation for All Ages

The SMPL [23] and SMIL [9] models are developed to parameterize 3D body meshes of adults and infants into

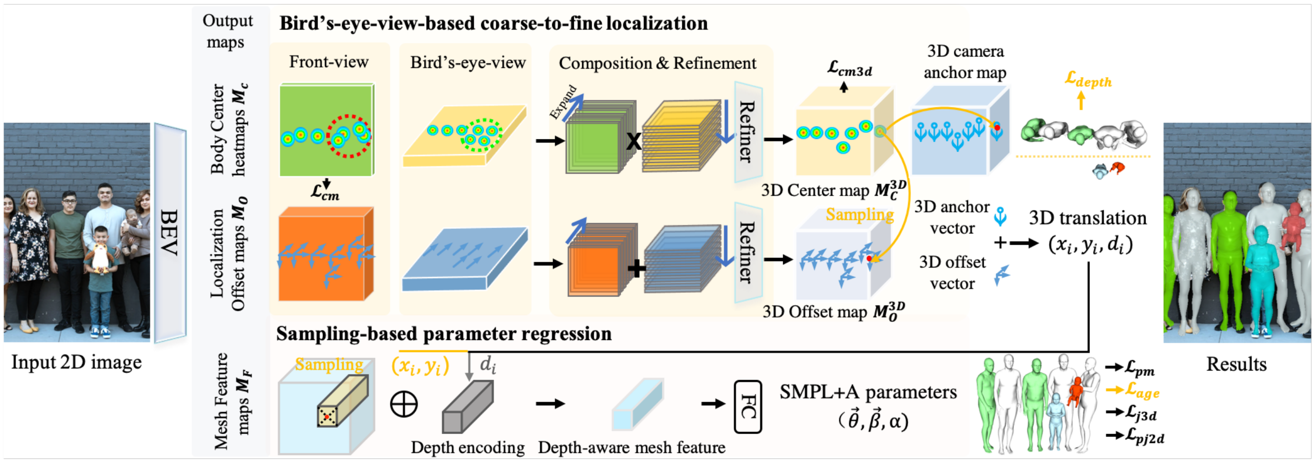


Figure 2. Overview. Given an RGB image, BEV first estimates the 3D translation of all people in the scene via compositing the front-view and the bird's-eye-view predictions. Then guided by the 3D translation, we sample the mesh feature of each person to regress their age-aware SMPL+A parameters. See Sec. 3.1 for details.

low-dimensional parameters. Recently, AGORA [28] further extends SMPL to support children by linearly blending the SMIL and SMPL template shapes with a weight $\alpha \in [0, 1]$, which we refer to as an "age offset." While blending the templates to address scale and proportion differences between adults and children, AGORA uses the adult shape space regardless of age. Additionally, AGORA does not address the representation of infants. We make a small, but important, change to better support all ages.

Following the notation of SMPL [23], the SMPL+A model defines a piece-wise function $\vec{B} = \mathcal{M}(\vec{\theta}, \vec{\beta}, \alpha)$ that maps 3D pose $\vec{\theta}$, shape $\vec{\beta}$, and age offset α to a 3D body mesh $\vec{B} \in \mathbb{R}^{6890 \times 3}$. The pose parameters, $\vec{\theta} \in \mathbb{R}^{6 \times 22}$, correspond to the 6D rotations [50] of the first 22 body joints of SMPL. The shape parameter $\vec{\beta} \in \mathbb{R}^{10}$ are the top-10 PCA coefficients of either the SMPL gender-neutral shape space or the SMIL shape space.

The adult shape space of AGORA produces shape deformations that are too large for an infant body, resulting in a distorted mesh when posed. Therefore, we use SMIL for infants when the age offset α is above a threshold t_α. When $\alpha > t_\alpha$, $\mathcal{M}(\vec{\theta}, \vec{\beta}, \alpha)$ is the SMIL model $\mathcal{M}_\mathcal{I}(\vec{\theta}, \vec{\beta})$. When the age offset $\alpha \leq t_\alpha$, we use the AGORA formulation

$$\begin{aligned} \mathcal{M}(\vec{\theta}, \vec{\beta}, \alpha) &= W(T_A(\vec{\theta}, \vec{\beta}, \alpha; \overline{\boldsymbol{T}}, \overline{\boldsymbol{T}}_{\boldsymbol{I}}), J(\vec{\beta}), \vec{\theta}, \boldsymbol{\mathcal{W}}), \\ T_A(\cdot) &= (1-\alpha)\overline{\boldsymbol{T}} + \alpha \overline{\boldsymbol{T}}_{\boldsymbol{I}} + B_S(\vec{\beta}) + B_P(\vec{\theta}), \end{aligned} \quad (1)$$

where $W(\cdot)$ performs linear blend-skinning with weights $\boldsymbol{\mathcal{W}}$ to convert the T-posed mesh $T_A(\cdot)$ to the target pose $\vec{\theta}$ based on the skeleton joints $J(\cdot)$. The T-posed mesh $T_A(\cdot)$ is the weighted sum of the templates $(\overline{\boldsymbol{T}}, \overline{\boldsymbol{T}}_{\boldsymbol{I}})$, shape-dependent deformation $B_S(\cdot)$, and pose-dependent deformation $B_P(\cdot)$. The age offset $\alpha \in [0, 1]$ is used to interpolate between the adult SMPL template $\overline{\boldsymbol{T}}$ and the infant SMIL template $\overline{\boldsymbol{T}}_{\boldsymbol{I}}$. The larger the α, the lower the mesh template height.

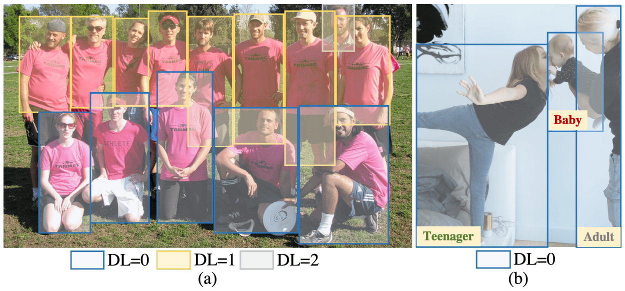


Figure 3. Example images from the Relative Human (RH) dataset with weak annotations: depth layers (DLs) and age group classification. Examples are a) adults at different DLs, and b) people of different age groups at the same DL.

The 3D joints $\vec{J}$ of the output mesh are derived via $\mathcal{J}\vec{B}$, where $\mathcal{J} \in \mathbb{R}^{K \times 6890}$ is a sparse weight matrix that linearly maps the vertices $\vec{B}$ to the K body joints. To supervise 3D joints $\vec{J}$ with 2D keypoints, regression methods [15,34] typically adopt a weak-perspective camera model to project $\vec{J}$ into the image plane. For better depth reasoning, we employ a perspective camera model to perform projection; see Sup. Mat. for the details of our camera model.

3.3. Relative Human dataset

Existing in-the-wild datasets lack groups of overlapping people with annotations. Since acquiring 3D annotations of large crowds is challenging, we exploit more cost-effective weak annotations. We collect a new dataset, named Relative Human (RH), to support in-the-wild monocular human depth reasoning.

The images are collected from multiple sources to ensure diversity in age, ethnicity, gender, and scene. Most images are collected from the existing 2D pose datasets [20,21,46].

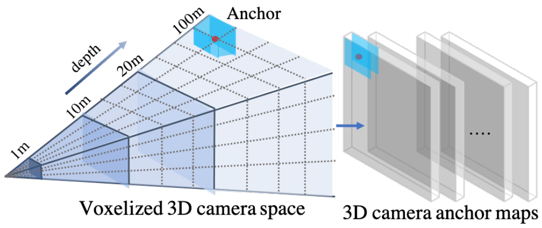


Figure 4. Pre-defined 3D camera anchor maps.

They contain few infants so we collect additional open-source family photos from Pexels [1] and then annotate their 2D poses. As shown in Fig. 3, we annotate the relative depth relationship between all people in the image. We treat subjects whose depth difference is less than one body-width ($\gamma = 0.3m$) as people in the same layer. We then classify all people into different depth layers (DLs). Unlike prior work, which labels the ordinal relationships between pairs of joints of individuals [5], DLs capture the depth order of multiple people. Additionally, we label people with four age categories: adults, teenagers, children, and babies.

In total, we collect about 7.6K images with weak annotations of over 24.8K people. More than 21% of the subjects are young people (5.3K), including teenagers, children, and babies. For more analysis, please refer to Sup. Mat.

3.4. Representations

Figure 2 gives an overview of BEV's representations.

Heatmaps: We build on the body-center heatmap representation from ROMP [34]. The front-view heatmap of size $\mathbb{R}^{1\times H\times W}$ is aligned with the pixel space and represents the likelihood of a body being centered at a 2D location using Gaussian kernels. We go beyond ROMP to add a second 2D heatmap of size $\mathbb{R}^{1\times D\times W}$ that represents an *unseen* bird's-eye-view. This heatmap represents the likelihood of a person being at some point in depth; this map, however, does not represent metric depth. BEV composes and refines these two maps into a 3D heatmap, $\boldsymbol{M_C^{3D}} \in \mathbb{R}^{1\times D\times H\times W}$, which represents the 3D position of the detected human body centers with 3D Gaussian kernels.

Offset maps: The discretized Center Heatmaps coarsely localize the body but we want the network to produce more precise estimates. To improve the granularity of 3D localization, we use additional maps that, at each position, add an estimated offset vector to refine the coarse detection. The front-view Offset map of size $\mathbb{R}^{3\times H\times W}$ contains 3D offset vectors. The bird's-eye-view Offset map of size $\mathbb{R}^{1\times D\times W}$ contains 1D offset vectors for depth correction. $\boldsymbol{M_O^{3D}} \in \mathbb{R}^{3\times D\times H\times W}$ corresponds to the 3D Center map and contains a 3D offset vector at each 3D position.

3D camera anchor maps: Each discretized coordinate in the 3D Center map corresponds to a set of camera parameters, representing its 3D position in the world. The anchor map serves as a mapping function to transform the coordinates of the 3D Center map to the 3D position in a predefined perspective camera space. To establish a one-to-one mapping from the square Center map to a pyramidal camera space, as shown in Fig. 4, we voxelize camera space. Each voxel center corresponds to a discretized 3D coordinate in the Center map. The 3D position vector (x, y, d) of voxel center is the anchor value of 3D camera anchor map. Voxels of equal depth form a depth plane, corresponding to a 2D (x-y) slice of the 3D camera anchor map. During inference, the 3D camera anchor map is sampled at the same coordinate of 3D Center map to obtain the coarse 3D translation of the corresponding detection.

Mesh feature map: $\boldsymbol{M_F} \in \mathbb{R}^{128\times H\times W}$ contains a 128-D mesh feature vector at each 2D position. These features are aligned with the input 2D image at the pixel level. After a 3D-center-based sampling process, the relevant features are used for the regression of SMPL+A parameters.

3.5. BEV

To effectively establish the 3D representation, the front-view and the bird's-eye-view must work together to estimate the image position and depth of corresponding subjects. Independently estimating the map of two views in parallel would inevitably cause misalignment, leading to the failure of 3D heatmap-based detection. To connect the two views, we estimate the bird's-eye-view maps conditioned on the front-view maps (i.e. Center and Offset maps). Specifically, to estimate the bird's-eye-view maps, we take the concatenation of the front-view maps and the backbone feature maps as input. The front-view 2D body-centered heatmap is used as a form of robust attention to people in the image, which helps the model focus on exploring depth during bird's-eye view estimation. Then we expand and composite the 2D maps from the front and BEV views to generate the 3D maps. To integrate 2D features from two views and enhance 3D consistency, we further perform 3D convolution on the composited 3D maps for refinement.

Next, we extract the 3D translation from the estimated 3D maps, $\boldsymbol{M_C^{3D}}, \boldsymbol{M_O^{3D}}$. High-confidence 3D positions of the 3D Center map are where we sample 3D offset vectors from the 3D Offset map. From the same 3D position in the 3D camera anchor maps (Fig. 4), we obtain the 3D anchor values, which are positions in camera space of the corresponding 3D center voxel. Adding the 3D offset vectors to the 3D anchor values gives the 3D translation as output.

Finally, we take the estimated 3D translation (x_i, y_i, d_i) and Mesh Feature maps $\boldsymbol{M_F}$ for parameter regression. We sample the pixel-level mesh feature vectors at (x_i, y_i) of $\boldsymbol{M_F}$. Inspired by positional embeddings [37], we learn an embedding space to differentiate people at different depths, especially for the overlapping cases. The predicted depth value d_i is mapped to a 128-dim encoding vector via an embedding layer. We sum up the depth encodings and the mesh feature vectors to differentiate the features of people at different depths, enabling individual estimates for different subjects. Then we estimate the SMPL+A parameters $(\vec{\theta}, \vec{\beta}, \alpha)$ via a fully-connected block. The output body meshes are obtained via $\mathcal{M}(\vec{\theta}, \vec{\beta}, \alpha)$.

3.6. Loss Functions

Our loss functions are divided into two groups illustrated in Fig. 2: relative losses (in gold) and the standard mesh losses (in black). BEV is supervised by the weighted sum of all loss items. First, we introduce two relative loss functions for weakly supervised training (WST).

Piece-wise depth layer loss $\boldsymbol{\mathcal{L}_{depth}}$. $\boldsymbol{\mathcal{L}_{depth}}$ is designed to supervise the predicted depth d_i, d_j of subject i, j by their depth layers r_i, r_j via

$$\begin{cases} (d_i - d_j)^2, & r_i = r_j \\ log(1 + e^{d_i - d_j}) \prod((d_i - d_j) - \gamma(r_i - r_j)), & r_i < r_j \\ log(1 + e^{d_j - d_i}) \prod(\gamma(r_i - r_j) - (d_i - d_j)), & r_i > r_j, \end{cases} \quad (2)$$

where $\prod$ is a binarization function that maps positive values to 1 and negative values to 0. $\prod$ is used to judge whether the BEV prediction is consistent with the depth relationship of the ground truth DLs. $\boldsymbol{\mathcal{L}_{depth}}$ is 0, if the predicted depth difference is within an acceptable range; that is, greater than the product of the DL difference and body-width γ. Otherwise, $\boldsymbol{\mathcal{L}_{depth}}$ will encourage the model to achieve it.

Previous ordinal depth losses [5, 30] encourage the model to enlarge the depth difference between people at different depth layers as much as possible. In contrast, the penalty in $\boldsymbol{\mathcal{L}_{depth}}$ is controlled within a range. This helps avoid pushing remote subjects too far away.

Ambiguity-compatible age loss $\boldsymbol{\mathcal{L}_{age}}$. The classification of age categories (infant, child, teenager, adult) is inherently ambiguous, especially for teenagers and children. Also, while height is correlated with age, one can easily find children who are taller than some adults. Consequently, we formulate an ambiguity-compatible mixed loss $\boldsymbol{\mathcal{L}_{age}}$.

Rather than supervise height directly, we supervise the α parameter that controls the blending between the SMIL infant body and the SMPL adult body. To do so, we define ranges of α values for each age group; i.e. (lower-bound, middle, upper-bound). We do this using the statistical data of heights for each age category that we then relate these to ranges of α values. Formally, the ranges are $(\alpha_l^k, \alpha_m^k, \alpha_u^k), k = 1 \cdots 4$ where k is the annotated age class number; see Sec. 4 for details.

BEV is then trained to predict the body shape as well as an α value for each person. Given the predicted α and ground truth age class k_g, the loss $\boldsymbol{\mathcal{L}_{age}}$ is defined as

$$\boldsymbol{\mathcal{L}_{age}}(\alpha) = \begin{cases} 0, & \alpha_l^{k_g} < \alpha \leq \alpha_u^{k_g} \\ (\alpha - \alpha_m^{k_g})^2, & \text{otherwise.} \end{cases} \quad (3)$$

Other losses. Following the previous methods [15, 34], we employ the standard mesh losses to supervise the output maps and regressed SMPL+A parameters. $\boldsymbol{\mathcal{L}_{cm}}$ is the focal loss [34] of the front-view Body Center heatmap. In the same pattern, we further use a 3D focal loss $\boldsymbol{\mathcal{L}_{cm3D}}$ to supervise the 3D Center map via converting $\boldsymbol{\mathcal{L}_{cm}}$'s 2D operation to 3D. $\boldsymbol{\mathcal{L}_{pm}}$ consists of three parts, $\boldsymbol{\mathcal{L}_{\theta}}$, $\boldsymbol{\mathcal{L}_{\beta}}$, and $\boldsymbol{\mathcal{L}_{prior}}$. $\boldsymbol{\mathcal{L}_{\theta}}$ and $\boldsymbol{\mathcal{L}_{\beta}}$ are L_2 losses of SMPL+A pose $\vec{\theta}$ and shape $\vec{\beta}$ parameters respectively. $\boldsymbol{\mathcal{L}_{prior}}$ is the Mixture of Gaussian pose prior [4, 23] on $\vec{\theta}$. To supervise the 3D body joints $\vec{J}$, we use $\boldsymbol{\mathcal{L}_{j3d}}$, which is composed of $\boldsymbol{\mathcal{L}_{mpj}}$ and $\boldsymbol{\mathcal{L}_{pmpj}}$. $\boldsymbol{\mathcal{L}_{mpj}}$ is the L_2 loss of 3D joints $\vec{J}$. To alleviate the domain gap between training datasets, we follow [34, 35] to calculate the L_2 loss $\boldsymbol{\mathcal{L}_{pmpj}}$ of the predicted 3D joints after Procrustes alignment with the ground truth. $\boldsymbol{\mathcal{L}_{pj2d}}$ is the L_2 loss of the 2D projection of 3D joints $\vec{J}$. Lastly, $w_{(.)}$ denotes the corresponding weight of these losses.

4. Experiments

4.1. Implementation Details

Training details. For basic training, we use two 3D pose datasets (Human3.6M [10] and MuCo-3DHP [24]) and four 2D pose datasets (COCO [21], MPII [3], LSP [12], and CrowdPose [20]). We also use the pseudo SMPL annotations from [14] and WST on RH. Most samples in RH are collected from 2D pose datasets [20, 21, 46]. For a fair comparison, we only use the samples that are also used for training in compared methods [11, 18, 19, 25, 34, 48]. To compare with [18, 28], we further fine-tune our model and ROMP on AGORA. The threshold for the age offset is set to $t_\alpha = 0.8$. The age offset ranges $(\alpha_l^k, \alpha_m^k, \alpha_u^k)$ are: adults $(-0.05, 0, 0.15)$, teenagers $(0.15, 0.3, 0.45)$, children $(0.45, 0.6, 0.75)$, and infants $(0.75, 0.9, 1)$. See Sup. Mat. for more details.

Evaluation benchmarks. We evaluate BEV on three multi-person datasets, RH, CMU Panoptic, [13] and AGORA [28], containing 257 child scans and significant person-person occlusion.

Evaluation matrix. To evaluate the accuracy of depth reasoning, we employ the Percentage of Correct Depth Relations (**PCDR**$^{0.2}$), and set the threshold for equal depth to $0.2m$. To evaluate the accuracy of projected 2D poses on RH, we also report the mean Percentage of Correct Keypoints (**mPCK**$_h^{0.6}$), setting the matching threshold to 0.6 times the head length.

Also, following AGORA [28], we evaluate the accuracy of 3D pose/mesh estimation while considering missing detections. To evaluate the detection accuracy, we report **Precision**, **Recall**, and **F1 score**. For matched detections, we report the Mean Per Joint Position Error (**MPJPE**) and Mean Vertex Error (**MVE**). To punish misses and false alarms in detection, we normalize the MPJPE and MVE by

Method	PCDR$^{0.2}$(%)↑					mPCK$_h^{0.6}$↑
	Baby	Kid	Teen	Adult	All	
3DMPPE† [25]	39.33	51.42	60.91	57.95	57.47	-
CRMH [11]	34.74	48.37	59.11	55.47	54.83	0.781
SMAP [48]	31.58	40.29	47.35	41.65	41.55	-
ROMP [34]	30.08	48.41	51.12	55.34	54.81	0.866
BEV w/o WST	34.27	50.81	54.34	57.43	57.17	0.850
BEV w/o $\mathcal{L}_{depth}$	43.61	51.55	50.88	57.27	55.97	0.794
BEV w/o $\mathcal{L}_{age}$	49.09	56.55	60.92	62.47	61.47	0.810
BEV	**60.77**	**67.09**	**66.07**	**69.71**	**68.27**	**0.884**

Table 1. Accuracy of relative depth relations (PCDR$^{0.2}$) and projected 2D poses (mPCK$_h^{0.6}$) on RH. † uses the ground truth bounding boxes.

Method	Haggl.	Mafia	Ultim.	Pizza	Mean
Zanfir et. al. [43]	141.4	152.3	145.0	162.5	150.3
MSC [42]	140.0	165.9	150.7	156.0	153.4
CRMH [11]	129.6	133.5	153.0	156.7	143.2
ROMP [34]	110.8	122.8	141.6	137.6	128.2
3DCrowdNet [6]	109.6	135.9	129.8	135.6	127.3
BEV	**90.7**	**103.7**	**113.1**	**125.2**	**109.5**

Table 2. Comparisons to the state-of-the-art methods on CMU Panoptic in MPJPE. Results are obtained from the original papers.

Method	Kid subset							Full set						
	Detection↑			Matched↓		All↓		Detection↑			Matched↓		All↓	
	F1 score	Precision	Recall	MVE	MPJPE	NMVE	NMJE	F1 score	Precision	Recall	MVE	MPJPE	NMVE	NMJE
PARE [17]	0.55	0.44	0.74	186.4	193.9	338.9	352.5	0.84	0.96	0.75	140.9	146.2	167.7	174.0
SPIN [28]	0.31	0.21	0.60	186.7	191.7	602.3	618.4	0.77	0.91	0.67	148.9	153.4	193.4	199.2
SPEC [18]	0.52	0.40	0.73	163.2	171.0	313.8	328.8	0.84	0.96	0.74	106.5	112.3	126.8	133.7
ROMP [34]	0.50	0.37	0.80	156.6	159.8	313.2	319.6	0.91	0.95	0.88	103.4	108.1	113.6	118.8
BEV w/o WST	**0.58**	**0.44**	**0.86**	146.0	148.3	251.7	255.7	0.93	0.96	0.90	105.6	109.7	113.5	118.0
BEV	0.55	0.41	0.85	**125.9**	**129.1**	**228.9**	**234.7**	**0.93**	**0.96**	**0.90**	**100.7**	**105.3**	**108.3**	**113.2**

Table 3. Comparison of SOTA methods on AGORA test set. All methods are fine-tuned on the AGORA training set or synthetic data [18] generated in the same way as AGORA. We fine-tune ROMP [34] using the public implementation; results from the AGORA leaderboard.

F1 score to get Normalized Mean Joint Error (**NMJE**) and Normalized Mean Vertex Error (**NMVE**).

4.2. Comparisons to the state-of-the-art methods

Monocular depth reasoning. We first evaluate BEV on monocular depth reasoning in Tab. 1 using the RH dataset. Results in Tab. 1 are obtained using the official implementations of compared methods. BEV uses the same training samples as [34] to perform WST. We first compare with the most competitive methods [11, 25, 48], which solve depth relations in monocular images. We also compare with ROMP [34], for one-stage multi-person mesh recovery. Their 3D translation results are obtained by solving the PnP algorithm (RANSAC [8]) between their 3D pose and projected 2D pose predictions. As shown in Tab. 1, BEV outperforms all these methods in the accuracy of both depth reasoning and projected 2D poses by a large margin.

Monocular detection and mesh regression. We also run BEV on AGORA and CMU Panpotic to evaluate the detection and 3D mesh accuracy. We compare with the state-of-the-art (SOTA) multi-stage methods [6, 11, 17, 18, 28, 42, 43] and the one-stage ROMP [34]. Benefiting from the superiority in recall, in Tab. 3, BEV outperforms SOTA methods on detection by 5.2% and 2.2% in terms of F1 score on the kid and full subset, respectively. This is evidence that the 3D representation helps alleviate depth ambiguity in crowded scenes. On the kid subset, BEV significantly outperforms previous methods in terms of mesh reconstruction. Especially, compared with ROMP [34], BEV reduces errors over 19.6% and 26.9% in terms of matched MVE and all NMVE on AGORA kids, indicating that BEV effectively reduces the age bias using WST. Also, as shown in Tab. 2, on CMU Panpotic, BEV significantly reduces 3D pose errors by 13.9% compared to multi-person SOTA methods. For qualitative results, see Fig. 1 and Fig. 5.

4.3. Ablation Studies

Bird's-eye-view representation & BEV w/o WST. To further test the effectiveness of BEV's 3D representation, we train it without performing WST on RH and compare it with SOTA methods on AGORA and RH. On RH in Tab. 1, compared with CRMH [11], the depth reasoning accuracy of BEV w/o WST is 4.1% higher (**PCDR**$^{0.2}$ of all). BEV w/o WST outperforms the 2D representation-based network ROMP [34]. These results point to the effectiveness of our 3D representation for dealing with monocular depth ambiguity. On AGORA, as shown in Tab. 3, BEV w/o WST significantly outperforms ROMP in all detection metrics. Additionally, the strong detection ability of the 3D representation makes BEV w/o WST outperform the SOTA methods [18, 28, 34] in terms of NMVE and NMJE.

Weakly supervised training (WST) losses, $\mathcal{L}_{depth}$ and $\mathcal{L}_{age}$. Results in Tab. 1 show that performing WST significantly improves the accuracy of depth reasoning, especially for the young groups. Also, Tab. 1 shows that separately using $\mathcal{L}_{depth}$ or $\mathcal{L}_{age}$ make BEV produce better depth reasoning than BEV w/o WST, and, when using both terms, BEV performs best.

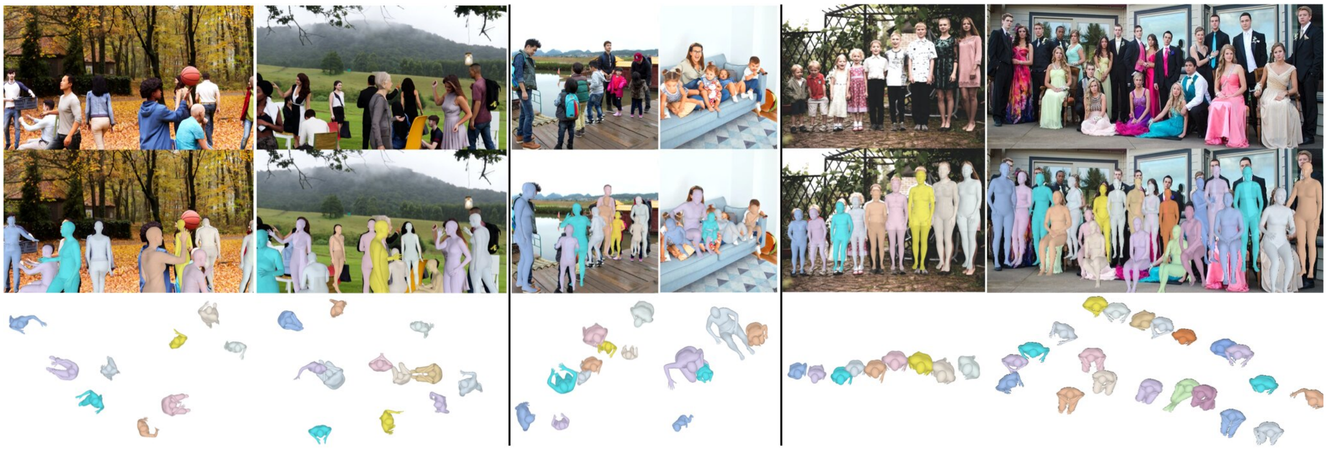


Figure 5. Qualitative results on AGORA, RH, and Internet images [1]. Note how children and adults are properly placed in depth.

Method	Relative Human		AGORA		
	PCDR$^{0.2}$	mPCK$_h^{0.6}$	F1	NMVE	NMJE
BEV	**68.27**	**0.884**	**0.93**	**108.3**	**113.2**
w/o FVC	67.99	0.880	0.89	118.9	123.0
w/o OM	60.76	0.620	0.87	126.6	130.7

Table 4. Ablation study of front-view condition (FVC) and 3D Offset map (OM) on RH and AGORA.

Method	Dist.↓	X↓	Y↓	Depth↓
Ordinal loss [38]	0.608	0.153	0.184	0.509
Piece-wise $\mathcal{L}_{depth}$ (ours)	**0.518**	**0.128**	**0.166**	**0.423**

Table 5. 3D translation error on AGORA validation set with different depth losses.

3D Offset map (OM) and Front-view condition (FVC) for 3D localization. FVC is taking the front-view 2D body-centered heatmap as a robust attention signal to explore the depth of detected persons during bird's-eye view estimation. Results in Tab. 4 verify that OM and FVC significantly improve the granularity of 3D localization.

Piece-wise depth layer loss $\mathcal{L}_{depth}$ v.s. ordinal depth loss [38]. Unlike an ordinal depth loss, $\mathcal{L}_{depth}$ keeps the penalty within a reasonable range (see Sec. 3.6). As shown in Tab. 5, on AGORA validation set, training with $\mathcal{L}_{depth}$ reduces the 3D translation error, especially in depth.

5. Conclusion, Limitations, Ethics, Risks

In this paper, we introduce BEV, a unified one-stage method for monocular regression and depth reasoning of multiple 3D people. By introducing a novel bird's eye view representation, we enable powerful 3D reasoning that reduces the monocular depth ambiguity. Exploiting the correlation between body height and depth, BEV learns depth reasoning from complex in-the-wild scenes by exploiting relative depth relations and age group classification. We make available an in-the-wild dataset to promote the training and evaluation of monocular depth reasoning in the wild. The ablation studies point to the value of the 3D representation and the fine-grained localization in the network, the importance of our training scheme, and the value of the collected dataset. BEV is a preliminary attempt to explore complex multi-person relationships in the 3D world, and we hope the framework will serve as a simple yet effective foundation for future progress.

Limitations. While BEV goes beyond current methods to cover more diverse ages, it is not trained to capture diverse weights, gender, ethnicity, etc. BEV also assumes a constant focal length. Our labeling approach, however, suggests that weak labels can produce strong results; i.e. improved metric accuracy. Note that BEV is not trained or designed to deal with large "crowds" (e.g. 100's of people).

Ethics and data. We collected RH images from a free photo website [1] under a Creative Commons license that enables sharing. We strove to have a dataset that is diverse in age, ethnicity, and gender. Also, our weak annotations do not contain any personal information and the annotators, themselves, are anonymous and were not studied.

Potential Negative Societal Impacts. Methods for monocular 3D pose and shape estimation might be used for automated surveillance, tracking, and behavior analysis, which may violate people's privacy. To help prevent this, BEV is released for research only.

Acknowledgements: This work was supported by the National Key R&D Program of China under Grand No. 2020AAA0103800.

Disclosure: MJB has received research funds from Adobe, Intel, Nvidia, Facebook, and Amazon and has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While he was part-time at Amazon during this project, his research was performed solely at Max Planck.

References

[1] Pexels. https://www.pexels.com. 5, 8

[2] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *CVPR*, pages 7617–7627, 2021. 3

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 6

[4] 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 *ECCV*, pages 561–578, 2016. 2, 3, 6

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NeurIPS*, pages 730–738, 2016. 5, 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 *CVPR*, 2022. 3, 7

[7] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *ICCV*, pages 11250–11259, 2021. 3

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[9] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *MICCAI*, pages 792–800, 2018. 3

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 6

[11] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 2, 3, 6, 7

[12] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472, 2011. 6

[13] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 3, 6

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *ECCV*, 2020. 6

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3, 4, 6

[16] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2, 3

[17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 7

[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 *ICCV*, pages 11035–11045, 2021. 2, 6, 7

[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 *ICCV*, pages 2252–2261, 2019. 2, 3, 6

[20] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 4, 6

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 4, 6

[22] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*, 2022. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 2, 3, 4, 6

[24] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130, 2018. 3, 6

[25] 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 *CVPR*, pages 10133–10142, 2019. 2, 3, 6, 7

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

[27] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, pages 9990–9999, 2021. 3

[28] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, 2021. 3, 4, 6, 7

[29] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, pages 803–812, 2019. 2, 3

[30] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, pages 7307–7316, 2018. 6

[31] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape

from a single color image. In *CVPR*, pages 459–468, 2018. 3

[32] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3d human motion model for robust pose estimation. In *ICCV*, pages 11488–11499, 2021. 3

[33] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, pages 5340–5348, 2019. 3

[34] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 2, 3, 4, 5, 6, 7

[35] 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 *ICCV*, pages 5348–5357, 2019. 2, 6

[36] Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *3DV*, pages 53–63, 2021. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 5

[38] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. HMOR: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259, 2020. 2, 3, 8

[39] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 2022. 3

[40] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *CVPR*, 2022. 3

[41] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3

[42] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, pages 2148–2157, 2018. 3, 7

[43] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *NeurIPS*, pages 8410–8419, 2018. 3, 7

[44] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *CVPR*, pages 7054–7063, 2020. 2

[45] 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 *ICCV*, pages 11446–11456, 2021. 2

[46] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *CVPR*, pages 889–898, 2019. 4, 6

[47] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *CVPR*, pages 5560–5569, 2021. 2

[48] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3D pose estimation. In *ECCV*, pages 550–566, 2020. 2, 3, 6, 7

[49] Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, and Qixing Huang. Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In *ECCV*, pages 137–153, 2018. 2

[50] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 4