Aligned Dense Supervision towards Physical-space Monocular Localization of Human Body Meshes

Anonymous CVPR submission

Paper ID 7982

Abstract

*Monocular RGB cameras may easily cover unconstrained daily scenarios under diverse focal lengths and viewpoints. However, due to the lack of explicit depth measurement from monocular cameras, it is extremely challenging to localize human bodies in physical size and location for automatic visual analytics. We consider this fundamental body mesh detection problem in the standard multi-task deep learning framework, in which local pelvis-centered meshes and global body-to-camera translations are respectively estimated by parallel network branches, and the branch-wise losses are separately aggregated across bodies. Such framework places unrealistic faith in (a) the training data diversity of both the mesh poses and locations and (b) the across-body-aggregated losses on supervising the body-wise pose and location estimation. In this paper, we present **Aligned Dense Supervision (ADS)**, a bag of techniques for physical-space body mesh detection. ADS creates dense supervision by locally body-aligned ROIs and the globally augmented locations, and boosts the local and global tasks by serially-fused regression heads. We showcase ADS in the classic MaskRCNN-based architecture, and report superior results than the state-of-the-arts over a physical 3D space of [0, 27.4]-by-[-7.6, 9.0] meters.*

1. Introduction

In unconstrained daily scenarios such as shopping malls, parking lots, rehabilitation centers and sports ground, monocular RGB cameras are probably the most accessible sensor to cover human bodies in diverse distances and orientations. Unfortunately, the lack of explicit distance measurement by monocular cameras leads to strong ambiguity in estimating physical sizes, shapes and locations of body meshes from monocular images.

Early monocular mesh recovery approaches map single-person image patches to SMPL meshes in a pelvis-centered local coordinate system [12, 14, 16, 17, 26, 29, 30, 33], then addresses the mutual occlusion and duplicated detection issues by fusing the local SMPL regressor with 2D person detectors [9, 34–36]. These approaches do not solve the global 3D mesh localization problem. Instead, they usually leverage a by-product of local mesh regression, i.e. the weak-perspective camera parameters (2D offset + 2D scale), to derive the Pseudo 3D location of the meshes in post-processing. A locally estimated mesh is shifted by the 2D offset in the image plane and shifted by [2D scale $\times$ Pseudo-focal-length] in the perpendicular direction to the image plane. Such a Pseudo 3D localization approach does not produce physical space measures of body locomotion and body-scene contact.

To estimate the physical space mesh size and location, one could combine a local mesh parameter regressor [9, 34–36] and a global depth regressor [5, 21, 39] in one Deep Neural Network (DNN) (see our ROI-Aligned baseline in Sec. 3.1) or Fixed-ROI mesh detector (BEV [28]). In existing methods, the Region-of-Interests of regressors are mis-aligned with bodies under aspect-ratio deformation, occlusion, truncation or scaling. Learning from the mis-aligned samples requires an unrealistic rich combination of

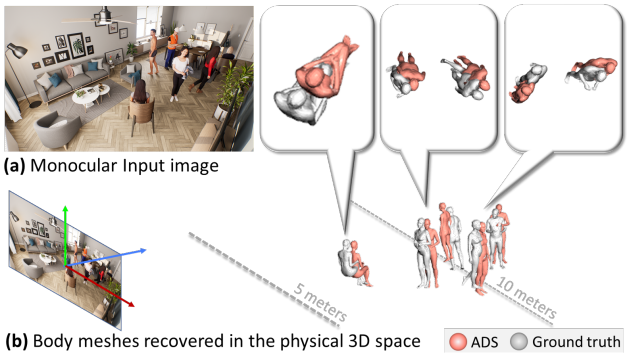

Figure 1. **Aligned Dense Supervision (ADS)** achieves physical-space monocular localization by boosting the overall supervision density and encouraging the body-wise collaborative regression of local-mesh&global-translation. See examples of ADS meshes v.s. Ground Truth meshes in physical sizes, shapes and 3D locations.

body poses and body locations in training data. Moreover, their local and global regressors originate separately from the global feature maps of the backbone network, disregarding the local-global dependency of the same body.

To address these issues, we present **Aligned Dense Supervision (ADS)** to simultaneously boost the overall density of supervision and the body-wise collaboration of local&global tasks. We first build a baseline: a Mask-RCNN-based network that conducts separated regression of local SMPL parameters and global translation. Then the baseline is upgraded by ADS training techniques and network modules for two main goals: **(a)** Denser combination of local poses and global locations from training data, where we align body-wise ROI features among all ranges preserving aspect-ratio, scale and body-to-ROI layout, we also augment mesh locations in 3D space. **(b)** Aligned supervision on the local and global tasks, where we share the body-wise internal features between the local and global tasks, and merge the global task as a residual regression with the local task. In short, ADS aligns features and distributes the regression burden among bodies and tasks. Figure 1 shows ADS results in a wide distance range.

2. Related Work

We only revisit the monocular visual solutions for 3D body mesh estimation and localization.

2.1. Local 3D Mesh Estimation

Mesh Representation. Most recent work represent human body meshes as the UV-maps such as DensePose [6], the SMPL coefficients such as SMPLify [4] or Generative Human Model (GHUM) [32].

Robustness. OOH [29] and PARE [15] incorporate specialized sub-networks to detect visible body parts, and improve single-person mesh regression under occlusion. THUNDR [37] estimates single-person GHUM mesh and Pesudo 3D locations using a intermediate marker representation. VIBE [14] and HuMoR [26] leverage temporal consistency. HuMoR [26] also impose constraint of ground contact and motion priors learned from nearby frames. ROMP [34] imposes the Collision-Aware loss to repel mutually occluded body centers in 2D space. CRMH [9] and ROMP [34] also regularizes multi-person regression with the interpenetration loss and depth order-aware loss.

By contrast, we improve the local SMPL mesh regression robustness by denser supervision in both the 2D pixel space and the 3D space. We also optimize the local and global tasks in a body-wise manner. **Fusion with Object Detectors.** skeleton with the 3D meshes. Zanfir *et al*. [35] further employ multiple scene constraints to optimize the multiperson 3D mesh results. Jiang *et al*. [9] conduct local SMPL regression using the RoI-aligned features of a Fast-RCNN [27] based network. For fast inference, researchers add additional output channels for local SMPL coefficients in single-stage objectors. For instance, BMP [38] and ROMP [34] leverage CenterNet to conduct grid-wise bbox and SMPL regression, which improves truncated meshes over the ROI-Align-based methods [9].

In our approach, we simultaneously address the mesh regression challenges on close-range truncated bodies and long-range small bodies in the Mask-RCNN architecture using advanced alignment and supervision techniques.

2.2. Global 3D Localization

Global mesh localization. Most of the existing multi-person SMPL mesh estimation approaches [9, 34–36] allow us to compute the pseudo 3D mesh coordinates by shifting local meshes in 3D with weakly-perspective camera parameters (2D offset + 2D scale). SPEC [16] conducts single-person localization in the world coordinates by estimating camera poses through contextual clues (e.g. horizontal line) in the images. There are still very few work for global 3D localization of multiple body meshes, mostly following the architectures in [35, 36]. They solve the problem indirectly with multiple separated stages and networks, such as single person 3D-joints $\rightarrow$ single 3D shape fitting in [36] and body part detection$\rightarrow$ skeleton grouping $\rightarrow$ 3D shape fitting [35]. BEV [28] conducts pixel-wise multi-branch regression of body 2D centers, local SMPL parameters and body depths, where the depth regression is based on an anchor-offset structure supervised in the top view (Bird-Eye-View).

General 3D object depth and pixel-wise depth estimation. Most of these work focus on the autonomous driving scenarios. In [5, 21, 39], 3D bounding boxes are estimated in physical bbox center depth, size, and orientation. BTS [18], DAV [7] and AdaBins [3] estimate visible surface depth leveraging the transformer-based mechanism to model the dense spatial dependency.

In our work, we directly estimate the multi-person body meshes in physical sizes and locations. Our work is inspired by [3, 5, 21, 39] to estimate the 3D global translation. By fusing the local and global task, our method decouples the body sizes and 3D translations that were mixed-up in existing mesh recovery methods.

3. Our Approach

Notations: The model input $\mathbf{I} \in \Re^{W\times H\times 3}$ is an RGB image [1]. The outputs are the body meshes in pelvis-centered local coordinates and their body-wise 3D translation vector in the global 3D coordinates. Each body mesh contains 6890 vertices with their 3D coordinates $\mathbf{M} \in \Re^{6890\times 3}$ and neural adult SMPL coefficients $\{\theta, \beta\}$ [22]. The pelvis of

[1] All non-bold letters represent scalars. Bold capital letter $\mathbf{X}$ denotes a matrix; Bold lower-case letters $\mathbf{x}$ is a column vector. $\mathbf{x}_i$ represents the i^{th} column vector of the matrix $\mathbf{X}$. x_j denotes the j^{th} element of $\mathbf{x}$. $\langle \mathbf{x}, \mathbf{y}\cdot\rangle = \mathbf{x}^T\mathbf{y}$ denotes the inner-product between two vectors or metrics.

$\mathbf{M}$ is at the origin of the local 3D coordinates, with their x-y-z coordinates all in the range of $[-1, 1]$ meters. $\beta \in \Re^{10\times 1}$ is the top-10 PCA coefficients of the SMPL statistical shape space. $\theta \in \Re^{6\times 24}$ is the 3D rotation of the 24 body joints in a 6D representation. The body-wise 3D translation vectors $\mathbf{t} \in \Re^{3\times 1}$s are in meters. The *origin* of the global 3D coordinates system is placed at the center of the image plane with its Z-axis perpendicular to the image plane.

Problem Description: We solve a regression problem from an RGB image $\mathbf{I}$ to multi-person mesh coefficients and translations $\{\theta, \beta, \mathbf{t}\}$. Upon a multi-branch baseline composed from [5, 9, 21, 39], we add Aligned Dense Supervision(ADS) components to achieve the state-of-the-art mesh localization results in the global 3D coordinates system.

3.1. Baseline Architecture

Our baseline network (the blue components in Fig. 2) conduct four sub-tasks: (1) 2D person detection ("Bounding box head" and "Classification head"); (2) local SMPL coefficient regression ("Local SMPL head"); (3) Weak-perspective camera parameter estimation ("SMPL $\mathbf{M}(\theta, \beta) \rightarrow \{\mathbf{J}_{2D}, \mathbf{J}_{3D}\}$"); and (4) Global 3D body center regression ("$\pi \rightarrow \mathbf{t}$"). Without loss of generality, we incorporate these sub-tasks under a classic ResNet50-FPN-Faster-RCNN [27] framework. In training the baseline, each of the sub-task loss functions is summed over all body instances.

We first borrow the 2D detection losses and local SMPL regression losses from [9, 34]:

$$\mathcal{L}_{Detection} = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{cls}\mathcal{L}_{cls}, \tag{1}$$

includes the bound box regression and person classification.

$$\begin{aligned}\mathcal{L}_{SMPL} \quad = \quad & \lambda_{shape}\mathcal{L}_{\beta} + \lambda_{pose}\mathcal{L}_{\theta} + \lambda_{j3d}\mathcal{L}_{j3d} + \lambda_{j2d}\mathcal{L}_{j2d} \\ & + \lambda_{JPE}\mathcal{L}_{JPE}, \end{aligned} \tag{2}$$

includes the Mean Square Error(MSE) losses of the SMPL coefficient $[\theta, \beta]$, local 3D joints $\mathbf{J}_{3D} \in \Re^{24\times 3}$ (linear projection from $\mathbf{M}(\theta, \beta)$), and 2D joints $\mathbf{J}_{2D} \in \Re^{24\times 2}$ on the 2D image plane. $\mathcal{L}_{JPE}$ includes the MPJPE&PA-MPJPE losses [34]. In calculating $\mathcal{L}_{j2d}$ in existing work such as [9, 34, 38], the 2D joints $\mathbf{J}_{2D} \in \Re^{24\times 2}$ are computed from the local 3D joints $\mathbf{J}_{3D}$ of the same person. Their relationship is defined by weak-perspective projection:

$$\mathbf{J}_{3D} = \big[s(\mathbf{J}_{2D} - [t_x, t_y]), sf\big], \tag{3}$$

where $[t_x, t_y]$ contains the 2D pixel coordinates of the body center on the image, f is a manually-set focal length, s is the scaling factor between the 2D enclosure box $\mathbf{J}_{2D}$ and the detected 2D bounding box. In optimizing $\mathcal{L}_{j2d}$, the weak-perspective camera parameters $\pi = [t_x, t_y, s]$ are estimated as by-products. Using the weak-perspective parameters π and the focal length f, one can produce a Pseudo 3D location of each mesh by shifting $\mathbf{M}$ along a Pseudo 3D translation vector according to [9].

$$\mathbf{t} = \begin{bmatrix} 2(t_x\alpha + c_x - W/2)/(s\cdot\alpha) \\ 2(t_y\alpha + c_y - H/2)/(s\cdot\alpha) \\ 2f/(s\cdot\alpha) \end{bmatrix}, \tag{4}$$

where α is the longer length of the 2D bounding box. $[c_x, c_y]$ is the 2D center coordinates of the bounding box. Under the weak-perspective assumption above, the body sizes and the depths are coupled. Meshes of different sizes may be interpreted as the same size at different depths.

Instead, our baseline directly estimates the person-wise 3D translation vectors $\mathbf{t} \in \Re^{1\times 3}$ by adding a global pelvis head after the ROI-Align features. The body mesh $\mathbf{M} + \mathbf{1t}$ now contains vertices in global 3D coordinates ($\mathbf{1} \in \Re^{6890\times 1}$ is the vector of 1s). This additional global task is one step forward from the Pseudo 3D localization in existing work. In more detail, $\mathbf{t}$ is estimated with direct supervision of the ground truth body center in the global (camera) 3D coordinate, denoted by $\mathbf{t}_{gt} \in \Re^{1\times 3}$. Formally, the loss for the global 3D localization task is:

$$\mathcal{L}_{Global} = \lambda_{translate}\|\mathbf{t} - \mathbf{t}_{gt}\|_2^2. \tag{5}$$

The overall training loss of our baseline hence aggregates all the above losses:

$$\mathcal{L}_{Baseline} = \mathcal{L}_{Detection} + \mathcal{L}_{SMPL} + \mathcal{L}_{Global}. \tag{6}$$

Moreover, to push the limit of the existing techniques on the Baseline, we include the up-to-date training tricks from [9, 34], occlusion-aware data augmentation from [38], and adversary training of SMPL coefficients from [9]. To save training time, we did NOT use the Depth Ordering-Aware and Interpenetration loss in [9, 34].

Baseline Limitation (1): Mis-aligned supervision. In the case of occlusion and truncation, the ROI captures body-wise features in different aspect-ratio and body-to-ROI layouts among different bodes. Fig. 3 illustrate the different alignment strategies of our anchor-based baseline and an anchor-free baseline(BEV [28]). Moreover, in estimating the body mesh of the same person, the Local SMPL head and Global pelvis head originate separately from the ROI features. The lack of cooperation in joint feature selection and error fitting leads to a redundant learning burden of two network branches and potentially low performance. In particular, the Global pelvis head does not leverage the fine-grained local features within the local SMPL head.

Baseline Limitation (2): Sparse supervision. (*i*) The local loss $\mathcal{L}_{SMPL}$ (Eq. 2) is supervised by the SMPL coefficients and coordinates of 24 body joints. The changes of individual coefficients/joints tend to sparsely affect more on some mesh vertices than others. (*ii*) The global loss $\mathcal{L}_{Global}$ is

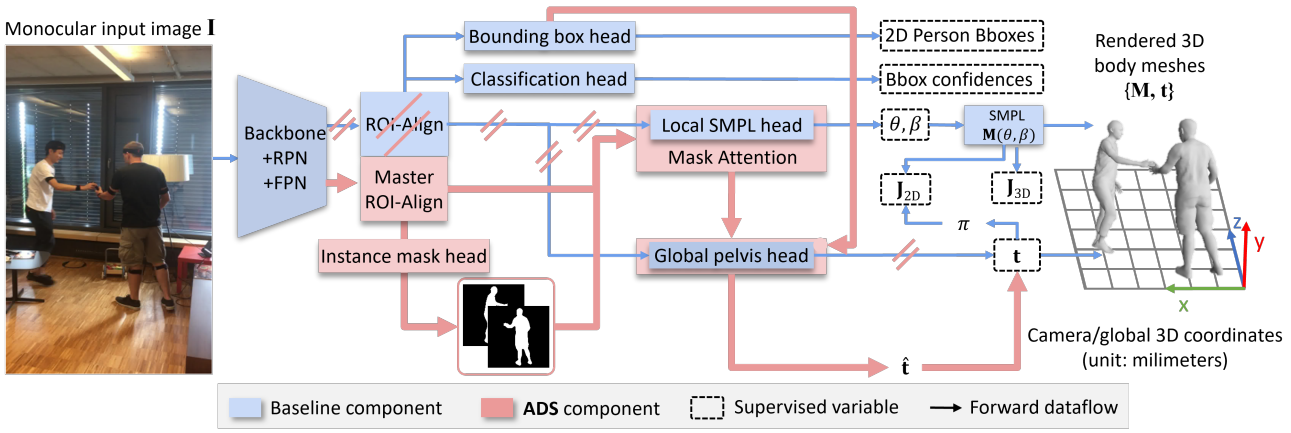

Figure 2. **Our Framework**. Baseline components (Section 3.1) are in blue. Our Aligned Dense Supervision(ADS) components (Section 3.2) are added upon baseline in pink, which include master ROI-Align, instance mask head, serially fused local-SMPL-Global-pelvis head and removing a few baseline components by "//". Training losses are computed on the dashed-box "Supervised Variable".

supervised with the GT 3D translation vectors. Compared with the dense body occurrence in 2D image databases (the sum of person mask area over image area in MS-COCO is over 60%), in the datasets for 3D tasks, the occurrence of the training instances in the 3D space is very sparse. (i.e. the overall mesh volume over the scene volume in the 3DPW training set is only 0.12%). The loss $\mathcal{L}_{Global}$ hence is naturally biased to the densely occupied regions of the 3D space. These sparse supervision issues make the local and global regressors very difficult to maintain consistent performance over the full 3D physical space (Fig. 3).

3.2. ADS upon the Baseline

To address the above limitations, we (*i*) align the person-wise local and global tasks by shared internal features and fused regression heads, (*ii*) increases the density of supervision in the pixel-wise, vertex-wise, and depth-wise perspectives. ADS components are added upon the baseline as pink components in Fig. 2 and qualitatively illustrated in Fig. 3).

(1) Master RoI-Align. In the standard ROI-Align operation in our baseline (as well as in CRMH), only the person-wise ROI enclosing the visible body parts is cropped from the backbone output features and resized to a fixed resolution. Under truncation, occlusion and low resolution, the standard RoI-Align deforms and mis-aligns the body layouts among persons. We introduce Master RoI-Align that pads out-of-scene and occluded features of RoI, preserves the body aspect ratios and keeps the mesh-to-ROI layout. To establish bbox reference of the whole body instead of only the visible body parts, we also re-trained the bbox detectors using the enclosing box over the 2D projection of GT meshes. Moreover, we share the Master RoI-Aligned features to both the local SMPL regression and the global translation regression tasks. See Fig. 3 for an example of our Master RoI-Align comparing to existing ROI operations. Implementing Master ROI-Align operation for a batch of mesh-wise feature maps is *non-trivial*. We developed an efficient CUDA implementation.

(2) Mask-attention Local SMPL Head. In the input feature tensor to the Local SMPL head, the feature elements originating from occluded image pixels are toxic to the regressor, which ideally can be improved with a perfect 2D binary person mask that selects the visible features. Following Mask-RCNN [1], an instance mask branch can be easily added to the baseline, but the predicted masks cannot be directly used for two reasons: **(a)** The estimated binary person mask via [1] is noisy and lacks spatial details. **(b)** The predicted instance mask confidence map (the tensor after Softmax before binarization) is supervised by the cross-entropy loss that averages errors over pixels, therefore the masking accuracy is biased to the torso over the limbs.

To smoothly introduce the mask supervision, we introduce a Mask-attention Local SMPL head(See the middle container in Fig. 4 for the head structure.). This head concatenates the predicted instance "mask" (the tensor after Softmax *before* binarization) with the RoI feature, where the mask branch and the SMPL head are trained jointly. This structure actually makes the mask branch a dense self-attention block to the SMPL regression branch. This head introduces an additional instance mask loss $\mathcal{L}_{Mask}$ (the same as in [1]) onto the baseline.

(3) Global pelvis head. We upgrade the estimator of the 3D translation vector $\mathbf{t}$ in Eq. (5) by (**a**) fusing the global task with the local SMPL regression head, such that the global translation regression task is reduced to a to residual translation regression associated with the 2D bbox task of the same mesh. The fused task is supervised by a global residual loss $\mathcal{L}_{GRes}$. (**b**) Densely supervising the global task

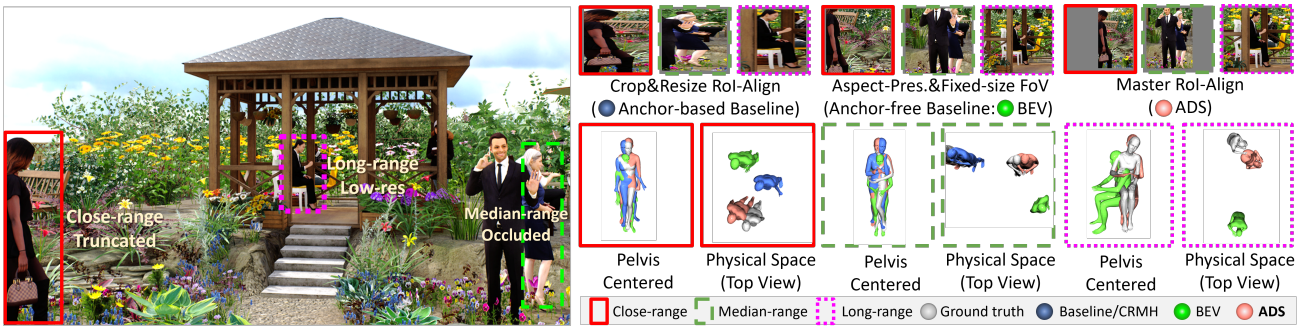

Figure 3. ADS benefits the close-to-far bodies over the Anchor-based Baseline (Section 3.1) and Anchor-free Baseline (BEV [28]). **Upper-right row:** Master RoI-Align(ADS) preserve body-to-ROI layouts and aspect-ratios. **Lower-right row:** ADS balances body-wise local and global tasks, resulting in the closest meshes to GTs in both the Pelvis-centered(local) space and the physical(global) space.

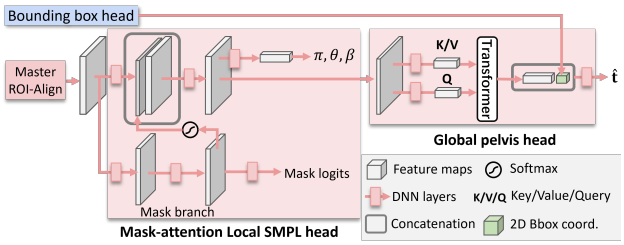

Figure 4. The local and global tasks of ADS share the body-wise ROI features in a serially fused regression head (Fig. 2), which consists of a Mask-attention SMPL regression head (middle container) and a Global pelvis regression head (top-right container).

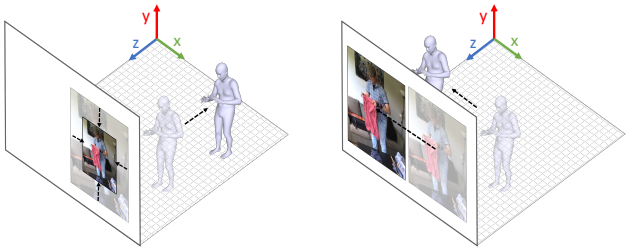

Figure 5. Global 3D Padding Augmentation: translating and scaling of the GT meshes in global 3D coordinate in response to the scaling and shifting of the input images respectively.

jointly with the local task by mesh vertices (with the dense vertex loss $\mathcal{L}_{GVertex}$). See the top-right container in Fig. 4 for the global pelvis head structure. The details of the above two new losses are described below.

(a) $\mathcal{L}_{\mathbf{GRes}}$. Independently supervising the global translation vector $\mathbf{t}$ in 3D as in Eq. (5) is very challenging. Thanks to the advances in 2D bounding box detection and a large amount of 2D training data, the 2D bounding box coordinates can help global 3D localization. We align the x-y translation of the 3D localization with the x-y translation from the 2D bounding boxes (i.e. t_x, t_y) for initialization, and reduce the global prediction on $\mathbf{t}$ to a residual prediction of $\hat{\mathbf{t}} = [\hat{t}_x, \hat{t}_y, d]$. Here $\hat{\mathbf{t}}$ is a translation vector from the weak-perspective projected body mesh center to the actual global 3D coordinates and d is the body-center depth in global 3D coordinates from the image plane. Formally, our global translation vector $\mathbf{t}$ is re-written as:

$$\mathbf{t}(\hat{\mathbf{t}}) = \begin{bmatrix} d \cdot (t_x \alpha + c_x - W/2) + \hat{t}_x \\ d \cdot (t_y \alpha + c_y - H/2) + \hat{t}_y \\ d \end{bmatrix}, \tag{7}$$

where α denotes the length of the longer side of the bounding box and W, H are the image width and height.

Recall that the input ROI features to the Mask-attention local SMPL head are enhanced by the mask attention, as shown in the middle container in Fig. 4. The serial fusion of global heads to the local head focuses the global head on the locally enhanced ROI features. In the global head, we further explore the depth-sensitive features by a transformer layer, inspired by [3]. The transformer layer learns a weight-averaged global descriptor to query the most persuasive features within the ROI. This significantly simplifies the network structure, and benefits the local and global tasks of the same person. The global loss between $\mathbf{t}(\hat{\mathbf{t}})$ and the GT translation $\mathbf{t}_{gt}$ is defined as:

$$\mathcal{L}_{GRes} = \lambda_{translate} \|\mathbf{t}(\hat{\mathbf{t}}) - \mathbf{t}_{gt}\|_2^2. \tag{8}$$

In the supplementary materials, we provide detailed Pseudo-algorithm on how to calculate $\mathbf{t}_{gt}$.

(b) $\mathcal{L}_{\mathbf{GVertex}}$. For a deeper bond between the local and global tasks, we further supervise them densely with GT vertices in the global 3D coordinates. From the ground truth $[\theta_{gt}, \beta_{gt}]$, we first reconstruct vertices $\mathbf{M}_{gt|pelvis} \in \Re^{6890\times3}$ in the pelvis-centered local coordinate. In the datasets that provide global GT SMPL, the reconstructed GT mesh $\mathbf{M}_{gt}$ is in the world coordinate orientation with a world translation $\mathbf{t}_w \in \Re^{1\times3}$. To transform the world coordinate mesh $\mathbf{M}_{gt} + \mathbf{1}\mathbf{t}_w$ to the global camera coordinates GT mesh $\mathbf{G} \in \Re^{6890\times3}$, we compute a 4d homogeneous

coordinates $\mathbf{H} = (\mathbf{M}_{gt} + \mathbf{1}\mathbf{t}_w)\mathbf{E}^T \in \Re^{6890\times 4}$, normalizing the first three dimensions and remove the 4th dimension. Here $\mathbf{1} \in \Re^{6890\times 1}$ is the vector of 1s and the extrinsic matrix $\mathbf{E} \in \Re^{4\times 4}$ is computed from the GT frame-wise camera pose parameters. The global vertex loss is computed as

$$\begin{aligned}\mathcal{L}_{GVertex}(\theta,\beta,\mathbf{t}) &= \|\mathbf{M}(\theta,\beta) - \mathbf{M}_{gt|pelvis}\|_2^2 \\ &+\|(\mathbf{M}(\theta,\beta)+\mathbf{1}\mathbf{t}) - \mathbf{G})\|_2^2, \quad (9)\end{aligned}$$

where θ and β are the estimated SMPL coefficients. $\mathbf{M}(\theta,\beta) \in \Re^{6890\times 3}$ contains the vertex coordinates computed from θ and β. $\mathbf{t} \in \Re^{1\times 3}$ again is the estimated translation vector from the local 3D coordinates to the 3D global camera coordinates. Comparing to Eq. (5) that learns from a single translation vector, Eq. (9) imposes vertex-wise supervision on both the vertices and the translation vectors.

(4) Global 3D Padding Augmentation. To generate dense 3D mesh locations, we augment the input images with random scaling and shifting while keeping the original image size ratio by padding the background pixels. Accordingly, we augment 3D GT meshes by shifting in the global 3D coordinate. Fig. 5 shows examples of the scaling and shifting augmentation respectively. This augmentation provides denser supervision in the global 3D coordinate system.

Aggregating all components above, the overall ADS training loss is:

$$\begin{aligned}\mathcal{L}_{ADS} =& \mathcal{L}_{Detection} + \mathcal{L}_{SMPL} + \lambda_{Mask}\mathcal{L}_{Mask} \\ &+ \lambda_{GRes}\mathcal{L}_{GRes} + \lambda_{GVertex}\mathcal{L}_{GVertex}. \quad (10)\end{aligned}$$

4. Experiments

Training protocols: Recent work on SMPL mesh estimation used very different extra training datasets and different GT SMPL parameters. Many of them were not released for license issues. For instance, VIBE, SPIN, CRMH, and BMP used the licensed codes from [23] to generate high-quality GT SMPL meshes on Human3.6M. This makes it nearly impossible to conduct a strictly fair comparison among all methods. We align our experiments with the most recent training datasets mentioned in BEV. (See supplementary material for method-specific training datasets).

Training configuration: As most previous approaches on multi-person mesh detection, our Baseline and ADS training consists of three steps. The training databases include the training sets of 3DPW [31], Human3.6M [8], AGORA [25], MPI-INF-3DHP [24], MS-COCO [20], LSP [10], LSP Extended [11], MPII [2], and CLIFF(pseudo-GT) [19]. The RAdam optimizer with weight_decay=$1e^{-4}$ is used in all steps. In *Step-1*, the framework is trained from scratch using the cropped single-person patches(cropped, resized, and padded to 256×256). This step mainly initializes the local SMPL regression head. All losses except the global losses are activated in this step. This is because the global distance cues are destroyed in the cropped single-person images. This step is trained for 76 epochs with a batch size of 128 and a learning rate of $1e^{-4}$. In *Step-2*, the framework is fine-tuned on multi-person images (resized and padded to 512×832) for multi-person mesh detection. Note that the global losses are only activated in the Step-2. This step is trained for 55 more epochs with a batch size of 48 and a learning rate of $1e^{-5}$. The global padding augmentation was activated with a 50% probability with random scaling of $(\times 0.8, \times 1.2)$ and translation to random but valid 3D positions. In *Step-3*, the framework is further fine-tuned with higher sampling probability on the training set of each evaluation database. More details on the loss weights and database sampling probabilities are provided in supplementary materials.

Evaluation datasets: We used three databases contains both the ground truth of SMPL parameters and extrinsics.

- **3DPW** [31] contains (24 train, 24 test, 12 validation) image sequences captured by a hand-hold camera with annotations of SMPL coefficients, 3D joints, 2D joints and frame-wise camera poses. Subjects are mostly walking and captured in the horizontal view.
- **Human3.6M** [8] contains 3.6 million RGB images and 3D body joint coordinates of 11 actors in the indoor environment. Subjects are mostly dancing and captured in the horizontal view. We use the SMPL GT generated by [34].
- **AGORA** [25] is a synthetic dataset with ground truth of body meshes and 3D translations, rendered using 4240 textured body scans in diverse poses and clothes. It contains 14K training and 3K validation images in the pelvis-to-camera distance of $[1.8, 27.4]$ meters and pelvis altitude of $[-7.6, 9.0]$ meters.

Metrics: Four metrics are reported in millimeters(**mm**): (a) Procrustes-Aligned Mean Per Joint Position Error **(PA-MPJPE)**; (b) Procrustes-Aligned Per-Vertex Error **(PA-PVE)**; (c) Global Pelvis Error **(GPE)** computed as the Euclidean distance error of the pelvis joints in physical space; (d) Global Per-Vertex Error **(GPVE)** computed as the Euclidean distance error of the mesh vertices in physical space. In short, **PA-MPJPE** and **PA-PVE** measure the local mesh errors, **GPE** measures the global location error of the pelvis, and **GPVE** measures the global location error of mesh. For fair comparisons, we unified the evaluation process by adopting a canonical focal length of 1000 on all datasets, and adjusted ground truth translations accordingly.

4.1. Comparing with State-of-the-arts

In Table 1, we compare ADS with the state-of-the-arts on three evaluation datasets. For fairness, we list results with the ResNet50 backbone. We copied all the local metrics

Table 1. Comparison with the state-of-the-arts on the 3DPW test, Human3.6M test and AGORA val sets. All methods used extra training data (see supplementary materials) beyond the evaluated datasets. The metrics of previous methods were either from their papers or their released models. All metrics are in millimeters (mm) and are the smaller "↓" the better. All methods use the neutral SMPL model ***except*** BEV [28], which uses the SMPL-Age model and GT age annotations to disentangle the child mesh shapes from adults.

Method	Evaluation Dataset: 3DPW test set				Evaluation Dataset: Human3.6M test set				Evaluation Dataset: AGORA val set			
	Local Metrics		Global Metrics		Local Metrics		Global Metrics		Local Metrics		Global Metrics	
	PA-MPJPE↓	PA-PVE↓	GPE↓	GPVE↓	PA-MPJPE ↓	PA-PVE↓	GPE↓	GPVE↓	PA-MPJPE ↓	PA-PVE↓	GPE↓	GPVE↓
Bbox-Cropped, Single Person												
SPIN [17]	68.1	-	-	-	-	-	-	-	-	-	-	-
PARE [15]	51.2	-	-	-	-	-	-	-	-	-	-	-
VIBE [14] (video)	51.9	-	-	-	-	-	-	-	-	-	-	-
HMR [13](video)	-	-	-	-	56.8	-	-	-	-	-	-	-
MotionBERT [40](video)	49.1	-	-	-	-	-	-	-	-	-	-	-
Detected Bbox, Multi-Person												
BMP [38]	63.8	-	-	-	51.3	-	-	-	-	-	-	-
CRMH [9]	62.3	80.4	420.0	442.0	52.7	68.3	279.2	295.4	77.6	105.7	3273.6	3295.2
SPEC [16]	52.2	81.0	2759.8	2761.7	**49.7**	64.1	1507.3	1510.1	-	-	-	-
ROMP [34]	49.7	77.9	3134.9	3146.4	-	-	-	-	-	-	-	-
BEV [28](SMPL-Age)	46.9	**76.1**	1898.8	1919.0	51.7	**63.6**	540.1	522.3	61.4	84.4	3020.1	3005.6
Ours												
Baseline	54.1	80.7	286.7	313.4	65.3	85.9	270.9	297.3	64.5	85.2	815.1	842.9
ADS	49.2	77.9	251.1	283.8	53.9	72.3	**217.5**	**230.6**	**58.9**	**78.3**	**705.6**	**731.6**
ADS-stage-IV(from-stage3EP10)	46.0	76.75	306.5	337.9	53.9	72.3	**217.5**	**230.6**	**58.9**	**78.3**	**705.6**	**731.6**
ADS-stage-IV(from-stage3EP10)	**45.3**	78.3	397.9	427.4	53.9	72.3	**217.5**	**230.6**	**58.9**	**78.3**	**705.6**	**731.6**
ADS-stage-V	48.5	77.7	**249.8**	**282.9**	53.9	72.3	**217.5**	**230.6**	**58.9**	**78.3**	**705.6**	**731.6**

reported in papers, and compute missing metrics using their released models. For single-person methods that use bbox-cropped image patches, we only copied the local metrics reported in papers. For multi-person methods (CRMH and ROMP) that do not conduct explicit global 3D localization, global metrics can only be computed using their predicted weak-perspective parameters. All methods use the SMPL-neutral model ***except*** BEV [28], which uses the SMPL-Age model and GT age annotations to disentangle the child mesh shapes from adults. This gives BEV an extra advantage on local mesh metrics over all other methods.

Table 1 shows that ADS works comparable or better locally (PA-MPJPE, PA-PVE) than all SMPL-neutral models on 3DPW, and by far the best globally (GPE, GPVE) among all compared methods. On the AGORA val set, over the 3D physical space of $[1.8, 27.4]$-by-$[-7.6, 9.0]$ meters, ADS not only got best local metrics, but is the only method that brings the global metrics, GPE and GPVE, below 1000 millimeters. Although ADS always improves over baseline, it struggles a bit on the local metrics on Human3.6M for two reasons: (1) In our **common** Step-1&2 training, most datasets contains daily actions instead of the indoor dancing poses in Human3.6M. (2) the lack of truncations and occlusions reduced the benefit of our Master ROI-Align. This is amendable by a Human3.6M-specific training recipe.

We also closely compare ADS with CRMH and BEV which produce the next two best sets of global metrics. In Table 2, we compute GPVEs in evenly splits of pelvis-to-camera distance ranges and pelvis altitude ranges on the AGORA val set, respectively. ADS constantly outperforms others in all ranges. Among the rendered 3D meshes in Fig. 6, ADS meshes are indeed the closest meshes to the ground truth meshes in the global 3D coordinates.

Why CRMH, ROMP, SPEC and BEV have worse global metrics than ADS? Besides the differences in network structures, these methods have basic modeling issues on global localization. Based on weak-perspective projection, CRMH and ROMP align meshes in the 2D image space rather than in the global 3D space. CRMH pursues the image-space mesh alignment by translating "small" person away and "big" person close to the image origin. ROMP achieves the image space alignment by scaling the mesh sizes. SPEC leverages extra natural-scene training images to learn a regressor for camera parameters: camera-space translation, camera pose and camera translation, which do not generalize well on human images in unseen scenes. BEV explicitly scales mesh sizes based on the estimated ages, the image space mesh projection and the mesh-to-mesh relative depths. Such mesh-size-scaling operations introduce substantial regression errors to global mesh translation. Compared to ROMP and BEV, CRMH got better global metrics on 3DPW and Human3.6M, thanks to its two-stage bbox detector that provides more reliable close-range weak-perspective parameters. This advantage diminished on AGORA due to heavy truncation, occlusion and the larger distance ranges. In ADS, we closely integrate the local and global cues (Fig. 4) for mesh translation in the global 3D space, which leads to physically valid mesh sizes and global mesh locations.

Table 2. Global Per-Vertex Error (GPVE) on the AGORA val set organized by pelvis-to-camera distances (Close: [1.8, 10.3]m, Median: [10.3, 18.8]m, Far: [18.8, 27.4]m) and by pelvis altitudes (Low: [-7.6, -2.1]m, Median: [-2.1, 3.4]m, High: [3.4, 9.0]m).

Method	GPVE overall↓	GPVE by Distance↓			GPVE by Altitude↓		
		Close	Median	Far	Low	Median	High
CRMH [9]	3295.2	975.1	2257.1	5250.6	3530.7	3177.1	6439.5
BEV [28]	3005.6	1263.7	2030.1	4074.9	4922.8	2937.7	7163.6
ADS(ours)	**731.6**	**222.0**	**576.6**	**1623.0**	**698.1**	**594.8**	**919.2**

Table 3. Ablation study of ADS on the 3DPW test set(w/o AGORA train set). All metrics are the smaller the better. "gain" highlights the improvements over the previous row.

Method	Local Metrics				Global Metrics			
	PA-MPJPE↓	gain	PA-PVE↓	gain	GPE↓	rel.	GPVE↓	gain
Baseline	60.19	0	85.49	0	272.06	0	308.57	0
+ Mask-att. Local Reg.	55.44	↓**4.75**	79.84	↓ **5.65**	271.49	↓0.57	303.10	↓5.47
+ Master RoI-Align	52.33	↓3.11	**77.44**	↓2.40	273.81	↑2.32	301.82	↓1.28
+ Global-pelvis reg. head	51.63	↓0.70	78.55	↑1.11	256.25	↓ **17.6**	290.81	↓ **11.0**
+ Global padding aug.	**51.24**	↓0.39	78.07	↓0.48	**245.32**	↓10.9	**280.35**	↓10.5

4.2. Ablation Study on the 3DPW Test Set

We examine the impact of ADS components by adding them to the baseline sequentially. As reported in Table 3, the biggest gain to the local metrics comes from Mask-attention SMPL regression with *4.75 mm* less PA-MPJPE and *5.65 mm* less PA-PVE. The biggest gain on global metrics comes from the global body pelvis regression head and Global 3D padding augmentation with over *20 mm* combined improvements on both GPE and GPVE.

5. Conclusion and Discussion

In this work, we presented a physical space solution to the multi-person mesh localization problem from monocular RGB images. We address the mis-alignment and sparsity issues in training the local and global regression tasks with a bag of novel operators and techniques, called ADS. We learn from this work that: (1) Dense supervision information can be generated from finite training data for better performance. (2) Body-wise feature sharing and regressor fusing between the local and the global tasks is better than training these tasks in separate branches.

Robustness to truncation, occlusion and mesh collision. Our robustness in truncation and occlusion is mainly boosted by our Master ROI-align operation. We also use body segmentation masks (Fig. 4) and vertex losses $\mathcal{L}_{GVertex}$ to address occlusion in the 2D image space and collision in depth respectively. Our burden of robustness is distributed by the sharing of local visible features to global regression, and by partially shifting the global regression task as a residual task in the local SMPL head. When training time permits, we could add the Depth Ordering-Aware and Interpenetration loss from [9, 34] to boost robustness.

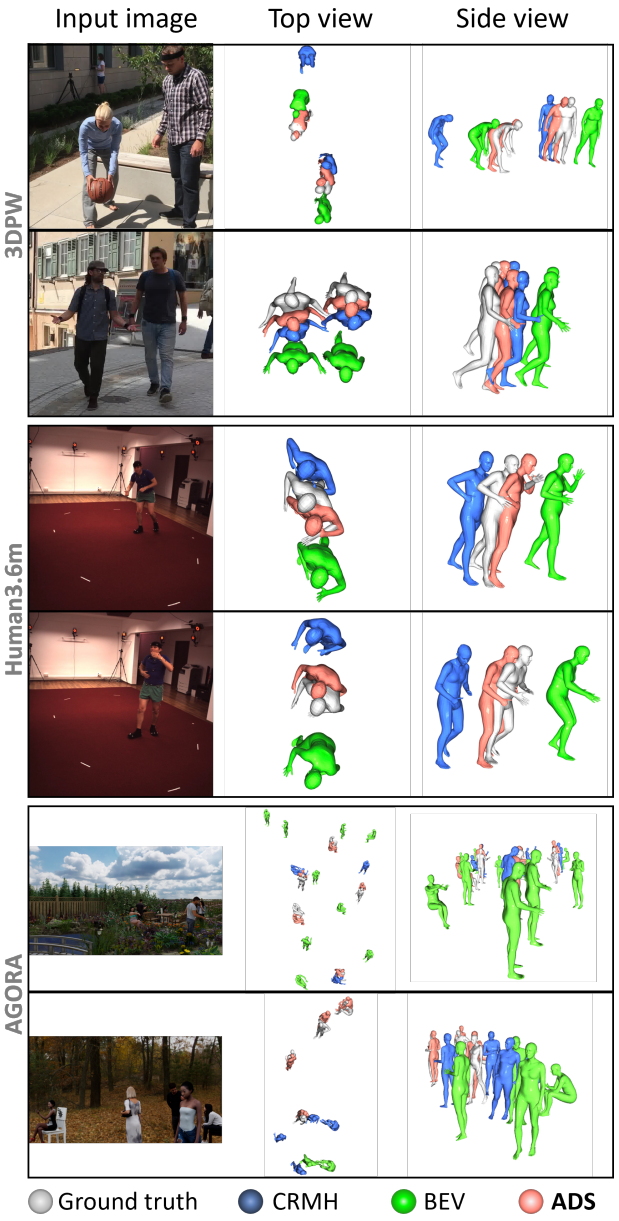

Figure 6. Qualitative results in the physical-space camera coordinate. **ADS**(pink) meshes are the closest to Ground Truth(gray) compared with CRMH(blue) and BEV(green). **More results on real-world images are in the supplementary materials.**

Additional context priors. Explicit priors may help but may also be limited in generalization. For instance, the ground plane assumption [16, 35] requires extra camera pose models and training data, which does not work on images from complex scenes (e.g. stairs in 3DPW, construction sites and garden bushes in AGORA) or uncommon body altitudes (e.g. jumping in Human3.6M, standing on the hills or on the 2nd floor in AGORA). Although there

are natural-image training data to provide camera priors, the lack of generalization (see SPEC [16] global metrics in Table 1) calls for fully annotated GT scene meshes and camera intrinsics/extrinsics in the human mesh datasets.

Social Impact. This work recovers body meshes from cameras, which facilitates many downstream applications and is only applicable to public or privacy-consented scenarios.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 4

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2, 5

[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 *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 2

[5] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 3

[6] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[7] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *ECCV*, 2020. 2

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 6

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 1, 2, 3, 7, 8

[10] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Citeseer, 2010. 6

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

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 7

[14] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2, 7

[15] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ArXiv*, 2021. 2, 7

[16] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11035–11045, October 2021. 1, 2, 7, 8, 9

[17] 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*, 2019. 1, 7

[18] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[19] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 6

[20] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[21] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *CVPR*, 2019. 1, 2, 3

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[23] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. 6

[24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 6

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, June 2021. 6

[26] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. 2021. 1, 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2, 3

[28] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 2, 3, 5, 7, 8

[29] Zhang Tianshu, Huang Buzhen, and Wang Yangang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[30] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1

[31] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 6

[32] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum "&" ghuml: Generative 3d human shape and articulated pose models. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[33] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-andcompare. In *ICCV*, 2019. 1

[34] Sun Yu, Bao Qian, Liu Wu, Fu Yili, Michael J. Black, and Mei Tao. Monocular, one-stage, regression of multiple 3d people. In *arxiv:2008.12272*, August 2020. 1, 2, 3, 6, 7, 8

[35] 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*, 2018. 1, 2, 8

[36] 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 *NerIPS*, 2018. 1, 2

[37] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12971–12980, 2021. 2

[38] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 2, 3, 7

[39] Xichuan Zhou, Yicong Peng, Chunqiao Long, Fengbo Ren, and Cong Shi. Monet3d: Towards accurate monocular 3d object localization in real time. In *ICML*, 2020. 1, 2, 3

[40] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: Unified pretraining for human motion analysis. *arXiv preprint arXiv:2210.06551*, 2022. 7