Physical-space Multi-body Mesh Detection Achieved by Local Alignment and Global Dense Learning

Haoye Dong[1], Tiange Xiang[2*], Sravan Chittupalli[1], Jun Liu[1], Dong Huang[1]
[1]Carnegie Mellon University
[2]Stanford University
{donghaoye,liujun,donghuang}@cmu.edu, schittup@andrew.cmu.edu, xtiange@stanford.edu

Abstract

*From monocular RGB images captured in the wild, detecting multi-body 3D meshes in physical sizes and locations is notoriously difficult due to the diverse visual ambiguity and lack of explicit depth measurement. Modern DNN approaches made numerous advances based on either two-stage Region-of-Interests(RoI)-Align or single-stage fixed Field-of-View (FoV) detector frameworks for two main subtasks: local pelvis-centered mesh regression and global body-to-camera translation regression. However, sub-meter-level physical-space monocular mesh detection is still out of reach by existing solutions. In this paper, we recognize two common drawbacks: (1) The local meshes are usually estimated without explicitly aligning body features under image-space scaling, occlusion, and truncation; (2) The global translations are estimated based on a weak-perspective assumption, which tricks the network into prioritizing image-space (front-view) mesh alignment and leads to inaccurate mesh depth. We introduce Physical-space Multi-body Mesh Detection (PMMD), in which (1) Locally, we preserve the body aspect ratio, align the body-to-RoI layout, and densely refine the person-wise RoI features for robustness; (2) Globally, we learn dense-depth-guided features to amend the body-wise local feature for physical depth estimation. With the cleaned local features and explicit local-global associations, PMMD achieves the **best** centimeter-level local mesh metrics and the **first** sub-meter-level global mesh metrics from monocular images in 3DPW and AGORA datasets.*

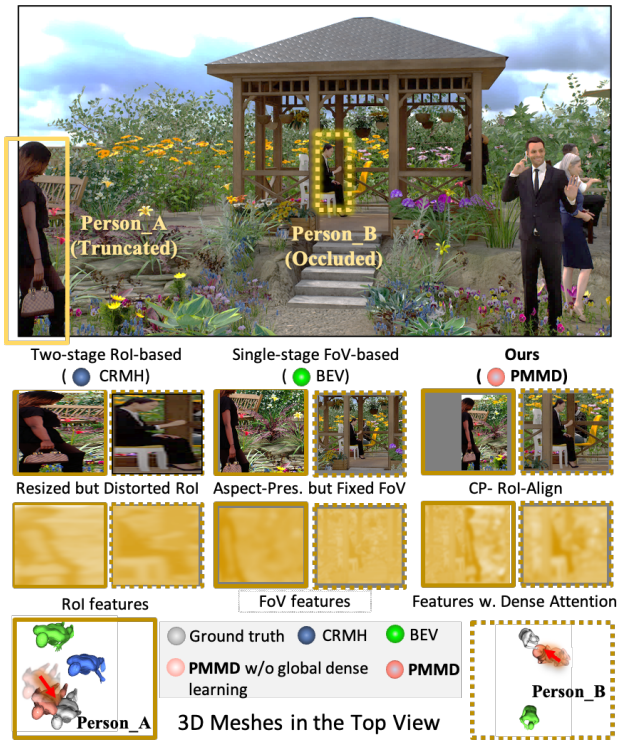

Figure 1. The proposed PMMD technical ideas, compared with CRMH [8] and BEV [28]. **(Local Alignment)** Center-Padding RoI-Align preserves the body aspect ratio and aligns body-to-RoI layout. **(Global Dense Learning)** Integrating local body-wise with global image-wise features to achieve precise translations.

1. Introduction

Multi-body 3D meshes in physical sizes and global locations provide human action / interaction / locomotion information at the surface level. Automatic perception tools for this task could enable boundless applications to promote our safety and well-being. Although many emerging depth sensors facilitate 3D sensing, in most daily scenarios, e.g., shopping malls, parking lots, and warehouses, the low-frame-rate and varying-focal-length monocular RGB cameras are still the most accessible sensors. Our hard journey in monocular-image-based 3D body detection is still ahead against diverse visual ambiguity in appearances, occlusion, truncation, and most importantly, depth measurement.

*Tiange Xiang conducted this work as a research intern at CMU.

Most body mesh detectors map a cropped single-person image patch to a SMPL [23] mesh in the pelvis-centered local coordinate system [11, 13, 15, 16, 19, 20, 27, 29, 30, 32]. More recent approaches address the mutual occlusion and duplicated detection issues by fusing the single-person SMPL regressor with 2D person detectors [8, 28, 34–36, 38]. Two representative frameworks are the two-stage Region-of-Interests (RoI) based methods ([8, 12]), and the single-stage fixed-Field-of-View (FoV) based methods ([28, 34, 38]). Most mesh detectors [8, 34–36] do not estimate body depth. One may derive the pseudo-3D location labels in post-processing using the weak-perspective camera parameters (2D offset + 2D scale). 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 couples the body sizes and depths, thus meshes of different 2D box sizes may be interpreted as meshes of the same size at different depths. BEV [28] introduces a dedicated global depth branch to estimate body depth but prioritizes image-space projection over absolute depth.

Despite advances in robustness and dedicated localization above, we see common drawbacks in solving the two fundamental regression subtasks of (1) local pelvis-centered meshes and (2) global body-to-camera translations. Figure 1 illustrates drawbacks with two example methods: the two-stage RoI-based method CRMH [8] and single-stage fixed-FoV-based method BEV [28]. For local single-body mesh regression, CRMH normalizes RoI features by resizing but distorts the aspect ratio and the body-to-RoI layout for occluded and truncated bodies. BEV produces mesh regression at each feature grid which corresponds to a fixed FoV size. The FoV covers bodies in the original aspect ratio and is robust to occlusion and truncation, but the fixed FoV size makes it burdensome to learn size-invariant regression. For global translation regression, no matter which depth estimators [4, 22, 39] are added to the detector framework as did in BEV [28], which unrealistically assumes the same body is equally optimized in different subtasks. The lack of cooperation in high-level feature selection and loss fitting leads to redundant learning burdens of two subtasks and potentially low performance. To address the common drawbacks above, we construct a novel Physical-based Multi-body Mesh Detection (PMMD), featuring the amalgamation of alignment, dense attention, and global dense learning. The contributions can be summarized as follows:

- We propose a simple yet effective cuda-based Center-Padding RoI-Align (CP-RoI-Align) module that normalizes the body-to-RoI layout of local features.
- We bridge the local body-wise and the global image-wise features, which are capable of predicting precise and physical-like global translation.
- We introduce a novel local dense-attention module to refine local features for robustness.
- We further enhance performance by employing global 3D padding augmentation and introducing global-local dense vertex supervision.

2. Related Work

Single-person mesh recovery. From pre-cropped resized single-person images(usually 224×224), OOH [29] and PARE [14] incorporate specified sub-networks to detect visible body parts and improve single-person mesh regression under occlusion. OOH [29] also augments occlusion data by randomly masking body pixels. THUNDR [37] estimates single-person mesh and Pesudo 3D locations using an intermediate marker representation. VIBE [13] and HuMoR [27] model temporal priors between video frames. HuMoR [27] also imposes constraints on ground contact and motion priors learned from nearby frames. Recent Tranformer-based methods [19, 20, 33] reach very high single-person metrics leveraging long-range dependency within the cropped image patch. However, each patch requires an HRNet-48 backbone, making it too heavy for multi-body cases. Most single-person methods are evaluated on image patches cropped with ground-truth bounding boxes instead of detected bounding boxes.

In our work, we solve a 3D multi-body mesh detection problem based on the monocular image, by creating clean local features and explicit local-global associations in a two-stage RoI-based detection framework while being capable of global translation estimation.

Global Multi-body mesh detection. Most multi-body mesh detection approaches [8, 34–36] only produce pseudo 3D mesh coordinates based on weakly-perspective camera parameters (2D offset + 2D scale). SPEC [15] locates single-person meshes in the world coordinates by estimating camera poses through contextual clues (e.g. horizontal lines) in the images. There is still very little 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] estimate global translation in the Bird-Eye-View regression branch. In [4, 22, 39], 3D bounding boxes are detected with a physical bbox center depth, size, and orientation. BTS [17], DAV [6], and AdaBins [3] estimate visible surface depth leveraging the transformers that model the dense spatial dependency.

In our work, we estimate the multi-body body meshes in physical sizes and locations. Our global translation regression employs a strong global branch in the two-stage RoI-based framework. Own to our aligned and densely refined

RoI features, we further improve performance by compensating global translation regression with local translation regression associated with the same body.

3. Our Approach

Notations: The model input is an RGB image $\mathbf{I} \in \Re^{W\times H\times 3}$ [1] captured under the camera intrinsics $C = [f, 0, W/2; 0, f, 0, H/2; 0, 0, 1]$ with focal length f and image center $[W/2, H/2]$. 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 $\{\theta, \beta\}$ [23]. The pelvis of $\mathbf{M}$ 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 multiple pelvis-centered body mesh coefficients $\{\theta, \beta\}$ and body-wise global 3D camera-to-body translations $\mathbf{t}$. To demonstrate our **explicit** body-wise alignment, feature refinement, and local-to-global association, we implement PMMD components under a two-stage RoI-based detection framework.

3.1. Framework

Delimma of RoI-based and FoV-based framework. Without loss of generality, we consider CRMH [8] and BEV [28] as typical examples of the two-stage RoI-based and single-stage FoV-based models, respectively. CRMH [8] has body-wise RoI features cropped and resized for local mesh regression but relies on the robustness of 2D bounding boxes to deal with occlusion and truncation. Lacking strong global features for depth estimation, CRMH [8] relies on weakly-perspective depth estimation based on the error-prone 2D bounding boxes. On the other hand, BEV [28] estimates body-wise mesh coefficients at the feature grid of pelvises, which is naturally robust to occlusion and truncation. However, each grid has a fixed FoV disregarding different body sizes. Without tight body-enclosed FoV features to provide additional local adjustment, BEV [28] relies on a standalone branch from global features to estimate the global translation.

[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.

Common limitations. (a) Either RoI-based or FoV-based framework does not simultaneously keep the aspect-ratio and body-to-RoI/FoV layouts among different bodes, leading to misaligned body-wise features. Fig. 1 illustrates the different alignment strategies. (b) Both frameworks are sparsely supervised, hence naturally biased to the densely sampled mesh shapes, poses, and locations. For instance, the changes of individual SMPL coefficients and body joint coordinates coefficients/joints tend to sparsely affect more on some mesh vertices more than others. Moreover, the occurrence of the training body 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%). (c) The local and global tasks originate separately from the backbone features (see BEV [28]). The lack of cooperation in high-level feature selection and loss fitting leads to a redundant learning burden of two network branches and potentially low performance.

Components from existing models. From [8, 28, 34, 38], we use the blue components in Fig. 2) for two subtasks: (1) 2D person detection ("Bounding box head" and "Classification head"); (2) local SMPL coefficient regression ("Local SMPL head"). The 2D detection losses and local SMPL regression losses :

$$\mathcal{L}_{2DBox} = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{cls}\mathcal{L}_{cls} \quad (1)$$

$$\mathcal{L}_{SMPL} = \lambda_{shape}\mathcal{L}_{\beta} + \lambda_{pose}\mathcal{L}_{\theta} + \lambda_{j3d}\mathcal{L}_{j3d} + \lambda_{j2d}\mathcal{L}_{j2d} \quad (2)$$

which include the losses of bounding box $\mathcal{L}_{bbox}$, person classification score $\mathcal{L}_{cls}$, the SMPL coefficients $\mathcal{L}_{\beta}, \mathcal{L}_{\theta}$, local 3D joints $\mathbf{J}_{3D} \in \Re^{24\times 3}$, and image-space 2D joints $\mathbf{J}_{2D} \in \Re^{24\times 2}$. We train Eq. 1-2 using up-to-date tricks from [8, 34], occlusion-aware data augmentation from [38], and adversary training of SMPL coefficients from [8]. To save training time, we did not use the Depth Ordering-Aware and Interpenetration loss in [8, 34].

For global 3D locations, we need a global translation loss $\mathcal{L}_{GTrans}$ on the translation vector $\mathbf{t} \in \Re^{1\times 3}$, naively

$$\mathcal{L}_{GTrans} = \lambda_{GTrans}\|\mathbf{t} - \mathbf{t}_{gt}\|_2^2. \quad (3)$$

There are two options to predict $\mathbf{t}$: (a) Pseudo depth estimation [8, 34, 38] based on weakly-perspective projection from the 3D joints $\mathbf{J}_{3D}$ to the 2D joints $\mathbf{J}_{2D}$ or bounding boxes. Under the weak-perspective assumption, the body sizes, poses and depths are coupled. Meshes of different sizes may be interpreted as the same size at different depths. (b) Direct regression of global translation following [4, 22, 25, 28, 39] using single-scale global features with grid-to-grid correspondence to the body center map. In Sec. 3.2, we will introduce **Global-Local Translation Heads** that benefit from both direct global regression and local mesh regression.

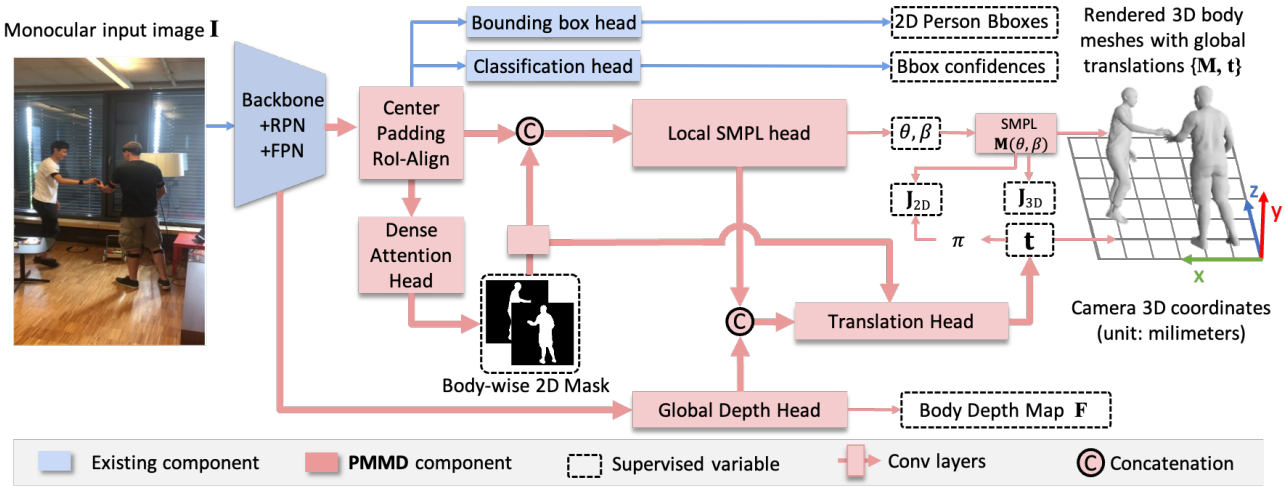

Figure 2. **PMMD based on the two-stage detector Framework**. Existing detector components (Section 3.1) are in blue. Our PMMD components (Section 3.2) are in pink. Training losses are computed on the "supervised variable" in the dashed boxes.

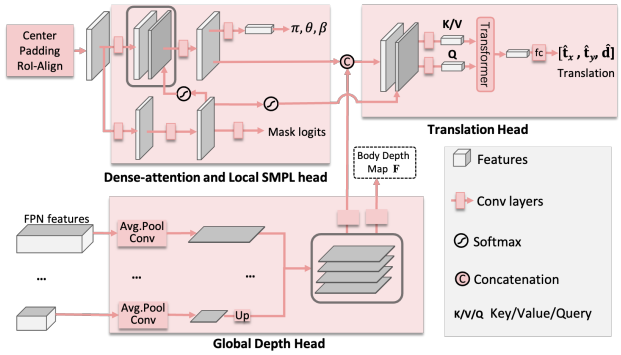

Figure 3. Detailed PMMD head structures. (Also in pink in Fig. 2)

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

$$\mathcal{L}_{Baseline} = \mathcal{L}_{2DBox} + \mathcal{L}_{SMPL} + \mathcal{L}_{GTrans}. \quad (4)$$

As a hybrid of the two-stage RoI-based and single-stage FoV-based frameworks, the common limitation remains. We address them by the PMMD components below.

3.2. The proposed PMMD Components

We introduce (1) Center-Padding RoI-Align to align bodies, (2) Dense-attention Local SMPL Head to refine local features, (3) Global and Local Translation Heads to associate body-wise local and global tasks, and (4) Global 3D Padding Augmentation to increase the density of 3D supervision. PMMD components are in pink in Fig. 2.

(1) Center-Padding RoI-Align. Center-Padding RoI-Align 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. See Fig. 1 for an example of our Center-Padding RoI-Align comparing to existing RoI and FoV operations. Implementing Center-Padding RoI-Align operation for a batch of mesh-wise feature maps is *non-trivial*. We developed an efficient CUDA implementation and will provide pseudo-codes in supplementary.

(2) Dense-attention Local SMPL Head. In the aligned RoI above, the feature elements originating from occluded image pixels are toxic to the local SMPL regressor. Ideally, we may select the features from visible pixels given a perfect 2D binary person mask. Following Mask-RCNN [1], an instance mask branch can be easily added to the baseline to potentially select the visible(unoccluded) features. However, the predicted binary masks by MaskRCNN are usually noisy and biased to the torso over the limbs.

We introduce a Dense-attention branch within the Local SMPL Head (See Fig. 3 top figure for the head structure.). This branch computes an instance mask confidence map (the [14, 14, 1] tensor after Softmax *before* binarization) and concatenates the features back to the RoI feature. The Dense-attention branch and the Local SMPL head are then trained jointly, such that the Dense-attention branch works as *a spatially dense self-attention block* to the SMPL regression branch. The effectiveness is evaluated in the ablation study (Table 4). The Dense-attention branch head introduces an additional instance mask loss $\mathcal{L}_{2DMask}$.

(3) Global and Local Translation Heads. In estimating body-wise global translations, the GT translations t_{gt}s are only available from sparsely located bodies, making it very challenging to supervise the network. Inspired by

the general 3D object detectors [5, 25], we introduce the dense depth supervision to support the translation learning. In Fig. 3 bottom part, we first construct a Global Depth Head originated from the multi-scale Feature Pyrimid Netowork(FPN) and supervise the head output $\mathbf{F}$ by a single-scale dense depth map $\mathbf{F}_{surf} \in \Re^{512\times832}$. Since we only focus on body mesh detection, we generate $\mathbf{F}_{surf}$ by projecting all body meshes to the common image space resulting in a depth map only containing body surface depth. We then concatenate the global depth-guided features with the local features from the SMPL head. Finally, the concatenated features are fed to the Translation Head for global translation vector $\hat{t} = [\hat{t}_x, \hat{t}_y, \hat{d}]$ estimation.

The translation heads above are supervised by:

$$\begin{aligned}\mathcal{L}_{GTrans} &= \lambda_{Ddepth}\|M_{F_gt} \odot (\mathbf{F} - \mathbf{F}_{surf})\| \\ &+\lambda_{Trans}\sum_{p=1}^{n}\|\hat{\mathbf{t}} - \mathbf{t}_{gt}\|_2^2. \end{aligned} \quad (5)$$

where $\mathcal{L}_{Ddepth}$ is surface dense depth loss and $\mathcal{L}_{Trans}$ is the translation head loss. In the supplementary materials, we provide a Pseudo-codes to calculate $\mathbf{t}_{gt}$.

(4) Global-Local Dense Vertex Supervision. In addition to the local-global bonds in the translation task above, we further introduce vertex-level local-global bonds to supervise local mesh and global translation. 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 $\mathbf{H} = (\mathbf{M}_{gt} + \mathbf{1}\mathbf{t}_w)\mathbf{E}^T$ normalizing the first three dimensions and remove the 4th dimension ($\mathbf{H} \in \Re^{6890\times4}$). Here $\mathbf{1} \in \Re^{6890\times1}$ is a vector in which every element is 1. and the extrinsic matrix $\mathbf{E} \in \Re^{4\times4}$ is computed from the GT frame-wise camera pose parameters. The global vertex loss is computed as

$$\begin{aligned}\mathcal{L}_{GVertex}(\theta,\beta,\mathbf{t}) = &\lambda_{loc}\|\mathbf{M}(\theta,\beta) - \mathbf{M}_{gt|pelvis}\|_2^2 \\ &+\lambda_{glo}\|(\mathbf{M}(\theta,\beta) + \mathbf{1}\mathbf{t}) - \mathbf{G})\|_2^2, \end{aligned} \quad (6)$$

where θ and β are the estimated SMPL coefficients. λ_{loc} and λ_{glo} are the loss weight of the local and global vertex respectively. $\mathbf{M}(\theta,\beta) \in \Re^{6890\times3}$ contains the vertex coordinates computed from θ and β. $\mathbf{t} \in \Re^{1\times3}$ again is the estimated translation vector from the local 3D coordinates to the 3D global camera coordinates.

Finally, the PMMD training loss is:

$$\begin{aligned}\mathcal{L}_{PMMD} =&\mathcal{L}_{2DBox} + \mathcal{L}_{SMPL}(\theta,\beta) + \lambda_{2DMask}\mathcal{L}_{2DMask} \\ &+ \mathcal{L}_{GTrans}(\mathbf{t}) + \mathcal{L}_{GVertex}(\theta,\beta,\mathbf{t}),\end{aligned} \quad (7)$$

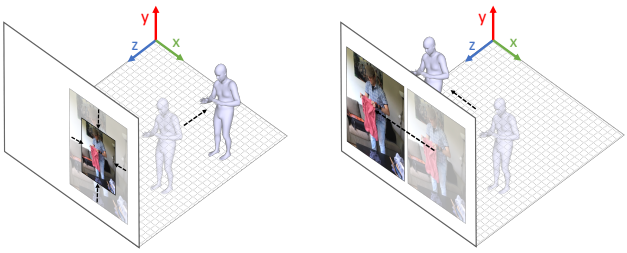

Figure 4. Mesh Translation Augmentation: 2D translating and depth translation of the GT meshes in global 3D coordinate in response to the shifting and scaling of the input images respectively.

where λ_{2DMask} is the loss weight of dense attention head.

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. 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. (Please refer to the supplementary material for method-specific training datasets).

Training configuration: As previous approaches on multi-body mesh detection, our PMMD training consists of three 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. 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 10 epochs with a batch size of 256 and a learning rate of $1e^{-4}$. In *Step-2*, the framework is continued to train on multi-body images (resized and padded to 512×832) for multi-body mesh detection. This step is trained for 19 more epochs with a batch size of 64 and a learning rate of $1e^{-5}$. In addition, to generate dense mesh translations in global 3D coordinates, we augment the mesh 2D locations by randomly shifting and padding the input images. We also augment the depthmap by scaling and padding the input images. (Refer to Fig. 4). The augmentation was activated with a 50% probability with random scaling of $(\times0.8, \times1.2)$ and translation to random but valid 3D positions.

In all steps, we use the Adam optimizer with weight_decay=$1e^{-4}$. Our training databases include the training sets of 3DPW [31], Human3.6M [7], AGORA [26], MPI-INF-3DHP [24], MS-COCO [21], LSP [9], LSP Extended [10], MPII [2], CLIFF-coco(pseudo-GT) [18], and ClIFF-mpii(pseudo-GT) [18]. More details on the loss weights and database sampling probabilities are provided in supplementary materials.

Evaluation datasets: We evaluated two databases con-

Table 1. Comparison with the state-of-the-art *multi-body detection models* on the 3DPW test set and AGORA val sets. All methods used extra training data. The metrics of existing methods were either copied from their papers (denoted as "xx.x") or computed using their released models if available, otherwise "-" is inserted. All methods use the neutral SMPL model ***except*** BEV [28], which uses the **SMPL-Age** model and age annotations. All metrics are in millimeters (mm) and the smaller "↓" the better. **PMMD is the only method that achieves sub-meter ($< 1000mm$) global metrics.**

Method	Backbone	Evaluation Dataset: 3DPW test set				Evaluation Dataset: AGORA val set			
		Local Metrics		Global Metrics		Local Metrics		Global Metrics	
		PA-MPJPE↓	PA-PVE↓	GPE↓	GPVE↓	PA-MPJPE ↓	PA-PVE↓	GPE↓	GPVE↓
Existing Model(GT bbox+Est. Camera-pose)									
SPEC [15]	ResNet50+ResNet50	52.2	81.0	-	-	-	-	-	-
Existing Model(2D Detector+Weak-persp. depth)									
OpenPose+SPIN [16, 34]	VGG-19+ResNet50	66.4	-	-	-	-	-	-	-
YoloV3+VIBE [13, 34] (video)	Darknet53+ResNet50	66.1	-	-	-	-	-	-	-
BMP [38]	ResNet50	63.8	-	-	-	-	-	-	-
Faster-RCNN+OCHMR [12]	ResNet50+HRNet32	58.3	-	-	-	-	-	-	
CRMH [8] (f=1000∗)	ResNet50	50.6	82.1	1080.6	1086.8	70.1	97.5	2298.6	2320.2
ROMP [34]	HRNet32	47.3/55.2	73.1	2823.7	2825.2	95.8	125.5	7682.5	7682.7
ROMP [34] (f=1000*)	HRNet32	47.3/55.2	73.1	376.2	392.8	95.8	125.5	6373.0	6377.5
Existing Model(2D Detector+Est. depth)									
BEV [28] (SMPL-Age)	HRNet32	46.9/51.1	70.5	2789.4	2792.2	84.9	114.1	6325.7	6331.8
BEV [28] (SMPL-Age, f=1000*)	HRNet32	46.9/51.1	70.5	359.8	373.3	84.9	114.1	4964.6	4974.1
Ours (f=1000*)									
PMMD **w/o.** dataset-specific fine-tuning	ResNet50	45.4	71.4	296.5	314.0	56.1	76.1	879.6	893.7
PMMD **w/.** dataset-specific fine-tuning	ResNet50	42.1	71.5	310.2	326.7	55.9	75.7	860.5	874.4

* Global metrics are computed against GT 3D translations under a canonical focal length specified by CRMH (f=1000).

taining the multi-body ground truth of SMPL parameters, global translation, and camera extrinsic.

- **3DPW** [31] contains (24 train, 24 test, 12 validation) image sequences captured by hand-held cameras with manual annotations of SMPL coefficients, 3D joints, 2D joints, and frame-wise camera poses. Subjects are mostly walking in the horizontal camera view.
- **AGORA** [26] is a synthetic dataset with ground truth body meshes and 3D translations. There are 14K training and 3K validation images. The human bodies are rendered using 4240 textured body scans in diverse poses and clothes. The pelvis-to-camera distances are in $[1.8, 27.4]$ meters and the pelvis-to-camera altitudes are in $[-7.6, 9.0]$ 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.

We followed CRMH to conduct person matching, and to compute global metrics under a canonical focal length of 1000 (f=1000). For both CRMH and PMMD, both the input images and the ground truth 3D translations are normalized accordingly. For fair comparisons with ROMP and BEV (fixed camera view 60^o), we compute the image-wise focal length and normalize the mesh translation of ROMP and BEV to f=1000. In Table 1, all the normalized results are denoted by "(**f=1000***)".

4.1. Comparing with State-of-the-arts

In Table 1, we compare PMMD with the state-of-the-art on common evaluation datasets. We copied all the local metrics reported in papers (denoted by "xx.x"), and computed missing metrics using recently released models. For methods that have no explicit translation regression (CRMH and ROMP), global translations are computed using their predicted weak-perspective parameters. All methods use the SMPL-neutral model ***except*** BEV [28]. BEV uses the SMPL-Age model and GT age annotations which provide extra advantages in local metrics over all other methods.

Paper vs. released-model metrics: In Table 1, many paper-metrics ("xx.x") are lower than the released-model-metrics (separated as "xx.x/yy.y"). We suspect the paper-metrics are from **data-specific fine-tuned models** on each evaluation dataset, while the released model is general to all datasets[2]. For this reason, we report two sets of PMMD metrics: "PMMD w/o. dataset-specific fine-tuning" and "PMMD w/. dataset-specific fine-tuning".

Table 1 shows that PMMD works better locally (PA-MPJPE, PA-PVE), and by far the best globally (GPE, GPVE) among all compared methods. On the AGORA

[2]Our suspicion is supported in Table 2-3 of the ROMP paper https://arxiv.org/pdf/2008.12272.pdf, and the github link: https://github.com/Arthur151/ROMP/blob/master/docs/romp_evaluation.md

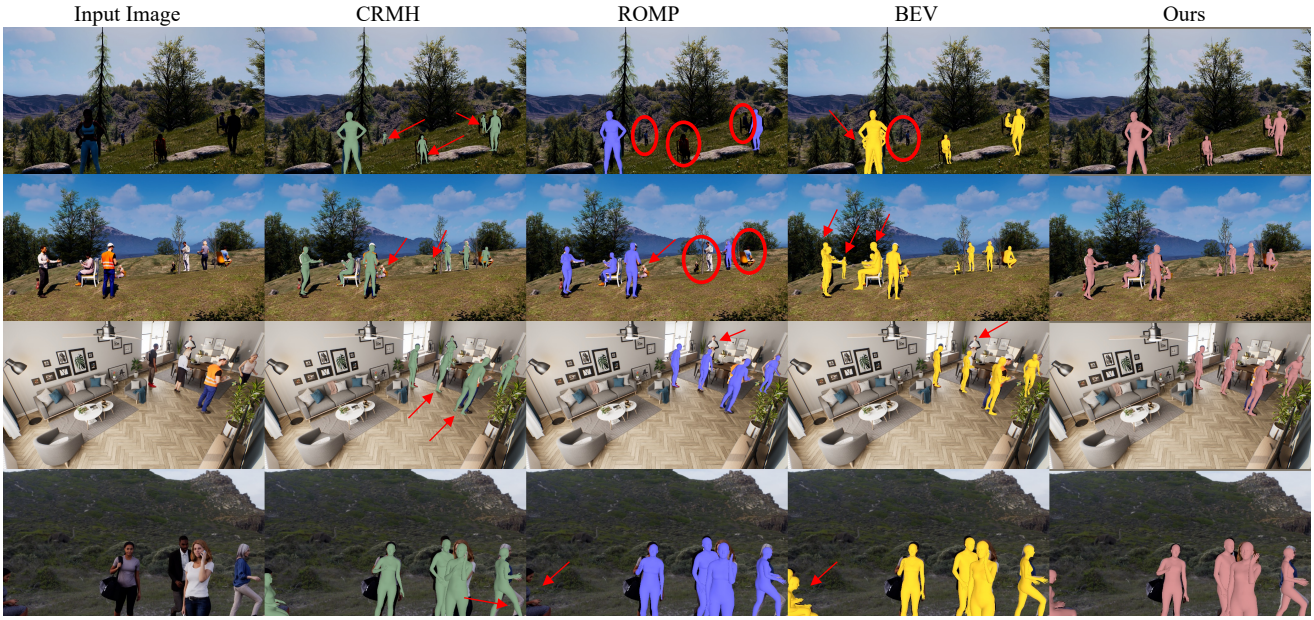

Figure 5. Qualitative results in the frontal view. **Ours**(pink) is compared with CRMH(green), ROMP(blue), and BEV(yellow). Better mesh-to-person intersection in pixels indicates better local metrics. The predicted meshes of significant defects are highlighted in red circles and arrows. The compared methods typically struggle with *side-oriented bodies* and *missing or redundant detection*. Our approach produces more robust detection and accurate meshes.

val set, over the physical 3D space $[1.8, 27.4]$(depth)-by-$[-7.6, 9.0]$(altitude) meters, the proposed PMMD not only gets the best local metrics but is the only method that brings the global metrics, GPE and GPVE, below 1 meter (1000 millimeters). In the frontal views in Fig. 5, the proposed PMMD shows better mesh-to-person intersection in pixels indicating better local metrics. In the top view in Fig 6, the proposed PMMD meshes show better intersection with the GT meshes indicating better global metrics, in particular, better depth estimation.

We also closely compare PMMD with CRMH, ROMP, and BEV which produce the next best sets of global metrics. In Table 2, we compute GPVEs on the evenly split subsets of pelvis-to-camera distance ranges and pelvis altitude ranges in the AGORA val set, respectively. PMMD constantly outperforms other methods in all ranges.

Why CRMH, ROMP, and BEV have worse global metrics than PMMD? 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 the "small" person away and the "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, 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 PMMD, we closely integrate the local and global cues (Fig. 3) 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 distances (Close: [1.8, 10.3]m, Median: [10.3, 18.8]m, Far: [18.8, 27.4]m) and by pelvis-to-camera altitudes (Low: [-7.6, -2.1]m, Median: [-2.1, 3.4]m, High: [3.4, 9.0]m). PMMD is the most robust method in the 3D space.

Method	GPVE overall↓	GPVE by Distance↓			GPVE by Altitude↓		
		Close	Median	Far	Low	Median	High
CRMH [8]	2320.2	944.0	1955.2	4085.9	3746.3	2177.8	5291.7
ROMP [34]	7682.7	2618.5	5519.5	14549.6	13626.4	7130.2	19167.6
ROMP [34] (f=1000)	6377.5	2732.4	3953.9	11497.5	11192.8	5906.2	16201.9
BEV [28]	6331.8	3261.8	5270.1	9513.2	9798.5	5976.2	13763.4
BEV [28] (f=1000)	4974.1	3477.1	4354.7	6065.5	5220.5	4735.9	10176.5
PMMD (ours)	893.7	389.3	784.8	1414.4	1683.4	840.4	1977.3

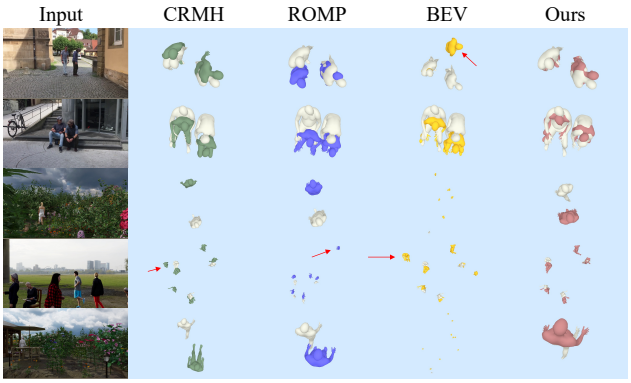

Figure 6. Qualitative results in the top view. **Ours**(pink), CRMH(green), ROMP(blue), and BEV(yellow) are compared with the GT meshes (gray). Better intersection with the GT meshes in the top view indicates better global metrics, in particular, better depth estimation. The predicted meshes of significant defects are highlighted in red arrows. The results demonstrate that PMMD meshes closely match the GT meshes in terms of 3D global translation, which verifies the robustness and effectiveness of our proposed global and local head modules.

Robustness to truncation, occlusion, and collision. PMMD robustness is mainly boosted by Center-Padding RoI-align which aligns the body-to-RoI layouts and our Dense Attention head which refines the feature based on body segmentation mask confidences. These techniques select robust features for local SMPL regression. Our body translation robustness is distributed between the local and global tasks as a translation residual problem. PMMD does not include but could further benefit from the Depth Ordering-Aware and Interpenetration loss as [8, 28, 34].

Context priors. Context priors may help but may also be limited in generalization. For instance, the ground plane assumption [15, 35] requires extra camera pose models and training data, which does not work on images with elevated body altitudes (e.g. stairs in 3DPW, construction sites and garden bushes in AGORA). Although using natural-image training data to provide camera priors, SPEC [15] still produce inferior metrics in Table 1), which indicates poor generalization in the human mesh datasets.

Model Overheads: PMMD enables the new capability in detecting 3D meshes in the physical size and locations. The overhead is low compared with its two-stage baseline Mask-RCNN and CRMH (See Table 3). Moreover, it is worth exploring PMMD in the single-stage detector framework (e.g., ROMP, BEV) for fewer parameters.

Table 3. Model overheads due to new capabilities in 2D mesh detection, and 3D physical mesh scale&location detection.

Method	Detector Framework	2D Mesh Detection	3D Physical scale& location	Parameters
MaskRCNN	Two-Stage	No	No	43.16M
CRMH	Two-Stage	**Yes**	No	45.35M
PMMD(ours)	Two-Stage	**Yes**	**Yes**	50.14M

Table 4. Ablation study of critical PMMD components on the AGORA validation set. All metrics are computed at the same training Epoch=3. All metrics are the smaller the better. By excluding each component, all local and global metrics degrade.

Ablation (epoch40)	Local Metrics				Global Metrics			
	PA-MPJPE↓	degrad.	PA-PVE↓	degrad.	GPE↓	degrad.	GPVE↓	degrad.
PMMD	58.3	0	78.7	0	853.5	0	871.4	0
w/o. Dense Attention	87.4	-29.1	121.1	-42.4	17884.7	-17031.2	17949.9	-17078.5
w/o. Center-Padding RoI-Align	59.6	-1.3	80.4	-1.7	837.8	+15.7	853.9	+17.5
w/o. Global Vertex Loss	58.5	-0.2	79.1	-0.4	859.6	-6.1	878.6	-7.2
w/o. Global Trans. Feature	58.4	-0.1	79.0	-0.3	861.2	-7.7	878.6	-7.2
w/o. Local Trans. Feature	58.1	+0.2	78.4	+0.3	901.2	-47.7	919.2	-47.8

4.2. Ablation Study

We examine the impact of the five most critical PMMD components by showing the metric degradation when excluding each component. We evaluate each case after training the same 10 epochs in Step-1 training and 3 epochs in Step-2 training. In Table 4, the local and global metrics suffered when critical components were excluded during training. The biggest to the smallest degradations result from excluding (w/o.) Dense Attention, Center-Padding RoI-Align, Global Vertex Loss, Global Translation Features, and Local Translation Features, respectively. In particular, w/o. Center-Padding RoI-Align hurts local metrics. Without the Local Translation (Trans.) Feature, local metrics only slightly decrease but global metrics significantly suffer. The Global Vertex Loss and the Global Translation (Trans.) Feature constantly helps both local and global metrics. The Dense-Attention module improves both local and global metrics significantly. Overall, Table 4 clearly demonstrates the effectiveness of each proposed module.

5. Conclusion and Discussion

In this work, we propose a new Physical-space Multi-body Mesh Detection framework, called PMMD, which addresses the misalignment, sparse supervision, and inaccurate global translation issues in achieving precise 3D physical space mesh detection from monocular RGB images. The proposed method significantly improves the training of the local mesh and global translation regression tasks with the local alignment and global dense learning. Compared with existing methods, we learn from this work that: (1) Local features alignment benefits the local mesh reconstruction. (2) Integrating local body-wise and global dense-depth-guided features that improve the physical global translation estimation. Furthermore, we also introduced a dense-attention module and global 3D padding augmentation to enhance the robustness and diversity. Extensive experiments on challenging multi-body benchmarks show the superiority of the proposed PMMD over other methods.

Social Impact. This work recovers 3D person shaps from cameras, which facilitates many downstream applications requiring human motions. This technique should only be applied 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 *CVPR*, pages 3686–3693, 2014. 5

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

[4] 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. 2, 3

[5] Yilun Chen, Shijia Huang, Shu Liu, Bei Yu, and Jiaya Jia. Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

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

[7] 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. 5

[8] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020. 1, 2, 3, 6, 7, 8

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

[10] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 5

[11] 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. 2

[12] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, pages 1715–1725, June 2022. 2, 6

[13] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2, 6

[14] 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

[15] 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. 2, 6, 8

[16] 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. 2, 6

[17] 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

[18] 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. 5

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2

[20] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 2

[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 *European conference on computer vision*, pages 740–755, 2014. 5

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

[23] 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, 3

[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. 5

[25] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 3, 5

[26] 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. 5, 6

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

[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, 6, 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. 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. 2

[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. 5, 6

[32] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-andcompare. In *ICCV*, 2019. 2

[33] Sen Yang, Wen Heng, Gang Liu, GUOZHONG LUO, Wankou Yang, and Gang YU. Capturing the motion of every joint: 3d human pose and shape estimation with independent tokens. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[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. 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. 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. 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, 6

[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. 2, 3