Divide and Fuse: Body Part Mesh Recovery from Partially Visible Human Images

Tianyu Luan[⋆,2,1], Zhongpai Gao[1], Luyuan Xie[3], Abhishek Sharma[1], Hao Ding[4,1], Benjamin Planche[1], Meng Zheng[1], Ange Lou[5,1], Terrence Chen[1], Junsong Yuan[2], and Ziyan Wu[1]

[1] United Imaging Intelligence, Boston MA, USA
[2] State University of New York at Buffalo, Buffalo NY 14260, USA
[3] Peking University, Beijing, China
[4] Johns Hopkins University, Baltimore MD 21218, USA
[5] Vanderbilt University, Nashville TN 37235, USA
{tianyulu, jsyuan}@buffalo.edu, {first.last}@uii-ai.com
2201110745@stu.pku.edu.cn, hding15@jhu.edu, ange.lou@vanderbilt.edu

Abstract. We introduce a novel bottom-up approach for human body mesh reconstruction, specifically designed to address the challenges posed by partial visibility and occlusion in input images. Traditional top-down methods, relying on whole-body parametric models like SMPL, falter when only a small part of the human is visible, as they require visibility of most of the human body for accurate mesh reconstruction. To overcome this limitation, our method employs a "Divide and Fuse (D&F)" strategy, reconstructing human body parts independently before fusing them, thereby ensuring robustness against occlusions. We design Human Part Parametric Models (HPPM) that independently reconstruct the mesh from a few shape and global-location parameters, without inter-part dependency. A specially designed fusion module then seamlessly integrates the reconstructed parts, even when only a few are visible. We harness a large volume of ground-truth SMPL data to train our parametric mesh models. To facilitate the training and evaluation of our method, we have established benchmark datasets featuring images of partially visible humans with HPPM annotations. Our experiments, conducted on these benchmark datasets, demonstrate the effectiveness of our D&F method, particularly in scenarios with substantial invisibility, where traditional approaches struggle to maintain reconstruction quality.

1 Introduction

Human body mesh recovery has applications in various fields including augmented and virtual reality (AR/VR), film production, human-computer interaction (HCI),

[⋆] This work was carried out during the internship of Tianyu Luan at United Imaging Intelligence, Boston MA.

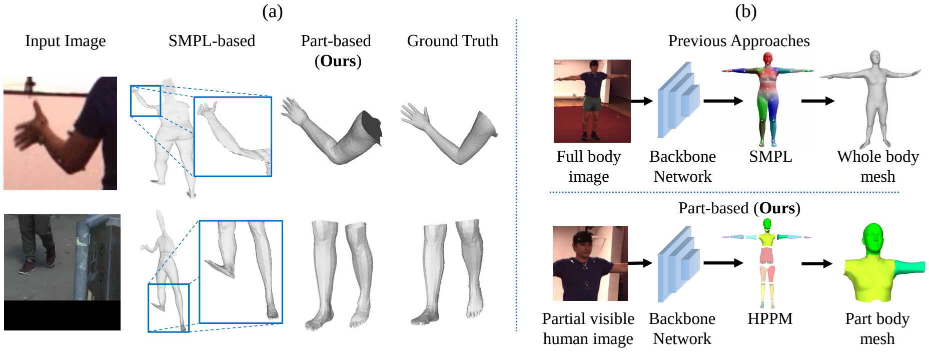

Fig. 1: Traditional top-down method vs. Divide and Fuse. (a) When the input image only shows a few body parts (1st column), top-down SMPL-based methods may easily fail (2nd column) due to the lack of whole-body information. Our part-based D&F method is designed for partially visible human reconstruction (see results in the 3rd column). (b) Primary framework of SMPL-based prior art versus our proposed model.

and sports. In specific applications such as movies, video games, or medical in-bore cameras, there are instances when major portions of the human body are outside the camera's field of view, leaving only small parts of the body visible. This scenario poses a significant challenge in accurately reconstructing the human mesh.

Previous methods, such as [19, 21, 27, 30, 50, 54, 55, 63], are effective when the majority of the human body is visible, but their performance significantly decreases when the human body is substantially invisible. Previous human mesh reconstruction methods mainly follow top-down designs, utilizing whole-body parametric models such as SMPL [34] or STAR [32]. These models extract global features from input images and transform them into parameters to reconstruct the human mesh, and are successful when the entire body is nearly fully visible. However, when the input human body is largely incomplete, the human body may not be well recognized by the network. Moreover, when using top-down parametric models, the parameters are entangled among different body parts, making the reconstruction of visible sections dependent on the information from the entire body. The absence of invisible parts will compromise this reliance and entanglement, thereby influencing the visible parts' reconstruction quality. Both the misrecognition and unnecessary entanglement issue would cause a drop in the body mesh accuracy. As shown in Fig. 1a, the quality of the human mesh reconstruction is poor when only the legs are visible.

To address these problems, we propose a "Divide and Fuse" (D&F) bottom-up human body mesh reconstruction approach (shown in Fig. 1b). When only a few human body parts are visible, capturing the entire human body can be challenging for the network, but capturing the visible human body parts is more feasible. Besides, when there are only a few visible parts in the input, the interdependence of top-down approaches brings less knowledge and more noise to each other. If we do the reconstruction part by part, we can naturally avoid both information capturing and inter-part interference problems. However, there are also challenges of part-by-part reconstruction. Two main challenges are: a) independent reconstruction of all mesh parts and b) adjacent part fusion when more than one part is visible. To address the first challenge, we design a set of parametric models for each body part. Each parametric model can take a few shape parameters along with part global transformations as input to reconstruct the mesh of that body part without relying on other parts. For the second challenge, we design a fusion module to connect the adjacent visible parts together. We also design overlapping areas in the parametric models between two adjacent parts, which makes the fusion easier.

Specifically, our work is divided into the following steps. We first design and train the Human Part Parametric Models (HPPM). We generate the part templates using the template mesh of SMPL. The SMPL mesh is divided into 15 parts, and each part is trained using a large volume of SMPL ground truth. By regressing a few shape parameters and global transformations, we can obtain various shapes of each human part mesh. Second, we build up a network that takes monocular images as input and reconstructs human parts independently. We use a transformer-based backbone to get image features, and use them to regress the parameters of HPPM to generate the part meshes. Then, we design a fusion module that connects the adjacent parts using a gradually-changed weighted-sum strategy, which could connect the mesh part seamlessly. Additionally, to evaluate our Divide and Fuse method, we constructed a benchmark comprised of images crops with partially visible human bodies and corresponding HPPM annotations. We use existing public datasets Human36m and 3DPW to generate our benchmark *Partially Visible Human3.6M* and *Partially Visible 3DPW*, and use both datasets for evaluation. In training, we use a similar image cropping strategy as augmentations on existing public datasets to get similar input image domains as in testing benchmarks, hence obtaining better performance.

To summarize, our contributions are as follows:

- We designed a Divide and Fuse bottom-up solution that can reconstruct the human body mesh part by part. Different from SMPL-based top-down approaches, our method can independently reconstruct visible body parts when a large portion of the human body is not visible.
- We designed a set of parametric models representing every body part. Different from SMPL, these parametric models can independently represent each body part, and be easily connected together when needed.
- We design a fusion module to smoothly connect each part together when multiple visible parts exist.
- In order to evaluate our method, we established 2 benchmark datasets with partially visible human image croppings and HPPM annotations. We also design a similar augmentation strategy on the training dataset to improve our mesh reconstruction quality.

Our experiments validate the expressive ability of our parametric model and show that our reconstruction method outperforms the state-of-the-art on multiple datasets for partially visible human body input.

2 Related Works

Human body mesh reconstruction. Human body mesh reconstruction has been a popular research area. Previous works [2, 9, 12, 13, 16, 18, 25, 26, 31, 43, 53] apply whole-body parametric model SMPL [24] for mesh reconstruction, and achieve good results when the whole body is visible. As top-down approaches, they rely on capturing whole-body information and their performance strongly degrades when the extraction of whole-body features fails. A number of past studies [6,14, 17,22,39,51,52,57,61] have concentrated on the recovery of comprehensive human body meshes using occluded monocular inputs. These techniques strive to infer obscured parts based on visible body parts, which would place a stronger emphasis on capturing entire body information over the accuracy of each visible body part. Thus, when the entire body is not readily recognizable, these approaches tend to have results similar to non-occlusion approaches. Other methods like those proposed by [10, 14, 20, 49, 59] employ temporal inputs or other modalities (*e.g.*, radio signals) to guide the reconstruction process, but their designs still focus on the body as a whole rather than individual parts. [64] focused on a bottom-up mesh reconstruction strategy, but it is not designed for learning architectures and partially visible inputs. In this paper, we design a bottom-up learning-based approach that can tackle few-part-visible inputs.

Human pose estimation. Different from human mesh reconstruction, many human pose estimation approaches [3, 7, 29, 35, 56, 58] are using a bottom-up design as we do on mesh in this paper. They use a lifting strategy to estimate 3D body joints from 2D key points. However, inferring a complete 3D human mesh from a sparse set of 3D joints is an ill-posed and challenging task. Existing works would use 3D joints as a byproduct and supervision. Previous works [1,4,5,8,36–38,44,46,48,60,61] can tackle slight occlusions, but they did not focus on the largely invisible cases. Since even 3D or 2D pose estimation would rely on inter-joint/key point correlation, the largely invisible case would compromise the basic 2D key points, resulting in unsatisfying results.

3 Divide and Fuse

3.1 Overview

Problem formulation. In this paper, our task is to address the challenge of human mesh reconstruction from an image where only a limited portion of the human body is visible. Specifically, the input for this task is a monocular image I containing a partially visible human body, and our objective is to develop a method $F(\cdot)$ capable of reconstructing the visible human body part mesh v from

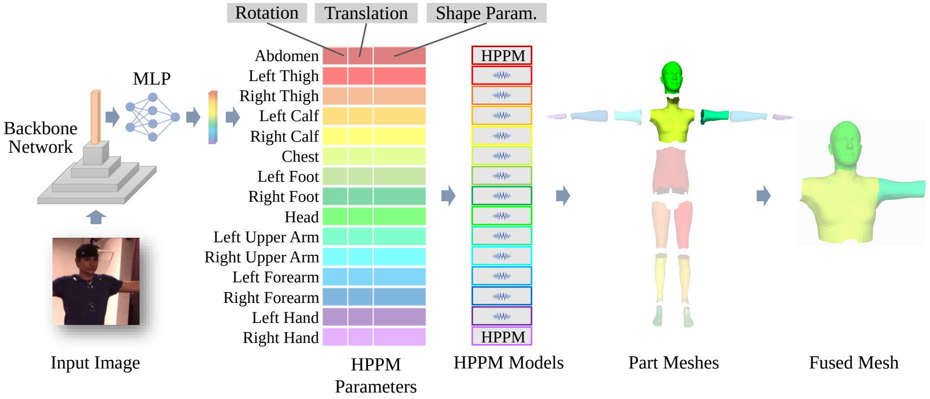

Fig. 2: Our Divide and Fuse (D&F) method takes a monocular partially visible human image as input and generates the human mesh of visible parts. The input image first goes through a backbone and an MLP network to get the parameters of HPPM. Then, these parameters are used to generate part meshes through each part-specific HPPM. Finally, a fusion module connects adjacent visible parts. Details are provided in Sec. 3.

the given image, *a.k.a.* $v = F(I)$. In the following sections, we elaborate on how $F(\cdot)$ is designed.

Framework overview. As illustrated in Fig. 2, our pipeline starts with the extraction of image features through a backbone network, followed by the use of a multi-layer perceptron (MLP) to obtain the features required by the Human Part Parametric Models (HPPM). In HPPM, we divide the human body into 15 parts. For each part, HPPM requires 3 inputs: translation, rotation, and shape parameters. Upon the shape adjustments made by the shape parameters, HPPM employs simple rigid transformations to generate the mesh of each body part without explicitly incorporating the human pose. HPPM outputs body part meshes and joints, and is supervised by corresponding ground truths via multiple loss functions, ensuring that each mesh part can be reconstructed independently. Moreover, a fusion module is proposed that seamlessly integrates multiple visible mesh parts. In Sec. 3.2, we introduce the design and training of the HPPM. Sec. 3.3 elaborates on the design of our independent reconstruction network, and Sec. 3.4 introduces our fusion module.

3.2 Human Part Parametric Models

We design Human Part Parametric Models (HPPM) to facilitate the independent reconstruction of each body part. Unlike the widely-used holistic approach SMPL, HPPM allows for the decoupling of body parts, eliminating the reliance on inter-part correlations within the model, *i.e.*, to improve performance when the inter-part correlation is not stable in input images. The HPPM design consists of two stages. First, a mesh template is defined for each body part; then a linear function is trained to the shape parameters into part meshes.

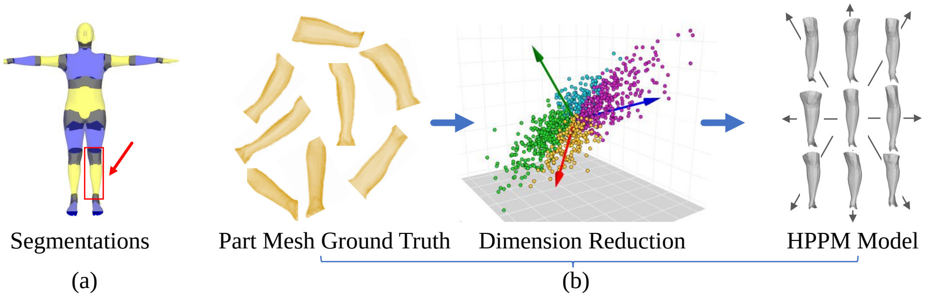

Fig. 3: **(a)** HPPM template segmentation. We segment the SMPL template to generate HPPM templates. The joint areas are covered by both adjacent parts (overlap). This design allows HPPM to naturally cover near-joint distortions using shape parameters, while also facilitating the fusion of parts together. **(b)** HPPM training process. We segment part ground truths from Human3.6M [11], 3DPW [47], and AMASS [28]. For each part, we use a dimension-reduction strategy to train a matrix that maps the high-dimensional part meshes into a few shape parameters. Shape parameters are estimated by the network to recover part meshes.

Part template design. We craft part-specific template meshes utilizing the template mesh from SMPL. We segment this mesh into 15 distinct parts based on the blend weight parameter W of SMPL, where each entry of blend weight matrix W_{ij} indicates how vertex v_i is influenced by bone b_j when the limbs move. Here, we first group the vertices using the following strategy:

$$p_i = \arg\max_j W_{ij}, \tag{1}$$

where p_i is the part index of vertex v_i. *I.e.*, we first assign each vertex to the bone with the highest blend-shape contribution. This grouping strategy ensures that vertices assigned to the same section are attached to the same body limb, allowing us to treat each part as a nearly rigid mesh, so that the model would not need explicit inner part poses. Obtaining the grouping as raw segmentation, we manually fuse some raw segments to create a cohesive set of 15 parts. This strategy minimizes unnecessary divisions so that body parts that are empirically considered nearly rigid when put together will be combined into a single part. *E.g.*, we combine the shoulders and neck with the torso section as they collectively form a nearly rigid structure. However, we do not merge the thigh and calf to be a single part because the knee joint allows significant bending. Finally, we perform dilation for each part to create some overlapping between adjacent parts. This design not only incorporates near-joint mesh deformation into the deformation of each part, but also simplifies the fusion between adjacent parts. More intricate details of HPPM design are further explained in Supplementary Material Sec. A. The final part templates are visualized in Fig. 3a.

Parameter-adjustable HPPM training. Having the template of each part, we train our HPPM on Human3.6M [11], 3DPW [47], and AMASS [28] datasets. Our training process is visualized in Fig. 3b. First, we segment the ground truth part meshes using the template part mesh vertex indices obtained above. Then we use a dimension reduction method to train our linear mapping. For part p, we form the mesh data in the training datasets into a matrix $X_p \in \mathbb{R}^{n \times m}$, where m is the number of samples and $n = 3N$ is the number of vertices times their dimension. We reduce the matrix dimensionality from n to k using principal component analysis. We denote as $\mathcal{U}_p \in \mathbb{R}^{k \times n}$ the matrix that maps the k dimensional shape parameters to part meshes. We use an adjustable number of parameters for each part, so that we can adapt the fitting capacity of different parts given different shape variances. As demonstrated in Fig. 4a, we show the relationship between fitting accuracy and parameter dimensions for each part of the training set. Here, we set a maximum joint and vertex fitting error for each part while maintaining a minimal dimension. We empirically set the maximum error allowance for both joint and vertex to 2mm and the minimum dimension to 16. In Tab. 1 we show the training errors of each part. The experiment in Tab. 3 "w/ fixed #parameters" row reveals that our model, with a total parameter count of 360, outperforms the method that uses a fixed dimension of 24 per part, which has the same total number of parameters.

Joint regressor. HPPM is also designed to generate body joints so that 3D joints can be estimated along with part meshes. We train a joint regressor matrix for each body part as

$$\mathcal{J}_p = \underset{\mathcal{J}'_p}{\arg\min} \|J - \mathcal{J}'_p v_0\|_F, \tag{2}$$

where J denotes the ground truth joints, and $\|\cdot\|_F$ is the Frobenius normal. Note that for every part, we only regress the joint close to that part. The part-joint correspondence is shown in Supplementary Material Sec. B.

Our HPPM model is finally defined as $\{\mathcal{U}_p, \mathcal{J}_p\}$, with p for the p-th part. Compared with SMPL, HPPM is capable of determining a human part mesh independently through overall translation, rotation, and a limited set of deformation parameters. This design enables precise reconstruction of body segments.

3.3 Divide: Part Independent Reconstruction

The divide stage provides a solution for reconstructing each part independently from images. Input images are processed through a Swin Transformer [23] backbone to obtain image features, which are then passed through a multi-layer perceptron (MLP) model to derive the HPPM shape parameters $\hat{S}_p$, global rotation $\hat{R}_p$, and global translation $\hat{T}_p$ for part p. As multiple rotations are being estimated, our network outputs 6D rotation [62] to improve convergence on rotation estimation. We write the global transformation in one matrix $\hat{M}_p$ and obtain the estimated mesh for part p as:

$$\hat{v}_p = \hat{M}_p(\mathcal{U}_p \hat{S}_p + \mathcal{M}_p), \quad \text{with } \hat{M}_p = \begin{pmatrix} \hat{R}_p & \hat{T}_p \\ \mathbf{0} & 1 \end{pmatrix}, \tag{3}$$

where $\mathcal{U}_p$ and $\mathcal{M}_p$ is the shape matrix and mean shape in HPPM, and $\hat{v}_p$ is in homogeneous coordinates. From these results, we calculate the part joints using regressors:

$$\hat{J}_p = \mathcal{J}_p \hat{v}_p, \tag{4}$$

thus inferring each part mesh and corresponding joint independently.

Training losses. We design the following losses to ensure the independent reconstruction of every part. In order to achieve part independence, all the supervisions in this section are defined on body parts.

First, we use the mesh and joint annotation of each part to directly supervise part mesh and joint. Specifically, we define part vertex loss $\mathcal{L}_v$ and part 3D joint location loss $\mathcal{L}_{j3d}$ as:

$$\begin{aligned} \mathcal{L}_v &= \textstyle\sum_p \sum_i \delta_p \|\hat{v}_{pi} - v_{pi}\|, \\ \mathcal{L}_{j3d} &= \textstyle\sum_p \sum_i \delta_p \|\hat{J}_{pi} - J_{pi}\|, \end{aligned} \tag{5}$$

where $\hat{v}_{pi}$ and v_{pi} are the i-th vertex of the p-th estimated part mesh and ground truth part mesh, respectively. $\hat{J}_{pi}$ and J_{pi} are estimated 3D joint location and ground truth joint location, respectively. δ_p indicates the visibility. $\delta_p = 1$ when that part is visible and $\delta_p = 0$ when its not. $\| \cdot \|$ indicates the L2-norm of a vector. Besides mesh and 3D joint loss, we also include 2D joint projection loss to enhance 2D projection accuracy as

$$\mathcal{L}_{j2d} = \sum_p \sum_i \delta_p \|\mathit{\Pi} \hat{J}_{pi} - \mathit{\Pi} J_{pi}\|, \tag{6}$$

where $\mathit{\Pi}$ is the projection matrix from camera to image coordinate.

Apart from the above losses, we also directly supervise the part parameters and their global transformations. Specifically, we define part shape parameter loss $\mathcal{L}_s$, part rotation loss $\mathcal{L}_r$, and part translation loss $\mathcal{L}_t$ as

$$\begin{aligned} \mathcal{L}_s &= \textstyle\sum_p \delta_p \|\hat{S}_p - S_p\|, \\ \mathcal{L}_r &= \textstyle\sum_p \delta_p \|\hat{R}_p - R_p\|, \\ \mathcal{L}_t &= \textstyle\sum_p \delta_p \|\hat{T}_p - T_p\|, \end{aligned} \tag{7}$$

where $\hat{S}_p$, S_p, $\hat{R}_p$, R_p, $\hat{T}_p$, and T_p are defined similar to Eq. (5).

In total, the loss for part independent reconstruction is defined as:

$$\mathcal{L}_{div} = \lambda_v \mathcal{L}_v + \lambda_{j3d} \mathcal{L}_{j3d} + \lambda_{j2d} \mathcal{L}_{j2d} + \lambda_s \mathcal{L}_s + \lambda_r \mathcal{L}_r + \lambda_t \mathcal{L}_t, \tag{8}$$

where λ_v, λ_{j3d}, λ_{j2d}, λ_s, λ_r, and λ_t are loss weights.

3.4 Fuse: Adjacent Part Fusion

The fuse part is designed to combine the visible part meshes into a single mesh when multiple parts are visible. Before applying the connection, we use two self-supervision fusion losses—namely overlapping loss and depth consistency loss—to bring the part meshes closer to each other. The overlapping loss is defined:

$$\mathcal{L}_{ol} = \sum_p \sum_{v \in O_p} \delta_p \|\hat{v}_p - \bar{v}\|, \tag{9}$$

where O_p is the vertex set that is in the overlapping area between 2 adjacent parts, δ_p is defined the same as in Eq. (5), and $\bar{v}$ is the average vertex location of the overlapping area of 2 adjacent parts. It is computed as:

$$\bar{v} = \frac{\sum_{v \in O_p \cap O_{p'}} \hat{v}_p}{2|O_p \cap O_{p'}|}, \tag{10}$$

where $O_{p'}$ is the overlapping area on the adjacent part mesh. This way, we can ensure that the junction vertices of adjacent parts are both closer to their average, thereby closer to each other. The depth consistency loss is designed to constrain the parts that occur in the same input image but are not directly connected. A regularization term is applied to those parts as:

$$\mathcal{L}_{dc} = \sum_p \sum_i \delta_p \|\hat{v}^z_{pi} - \bar{v}^z\|, \tag{11}$$

where $\hat{v}^z_{pi}$ is the z coordinate of the i vertex location of p-th estimated part, and $\bar{v}^z = \frac{\sum_p \sum_i \hat{v}^z_{ip}}{\sum_p N_p}$ is the average z coordinate of all vertices in all visible parts.

Then our self-supervised fusion loss is defined as:

$$\mathcal{L}_{fu} = \lambda_{ol}\mathcal{L}_{ol} + \lambda_{dc}\mathcal{L}_{dc}, \tag{12}$$

where λ_{ol}, λ_{dc} are loss weights. Our total loss function is:

$$\mathcal{L} = \mathcal{L}_{div} + \mathcal{L}_{fu}. \tag{13}$$

Gradual part connecting. Besides the aforementioned training supervision, we also design a post-processing connecting module to seamlessly attach adjacent parts together during inference, based on a weighted-sum strategy. We identify two types of output vertices v^c_k when connecting two adjacent parts into one mesh: the final vertices that belong to the overlapping region shared by both part meshes, and those that do not belong to this region. In non-overlapping regions, the final vertices are computed:

$$v^c_k = \hat{v}_{pi}, \tag{14}$$

where $\hat{v}_{pi}$ is the corresponding vertex of v^c_k in part p. Here, "corresponding" means the i-th vertex in part p is topologically the k-th vertex in the SMPL template. If v^c_k is in the overlapping area, then:

$$v^c_k = v^c_{p_1 i} = v^c_{p_2 j} = \frac{\hat{v}_{p_1 i} d_{2j} + \hat{v}_{p_2 i} d_{1i}}{d_{1i} + d_{2j}}. \tag{15}$$

Here, d_{1i} is the shortest topology distance from $\hat{v}_{p_1 i}$ to the nearest non-overlapping vertex in p_1th part mesh. That means, if $d_{1i} = 2$, $\hat{v}_{p_1 i}$ needs to go through one other vertex to connect to the nearest non-overlapping vertex in the p_1th part mesh. d_{2j} is defined similar to d_{1i}. In this fusion process, the overlapping loss is used to ensure the vertices in the overlapping areas are closely aligned, and the gradual part connecting process can further eliminate the small vertex difference. We use this design to avoid undesirable deformations in the fused mesh.

Table 1: Per-part parameters and registration errors. For each part p, we first provide their number of shape parameters k_p, number of vertices N_p, and number of body joints $|\mathcal{J}_p|$, followed by their mean ℓ_2 errors (in mm) w.r.t. GT vertices and joints. For both joint and vertex errors, we report the training error in the HPPM optimization process (red boxes), along with the error in the ground truth fitting process for both datasets (blue boxes for *PV-Human3.6M* and cyan boxes for *PV-3DPW*).

Part Names	Hyperparameters			Vertex Errors			Joint Errors		
	k_p	N_p	$\lvert\mathcal{J}_p\rvert$	Train.	Fit. PV-H36M	Fit. PV-3DPW	Train.	Fit. PV-H36M	Fit. PV-3DPW
Abdomen	33	839	4	1.95	1.00	1.81	1.67	1.06	1.22
Left Thigh	31	375	2	1.29	0.86	0.98	2.00	1.81	2.13
Right Thigh	31	370	2	1.25	0.87	0.92	2.00	0.96	1.93
Left Calf	16	284	2	1.07	0.62	0.92	0.79	0.70	0.24
Right Calf	16	284	2	1.10	0.67	0.85	1.23	1.42	1.25
Chest	42	1,490	4	2.00	0.90	2.09	0.95	0.47	0.93
Left Foot	16	283	1	0.53	0.27	0.45	0.48	0.57	1.78
Right Foot	16	283	1	0.57	0.28	0.41	0.87	1.32	1.28
Head	21	1,257	3	0.59	0.45	0.50	1.89	1.96	4.34
Left Upper Arm	36	381	2	0.81	0.37	0.64	1.96	1.08	0.80
Right Upper Arm	38	382	2	0.80	0.33	0.84	1.90	0.69	3.21
Left Forearm	16	316	2	1.08	0.52	0.83	1.13	0.66	1.19
Right Forearm	16	316	2	1.11	0.55	0.97	1.37	0.95	1.75
Left Hand	16	810	1	0.57	0.29	0.60	1.34	0.77	2.24
Right Hand	16	810	1	0.56	0.32	0.69	1.52	1.24	0.46
Average	-	-	-	1.11	0.59	1.05	1.46	1.05	1.69

4 Experiments

4.1 Partially Visible Benchmarks

To train and evaluate our approach, we created two benchmarks, *Partially Visible Human3.6M* (*PV-Human3.6M*) and *Partially Visible 3DPW* (*PV-3DPW*) based on existing public datasets, Human3.6M [11] and 3DPW [47], respectively. The input images for our benchmarks are partially visible human images. Additionally, we generated the corresponding HPPM annotations. In our experiments, we utilize the mean per-vertex error (MPVE) to assess the accuracy of the mesh reconstruction and the mean per-joint position error (MPJPE) to evaluate the precision of the joint positions, following [12]. The detailed definition of MPVE and MPJPE can be found in Supplementary Material Sec. C.

HPPM annotations. We generate HPPM annotations from SMPL ground truth of Human3.6M and 3DPW. The annotations consist of shape parameters S and global transformation M, which are used as annotations for training. From here and the following representations in this section, we omit the part subscript p for simplicity. Specifically, we first use the segmentation strategy in Sec. 3.2 to generate ground truth part meshes. For each ground truth body part mesh v and its corresponding part template v_0, we calculate the global transformation M from v_0 to v as:

$$M = \arg\min_{M\prime} \|v - M' v_0\|, \tag{16}$$

where v and v_0 are in homogeneous coordinates and $M \in \mathbb{R}^{4\times 4}$. In practice, we use the least-square method to solve Eq. (16). Given the global transformation,

Table 2: Comparison of D&F with recent previous approaches on our *Partially Visible Human3.6M* and *Partially Visible 3DPW* benchmarks. MPVE and MPJPE are in millimeters, lower means better mesh and joint accuracy. The best results are shown in **bold**. The left 2 columns are the results directly tested on our benchmark using released model weights, among which "D&F (Ours)" are trained on public datasets. The 2 columns on the right are the results finetuned using our partially visible augmentation and part mesh pseudo ground truth. The results show our method outperforms recent previous methods, and our partially visible augmentation and part mesh pseudo ground truth can contribute to the improvement on partially visible human image inputs.

Methods	Directly Tested				Finetuned on Our Augmentation			
	PV-Human3.6M		*PV-3DPW*		*PV-Human3.6M*		*PV-3DPW*	
	MPVE↓	MPJPE↓	MPVE↓	MPJPE↓	MPVE↓	MPJPE↓	MPVE↓	MPJPE↓
MotionBERT [63]	350.6	305.4	284.7	250.0	196.9	169.6	185.5	155.4
SEFD [50]	336.7	259.6	275.0	224.6	276.0	198.9	241.1	203.0
GLoT [42]	325.2	298.5	297.2	255.7	214.1	199.0	235.0	213.5
CycleAdapt [30]	367.0	318.1	268.2	230.8	249.4	231.9	189.0	137.1
D&F (Ours)	**155.7**	**156.3**	**208.6**	**194.9**	**63.3**	**55.9**	**109.9**	**102.7**

Table 3: We do ablation studies on designing necessity and loss functions on *PV-Human3.6M* and *PV-3DPW* benchmarks. As we remove some losses and module designs, the performance suffers decreases to different extents. The "Directly Tested" result of D&F is when we train on public datasets and test on our partially visible benchmarks. The best results are shown in **bold**.

Experiment settings	*PV-Human3.6M*		*PV-3DPW*	
	MPVE/mm↓	MPJPE/mm↓	MPVE/mm↓	MPJPE/mm↓
w/o part 2D projection loss	64.2	56.7	111.8	104.9
w/o part 3D joint loss	63.9	56.2	112.4	105.8
w/o part 3D per-vertex loss	68.4	63.5	120.2	108.3
w/o HPPM shape-parameter loss	70.1	57.0	119.7	103.9
w/o HPPM 6D rotation loss	95.6	87.5	138.5	127.4
w/o HPPM translation loss	75.1	69.9	123.0	115.3
w/o overlapping loss	74.5	64.2	125.2	111.8
w/o depth consistency loss	65.4	58.5	113.9	104.1
w/ fixed #parameters	67.5	61.7	114.0	105.4
D&F(Ours)	**63.3**	**55.9**	**109.9**	**102.7**

we can transform ground truth mesh v to the canonical space of template mesh as $v' = M^\top v$. Then we calculate the HPPM shape parameter S as:

$$S = \mathcal{U}^\top (v' - \mathcal{M}), \tag{17}$$

where $\mathcal{U}^\top$ is the transpose matrix of HPPM shape matrix, $\mathcal{M}$ is the mean shape of HPPM, and $v' \in \mathbb{R}^{3N}$ is the flattened vertex vector, given N the vertex number of this part. Having shape parameters and global transformations, we can recover the part mesh as:

$$v_{HPPM} = M(\mathcal{U}S + \mathcal{M}), \tag{18}$$

In Tab. 1, we evaluate the HPPM annotation error for each part in both benchmarks. We observe that for all body parts, the error is negligible and acceptable. We also evaluate the necessity of the generated supervisions in Tab. 3. The ex-

Table 4: Results on *PV-3DPW* images with more visible parts. The best results are shown in **bold**. As the number of visible parts increases to 5-10, our method still works well and outperforms previous approaches.

#Parts Visible	1-4		5-10	
Methods	MPVE/mm↓	MPJPE/mm↓	MPVE/mm↓	MPJPE/mm↓
CycleAdapt [30]	189.0	137.1	169.9	132.8
GLoT [42]	185.5	155.4	142.3	131.9
D&F (Ours)	**109.9**	**102.7**	**117.4**	**107.5**

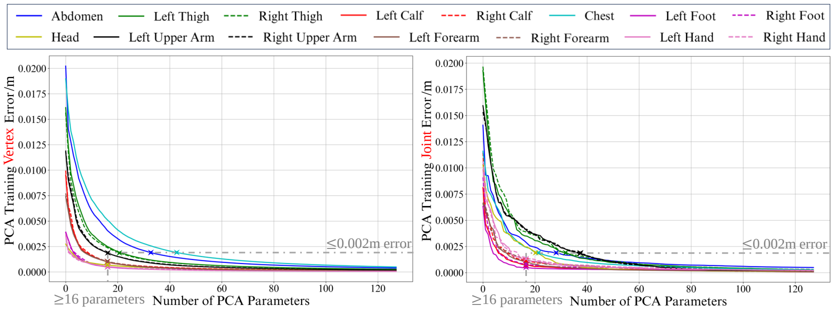

Fig. 4: HPPM training error of each part changes with the number of shape parameters used. We consider an adjustable number of parameters for each part. We set the maximum joint and vertex training errors to be 2mm, and a minimum number of parameters to 16. Left: vertex training error. Right: joint training error.

periments show that the global transformation and shape parameter annotations are helpful to improve performance.

Partially visible human images. We use a random cropping strategy to generate partially visible human images from Human3.6M and 3DPW. First, we project the HPPM ground truth to the image, to determine which part of the image corresponds to which body part(s). Next, a center point is randomly selected within the human bounding box, as well as a random side length for the square cropping. For every image in the original dataset, this process is performed 20 times, and cropped results with 1 to 4 body parts visible are short-listed. For visible parts, we require $\geq 50\%$ area of the part bounding box to be inside the cropped image. Resulting cropped images are shown in Fig. 1 and Fig. 6.

Partially visible augmentation. We use the similar image cropping strategy mentioned in the previous paragraph as training augmentation. This operation improves the generalizability of trained methods w.r.t. visibility distributions. Our experiment in Tab. 2 shows that our partially visible augmentation strategy not only increases the performance of our framework but also other previous methods on our partially visible datasets.

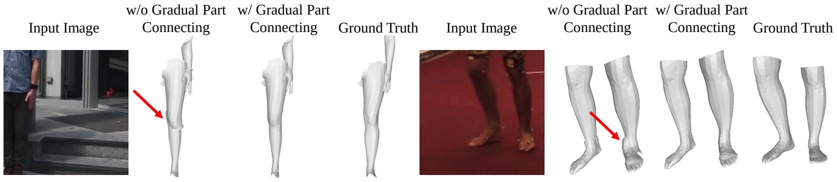
Input Image
w/o Gradual Part Connecting
w/ Gradual Part Connecting
Ground Truth
Input Image
w/o Gradual Part Connecting
w/ Gradual Part Connecting
Ground Truth

Fig. 5: Visual ablation on gradual part connecting. When this module is removed, the connection points between two adjacent parts become misaligned, as indicated by the red arrow. This alignment issue is resolved using the gradual part connecting.

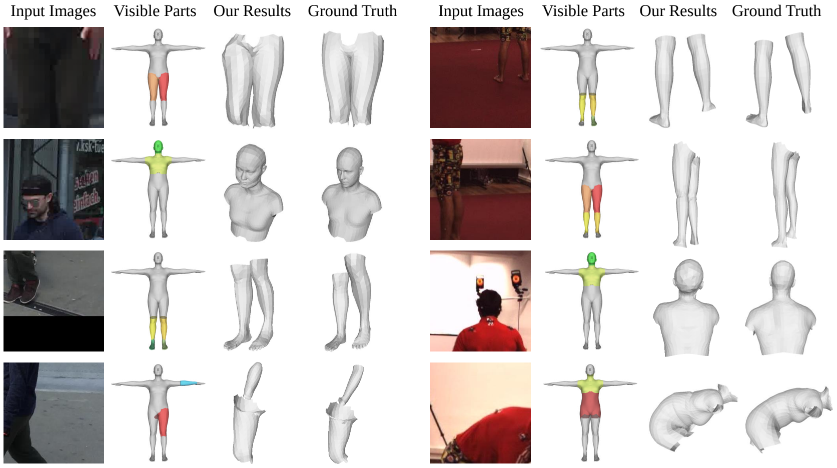

Fig. 6: Qualitative results from *PV-3DPW* benchmark (left) and *PV-Human3.6M* benchmark (right). Each benchmark features an input image (1st column), selected visible parts (2nd column), our mesh results (3rd column), and the ground truth mesh for visible parts only (4th column). Our method successfully generates accurate meshes from partially visible inputs.

4.2 Implementation Details

We end-to-end train our network on a single NVIDIA A100 40GB GPU. We optimize our model on Human3.6M [11], 3DPW [47], and SURREAL [45], with the augmentation introduced in Sec. 4.1. We evaluate our model on the *PV-Human3.6M* and *PV-3DPW* benchmarks separately. The batch size is set to 92. The training weights are set to $\lambda_v = 2.5$, $\lambda_{j3d} = 1,250$, $\lambda_{j2d} = 2,500$, $\lambda_s = 100$, $\lambda_r = 200$, $\lambda_t = 500$, $\lambda_{ol} = 100$, and $\lambda_{dc} = 1$. We use Adam [15] for optimization and the learning rate is set to 1×10^{-4}. During inference, the dataset annotations for part visibility are leveraged. The code is implemented using PyTorch [33].

4.3 Results

Comparison with state-of-the-art. We compare our method with recent previous methods on our *PV-Human3.6M* and *PV-3DPW* benchmarks on visible parts in Tab. 2. Our D&F outperforms recent previous methods on both benchmarks in terms of both mesh and joint accuracy (MPVE/MPJPE), regardless of whether they are finetuned with our partially visible augmentation.

HPPM parameter numbers. To determine the number of shape parameters used in each part, Fig. 4 highlights how HPPM training error w.r.t. part vertices and joints are impacted by the number of shape parameters. *I.e.*, the error drops as the number of parameters increases. The larger the number of parameters, the more exact the part meshes are, but the more challenging the prediction task becomes for D&F. Therefore, we propose the trade-off by setting the maximum joint and vertex training errors to be 2 millimeters, and a minimum number of parameters to 16 in HPPM design. In total, we use 360 shape parameters.

Ablation studies. We show the necessity of our module and loss function design in Tab. 3. The performance of our method drops when removing some necessary modules or loss functions. We observe that the part 3D per-vertex/3D joint/2D projection loss generally helps part mesh recovery and part joint accuracy, which is similar to the whole body SMPL-based frameworks. The HPPM pseudo-ground-truth annotations including 6D rotations are also helpful. The overlapping loss also increases the result by a large margin. Apart from these ablation studies, We also show the effectiveness of our gradual part connecting in Tab. 3. Without this module, the connection vertices between 2 adjacent parts can be misaligned (red arrow); an issue solved with the proposed connecting scheme.

Higher-visibility scenarios. We show some results of our method when the input contains a higher number of visible parts in Tab. 4. On *PV-3DPW* dataset, we increase the number of visible parts to 5-10. We observe that, even when there are more visible portions of the input human body, our method still works well and outperforms previous approaches.

Visualizations. We show some additional qualitative results on *PV-Human3.6M* and *PV-3DPW* benchmarks in Fig. 6. We observe that our method can generate valid mesh results with partially visible human image inputs.

5 Future Work and Conclusion

Though our HPPM can express some deformation on hands and faces, its design could be extended, *e.g.*, by leveraging the SMPL-X body template [34]. This extension would increase the accuracy of hand poses and facial expressions. Besides, a part-detection or segmentation network such as [40,41] could be added to automatically detect which part is visible.

In conclusion, our Divide and Fuse method successfully addresses the limitations of existing top-down human mesh reconstruction techniques in the presence of occlusions. Through Human Part Parametric Models, independent part reconstruction, and strategic fusion, our approach consistently delivers accurate meshes with partially visible bodies, as validated by our provided benchmarks *PV-Human3.6M* and *PV-3DPW*. These advancements represent a considerable improvement in mesh reconstruction accuracy and reliability when dealing with partially visible human images.

References

1. Banik, S., Gschoßmann, P., Garcia, A.M., Knoll, A.: Occlusion Robust 3D Human Pose Estimation with StridedPoseGraphFormer and Data Augmentation (2023), arXiv:2304.12069 [cs]
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
3. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: ICCV (2019)
4. Chen, X., Zhang, J., Wang, K., Wei, P., Lin, L.: Multi-Person 3D Pose Esitmation With Occlusion Reasoning. IEEE Transactions on Multimedia pp. 1–13 (2023)
5. Cheng, Y., Yang, B., Wang, B., Wending, Y., Tan, R.: Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In: ICCV. pp. 723–732 (2019)
6. Choi, H., Moon, G., Park, J., Lee, K.M.: Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes. In: CVPR. pp. 1465–1474 (2022)
7. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: ICCV (2019)
8. Ghafoor, M., Mahmood, A.: Quantification of Occlusion Handling Capability of a 3D Human Pose Estimation Framework (2022), arXiv:2203.04113 [cs]
9. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR. pp. 10884–10894 (2019)
10. Huang, B., Shu, Y., Ju, J., Wang, Y.: Occluded Human Body Capture with Self-Supervised Spatial-Temporal Motion Prior (2022), arXiv:2207.05375 [cs]
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2013)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
13. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR. pp. 5614–5623 (2019)
14. Khirodkar, R., Tripathi, S., Kitani, K.: Occluded Human Mesh Recovery. In: CVPR. pp. 1705–1715 (Jun 2022)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR. pp. 5253–5263 (2020)
17. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part Attention Regressor for 3D Human Body Estimation (2021), arXiv:2104.08527 [cs]
18. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
19. Li, J., Bian, S., Liu, Q., Tang, J., Wang, F., Lu, C.: NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In: CVPR (June 2023)
20. Li, T., Fan, L., Yuan, Y., Katabi, D.: Unsupervised Learning for Human Sensing Using Radio Signals. In: WACV. pp. 1091–1100 (2022)
21. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022)

22. Liu, G., Rong, Y., Sheng, L.: VoteHMR: Occlusion-Aware Voting Network for Robust 3D Human Mesh Recovery from Partial Point Clouds (2021), arXiv:2110.08729 [cs]
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
24. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (2015)
25. Luan, T., Wang, Y., Zhang, J., Wang, Z., Zhou, Z., Qiao, Y.: Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2269–2276 (2021)
26. Luan, T., Zhai, Y., Meng, J., Li, Z., Chen, Z., Xu, Y., Yuan, J.: High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16795–16804 (2023)
27. Ma, X., Su, J., Wang, C., Zhu, W., Wang, Y.: 3d human mesh estimation from virtual markers. In: CVPR. pp. 534–543 (2023)
28. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
29. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
30. Nam, H., Jung, D.S., Oh, Y., Lee, K.M.: Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In: ICCV. pp. 14829–14839 (2023)
31. Nikos Kolotouros, Georgios Pavlakos, K.D.: Convolutional mesh regression for single-image human shape reconstruction. CVPR (2019)
32. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: A sparse trained articulated human body regressor. In: ECCV. pp. 598–613 (2020)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
34. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR. pp. 10975–10985 (2019)
35. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
36. Qammaz, A., Argyros, A.: Occlusion-tolerant and personalized 3d human pose estimation in rgb images. In: ICPR. pp. 6904–6911 (2021)
37. Radwan, I., Dhall, A., Goecke, R.: Monocular Image 3D Human Pose Estimation under Self-Occlusion. In: ICCV. pp. 1888–1895 (2013)
38. Rafi, U., Gall, J., Leibe, B.: A semantic occlusion model for human pose estimation from a single depth image. In: CVPRW. pp. 67–74 (2015)
39. Ran, H., Ning, X., Li, W., Hao, M., Tiwari, P.: 3D human pose and shape estimation via de-occlusion multi-task learning. Neurocomputing p. 126284 (2023)
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
41. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
42. Shen, X., Yang, Z., Wang, X., Ma, J., Zhou, C., Yang, Y.: Global-to-local modeling for video-based 3d human pose and shape estimation. In: CVPR. pp. 8887–8896 (2023)
43. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: ICCV (2019)

44. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: Synthetic Occlusion Augmentation with Volumetric Heatmaps for the 2018 ECCV PoseTrack Challenge on 3D Human Pose Estimation (2018), arXiv:1809.04987 [cs]
45. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
46. Veges, M., Lorincz, A.: Temporal Smoothing for 3D Human Pose Estimation and Localization for Occluded People (2020), arXiv:2011.00250 [cs]
47. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV. pp. 601–617 (2018)
48. Wang, Y., Mori, G.: Multiple Tree Models for Occlusion and Spatial Constraints in Human Pose Estimation. In: ECCV, pp. 710–724 (2008)
49. Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., Su, L.: mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In: Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. pp. 269–282 (2021)
50. Yang, C., Kong, K., Min, S., Wee, D., Jang, H.D., Cha, G., Kang, S.: Sefd: Learning to distill complex pose and occlusion. In: ICCV. pp. 14941–14952 (2023)
51. Yang, K., Gu, R., Wang, M., Toyoura, M., Xu, G.: LASOR: Learning Accurate 3D Human Pose and Shape via Synthetic Occlusion-Aware Data and Neural Mesh Rendering. IEEE TIP pp. 1938–1948 (2022)
52. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras. In: CVPR. pp. 11028–11039 (2022)
53. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3d human mesh regression with dense correspondence. In: CVPR (2020)
54. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
55. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021)
56. Zhang, J., Wang, Y., Zhou, Z., Luan, T., Wang, Z., Qiao, Y.: Learning dynamical human-joint affinity for 3d pose estimation in videos. IEEE Transactions on Image Processing **30**, 7914–7925 (2021)
57. Zhang, T., Huang, B., Wang, Y.: Object-Occluded Human Shape and Pose Estimation From a Single Color Image. In: CVPR. pp. 7374–7383 (2020)
58. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. CVPR (2019)
59. Zhao, M., Liu, Y., Raghu, A., Zhao, H., Li, T., Torralba, A., Katabi, D.: Through-Wall Human Mesh Recovery Using Radio Signals. In: ICCV. pp. 10112–10121 (2019)
60. Zhou, L., Chen, Y., Gao, Y., Wang, J., Lu, H.: Occlusion-Aware Siamese Network for Human Pose Estimation. In: ECCV, pp. 396–412 (2020)
61. Zhou, Q., Wang, S., Wang, Y., Huang, Z., Wang, X.: Human De-occlusion: Invisible Perception and Recovery for Humans. In: CVPR. pp. 3690–3700 (2021)
62. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR. pp. 5745–5753 (2019)
63. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: ICCV (2023)

64. Zuffi, S., Black, M.J.: The stitched puppet: A graphical model of 3d human shape and pose. In: CVPR. pp. 3537–3546 (2015)