VISIBILITY-AWARE HUMAN MESH RECOVERY VIA DENSE CORRESPONDENCE AND PROBABILITY MODEL

Yanjun Wang[1] *Wenjia Wang*[2] *Jun Ling*[1] *Rong Xie*[1] *Li Song*[1⋆]

[1] Shanghai Jiao Tong University [2]The University of Hong Kong

ABSTRACT

Human mesh recovery from a single image is challenging due to self-occlusion, object occlusion, or human occlusion. However, existing methods failed to strike a balance between occlusion robustness and mesh-image alignment. We propose Visibility-aware Human Mesh Recovery (VisHMR), a dual-branch design that adaptively integrates strong 2D visual cues and robust human features based on visibility. We utilize a UV dense correspondence map as a visual cue for visual estimation and tackle the occlusion as a probability model. We further merge two results based on body parts visibility to derive robust and accurate results. We evaluate our methods on both common datasets and occlusion datasets. Extensive experiments show that our method clearly outperforms prior SOTA results on occluded scenarios and comparable results on the standard benchmarks (3DPW).

Index Terms— Occluded human pose and shape estimation, Dense correspondence map, Probability model

1. INTRODUCTION

Reconstructing human mesh from a single image has a broad range of applications, such as human motion analysis, digital human animation, and augmented reality. HMR (Human Mesh Recovery) has developed rapidly in recent years with the boost from new datasets and deep neural networks, but limited due to occlusion issues in real-world scenarios as it disrupts the correspondence between image and human pose.

Current HMR methods largely rely on the parametric model SMPL(A Skinned Multi-Person Linear Model) [1], which controls the human pose and shape with a set of vectors. Many of the methods seek to fit SMPL from existing 2D human cues such as keypoints, segmentations, or dense correspondence maps. These cues are easier to predict compared to the highly non-linear SMPL parameters, making these methods have great mesh-image alignment. But these cues are also affected by occlusion, so they often perform worse when human images are occluded or truncated. There are also direct regression methods that directly regress SMPL parameters from image features. As it has various solutions to utilize these features, such as attention, and inpainting, making them more robust to occlusion. But these methods often have poor mesh-image alignment, as there is no direct correspondence between SMPL parameters and image.

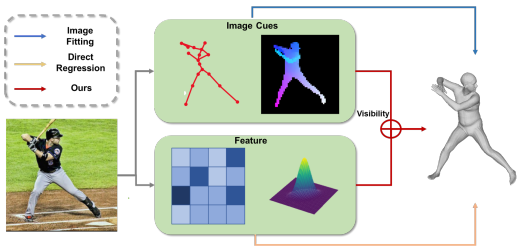

Fig. 1: **Illustration of our proposed method.** We merge two approaches for human mesh recovery to improve occlusion robustness and mesh-image alignment

As illustrated in Fig. 1, we resort to solving occlusion by adopting both 2D human cues and image features and combining them according to visibility in this work. Specifically, we propose a parallel architecture, one branch utilizes dense correspondence maps to obtain image-aligned cues and regresses image-aligned pose parameters based on visible cues. The other branch utilizes probability modeling to generate multiple whole-body hypotheses directly. We introduce a method to derive pseudo body parts visibility based on a dense correspondence map. Based on this pseudo visibility, we match the best hypotheses with visible prediction and utilize a kinematic-based spherical linear interpolation helping the model merge these two results based on visibility. Our main contributions are: 1) We introduce a novel way to utilize a dense correspondence map to obtain occlusion-awareness features for accurate SMPL estimation. 2) We propose an architecture to adaptively utilize strong image cues and robust global features for occlusion human mesh recovery. 3) We achieve SOTA results on occlusion benchmarks and comparable results on general benchmarks.

2. RELATED WORK

Direct-regression Methods: Direct regression methods [2, 3, 4, 5] tend to directly regress all SMPL parameters at once based on the input image. These methods might generate intermediate image-aligned features but they only serve as

⋆Corresponding author.

help for accurate regression. For instance, HMR [2] first utilizes CNN to extract the global features and directly regresses the SMPL parameters with MLPs. PARE [3] devises a part attention regressor to predict body-part-guided attention mask, then performs attention across body part features to obtain occlusion robust features for human mesh regression. ProHMR [6] model ambiguity as a probabilistic problem and utilize normalizing flow to derive the final result. However the map between the image to SMPL pose is highly non-linear, and direct-regression methods have lower accuracy compared to image-aligned methods on non-occluded scenes. **Image-aligned methods:** Many methods seek to derive SMPL parameters from other more direct image cues such as 2D keypoints [7] or dense correspondence [8]. Some methods like EFT [9] and SMPLify [7] utilize optimization to fit the SMPL model based on the 2D keypoints. HybrIK [10] proposes to obtain the pose parameters from estimated 3D keypoints via Inverse kinematics. DecoMR [8] proposes to estimate a dense correspondence map and warp the feature to UV space for coordinate regression. However, these cues might be disturbed by occlusion and create severe artifacts when fitting SMPL parameters. Compared to these methods, our model only utilizes image cues on visible body parts, and uses global regression results for invisible parts, making our methods more robust to occlusions.

3. METHODS

3.1. Overview

We introduce VisHMR to address occlusion challenges in human reconstruction, as shown in Fig. 2. Starting with cropped image I, our goal is to accurately estimate SMPL pose and shape parameters, along with a weak-perspective camera, to create a human mesh aligned with the input image. The model employs a ResNet-50 to extract a feature map for two branches. One branch is dedicated to precise pose estimation, aligning with the visible human body parts as observed in the image. This branch utilizes UV-dense correspondence maps, a strong image cue to extract visible human features and estimate the visibility of different body parts. In parallel, the other branch generates multiple whole-body pose assumptions from global image features. Given the definitive prediction for the visible human parts and multiple assumptions of the whole body, our model performs a matching process, comparing these hypotheses with the estimated visible joints' rotations. Finally, these two estimations are merged using Kinematic Slerp to derive the joint prediction.

SMPL Parametric Model: We utilize the parametric model SMPL [1] to generate human mesh. SMPL represents a 3D human mesh with $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$ where $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is the first 10 coefficients for the body shape. And $\boldsymbol{\theta} \in \mathbb{R}^{24\times 3}$ is the body pose defined by relative joint rotations along the SMPL kinematic tree and global orientation. A mapping function can derive the vertex mesh $M \in \mathbb{R}^{6890\times 3}$ based on the shape and pose parameters.

3.2. UV Based Visible Branch

Our design of the visible branch is aimed at efficiently extracting visible features, thereby improving the accuracy of image-aligned predictions for visible body parts. It first employs dense correspondence estimation to isolate visible human parts from the background and occlusion. Further, the model employs feature warpping and cross-attention mechanisms to gather pertinent features from the visible body, facilitating the estimation of rotation parameters for these visible body parts. To emphasize its image-aligned estimation, we change the relative rotation to parent joints to absolute rotation to the camera. We hope this will decouple each body part estimation.

UV Map Estimation: We utilize a dense correspondence UV map to extract features of the visible human parts. This UV map is generated from the SMPL model's UV texture map, where each vertex on the model corresponds to a designated 2D coordinate in the UV image space. By rendering the mesh's UV values into an image, we create an IUV image for a human image. This IUV image assigns a mesh location to each visible human pixel in the image. Using the IUV image, we align the image feature with the mesh locations and subsequently map the feature vectors to UV space.

Based on the extracted image feature $\mathbf{F_{img}}$, the model regresses the IUV map of the image, $M_{iuv} \in \mathbb{R}^{3\times H\times W}$, which the first channel $i \in \mathbb{R}^{H\times W}$ represents the segmentation of the human object, and the second and third channel are a two-dimensional coordinate $(u, v) \in \mathbb{R}^{2\times H\times W}$, which represents the warpped location of this pixel. The estimated IUV map provides two key features to extract accurate visible features. Firstly, the IUV map filtered background and occluders, obtaining only visible human area. And second, it provides a dense correspondence between the image and the human body.

UV based visibility calculation: To enable visibility-awareness, we calculate the pseudo body-part visibility $\boldsymbol{\omega}_{vis}$ based on the UV map, and supervise visibility prediction with it. Specifically, the IUV image will only show visible human mesh surfaces, it will only warp visible area to the UV map. Therefore, we render the ground truth body-parts segmentation image I_{seg}, and use the IUV map to warp them to the segmentation UV map UV_{seg}. We can obtain the total areas on UV_{seg}. Therefore, we can model the visibility as the ratio between the warpped areas and the total mesh areas on UV space. UV_{gt} is the total UV segmentation map un-warpped with full SMPL mesh. Note that the rendered image has a minimum 50 % of the body surface which is global, so we

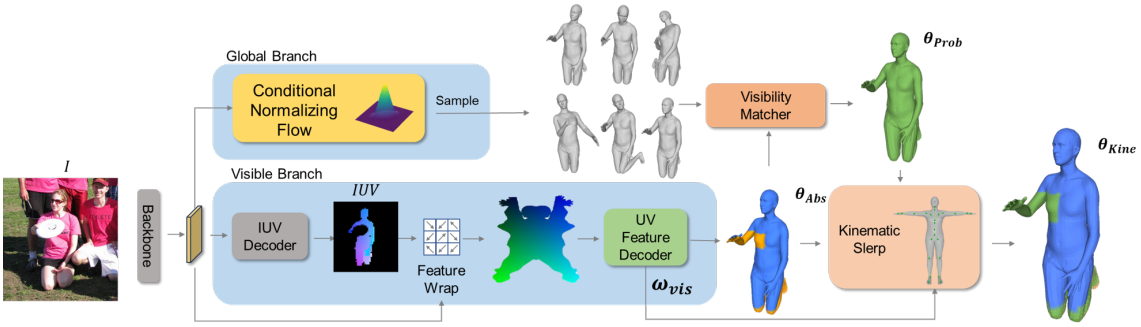

Fig. 2: **Pipeline overview** Our model consists of a UV-based visible branch and a probability-based global branch. For the visible branch, the IUV image warps the image feature to UV space and regresses visible human parameters from the visible feature. The global branch samples multiple whole-body hypotheses based on image features. The Visibility Matcher selects the best hypotheses. Kinematic Slerp merges two predictions to get the final results.

divide 0.5 for all the visibility. It can be formulated as:

$$\begin{aligned} \boldsymbol{\omega}_{vis} &= (\omega_0, \omega_1, \omega_2, \ldots, \omega_{23}) \\ \omega_i &= \frac{2 \times \text{area}\,(UV_{seg} = i)}{\text{area}\,(UV_{gt} = i)}, i \in (0, 23) \end{aligned} \tag{1}$$

Visible Parts Estimation: We inverse warp the feature map $\mathbf{F_{img}}$ into the UV space feature map $\mathbf{F_{UV}} \in \mathbb{R}^{D \times H' \times W'}$ to rearrange the human parts feature into a canonical and continuous UV space. Further, the module performs cross-attention on the warpped image feature and outputs the visible body parts features. We use an MLP to regress the part-wise rotation angle and part visibility based on the visible body parts features.

Loss Functions: The loss function of the visible branch is formulated as follows:

$$\mathcal{L}_{Visible} = \lambda_{IUV}\mathcal{L}^2_{IUV} + \lambda_{SMPL-abs}\mathcal{L}^1_{SMPL-abs} + \lambda_{vis}\mathcal{L}^1_{vis} \tag{2}$$

where $\mathcal{L}^2_{IUV}$ is the $\mathcal{L}2$ loss of the IUV image and λ_{IUV} is its weight, $\mathcal{L}^1_{SMPL-abs}$ is the $\mathcal{L}1$ loss of the SMPL parameters and $\lambda_{SMPL-abs}$ is its weight, and $\mathcal{L}^1_{vis}$ is the $\mathcal{L}1$ loss of body part visibility and λ_{vis} is the weight .

3.3. Probability Based Global Branch

Similar to ProHMR [6], we utilize conditional normalizing flow to model the distribution of the human pose, which transforms the complex pose distribution of into a normal distribution through a set of invertible transformations. This branch f is able to sample multiple poses given input condition $\mathbf{c}$ and sampled vector $\mathbf{z}$: $\boldsymbol{\theta_{prob}} = f(\mathbf{z}; \mathbf{c})$, and also able to calculate the likelihood of condition $\mathbf{c}$ and pose $\boldsymbol{\theta_{Prob}}$: $\mathbf{z} = f^{-1}(\boldsymbol{\theta_{Prob}}; \mathbf{c})$. As the occluded pose is highly ambiguous, it is intuitive to estimate multiple hypotheses. Also, since the branch will learn the ground truth pose distribution, it is unlikely to make out-of-distribution predictions compared to pure image-aligned methods.

Loss Functions: The loss function of the global branch is formulated as follows:

$$\mathcal{L}_{Global} = \lambda_{SMPL-exp}\mathcal{L}^1_{SMPL-exp} + \lambda_{nll}\mathcal{L}^1_{nll} \tag{3}$$

where $\mathcal{L}^1_{SMPL-exp}$ is the $\mathcal{L}1$ loss of the expectation of SMPL parameters of multiple hypotheses, and $\mathcal{L}^1_{nll}$ is used to minimize negative log-likelihood of ground truth samples.

3.4. Matching and Combine

Visibility-based matching: When a body part is visible, the difference between visible estimation and global estimation is expected to be similar. Thus, our model uses the visible branch's prediction to sample the best hypothesis of the global branch. Specifically, we calculate the cosine similarity of each body part rotation, and weighted sum them by visibility as the final similarity score and select the sample with the highest sim:

$$sim = \sum_{i=0}^{23} \omega_i \frac{\boldsymbol{\theta}^i_{Abs}\boldsymbol{\theta}^i_{Prob}}{|\boldsymbol{\theta}^i_{Abs}|\,|\boldsymbol{\theta}^i_{Prob}|} \tag{4}$$

where $\boldsymbol{\theta}_{Abs}$ is global estimation and $\boldsymbol{\theta}_{Prob}$ is the probability branch estimation.

Kinematic Slerp: Given the selected hypothesis and the visible part's estimation. we need a method, such as spherical linear interpolation (slerp), to reasonably integrate these two predictions. The visible branch emphasizes absolute alignment and the global branch emphasize reasonable relative pose, the slerp methods should perserve these features. Specifically, we perform a step-by-step merging based on the kinematic tree of the SMPL model called Kinematic Slerp. The process proceeds as follows.

Starting from the root node, based on the visibility of the root node, we interpolate between the predictions from both branches to obtain the absolute rotation angle of the root node.

For each interpolated parent node, we calculate its absolute rotation based on the relative rotation θ^i_r of the global

branch and the slerped absolute rotation of the parent joint $\hat{\theta}_a^{i-1}$: $\hat{\theta}_a^{i-1}\theta_r^i$. This is then slerped with the absolute rotation θ_a^i of the visible branch to get the slerped rotation. This operation is repeated for each node along the tree, yielding the final result $\hat{\theta}_a^i$. It can be formulated as follows:

$$\hat{\theta}_a^i = \begin{cases} SLERP\left(\omega_i\theta_a^i + (1-\omega_i)\,\hat{\theta}_a^{i-1}\theta_r^i\right), i > 0 \\ SLERP\left(\omega_i\theta_a^0 + (1-\omega_0)\,\theta_r^0\right), i = 0 \end{cases} \quad (5)$$

4. EXPERIMENTS

Datasets: We train VisHMR using the training sets of Human3.6M [11], COCO [12], MPII [13], Lspet [14] and 3DPW [15]. For Human3.6M [11] and 3DPW, we utilize ground truth SMPL parameters. For COCO, MPII and Lspet, we adopt the pseudo ground truth SMPL parameters by CLIFF [16]. Evaluation benchmarks are: (1) 3DPW, a general outdoor benchmark for human pose and shape estimation; (2) 3DPW-OC, the object-occluded subset of 3DPW; (3) 3DOH [17], a indoor object-occluded dataset.

Methods	3DPW-OC			3DOH	
	PA↓	MP↓	PVE↓	PA↓	MP↓
OOH [17]	72.2	-	-	58.5	-
PyMAF [4]	-	-	-	96.2	107.3
HMR-EFT [2]	60.9	94.9	111.3	66.2	101.9
SPIN [18]	60.8	95.6	121.6	68.3	104.3
HybrIK [10]	58.8	90.8	111.9	**31.2**	**40.4**
PARE [3]	56.6	90.5	107.9	44.3	63.3
VisHMR(Ours)	**55.6**	**83.9**	**99.96**	40.0	54.4

Table 1: **Evaluation on occlusion datasets 3DPW-OC and 3DOH. PA**: PA-MPJPE; **MP**: MPJPE.

Evaluation metrics. MPJPE: mean per joint position error, assess the accuracy of 3D joint rotation and body orientation. PA-MPJPE: Procrustes-aligned MPJPE, mainly assess the accuracy of 3D joint rotation. PVE:per-vertex error, evaluates the 3D surface error.

Occlusion Evaluation: Table 1 compares VisHMR's robustness against occlusion with other SOTA (state-of-the-art) occlusion handling methods on occlusion datasets. Fig. 3 has shown our qualitative results. Our model achieves SOTA performance on the 3DPW-OC dataset. Compared to other image correspondence methods such as HybrIK [10], our model is clearly better, as our model utilizes global image features for occluded human parts instead of image cues. On the 3DOH dataset, our model significantly outperforms SOTA occlusion handling methods PARE [3], OOH [17] and image correspondence method PyMAF [4].

General Comparison: As shown in Table 2, we benchmark our model on 3DPW. Our results indicate that VisHMR, achieves comparable performance when compared to SOTA methods. Compared to the occlusion-handling methods like PARE, OCHMR, our model significantly outperforms them. We attribute this success to our model's ability UV dense correspondence to achieve accurate image-aligned prediction, making them more accurate than direct regression methods.

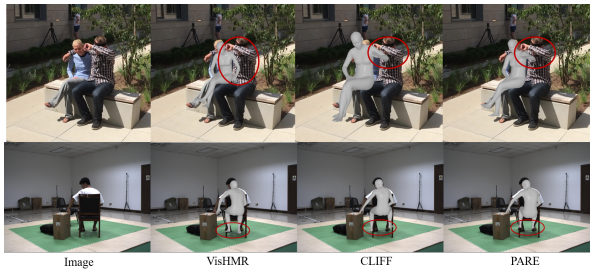

Fig. 3: **Qualitative Comparison**

	3DPW	PA-MPJPE↓	MPJPE↓	PVE↓
Align	DecoMR [8]	61.7	-	-
	PyMAF [4]	58.9	92.8	110.1
	METRO [19]	47.9	77.1	**88.2**
Direct	SPIN [18]	59.2	96.9	116.4
	OCHMR [20]	58.3	89.7	107.1
	HMR-EFT [9]	52.2	85.1	98.7
	ROMP [5]	53.3	85.5	103.1
	PARE [3]	50.9	82	97.9
	VisHMR (Ours)	**46.7**	**75.5**	91.4

Ablation on 3DPW	PA-MPJPE↓	MPJPE↓	PVE↓
Ours' visible branch	47.5	76.2	92.2
Ours' global branch	48.8	78.0	95.7
Ours' KS-Highest	46.8	75.8	91.4
Ours' KS-Matched	**46.7**	**75.5**	**91.3**

Table 2: **Quantitative comparison results and ablation studies**

Ablation Study: We first evaluate the result of using one of our branches. As the global branch outputs multiple samples, we take the highest log-likelihood one as the result. Further, we conduct Kinemtaic Slerp with these two results, which ignored the visibility matching stage. At last, we evaluate the results with visibility match. As shown in Table 2, the results of two separate branches are lower than the slerped results, proving the effectiveness of our model's fundamental design, combining image-aligned features and global features. Further, visibility matching boosts performance.

5. CONCLUSION

In this paper, we present a visibility-aware model for 3D human pose and mesh estimation. It leverages both 2D visual cues and global image features to give mesh-image aligned and occlusion-robust estimation. Further, it merges two different results based on body-part visibility, achieving accurate prediction in different scenarios. Qualitative and quantitative experiments on various datasets illustrate that our model shows promising performance and generalization ability.

6. REFERENCES

[1] 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)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7122–7131.

[3] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black, "PARE: Part attention regressor for 3d human body estimation," in *Int. Conf. Comput. Vis.*, 2021, pp. 11127–11137.

[4] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun, "PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11446–11456.

[5] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei, "Monocular, one-stage, regression of multiple 3d people," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 11179–11188.

[6] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis, "Probabilistic modeling for human mesh recovery," in *ICCV*, 2021.

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 561–578.

[8] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang, "3d human mesh regression with dense correspondence," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 7054–7063.

[9] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi, "Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation," in *Int. Conf. 3D. Vis.* IEEE, 2021, pp. 42–52.

[10] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 3383–3393.

[11] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2013.

[12] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[13] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[14] Sam Johnson and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR 2011*. IEEE, 2011, pp. 1465–1472.

[15] 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 *Eur. Conf. Comput. Vis.*, 2018.

[16] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," *arXiv: Comp. Res. Repository*, 2022.

[17] Tianshu Zhang, Buzhen Huang, and Yangang Wang, "Object-occluded human shape and pose estimation from a single color image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7376–7385.

[18] 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 *Int. Conf. Comput. Vis.*, 2019, pp. 2252–2261.

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1954–1963.

[20] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani, "Occluded human mesh recovery," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 1715–1725.