Complete 3D Human Reconstruction From a Single Incomplete Image Supplemental Material

Junying Wang[1] Jae Shin Yoon[2] Tuanfeng Y. Wang[2]
Krishna Kumar Singh[2] Ulrich Neumann[1]
[1]University of Southern California [2]Adobe Research
{junyingw,uneumann}@usc.edu {jaeyoon,yangtwan,krishsin}@adobe.com

This document contains additional information on the method and experiment presented in the main paper, specifically as an extension of Sec.3 and Sec.4. We add more quantitative and qualitative results, and in the final section, we demonstrate the potential application usages of our model.

A. Network Architecture

We demonstrate a novel coarse-to-fine human reconstruction framework for using both the coarse shape and explicit occupancy. Our design is motivated by both optimization considerations and computational efficiency. Explicit occupancy allows the generative features to be optimized in a way that learns the global ordinal relations across 3D voxels with discriminators (whereas implicit models are often local). The coarse shape with an implicit model enables continuous surface reconstruction with effective computational cost (whereas explicit volumes are often costly).

A.1. Network Details

2D image encoder Our image encoder $\mathcal{E}$ applies the Unet structure using a 2D Stacked Hourglass [4], with a stack number of 3 as the backbone, and we remove MaxPooling Layers and Residual Layers in our experiments. We take an image with a spatial resolution of 512×512 as input and output a 32 dimensional vector with a resolution of 128×128. Then, we take the image feature from the last stacked layer and unproject it into the 3D space as volume feature $\mathbf{F}$ for the 3D CNN.

3D Generator For 3D generator $\mathcal{G}_{\mathrm{3d}}$, we applies the Unet structure using a 3D Stacked Hourglass, with a stack number of 2, as the backbone. The 3D generator takes a raw explicit volume as input with a spatial resolution of $128 \times 128 \times 128$, and the channel number of the raw volume feature is 33, where 3D body pose $\mathbf{P}$ is a one dimensional vector and image volume feature $\mathbf{F} \in \mathbb{R}^{32}$. The 3D generator $\mathcal{G}_{\mathrm{3d}}$ consists of 3D convolution layers and upsample layers. The depth of 3D CNN is equal to 3, with numbers of channels $\{48, 64, 96\}$, kernel sizes $\{3, 3, 3\}$, and strides $\{2, 2, 2\}$. There are 3D batch normalization layers in between. To regress the explicit 3D volume feature, we keep the intermediate features of each stack, and then feed them into the next implicit function, and the losses from all the stacks are aggregated for parameter update.

3D Discriminator The discriminator $\mathcal{D}_{\mathrm{3d}}$ takes a $128 \times 128 \times 128$ volume as input, and outputs a real number in 0, 1. The discriminator consists of four 3D convolution layers, with numbers of channels $\{64, 128, 256, 1\}$, kernel sizes $\{4, 4, 4, 4\}$, and strides $\{2, 2, 2, 2\}$. There are leaky ReLU layers of parameter 0.2 and batch normalization layers in between.

Coarse and fine implicit function The coarse MLP $\mathcal{C}$ decodes an occupancy value for each query point $\mathbf{X}$, as well as an intermediate global features $\mathbf{F}^*$, where $\mathbf{F}^*_{\mathbf{X}} \in \mathbb{R}^{256}$. The inputs of the coarse MPL are 3D position $\mathbf{X} \in \mathbb{R}^{63}$, concatenate with explicit trilinear interpolated volume feature $\mathbf{F}^g_{\mathrm{3d},\mathbf{X}} \in \mathbb{R}^{33}$. Same as NeRF [3], we use the positional encoding for each 3D position, which can enhance the high-frequency details. The number of neurons of coarse MLP $\mathcal{C}$ is $\{96, 1024, 512, 256, 128, 257\}$, with non-linear activations leaky ReLU and a Sigmoid layer at the end. Our fine surface reconstruction is based on another multi-layer perceptron, with the number of neurons $\{447, 1024, 512, 256, 128, 1\}$, this fine MLP $\mathcal{C}^f$ is designed to decode the fine-grained occupancy field. We use 0.5 level-set occupancy field for our fine surface representation. Same as $\mathcal{C}$, we use the positional encoding for each 3D position. For $\mathcal{C}^f$, it takes 3D position $\mathbf{X}$, surface normal features $\mathbf{F}^n_{\mathbf{x}}$ and global intermediate features $\mathbf{F}^*_{\mathbf{X}}$ as inputs, where $\mathbf{X} \in \mathbb{R}^{63}$, $\mathbf{F}^n_{\mathbf{x}} \in \mathbb{R}^{32\times 4}$, $\mathbf{F}^*_{\mathbf{X}} \in \mathbb{R}^{256}$. And the output of $\mathcal{C}^f$ is fine-grained occupancy for each query point.

A.2. Surface Sampling

Same as [7], we sample points using a mixture of uniform volume samples and importance sampling around the surface using Gaussian perturbation. Importance sampling enhances the surface details, while background sampling within the uniform volume removes background ambiguity and noise. For fast convergence, for each object, we precompute and save the occupancy of each sampling point. The total number of sampling points per object is N_t, $N_t = 100000$. During each iteration, we do batch processing with the number of samples $N = 8000$ and the ratio of importance sampling to uniform volume sampling is 8:1.

B. Quantitative Results

B.1. The requirement of GT SMPL

In our experiment, we used GT SMPL to provide fair and quantitative evaluation results that can be compared with other baselines that rely on SMPL. We admit that the pose errors in SMPL estimate affect the final 3D reconstruction results, and we use GT SMPL to factor out such pose errors as shown in Fig. 1. To be further fair, we include Tab.1 which summarizes the reconstruction accuracy given the predicted SMPL where our method still demonstrates better reconstruction results than PIFuHD and ICON. We kindly note that GT SMPL is not a requirement for testing in-the-wild images, and we attain SMPL estimates with an existing method [10, 11].

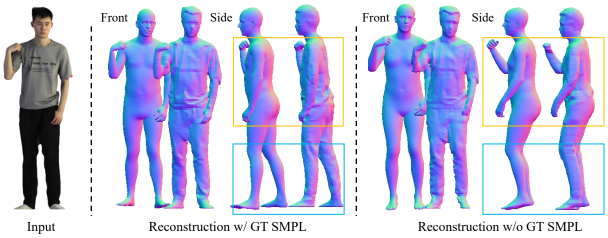

Figure 1. **3D reconstruction visualization quality with and without GT SMPL.** The main changes between them are colorized where the inaccurate 3D pose affects reconstruction accuracy.

Method	PIFuHD	ICON+PS	Ours+PS	ICON+GS	Ours+GS
Chamfer ↓	2.890	1.535	**1.224**	0.965	0.798
P2S ↓	2.631	1.479	**1.062**	0.848	0.808

Table 1. **Comparison with other baseline methods.** Here we show the quantitative comparison results for human reconstruction from an image where *GS* denotes the inference with ground-truth SMPL and *PS* with predicted SMPL [11].

B.2. Cross-dataset validation

We conduct the cross-dataset validation on MultiHuman-Dataset [12], this dataset includes the case with natural occlusion by objects and people and provides 3D surface ground truth and fitted SMPL for each person. We test on ten unseen objects with full body image as inputs, and calculate the average values of Chamfer distance, P2S, and normal PSNR over these testing subjects. For normal error, we render front, back, left and right side normal maps for each object and calculate the average normal error. Since ICON [8] and our model require SMPL [2] as human body prior, for evaluation, we use ground truth SMPL during the comparison for ours and ICON. The numerical results in Tab. 2 show that for full-body image reconstruction, our model has comparable performance to other models. Although our globally consistent volume feature slightly sacrifices local details, it still makes feasible reconstructions with better contour when given a full-body image. We also conduct an ablation study, testing on MultiHuman-Dataset. As shown in Tab. 3, our coarse-to-fine design can achieve better performance than other designs [6–8].

Cross-dataset Full Body Image Reconstruction				
Method	SMPL	Chamfer ↓	P2S ↓	Normal ↑
PIFu [6]	✗	3.768	3.989	10.879
PIFuHD [7]	✗	3.255	3.076	11.358
ICON [8]	✓	1.467	**1.389**	**11.986**
Ours	✓	**1.359**	1.402	11.932

Table 2. **Comparison with SOTA on Human Modeling**.

Method	Chamfer↓	P2S↓	Normal↑
Ours - coarse MLP - fine MLP	2.403	2.389	6.034
Ours - fine MLP	1.636	1.844	11.348
Ours w/o GT SMPL	1.977	1.802	11.455
Ours	**1.359**	**1.402**	**11.932**

Table 3. **Ablation study results.** We show the average values of Chamfer distance, P2S, and normal PSNR over 10 testing subjects on MultiHuman-Dataset [12]

B.3. Texture Inpainting

Our main contribution is 3D reconstruction, and texture inpainting is for application. Therefore, for the inpainting model, we adopt an existing human inpainting network [9] and designed two independent inpainting models: frontside inpainting and backside inpainting. Each inpainting model is based on partial body image and surface normal map. For the frontside inpainting model, we train with partial body image and conditioned on the surface normal map $\mathbf{N}^f$, which is supervised by the ground truth full body image. For the backside inpainting model, we train with the same partial body image as well as a mirrored backside normal map $\mathbf{N}^b$, and then supervised by the ground truth full-body image. During the inference time, we can get the complete full body texture by using view-progressive texture inpainting. In Fig. 3, we illustrate our method for generating a complete texture of

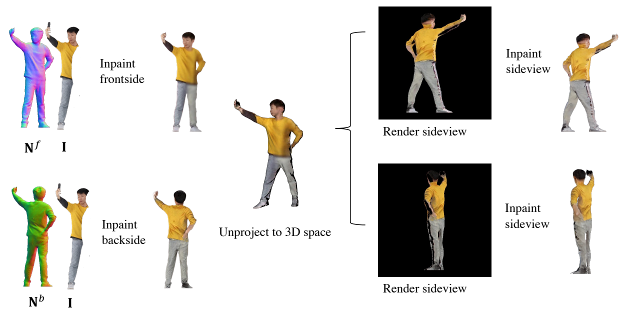

Figure 2. **Progressive texture inpainting** During the inference process, given our reconstructed model, we render the frontside normal $\mathbf{N}^f$ and the backside normal $\mathbf{N}^b$, and use the partial body image as input. We apply two independent inpainting models to inpaint the front and back sides. After obtaining the complete front and back images, we unproject them into 3D space and render the side views, and then inpaint them again. After 3D warping to other views, we can finally get the complete body texture.

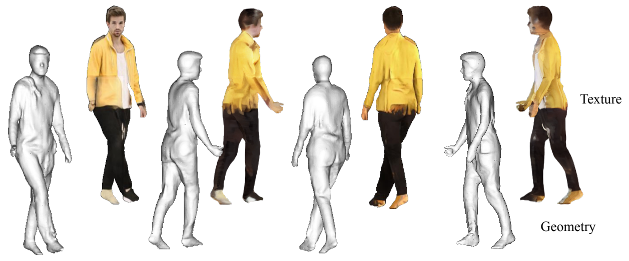

Figure 3. **Texture inpatining and 3D reconstruction results**

a human body from a partial input image and a complete geometry. Our approach involves synthesizing the image of a complete human body from multiple viewpoints in a progressive manner, using surface rendering, texture inpainting, and 3D warping techniques. 1) To perform texture inpainting, we first render the surface normal map for the front and back views of the human body, based on the complete geometry and partial input image. We then apply frontside and backside inpainting models to complete the front and back images, respectively. Next, we reproject the complete front and back images into 3D space and obtain the vertices colors from these views. 2) As the side view texture remains incomplete, we re-render the side view into 2D image space and perform inpainting again. 3) We iterate these steps to obtain a fully textured 3D model, which enables us to generate novel views of the human body from the partial input image. Finally, we can obtain the complete texture of the human body as shown in Fig.3.

B.4. Accumulative occlusion-to-reconstruction

As shown in Fig 7 (row 1), our model can achieve high quality reconstruction results comparable to other models when given a full-body image. Since other models. PIFu [6], PIFuHD [7] and ICON [8] only take into account fine-grained local image features and cannot enforce global consistency. Therefore, they can only achieve high quality reconstruction of visible parts, but are unable to reconstruct the whole human body from large-occluded image. As seen in Fig. 7 (row 2 to 5), we gradually change the occluded area of the image from 20% to 80%, and the reconstruction quality of other models decreases dramatically, while our model is able to reconstruct the whole body mesh with minor errors.

Human Body Reconstruction from 2D Inpainted Image			
Method	SMPL	Chamfer ↓	P2S ↓
PIFuHD [7]	✗	3.714	3.140
ICON [8]	✓	1.175	1.126
Ours	✓	**0.989**	**1.013**

Table 4. **Comparsion with SOTA on 3D reconstruction from 2D inpainted image.**

B.5. 2D inpainting first and then 3D reconstruction

We acknowledge that the inpainting-to-reconstruction approach could be a meaningful baseline method. To demonstrate this, we apply a diffusion-based 2D inpainting model [5] to the image of a partial body and reconstruct a complete 3D human with ICON and PIFuHD. Tab. 4 shows the quantitative evaluation results on the sub-sampled Thuman2.0 dataset, and Fig. 4 shows a sample of in-the-wild 2D inpainting which often struggles to complete realistic human structure for the bigger holes. The inpainting artifacts such as distortion are directly propagated to the 3D reconstruction results. Meanwhile, our 3D generative pipeline enables globally coherent and plausible 3D reconstruction.

B.6. Evaluation on Missing vs. Observed

For more concrete analysis, we break down of the errors based on pixels that are missing vs observed. We break down the graph in Fig. 8 of the main paper into missing and observed pixels as described in Fig. 5. Based on the observed pixels in Fig. 5-(left), our method shows comparable performance to ICON. However, this performance gap is magnified for the missing pixels in Fig. 5-(right) when the impact of occlusion is significant.

C. Applications

C.1. Reconstruction From a Group-shot Image

Obtaining full-body images of all individuals in a group-shot image can be challenging, as some individuals may be naturally obscured by others. In such cases, our model offers a solution for reconstructing multiple individuals from the

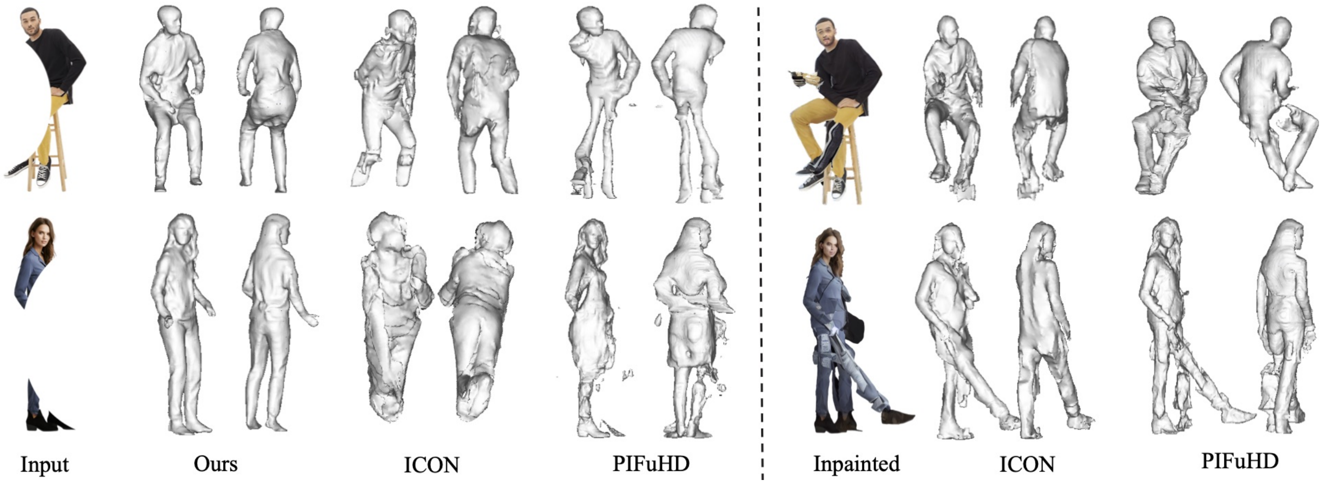

Figure 4. **3D reconstruction from incomplete images and inpainted 2D images.** We present comparative results with other baseline models. For a given incomplete image (First colume), our model robustly completes the full human body with high-fidelity local details, while others struggle to complete the invisible parts of the body. We demonstrate 2D inpainting-to-3D reconstruction (second column), which ensures completeness of the human body. However, the errors caused the 2D inpainting can directly propagate into the reconstructed model.

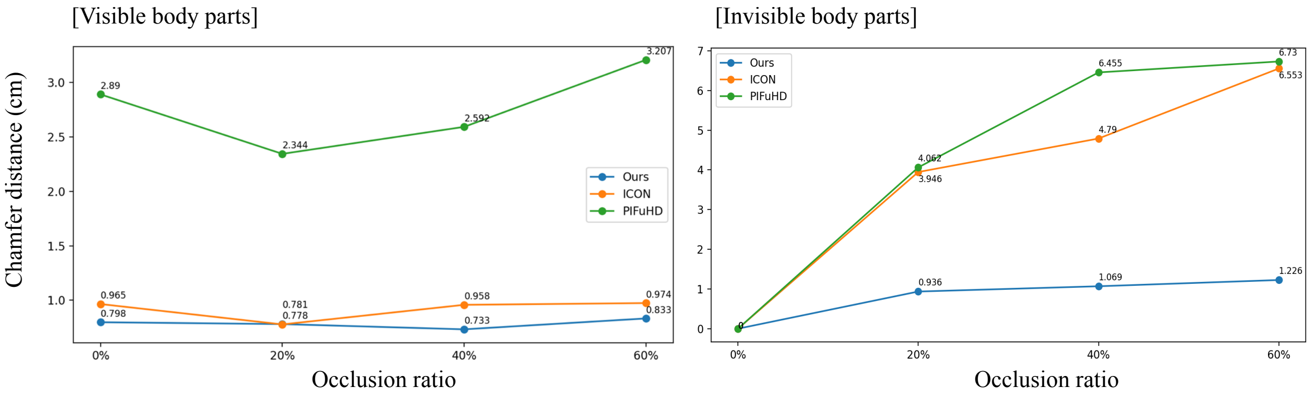

Figure 5. **Occlusion-to-accuracy graphs.** We show the observed (left) and missing (right) body parts measured by Chamfer distance.

group photo. As illustrated in Fig. 6, our model can reconstruct individuals even from occluded regions of the image, without requiring multi-view or camera calibration inputs that are needed for other multi-view reconstruction methods. We can apply our model to reconstruct each person in the same image, one by one. Our qualitative visualization results demonstrate that our model is capable of achieving feasible reconstructions from various challenging viewports, such as individuals facing different directions or being partially occluded by others. These findings underscore the potential of our model as a practical tool for reconstructing individuals from a single-view group-shot image.

C.2. In-the-wild Reconstruction

For in-the-wild testing, we use photos from Deepfashion [1] dataset, where we obtain the 3D body model by applying existing fitting method [10, 11]. We first estimate the SMPL pose from the input partial body image and render the estimated SMPL pose on the partial body image, and then we can remove the background to get the whole body mask. Since we apply a weak perspective projection, for the single-view camera matrix, we can directly crop and center the image without camera calibration. Shown in Fig. 4 are samples of our in-the-wild reconstruction results. Despite training our model on a limited number of objects, we demonstrate promising in-the-wild reconstruction results compared to other baselines.

References

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 4

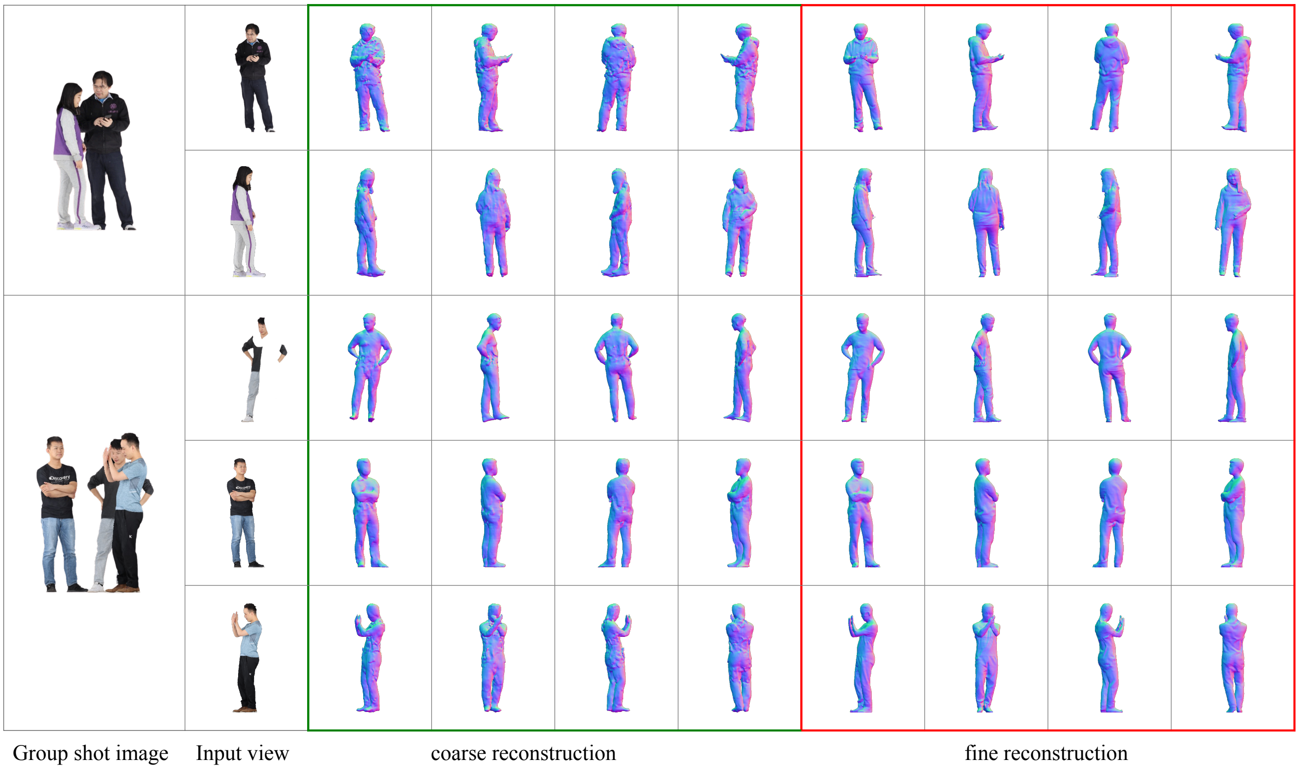

Figure 6. **Reconstruction from a group-shot image** Our model can be applied to reconstruct from the group image. The first column shows a single view of the group image; the second column shows the different persons in the group image. Then, we show our coarse reconstruction and fine reconstruction from four views: front, right, left and back. The group-shot image is rendered from MultiHuman [12] dataset

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

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[4] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3, 6

[7] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 2, 3, 6

[8] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13296–13306, 2022. 2, 3, 6

[9] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 2

[10] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 2, 4

[11] 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*, pages 11446–11456, 2021. 2, 4

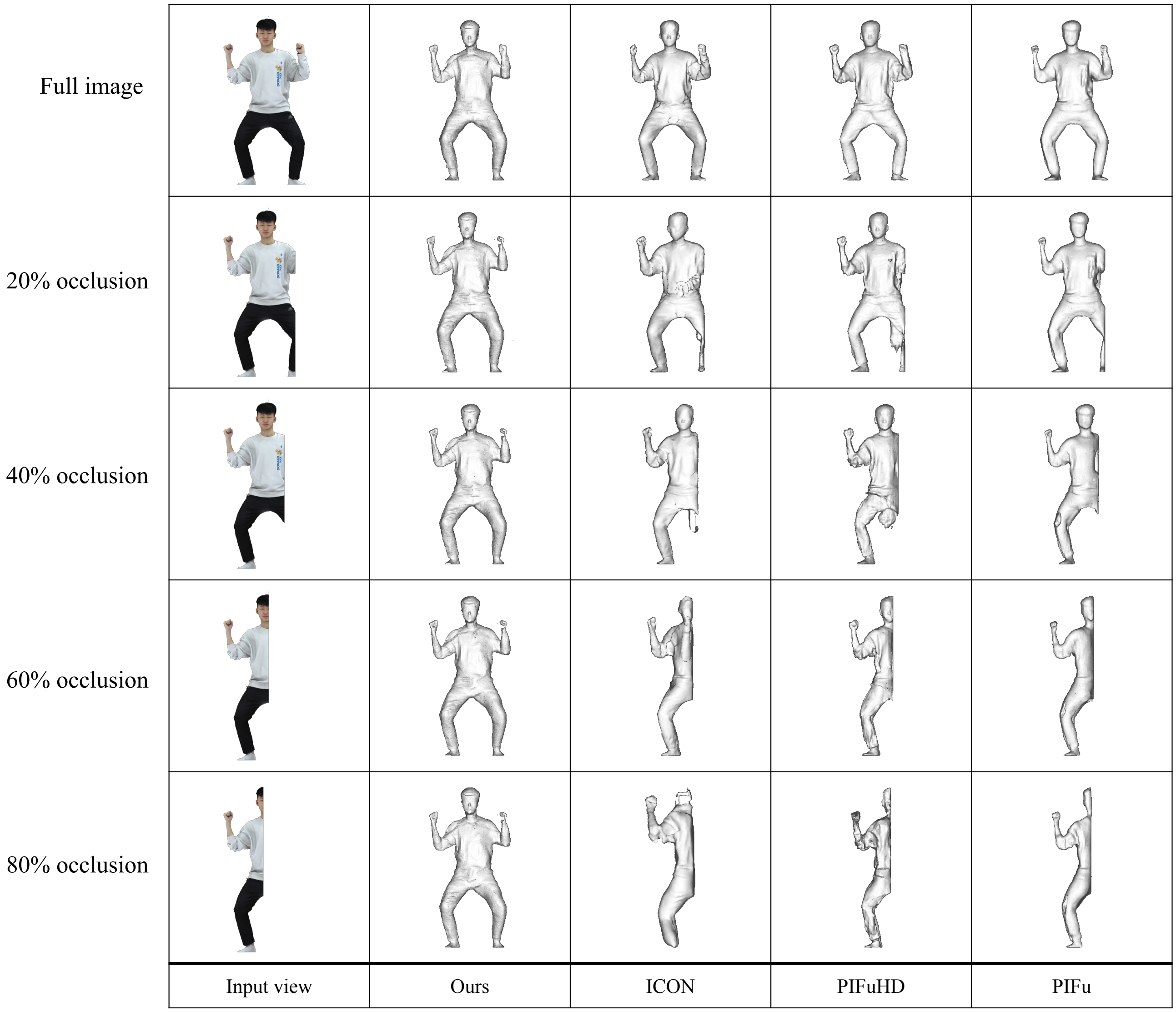

Figure 7. **Accumulative occlusion-to-reconstruction** This figure shows the results of the reconstruction of our model compared to the other baselines [6–8]. In our experiments, we gradually changed the occlusion area of the image from 0% to 80%. The visualization results indicate that our model has better performance even if the given image is large-area occluded. When the occlusion area increases, other models cannot handle the geometry completion from an incomplete image.

[12] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV 2021)*, 2021. 2, 5