3D Reconstruction of Interacting Multi-Person in Clothing from a Single Image -Supplementary-

In this supplementary material, we provide the details of the implementation and network architectures used in our pipeline; ablations study on 2D contact networks and 3D reconstruction results; and more results with in-the-wild images. Please also refer to the supplementary video.

1. Implementation Details

We design f^z with ResNet encoder [4] and one fully-connected layer. f^{3D} consists of 3 sequences of fully-connected layers, 3D convolution layers and last one fully-connected layer. The 2D image-aligned feature extractor f^{2D} consists of a stacked hourglass [10] with 1 stack. Aggregation network f^{agg} is designed with 5 fully-connected layers. Please see Supple. for the detailed architectures of our networks.

For training the networks except for f^{3D}, we use Adam optimizer [5] with learning rate $1e^{-3}$. To train f^{2D} and f^{agg}, we sample the 2,250 points with 2,000 near the ground-truth mesh surface and 250 uniformly sampled. We train f^z for 100 epochs, and f^{2D} and f^{agg} for 300 epochs. λ_c, λ_p, λ_r, and λ_g are set to 500, 10, 100, and 0.001, respectively.

2. Network Detail

2.1. Generation

Latent Feature Encoder f^z predicts latent feature vector z_{shape} from a single RGB image. It consists of ResNet encoder [4] and one fully-connected layer. The ResNet encoder outputs 2048-dim intermediate feature from image $\mathbf{I}_i \in \mathbb{R}^{256\times256\times3}$. Then, the fully-connected layer extracts 64-dim latent feature vectors from the intermediate feature.

3D Voxel Feature Generator $f^{\mathbf{3D}}$ generates 3D voxel feature $\mathbf{F}^{3D} \in \mathbb{R}^{64\times16\times64\times64}$ from estimated latent feature vector z_{shape}. We borrow its architecture from gDNA [2]. It is composed of three sequences consisting of fully-connected layers and 3D convolution layers, and one fully-connected layer at the end.

2D Image-aligned Feature Extractor $f^{\mathbf{2D}}$ extracts 2D image-aligned feature $\mathbf{F}^{2D} \in \mathbb{R}^{16\times512\times512}$ from normal map $\mathbf{N}_i \in \mathbb{R}^{1024\times1024\times3}$. We use a stacked hourglass [10] with 1 stack for its architecture.

Aggregation Network $f^{\mathbf{agg}}$ estimates an occupancy value at 3D point $\mathbf{p}$, from a concatenated feature [$\mathbf{F}^{3D}_{\mathbf{p}_c}$, $\mathbf{F}^{2D}_{\pi(\mathbf{p})}$, $\mathbf{p}_c$]. It consists of 5 fully-connected layers with SoftPlus activation function [15]. In the fourth layer, we use a skip connection. Its intermediate channel is 256.

Input type	IoU	
	Segm.	Signature
RGB	0.44	0.12
RGB + $\mathbf{h}$	0.66	0.15
RGB + $\mathbf{m}^{\text{part}}$	0.73	0.20
RGB + $\mathbf{h}$ + $\mathbf{m}^{\text{part}}$	**0.78**	**0.31**
RGB + $\mathbf{h}$ + part segm. [6]	0.75	0.22

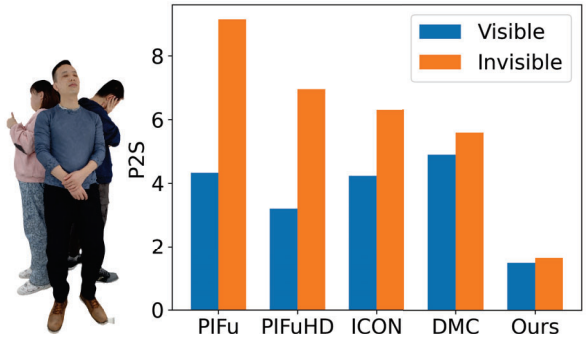
Table 1. Comparison the contact estimation performance on MultiHuman dataset [16] using different input types. We compare five input types: (1) RGB image only, (2) RGB image and keypoint heatmap $\mathbf{h}$, (3) RGB image and semantic part segmentation mask $\mathbf{m}^{\text{part}}$, (4) RGB image, $\mathbf{h}$, and $\mathbf{m}^{\text{part}}$, and (5) RGB image, $\mathbf{h}$, and part segmentation obtained from [6].

2.2. Contact Estimation

Contact Discriminator $f^{\mathbf{Contact}}$ predicts whether two people are contacted to each other or not, from a concatenated image [$\mathbf{I}$, $\mathbf{m}^{\text{part}}_1$, $\mathbf{m}^{\text{part}}_2$, $\mathbf{h}_1$, $\mathbf{h}_2$]. It consists of ResNet encoder [4] and two fully-connected layers with ReLU activation function [9]. We modify the input channel of the first 2D convolution layer in a way that it can use the concatenated images as input.

Contact Segmentation Estimator $f^{\mathbf{CS}}$ estimates contact segmentation $\mathbf{s} \in \mathbb{R}^{75\times1}$ from a concatenated image [$\mathbf{I}$, $\mathbf{m}^{\text{part}}_1$, $\mathbf{m}^{\text{part}}_2$, $\mathbf{h}_1$, $\mathbf{h}_2$]. It consists of ResNet encoder [4] and five fully-connected layers with ReLU activation function [9]. We modify the input channel of the first 2D convolution layer to make it able to use the concatenated images as input.

Contact Signature Estimator $f^{\mathbf{sig}}$ outputs $\mathbf{F}^{\text{sig}} \in \mathbb{R}^{75\times10}$ to estimate contact signature $\mathbf{C} \in \mathbb{R}^{75\times75}$. It consists of two fully-connected layers with ReLU activation function [9]. We compute contact signature $\mathbf{C}$ by $\mathbf{F}^{\text{sig}}_1 \times \mathbf{F}^{\text{sig}^T}_2$.



(a) Input (b) Result

Figure 1. Comparison on the difference between separate performance measuring on the visible and invisible body parts. (a) shows RGB input image. We compare our method with PIFu [11], PIFuHD [12], ICON [13], and DMC [16]. "Visible" denotes the performance measuring on the visible body parts. "Invisible" denotes the performance measuring on the occluded body parts.

		Method	
		Baseline	Ours
Input	Single	1.41	**1.37**
	Occluded single	1.66	**1.63**
	Two natural-inter	1.22	**1.17**
	Two closely-inter	1.91	**1.33**
	Three	1.82	**1.54**

Table 2. Comparison of our method with the baseline that predicts z_{shape} on MultiHuman dataset [16], based on Chamfer Distance metric.

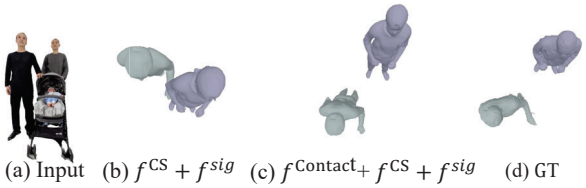
3. Ablation Study

3.1. Comparison on 2D Contact Network

We investigate the performance of the contact estimation using different input types. We use intersection over union (IoU) as our evaluation metric. Tab. 1 shows that using additional input types, such as keypoint heatmap $\mathbf{h}$ and semantic part segmentation mask $\mathbf{m}^{\text{part}}$, improves the performance of contact estimation compared to using only RGB images. Specifically, adding $\mathbf{h}$ to the RGB image improves the performance by 50% on contact segmentation and 25% on contact signature, while adding $\mathbf{m}^{\text{part}}$ improves the performance by 65% on contact segmentation and 67% on contact signature. Finally, using all three input types results in the best performance, with an improvement of 77% on contact segmentation and 158% on contact signature. We also compare with a previous approach [3] which uses 2D part segmentation. We use the method proposed by Lin et al. [6] for 2D part segmentation estimation. It improves the performance compared to using RGB and $\mathbf{h}$, but it does not outperform the performance of the method which uses our final input type (RGB+$\mathbf{h}$+$\mathbf{m}^{\text{part}}$).

Method	Scenario	initial	refined
DMC	Two natural-inter	2.53	2.50
Ours		1.19	**1.17**
DMC	Two close-inter	3.24	3.10
Ours		1.84	**1.34**
DMC	Three people	3.81	3.76
Ours		1.76	**1.54**

Table 3. Comparison with DMC on MultiHuman dataset [16], based on Chamfer Distance metric. 'initial' and 'refined' denote initial SMPL from [1] and SMPL refined by our refinement module.



(a) Input (b) $f^{\text{CS}} + f^{sig}$ (c) $f^{\text{Contact}} + f^{\text{CS}} + f^{sig}$ (d) GT

Figure 2. The 3D reconstruction results without and with f^{Contact}.

3.2. Comparison on Baseline Predicting z_{shape}

We compare our full model to the baseline, f^z, that predicts z_{shape}. Predicting only fz_{shape} cannot outperform the performance of original gDNA [2]. With Chamfer Distance error in Tab. 2, our proposed networks, such as 2D image-aligned feature extractor and aggregation network, as well as the refinement module, enhance the level of detail and global coherence of the human mesh.

3.3. Effectiveness of Refinement Module

Based on the comparison with DMC in Tab. 3, we highlight that our refinement module is compatible with any off-the-shelf mesh reconstruction method to improve its accuracy. Performance improvement through the refinement module is observed for all methods and scenarios.

3.4. Performance on the Occluded and Visible Body Parts

Fig. 1 shows the difference between separate performance measuring on the visible and invisible body parts for a representative example. We use point-to-surface (P2S) as our evaluation metric. "Visible" denotes the P2S measuring on the visible body parts. "Invisible" denotes the P2S measuring on the invisible body parts occluded by other people. While the state-of-the-art methods exhibit a significant performance gap between the visible and invisible parts, our method demonstrates a small difference in performance between the two areas. Furthermore, our method outperforms all the compared approaches in terms of overall performance.

4. More Qualitative Results

We visualize our contact prediction results on the front and back of the mesh as shown in Fig. 5. The columns show the contact estimator input, bounding boxes colored according to the contact prediction results, and contact prediction results on the mesh, respectively. (Note that bounding boxes are not estimated results, and just make it easier to find individuals in the input image who correspond to the contact result.) The contact estimator predict the contact region such as hand, head, back, arm, etc. Sometimes, the contact prediction results are not accurate due to depth ambiguity, as shown in the fourth row. The reconstruction results for each image in Fig. 5 are presented in Figs. 6 and 7.

We compare our method with PIFu [11] and DMC [16] on in-the-wild images and MultiHuman image [16] in Fig. 6 and Fig. 7. We use COCO dataset [7] to compare the qualitative results on in-the-wild images. In Fig. 6, the first to third rows of input images are in-the-wild images, while the fourth row shows result on a MultiHuman image. The first and fourth rows show the front-view and back-view results side by side. The second and third rows show the front-view and back-view results vertically. In Fig. 7, all results are presented in a vertical format, displaying both front-view and top-view results for each row on in-the-wild images. PIFu does not consider the depth location, and therefore, it can not represent the perspective distance(i.e. large for near and small for far). In addition, it can not reconstruct the 3D geometry for the invisible parts. DMC generates a relatively complete human mesh when compared to PIFu. Nevertheless, the meshes reconstructed by DMC are coarse and lack some parts. Conversely, our method reconstructs the complete and detailed human mesh, even in the presence of occlusions.

5. Inference speed

For comparison, PIFu [11] takes 5 seconds, and DMC [16] takes 41 seconds to generate a clothed human mesh. In contrast, our model, excluding the contact-based refinement, requires 6 seconds to create a clothed human mesh using the Marching Cubes algorithm [8] from an occupancy field. Additionally, our refinement module takes 58 seconds to refine two meshes based on contacts. These inferences were performed on an RTX Titan GPU.

6. Training data

We train our network using THuman2.0 [14] dataset. For the training images, we augment the images by masking them with the segmentation masks of humans to simulate occlusion, thereby enabling the model to learn how to handle partially occluded humans. This is illustrated in Fig. 3. Furthermore, we use color-jitter for data augmentation.

Figure 3. Training data. Left shows an original image and right shows a masked input image.

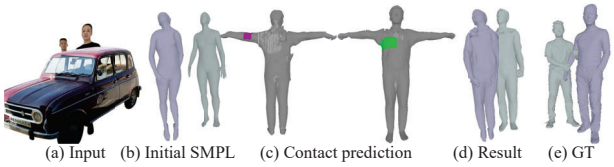


Figure 4. Failure case: a strong occlusion leads to the failure of contact prediction (left arm and right chest) and inaccurate 3D reconstruction results. Left to right: input, initial SMPL, contact prediction result, reconstruction result, and ground-truth.

7. Failure Case

In Fig. 4, our method demonstrates weak performance in the presence of strong occlusion which leads to the failure of contact estimation and inaccurate 3D mesh reconstruction.

References

[1] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3d pose and shape estimation via inverse kinematics and refinement. In *ECCV*, 2022.

[2] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *CVPR*, 2022.

[3] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[6] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

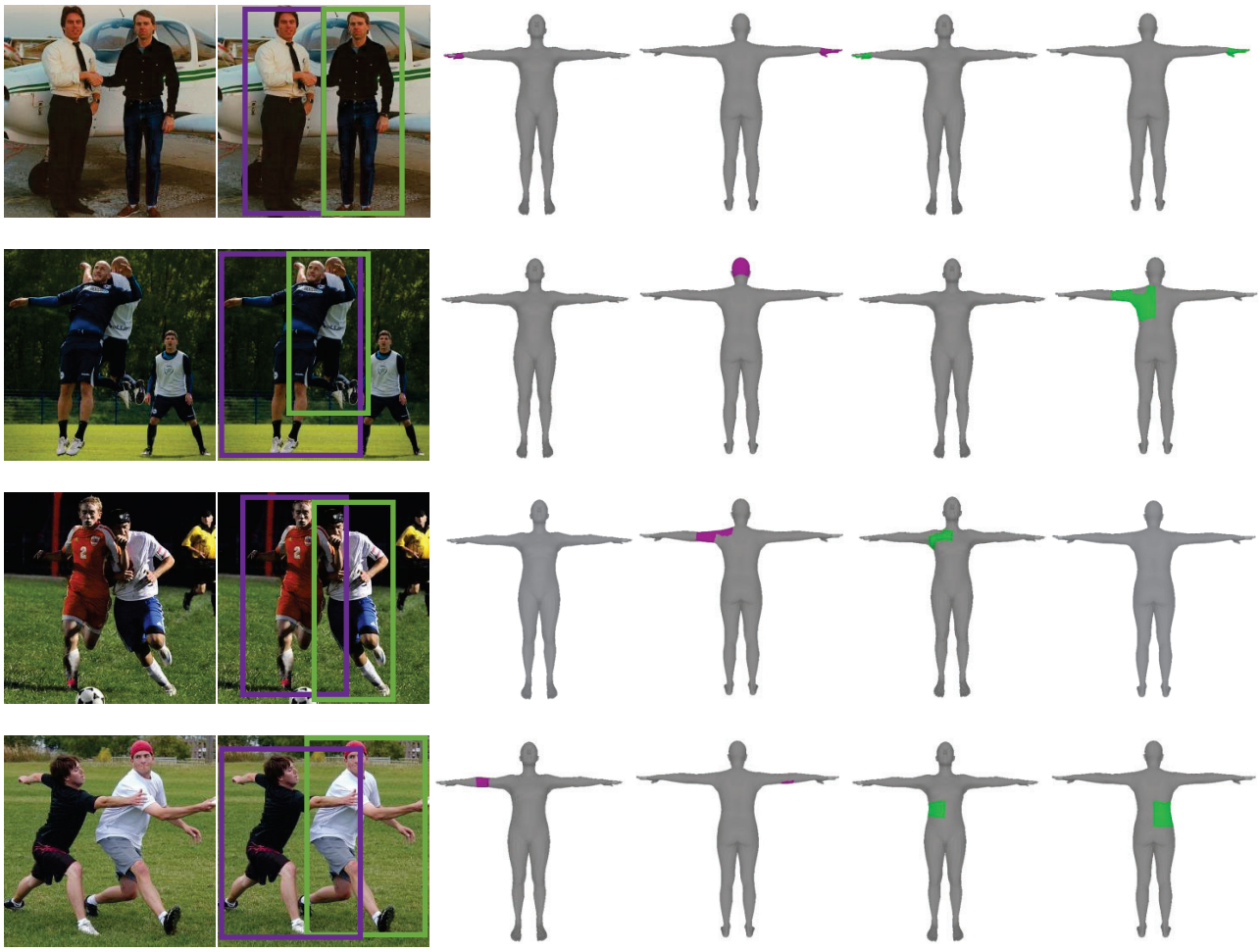


Figure 5. We show the contact prediction results for in-the-wild images from our contact estimator. (a) Input for contact estimator, (b) the image is marked with boxes of the same color corresponding to the contact prediction results, and (c) the contact prediction results on the front and back of the mesh.

Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[8] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 1987.

[9] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[11] 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 *ICCV*, 2019.

[12] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

[13] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *CVPR*, 2022.

[14] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021.

[15] Huizhen Zhao, Fuxian Liu, Longyue Li, and Chang Luo. A novel softplus linear unit for deep convolutional neural networks. *Applied Intelligence*, 2018.

[16] 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 *ICCV*, 2021.

Figure 6. Results comparison on in-the-wild images and MultiHuman image [16]. We compare our method with PIFu [11] and DMC [16]. We visualize the results in front-view and back-view.

Figure 7. Results comparison on in-the-wild images. We compare our method with PIFu [11] and DMC [16]. We visualize the results in front-view and top-view.