SeSDF: Self-evolved Signed Distance Field for Implicit 3D Human Reconstruction
– *Supplementary Material* –

Yukang Cao Kai Han Kwan-Yee K. Wong
The University of Hong Kong
{ykcao, kykwong}@cs.hku.hk kaihanx@hku.hk

Contents

1. Implementation details

1.1. SMPL-X optimization

Given a single-view RGB image or uncalibrated multi-view RGB images as input, we first apply PIXIE [2] to fit the SMPL-X model of the human subject after segmenting out the background with rembg[1]. For the multi-view case, we first fit the SMPL-X models for each input image and take the mean of the fitted shape, pose, and expression parameters $\bar{\theta}, \bar{\beta}, \bar{\phi}$ as the shared model for all views, while each model keeps its own global orientation parameters $R_i^{'}, T_i^{'}$. However, the resulting SMPL-X model is often misaligned with the image, degrading the quality of the final reconstruction. To remedy this problem, as described in Sec. 3.1, we refine the fitted SMPL-X model by maximizing the IoU between the projected SMPL-X silhouette and human body mask [1] and minimizing the 2D keypoint distances, to achieve optimized SMPL-X parameters, *i.e.*, $\theta, \beta, \phi, R_i, T_i$ (see Fig. S1).

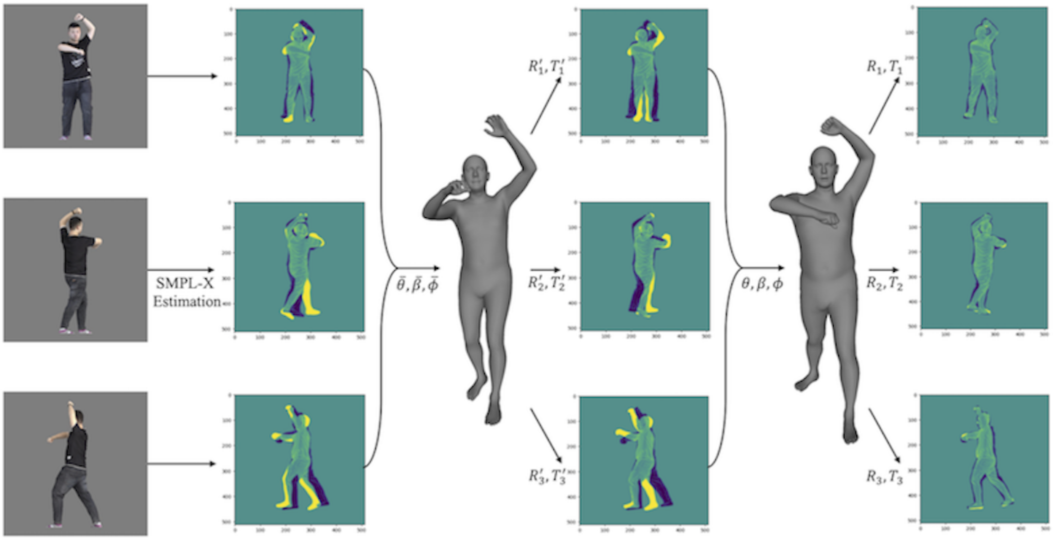

Figure S1. SMPL-X Optimization. By maximizing IoU between the body mask and the projected silhouette, and minimizing the 2D keypoint distances, we achieve optimized SMPL-X which is well-aligned with images.

Failure cases We initialize the SMPL-X model using all the information from multi-view images, which is generally robust. However, there may still be instances of failures when the occlusions are intense and cannot be resolved through multi-view images, particularly when the input viewpoints have short baselines.

1.2. Learnable parameters and inference latency

Our model has around 18M trainable parameters and the inference time is around 20s. As a reference, PIFu has about 15M trainable parameters and requires a similar inference time.

[1] https://github.com/danielgatis/rembg

2. Further analysis

2.1. Effects of SeSDF module

Our SeSDF successfully evolves the signed distance field derived from SMPL model, by leveraging the 2D pixel-aligned feature and 3D space-aligned feature. In addition to the quantitative analysis in Tables 1 and 2 in the main paper, we further provide qualitative comparisons on real-world images with loose clothing in Fig. S2 to analyze the visual improvement provided by our SeSDF module.

It can be observed that our SeSDF module can significantly improve the SDF computed from the 'naked' SMPL-X model, and successfully reconstruct loose clothing to some extent. By comparing 'Ours w/o DE' with 'Ours', we can see that distance encoding can help improve the clothing details. Overall, our full model produces the best results, outperforming the other methods on real-world images with easier poses.

'Ours w/o SeSDF' vs ICON 'Ours w/o SeSDF' still differ from ICON in several aspects: (1) we apply Distance Encoding (DE) to SMPL-X based SDF and compute 3D point features from SMPL-X, while ICON does not contain these components; (2) ICON uses a network to predict normals from the 2D image, while our method directly takes SMPL-X vertex normals.

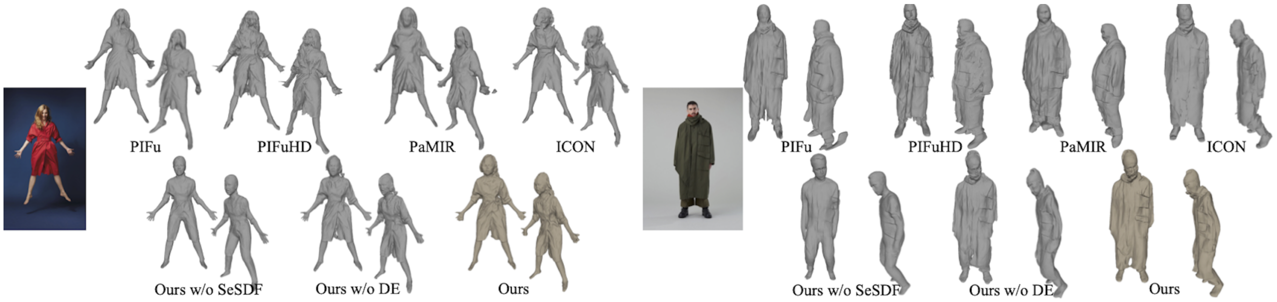

Figure S2. Comparisons on real-world images with loose clothing and easier poses. Best viewed in PDF with zoom.

2.2. Effects of feature design

SeSDF learns to predict the occupancy values for the 3D human reconstruction by Eq. (1) and Eq. (2):

$$f_{sd}(F_{2D}(X), F_{3D}(X), \mathcal{D}(d'(X)), \mathbf{n}'(X)) \mapsto (d, \mathbf{n}), \tag{1}$$

$$f_o(F_{2D}(X), F_{3D}(X), \mathcal{D}(d(X)), \mathbf{n}(X), Z(X)) \mapsto [0, 1]. \tag{2}$$

In our method, we use both 2D pixel-aligned ($F_{2D}(X)$) and 3D space-aligned ($F_{3D}(X)$) features for both self-evolved SDF learning (Eq. (1)) and occupancy prediction (Eq. (2)). In Fig. S3, we show results on three alternatives to further validate our choice of features. Namely, (a) instead of using 3D features from all transformed SMPL-X models by R_i, T_i, we only use the 3D feature from the model in the first view; (b) we exclude 3D features in Eq. (1) and Eq. (2) to demonstrate its effectiveness; (c) we eliminate the use of 2D and 3D features in Eq. (2) to validate the necessity of 2D and 3D features for occupancy prediction. As can be seen, using 3D features from the SMPL-X model under transformed views is effective, and our choice of using both 2D and 3D features for both self-evolved SDF learning and occupancy prediction appears to be optimal.

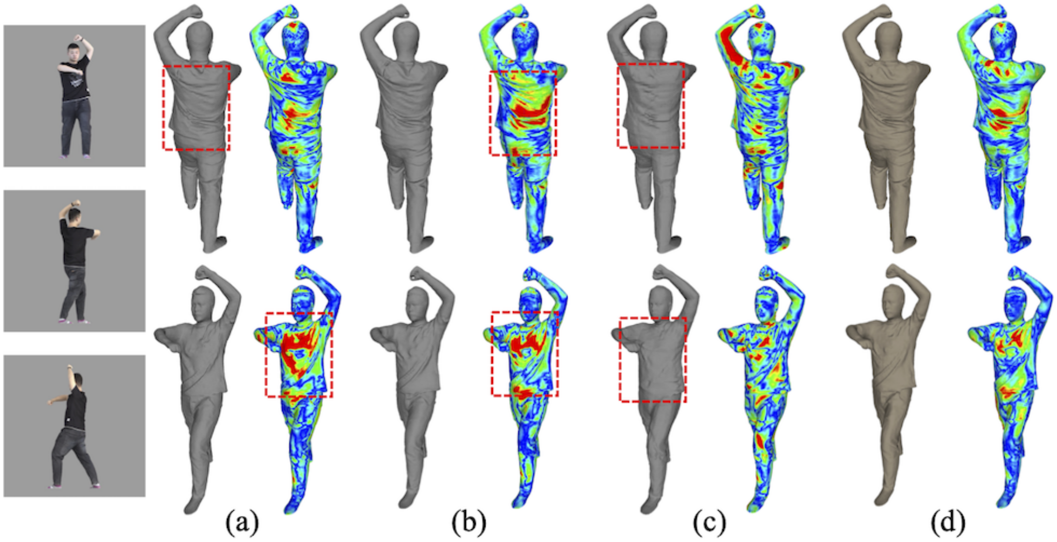

Figure S3. Further analysis of the design of our feature modules. (a) Ours w/ only the 3D feature from the first view; (b) Ours w/o 3D feature; (c) Ours w/o 2D and 3D features in Eq. (2); (d) Ours.

2.3. Effects of feature fusion strategy

In the multi-view reconstruction setting, the feature fusion method plays an important role. In Fig. S4, we provide more comparison of our occlusion-aware feature fusion method with average pooling (“AVG-Pool”) [4], SMPL-visibility (“SMPLX-Vis”) [5] and attention-based mechanism (“Attention”) [6]. In addition, we also provide an alternative to our occlusion-aware feature fusion method by leveraging the surface normals rather than the depth, which can be defined as:

$$N_i(X) = \tanh\left(\arccos\left(\frac{\vec{v_n} \cdot \vec{v_d}}{||\vec{v_n}|| \cdot ||\vec{v_d}||}\right)\right). \tag{3}$$

$$F_{fused}(X) = \frac{\sum_{i=1}^{n} N_i(X) F_i(X)}{\sum_{i=1}^{n} N_i(X)}. \tag{4}$$

We call this normal-based fusion strategy “Normal-Fusion”.

“AVG-Pool” treats different views equally, even the features from occluded views, leading to notable artifacts in the reconstruction (see Fig. 7 in the main paper). Meanwhile, it cannot handle the input images from non-evenly distributed views due to the presence of self-occlusion. either (see Fig. S4). “SMPLX-Vis” discards the features of a 3D point X, if its closest SMPL-X vertex is invisible along the camera view direction. However, this strategy will negatively affect the reconstruction by introducing severe artifacts across the surface (see Fig. 7 in main paper and Fig. S4). “Attention” tries to improve the features by learning the correlations across different views. Unfortunately, it also appears to suffer from the occluded features (see Fig. 7 in main paper), and has difficulty handling input images from non-evenly distributed views (see Fig. S4). “Normal-Fusion” appears to be more effective than others, except our depth-based occlusion-aware fusion method (“Ours”). However, the occluded 3D point X may also have a large $N_i(X)$ coefficient, which is undesired but often occurs in the boundary regions, leading to minor artifacts across the boundary of the 3D human reconstruction (see Fig. S4). We also provide the quantitative evaluation on “Normal-Fusion” in the last row of Table 2 in the main paper. As can be seen, our occlusion-aware fusion method based on depth consistently achieves the best performance.

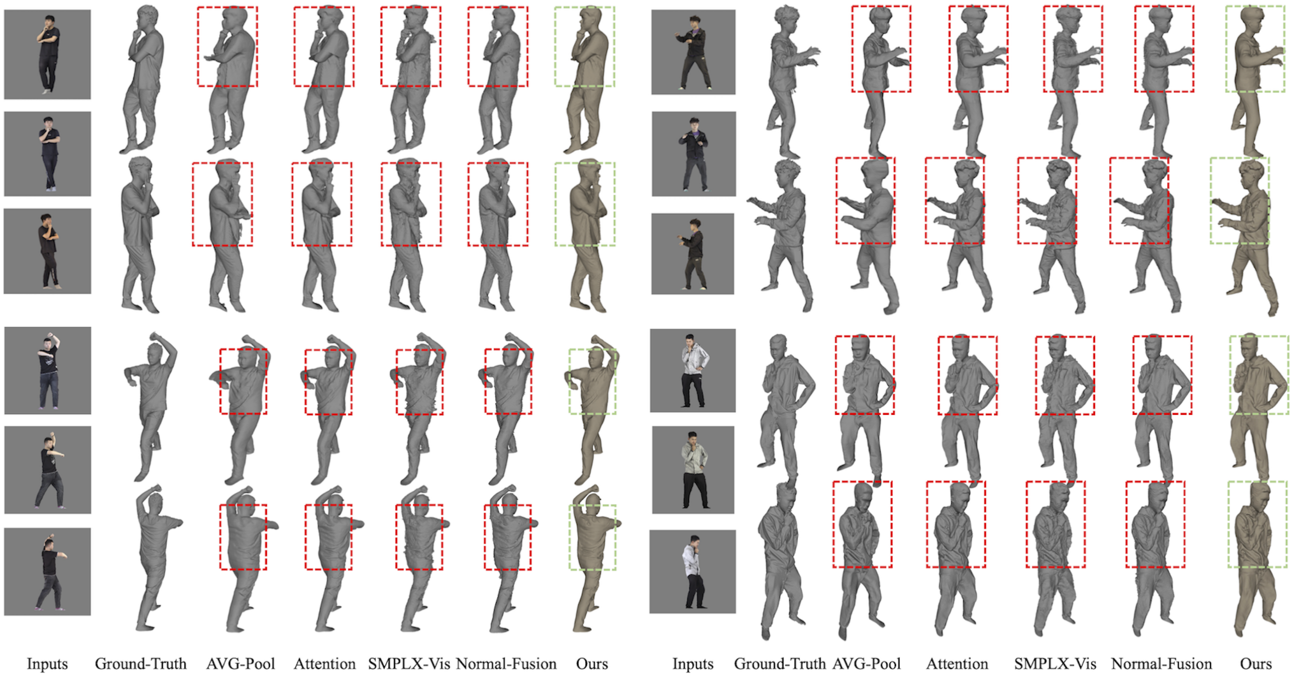

Figure S4. Comparing our occlusion-aware strategy and our normal-fusion strategy with other alternatives.

2.4. Analysis of occupancy prediction

Instead of directly applying marching cube [3] to the evolved signed distance field output by SeSDF module, we reconstruct the final 3D human model from the occupancy field (see Fig. S5). Through experimental results, we find that, without the occupancy prediction, the reconstruction appears to be grainy. The occupancy prediction can greatly help smooth the reconstruction.

2.5. Analysis of Normal

Like ICON [5] which predicts per-pixel normal to aid occupancy prediction, we choose to predict normal for 3D points. The intuition behind this is that directly deriving the normal from the signed distance field will not provide additional information to improve the occupancy prediction.

In Fig. S5, we further provide ablation analysis on the normal information. It demonstrates that the Eikonal loss is beneficial and the normal feature helps predict better occupancy value.

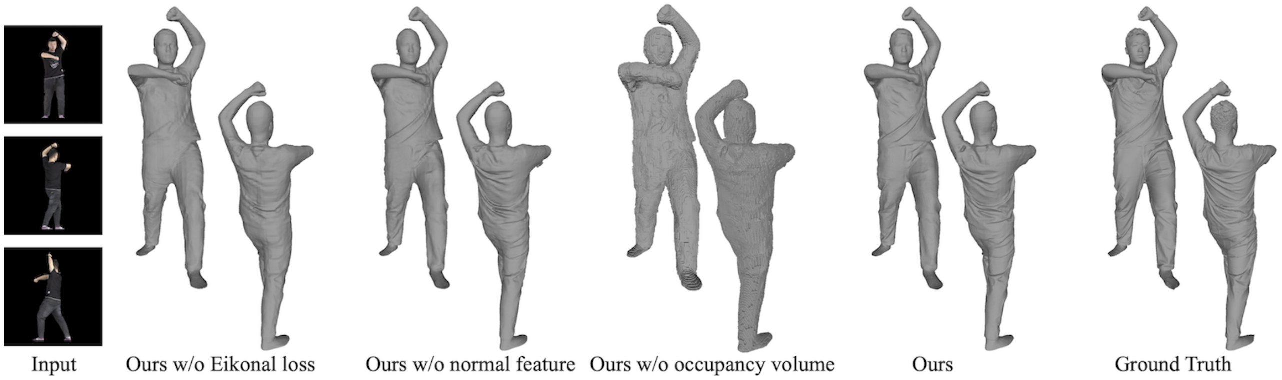

Figure S5. Ablation study for normal and occupancy volume. Best viewed in PDF with zoom.

3. More qualitative results

3.1. Single-view reconstrcution

3.1.1 Reconstruction with real-world images

Fig. S6-Fig. S9 show more qualitative comparisons between our SeSDF and other SOTA methods on real-world images. As can be observed, our method can robustly handle complicated poses, and faithfully reconstruct fine details.

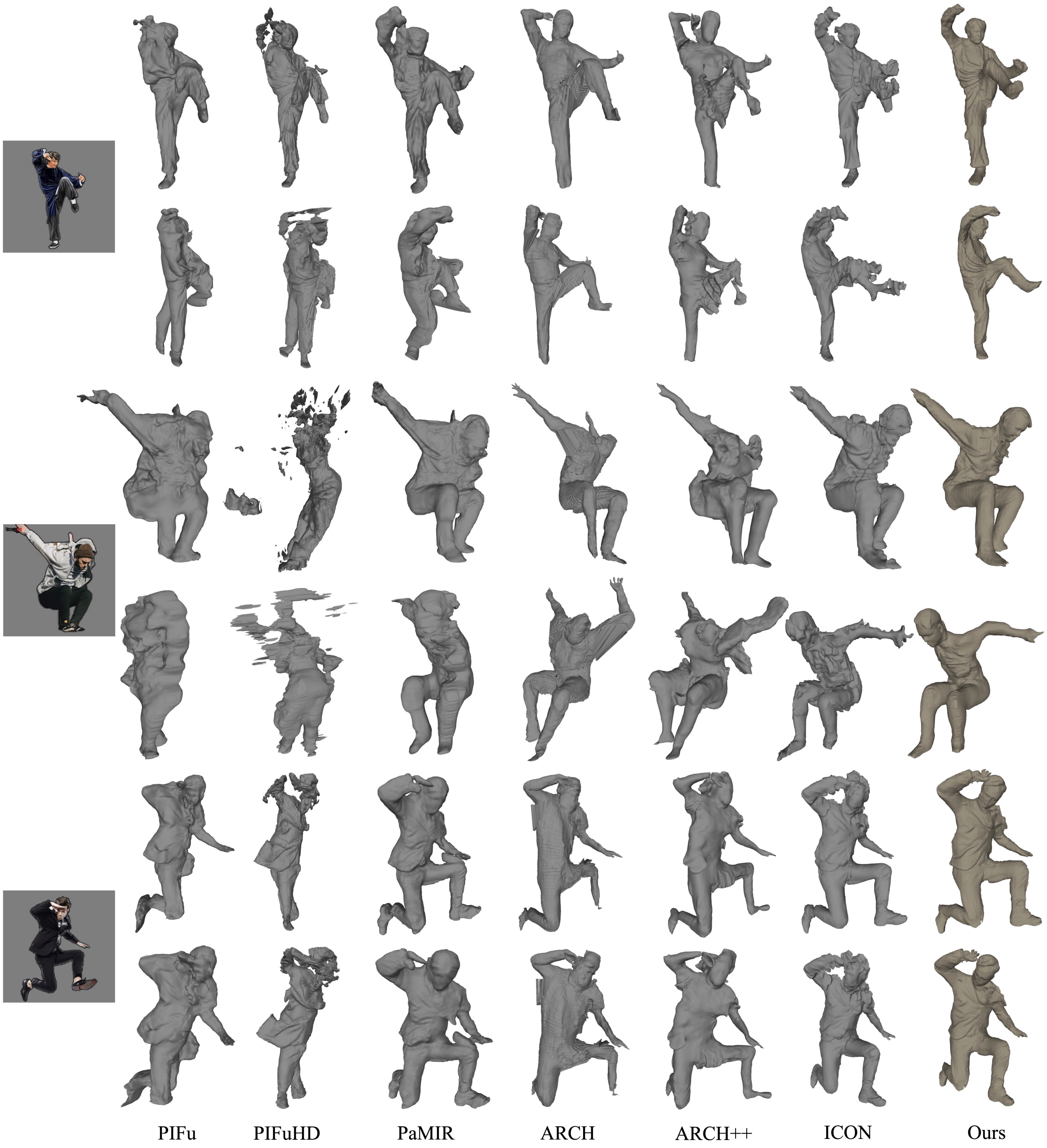

Figure S6. Qualitative comparison with SOTA methods on real-world images.

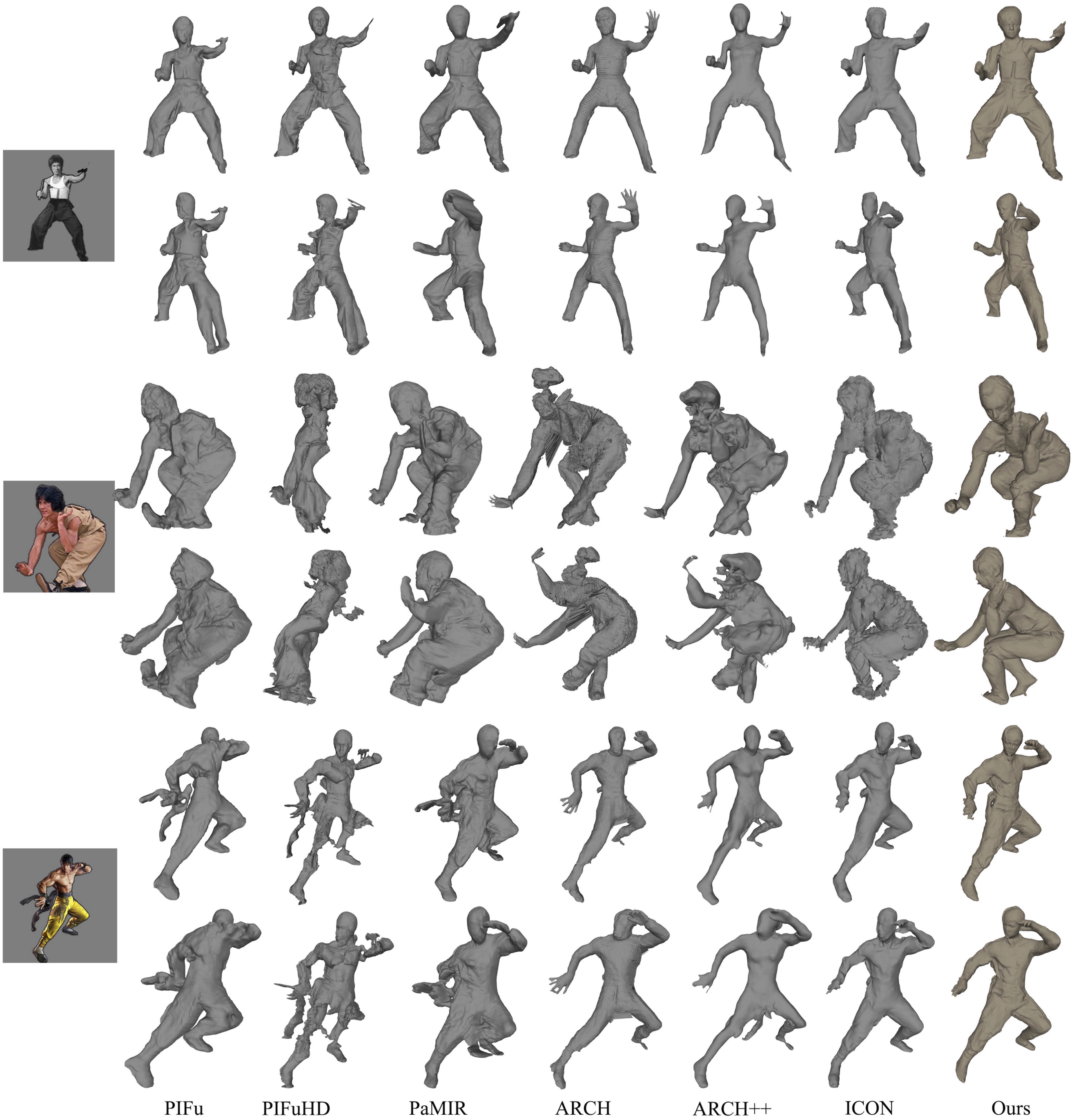

Figure S7. Qualitative comparison with SOTA methods on real-world images.

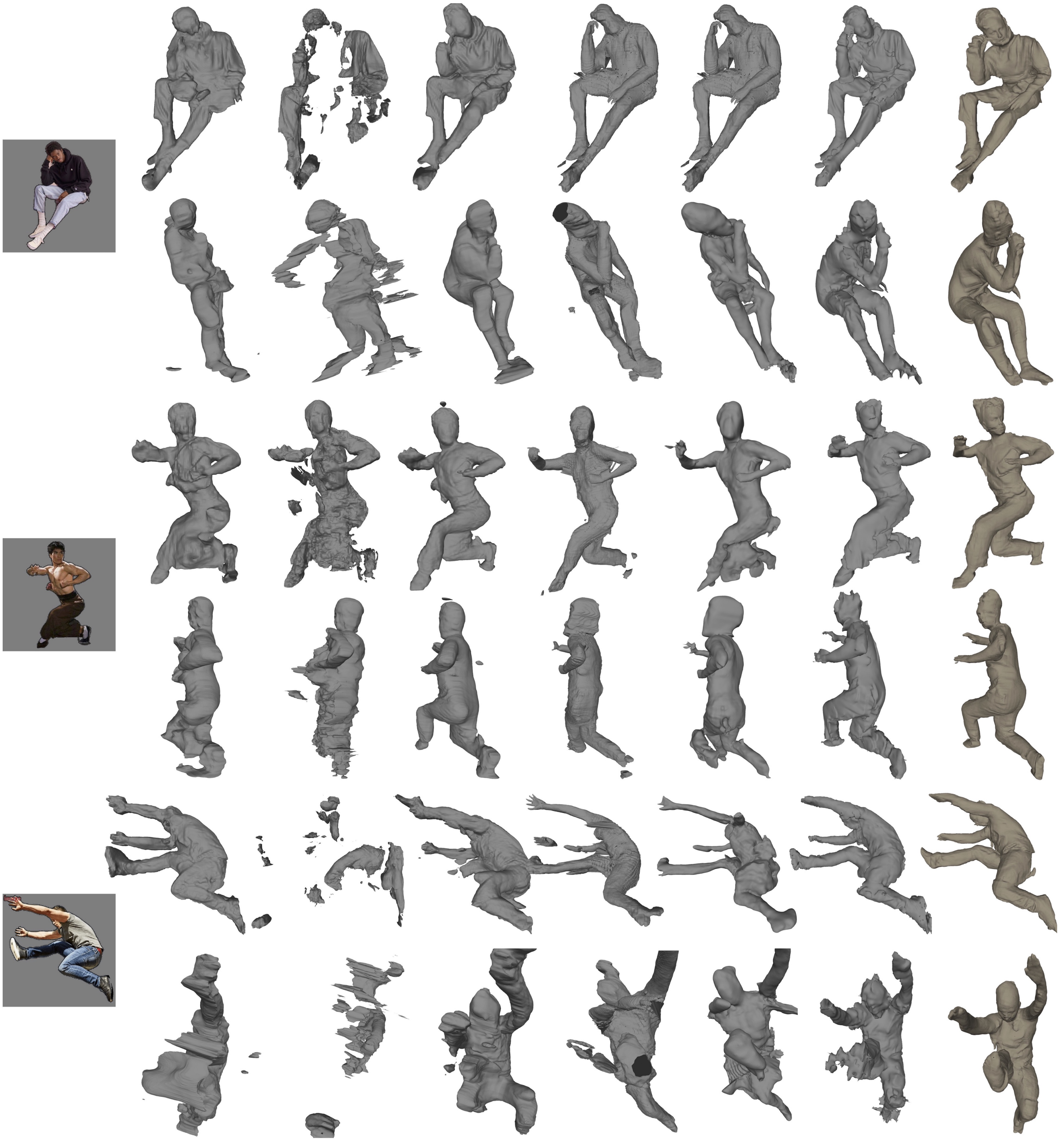

Figure S8. Qualitative comparison with SOTA methods on real-world images.

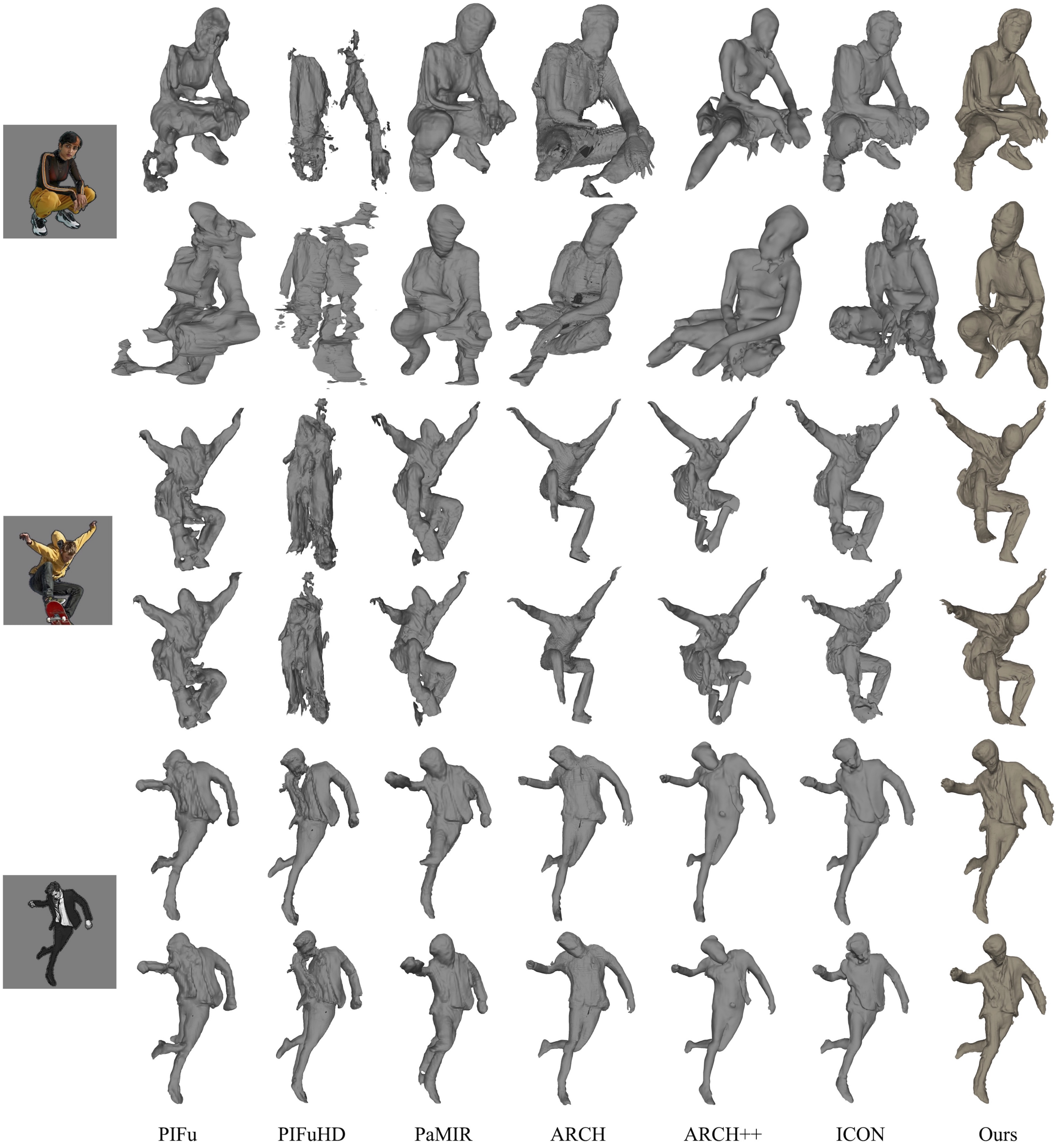

Figure S9. Qualitative comparison with SOTA methods on real-world images.

3.1.2 Reconstruction with THUman2.0 images

Fig. S10 provides more comparisons on THUman2.0 test data. It demonstrates that our method can consistently reconstruct better clothing topology than existing SOTA methods under different scenarios.

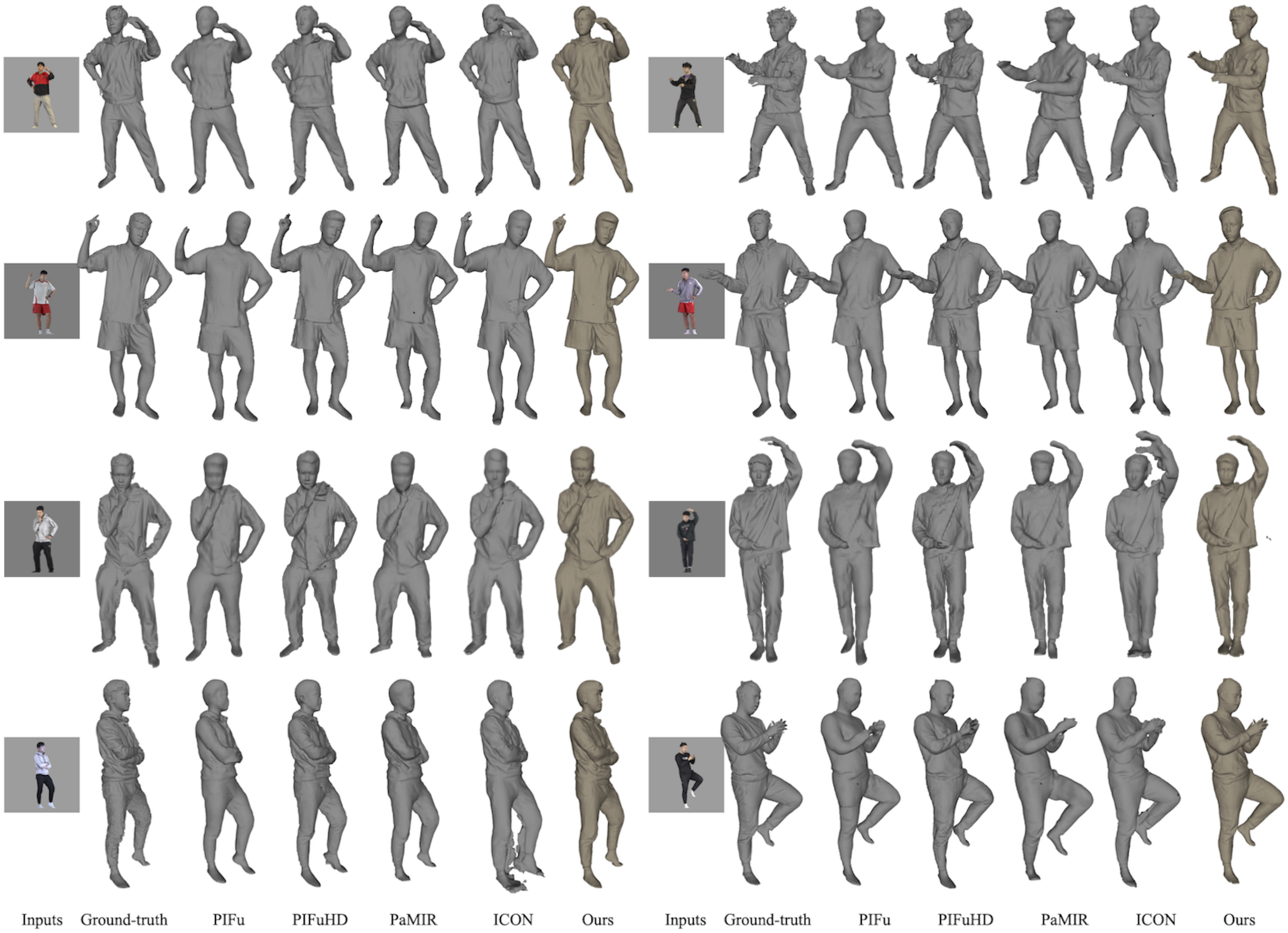

Figure S10. Qualitative comparison with SOTA methods on THUman2.0 test data with controlled poses.

3.2. Improvement from single-view to multi-view reconstruction

In Fig. S11, we show results of the same subject using a single image or uncalibrated multi-view images (we use three images) as the input for SeSDF. As can be seen, more details, especially those that are absent from the single-view input, can be reconstructed faithfully by our SeSDF framework.

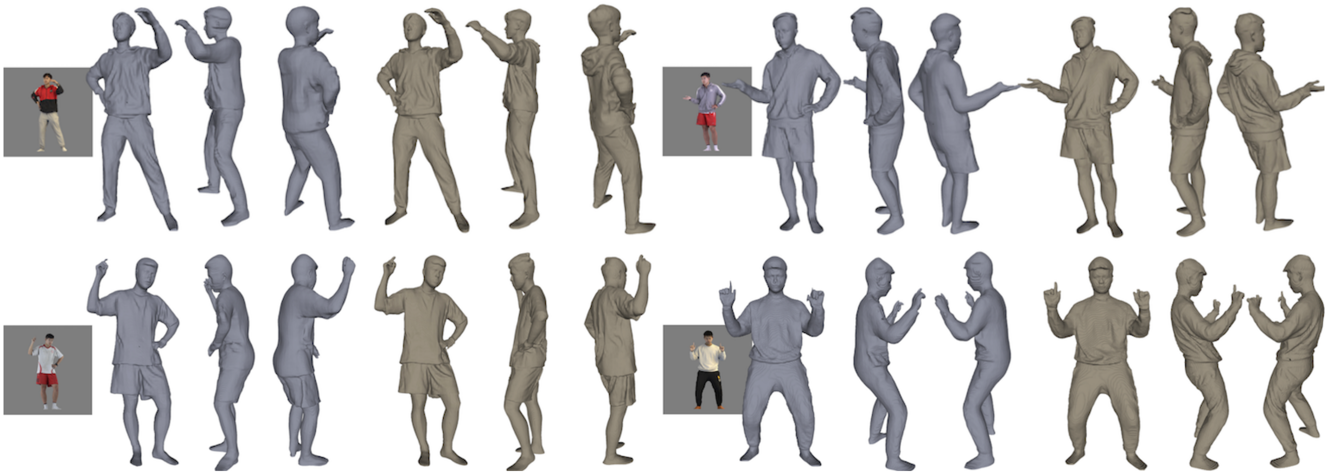

Figure S11. Single-view reconstruction vs multi-view reconstruction.

3.3. Multi-view reconstruction

3.3.1 Reconstruction on THUman2.0

We provide more multi-view reconstruction results in Fig. S12 to demonstrate the performance of our method.

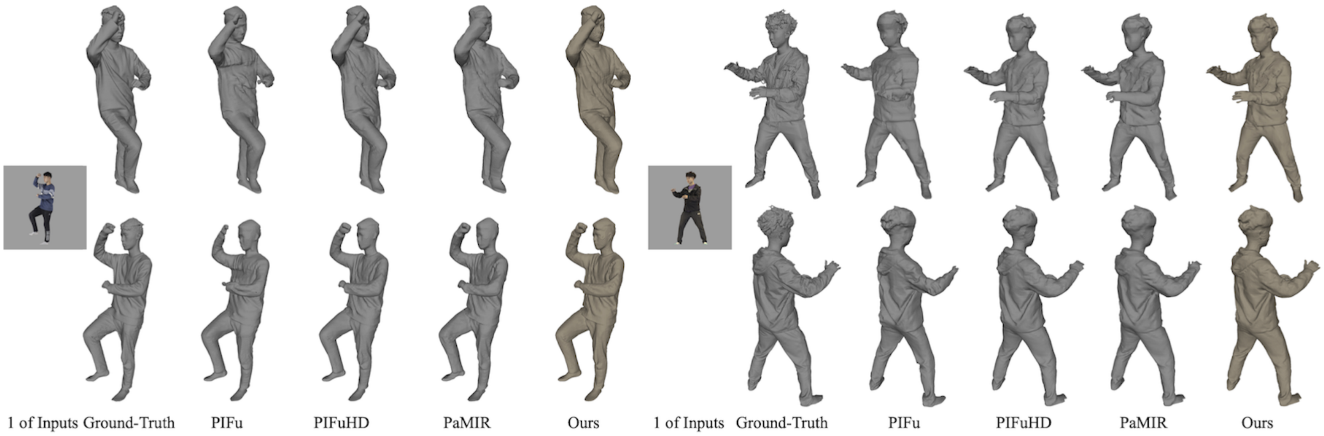

Figure S12. Qualitative comparison with SOTA methods on THUman2.0 test data for multi-view 3D human reconstruction.

3.3.2 Different number of views

Our SeSDF framework can be trained and tested on an arbitrary number of views and the views during the training and testing are not necessarily to be equal. In Fig. S13-Fig. S14, we train our network with 3-views and test with different numbers of views:

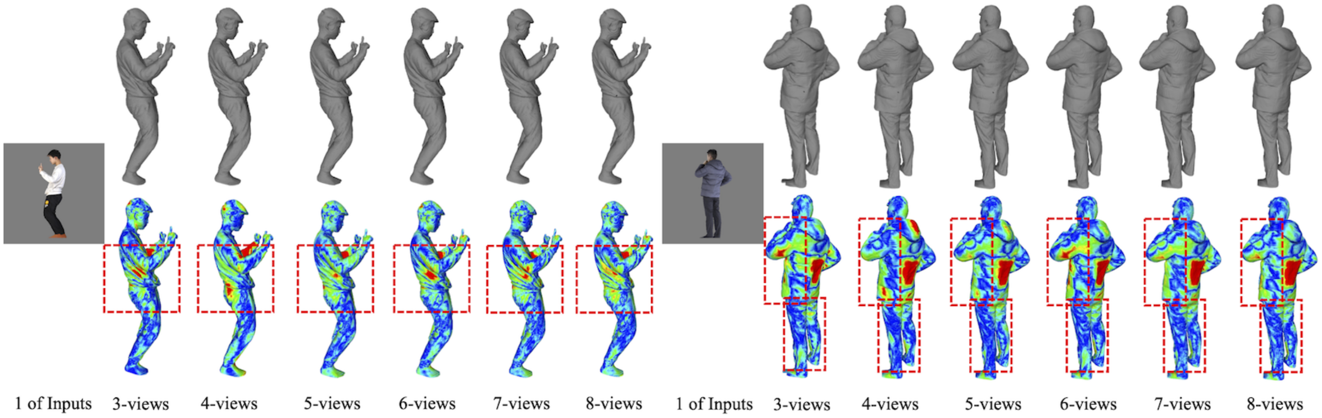

Figure S13. Reconstructions with different numbers of input views.

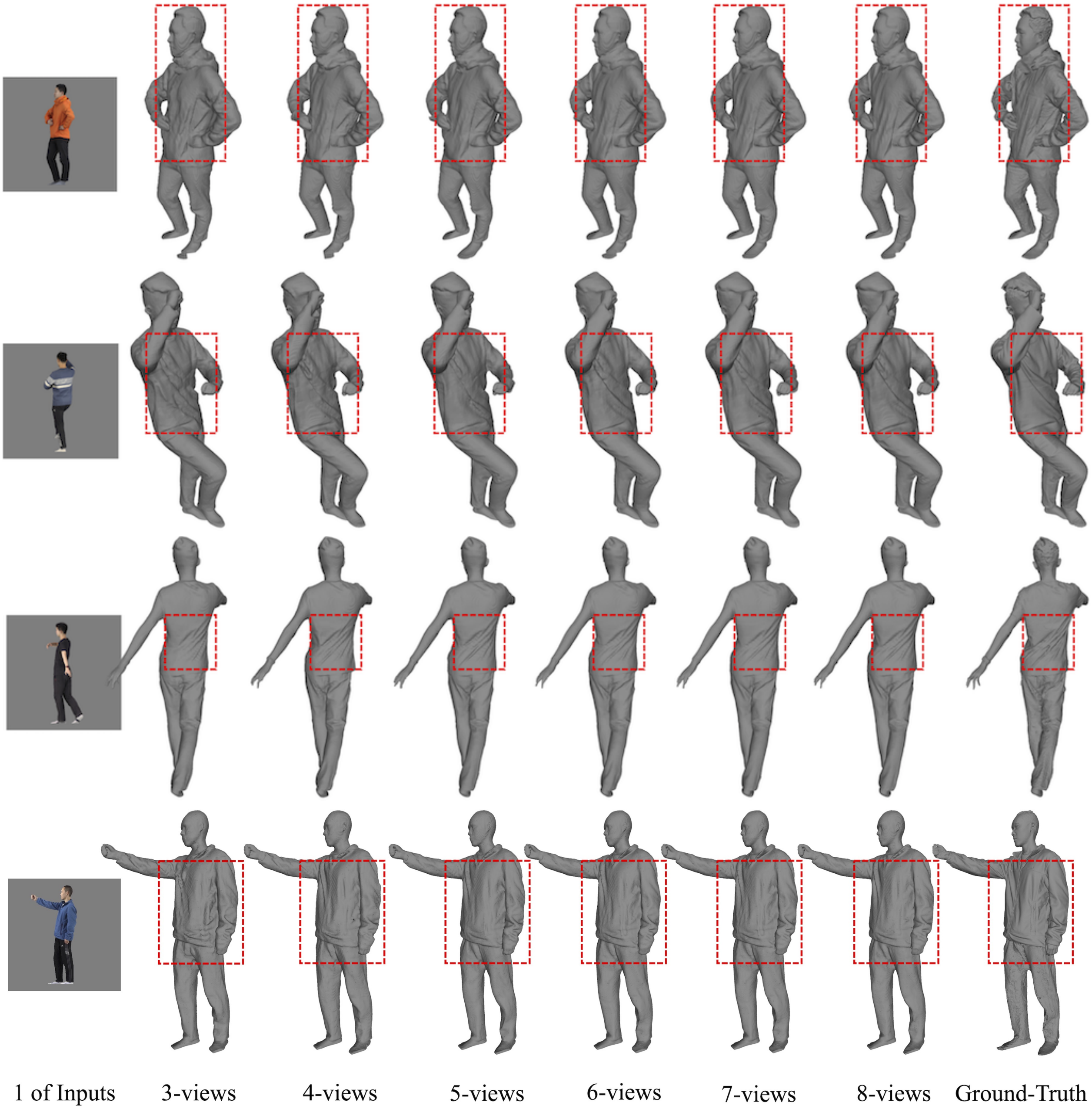

Figure S14. Reconstructions with different numbers of input views.

References

[1] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 2

[2] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision*, 2021. 2

[3] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics*, 2018. 6

[4] 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 *IEEE International Conference on Computer Vision*, 2019. 5

[5] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 5, 6

[6] 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. *arXiv preprint arXiv:2105.00261*, 2021. 5