JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction
– *Supplementary Material* –

Yukang Cao[1] Guanying Chen[2] Kai Han[1] Wenqi Yang[1] Kwan-Yee K. Wong[1]

[1]The University of Hong Kong [2]The Future Network of Intelligence Institute (FNii), CUHK-Shenzhen

Contents

1. More implementation details

1.1. Training data preparation

The preparation of our training dataset consists of three steps. First, we perform PRT computation and rendering processing following PIFu. Second, we detect the face region using MTCNN [7] and the estimate 3DMM with the method in [1], respectively. Last, we align the 3DMM mesh to the ground-truth face surface. In order to get better 3D face features, accurate alignment between the 3D face model and ground-truth face surface mesh is very important, especially when the estimated 3DMM has a normalized scale and locates in a predefined camera coordinate system. Specifically, we compute the ratio between the y-scale of the face and the body in the image, which is used to uniformly rescale the 3DMM mesh. We then adopt ICP algorithm to align the 3DMM mesh with the ground-truth mesh. Note that the paired point sets used for ICP are obtained by back-projecting 2D face landmarks to 3DMM mesh and the ground-truth mesh separately.

1.2. MLP architectures

In our proposed network, we reconstruct the face-region and non-face region with two different MLP networks. For points projected to non-face region, we use an MLP (with channel numbers of [257, 1024, 512, 256, 128, 1]) for occupancy prediction using a 2D pixel-aligned feature. Otherwise, for points located in the face region, we jointly utilize 2D pixel-aligned feature and 3D space-aligned feature with specially designed MLPs. We first process 2D and 3D features with MLPs (with channel numbers of [257, 1024, 512] and [128, 1024, 512], respectively). We then concatenate the transformed features, and feed them into another MLP (with channel numbers of [1024, 512, 256, 128, 1]) for occupancy prediction of the face region. Note that the texture network is the same as the shape network, except that the MLP output channel is changed from 1 to 3 with $tanh$ activation.

1.3. Inference procedures

Given a testing input image and the corresponding mask, we first estimate the 3DMM as 3D face prior after detecting the face region. We then use the original PIFu to get a rough reconstruction, as we need to align 3DMM mesh with ICP to better extract 3D face features. Note that we get the landmarks for ICP using the same back-projection method as illustrated in Section 1.1. Taking the 2D input image and the well-aligned 3DMM mesh as input, our JIFF can predict accurate occupancy value and vertex color.

2. More qualitative comparisons with previous methods

2.1. Comparisons on full body reconstruction

Figures S1-S4 show the qualitative comparisons between our method with existing methods on full body reconstruction. As can be observed, our method can faithfully recover fine face details in terms of both geometry and error color.

Our method performs with impressive fine face geometry details, which is the closet to ground truth with the smallest P2S error compared to the other competing methods. We implement PIFu [4] for body region, so that getting similar performance with it. Note that JIFF is capable of incorporating any other implicit representation for body region.

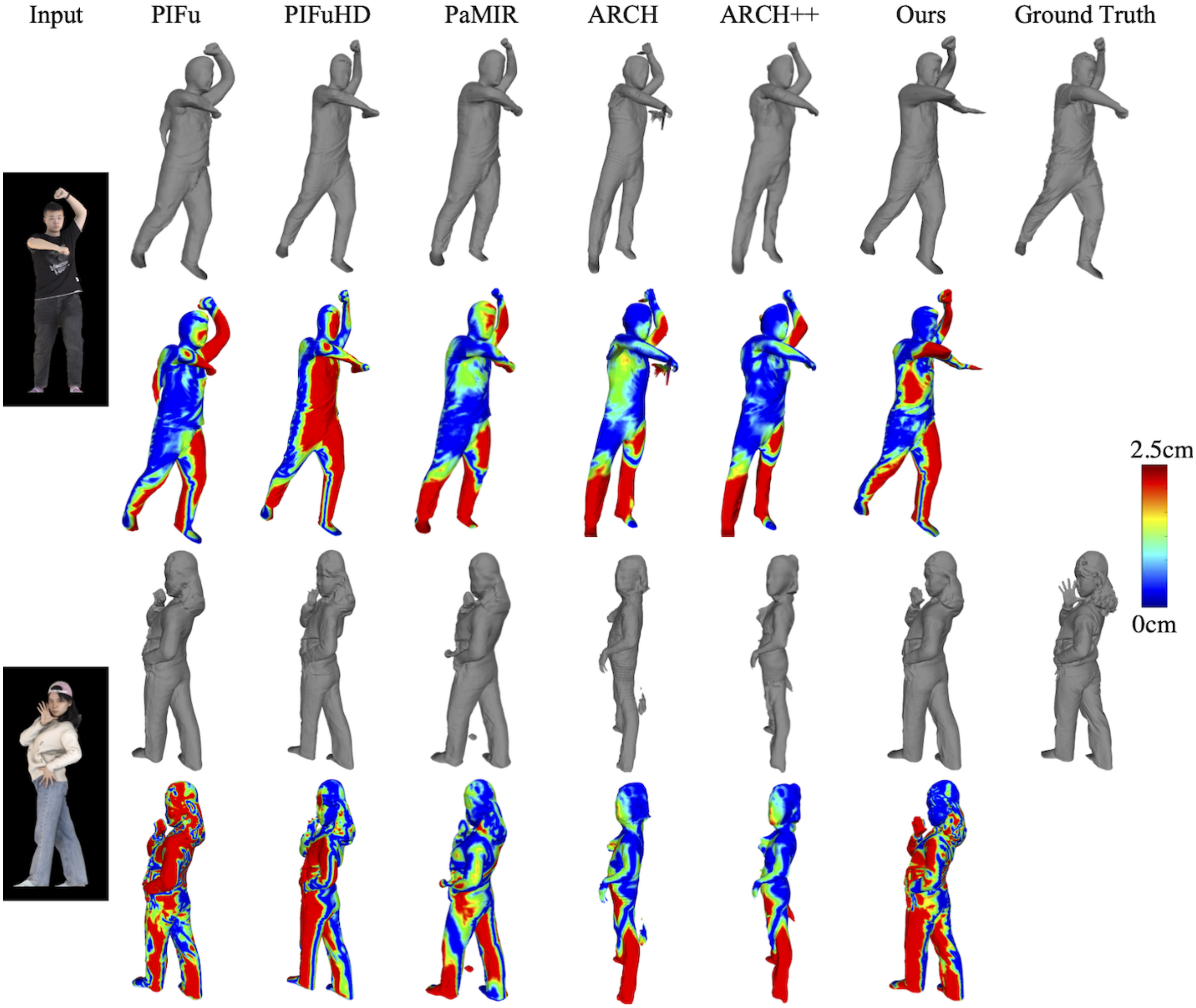


Figure S1. Qualitative comparisons between PIFu, PIFuHD, PaMIR, ARCH, ARCH++, and our method on full body reconstruction (part A).

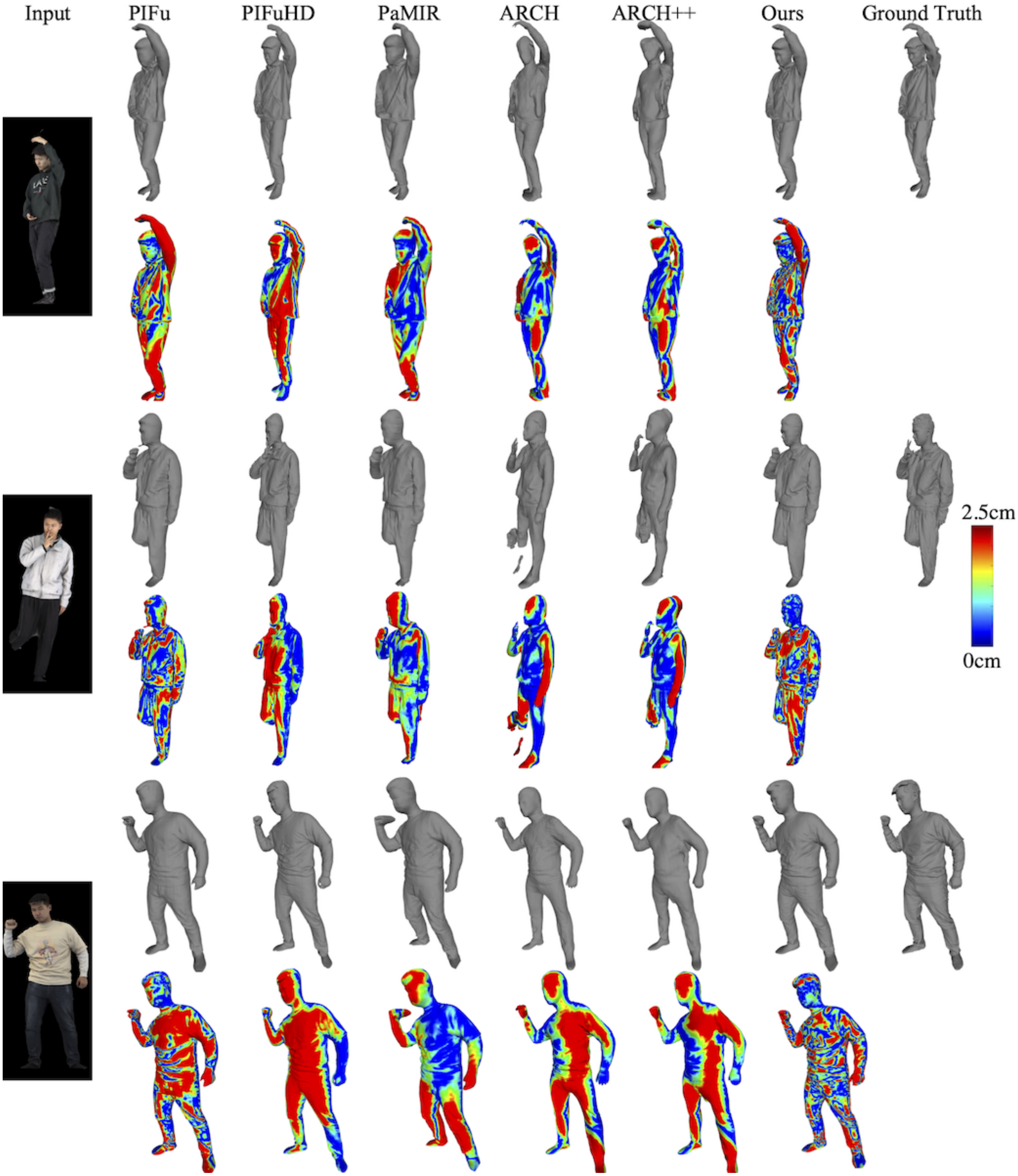


Figure S2. Qualitative comparisons between PIFu, PIFuHD, PaMIR, ARCH, ARCH++, and our method on full body reconstruction (part B).

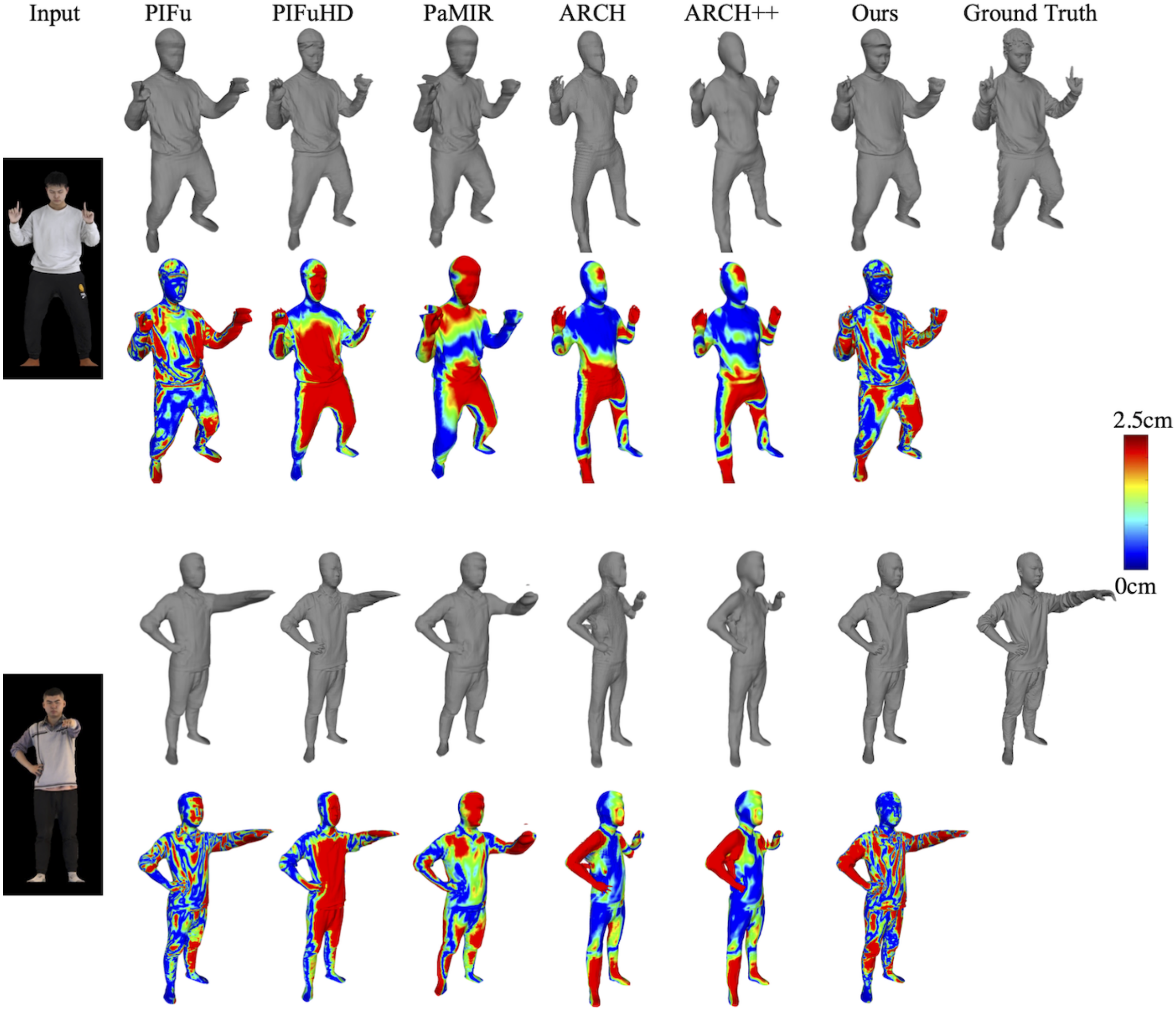


Figure S3. Qualitative comparisons between PIFu, PIFuHD, PaMIR, ARCH, ARCH++, and our method on full body reconstruction (part C).

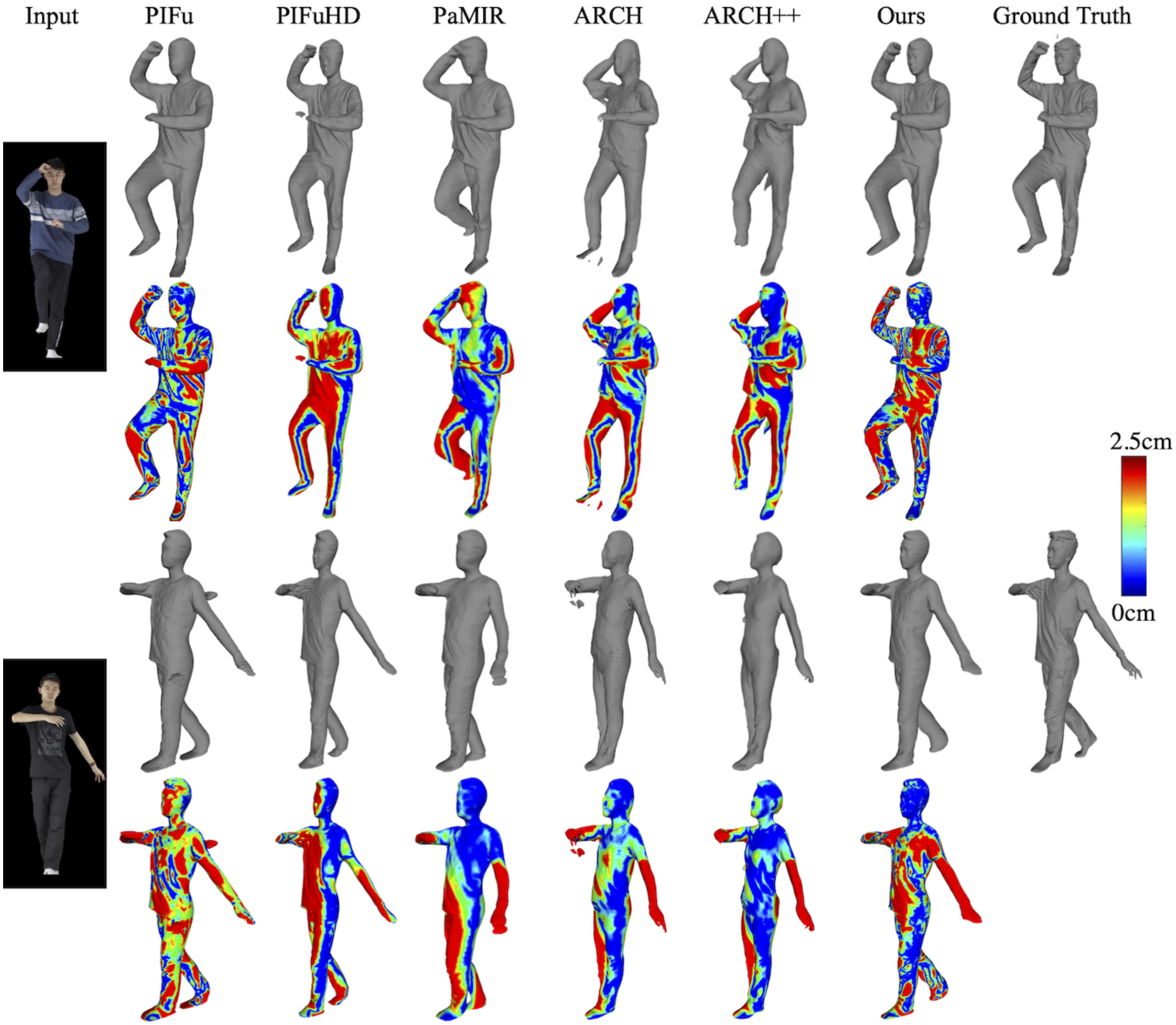


Figure S4. Qualitative comparisons between PIFu, PIFuHD, PaMIR, ARCH, ARCH++, and our method on full body reconstruction (part D).

2.2. Comparisons on face region reconstruction

As can be observed clearly, our method performs the best in terms of both geometry, error color and texture for different subjects under different poses. PIFu [4] and PaMIR [8] cannot generate face details. PIFuHD [5] is able to generate some details but the error is large compared with the ground truth. Our method also consistently outperforms ARCH and ARCH++ in terms of both geometry and texture.

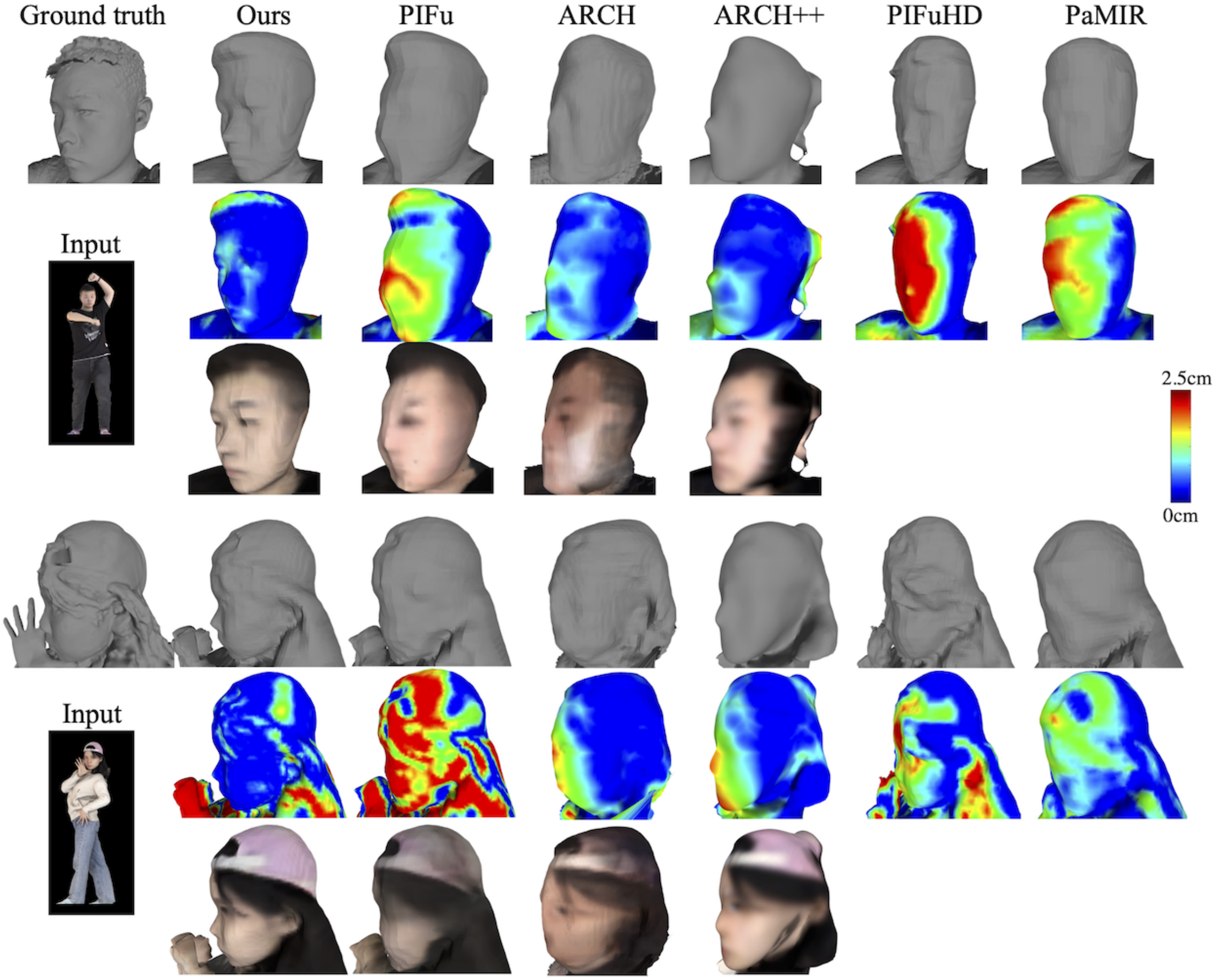


Figure S5. Qualitative comparisons on face reconstruction. Note that PIFuHD and PAMIR do not estimate surface colors (part A).

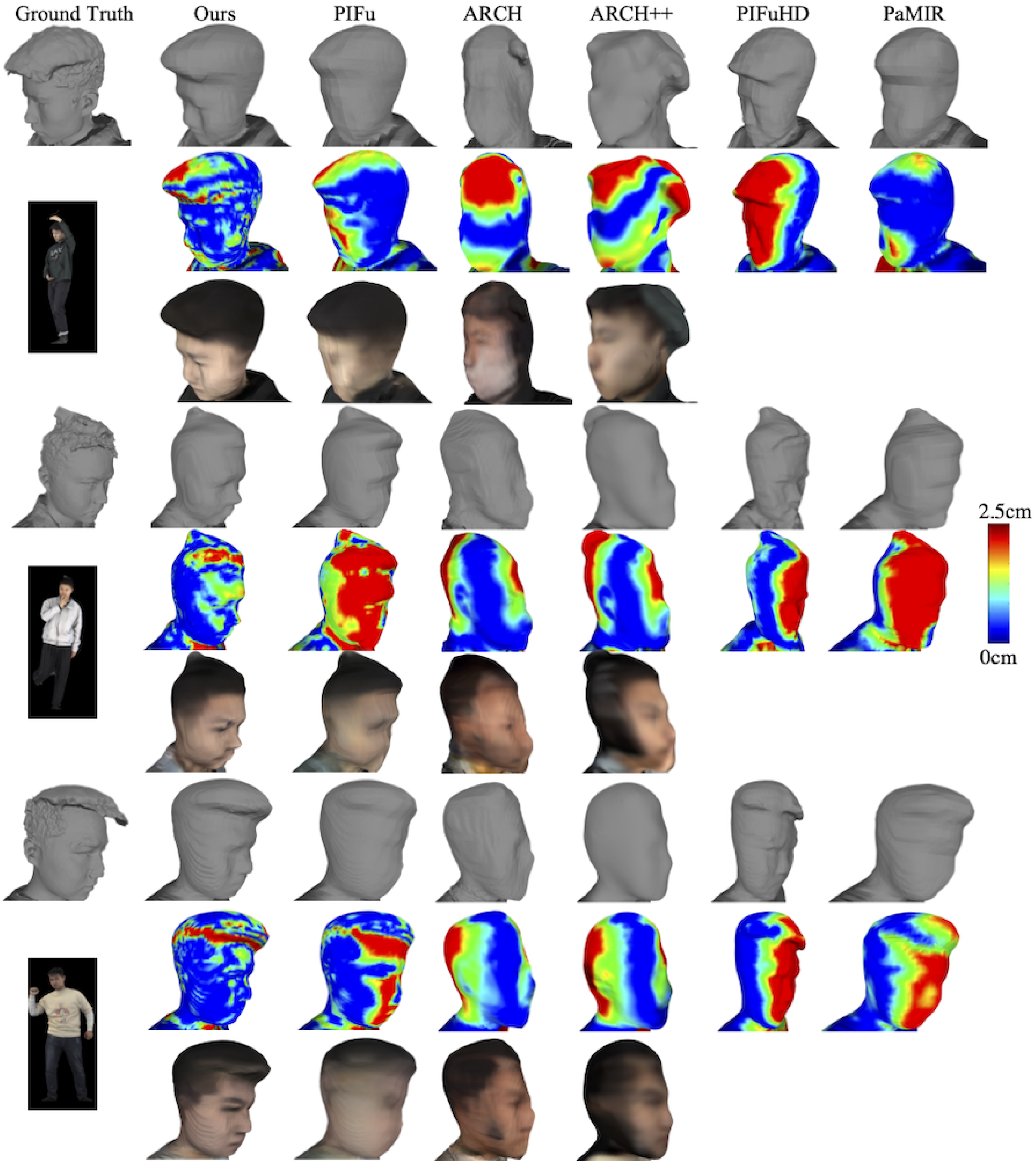


Figure S6. Qualitative comparisons on face reconstruction. Note that PIFuHD and PAMIR do not estimate surface colors (part B).

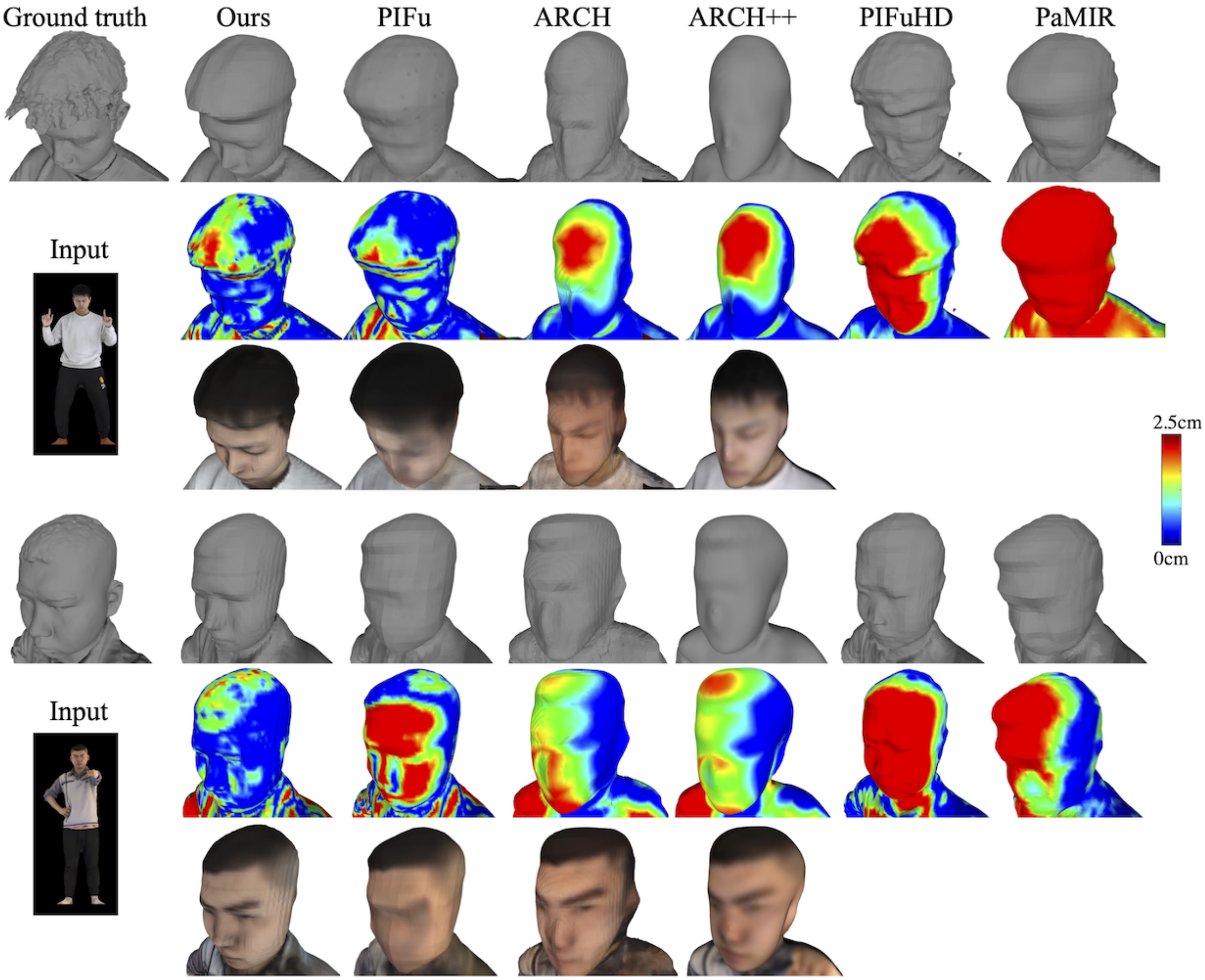


Figure S7. Qualitative comparisons on face reconstruction. Note that PIFuHD and PAMIR do not estimate surface colors (part C).

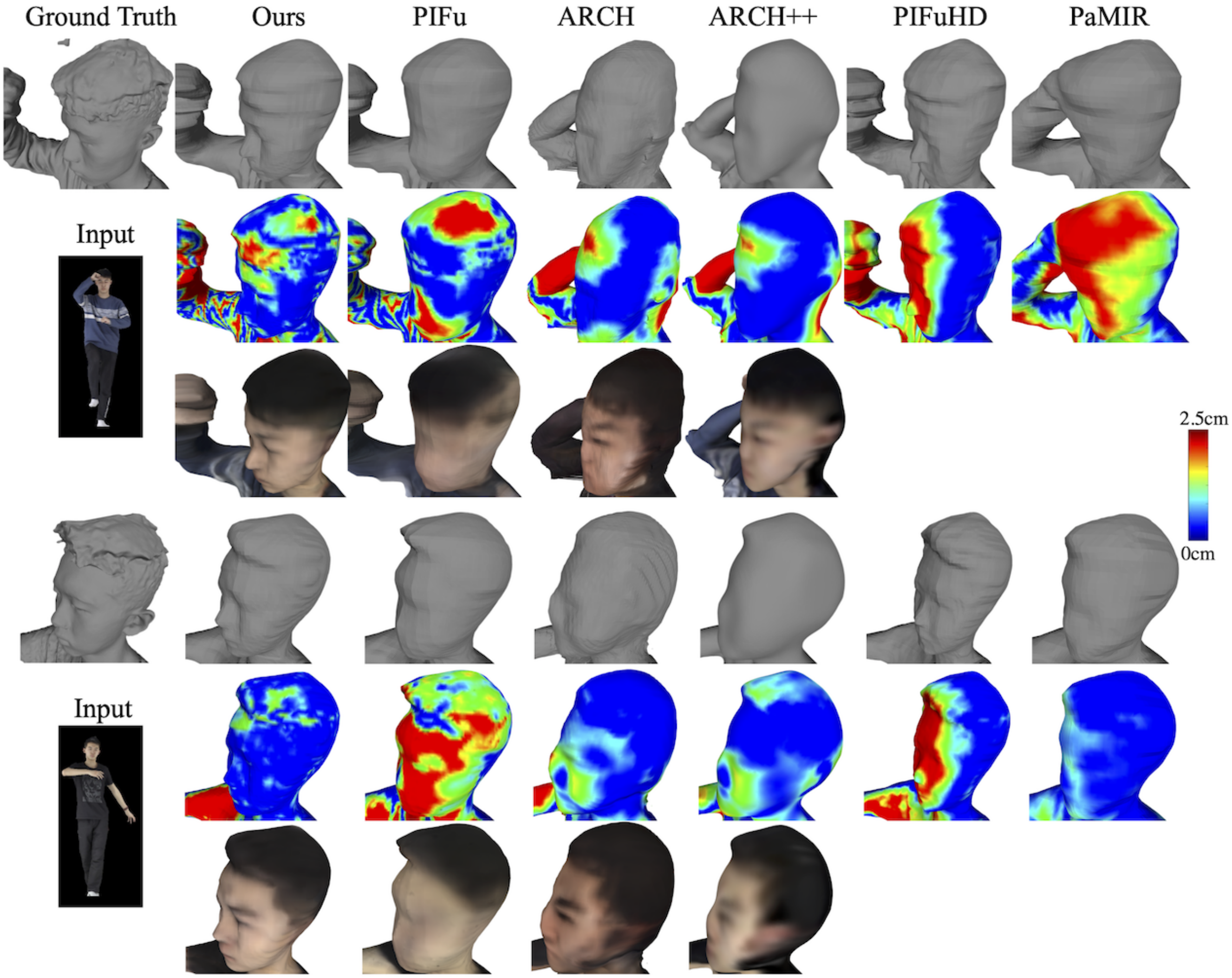


Figure S8. Qualitative comparisons on face reconstruction. Note that PIFuHD and PAMIR do not estimate surface colors (part D).

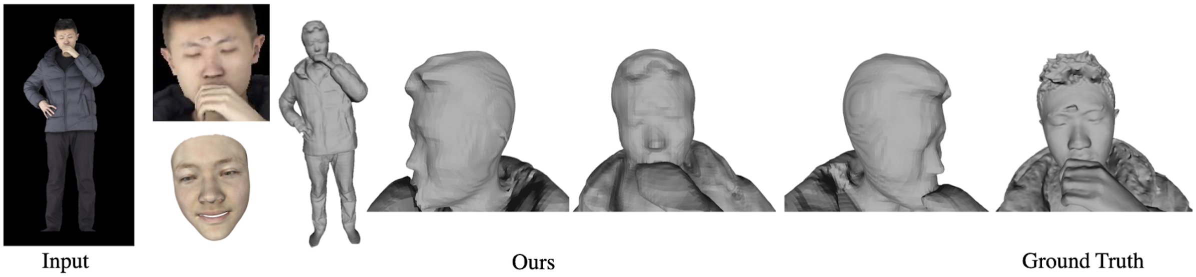


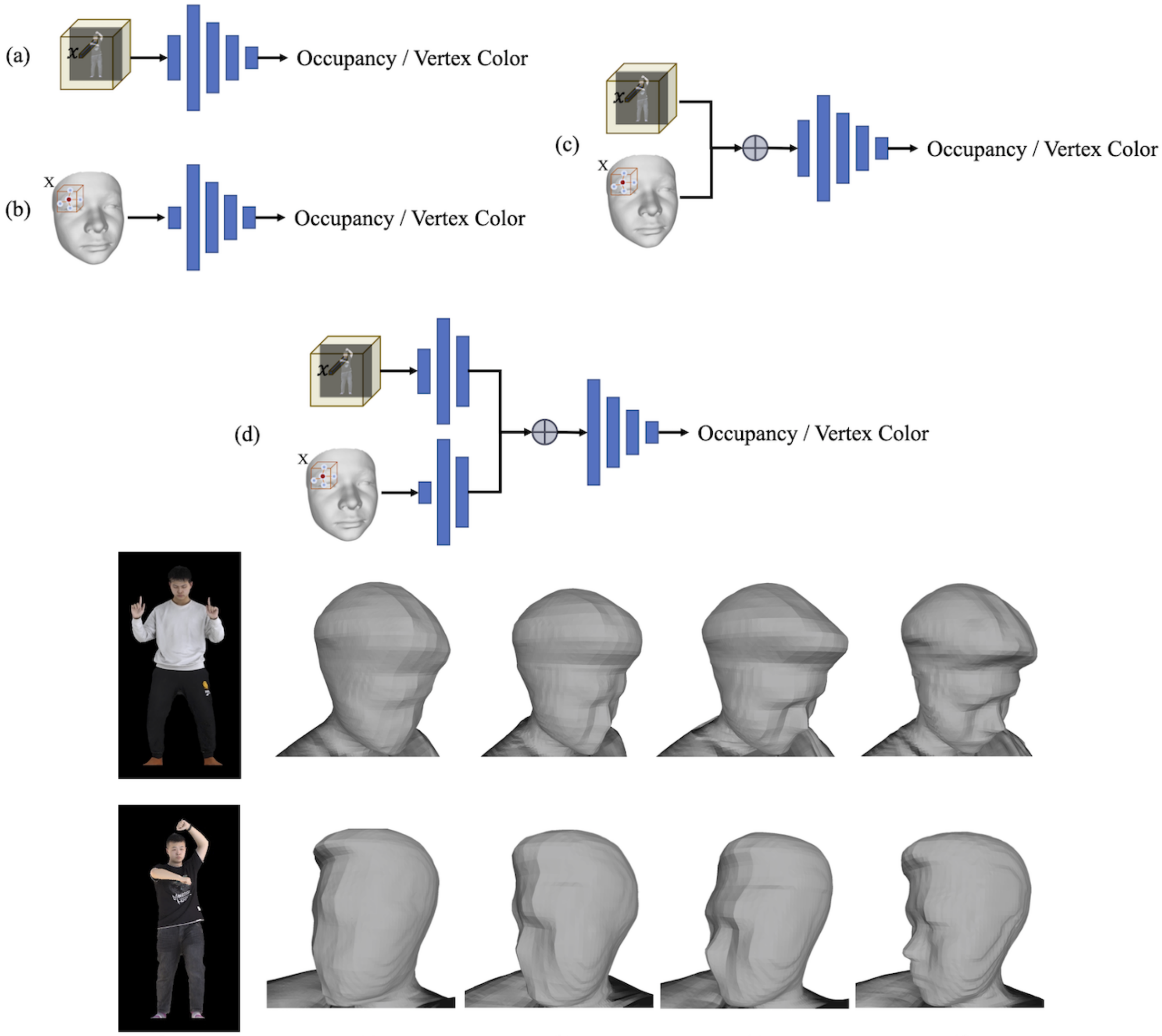
Figure S9. Example reconstruction result for a subject with self-occlusion.

3. More results for ablation study

3.1. Effects of different MLP architectures

Figure S10 presents the pipelines of different MLP settings, and example results produced by them. (a) It is the same setting as PIFu, using only 2D pixel-aligned feature as input for the MLP; (b) We extract the 3D space-aligned feature from the 3D face prior, and feed it into the MLP; (c) 2D pixel-aligned feature and 3D space-aligned feature are concatenated before going through the MLP; and (d) 2D pixel-aligned feature and 3D space-aligned feature are embedded by an MLP before concatenation, and the concatenated feature is sent to the final MLP.

We can observe 3D feature can help generate better shape details than 2D. Directly concatenated 2D and 3D feature can help gain face shape details, while our feature transformation further improve the performance.



Input (a) (b) (c) (d)

Figure S10. (a) 2D feature only; (b) 3D feature only; (c) concatenated 2D and 3D features; (d) jointly training with 2D and 3D feature.

3.2. Effects of sampled point numbers

Sampling more points within the face region can give better convexity / concavity, while the 3D face prior makes the reconstruction more consistent with the ground truth with better shape. Our final choice with more sampling and 3D face prior leads to the best performance.

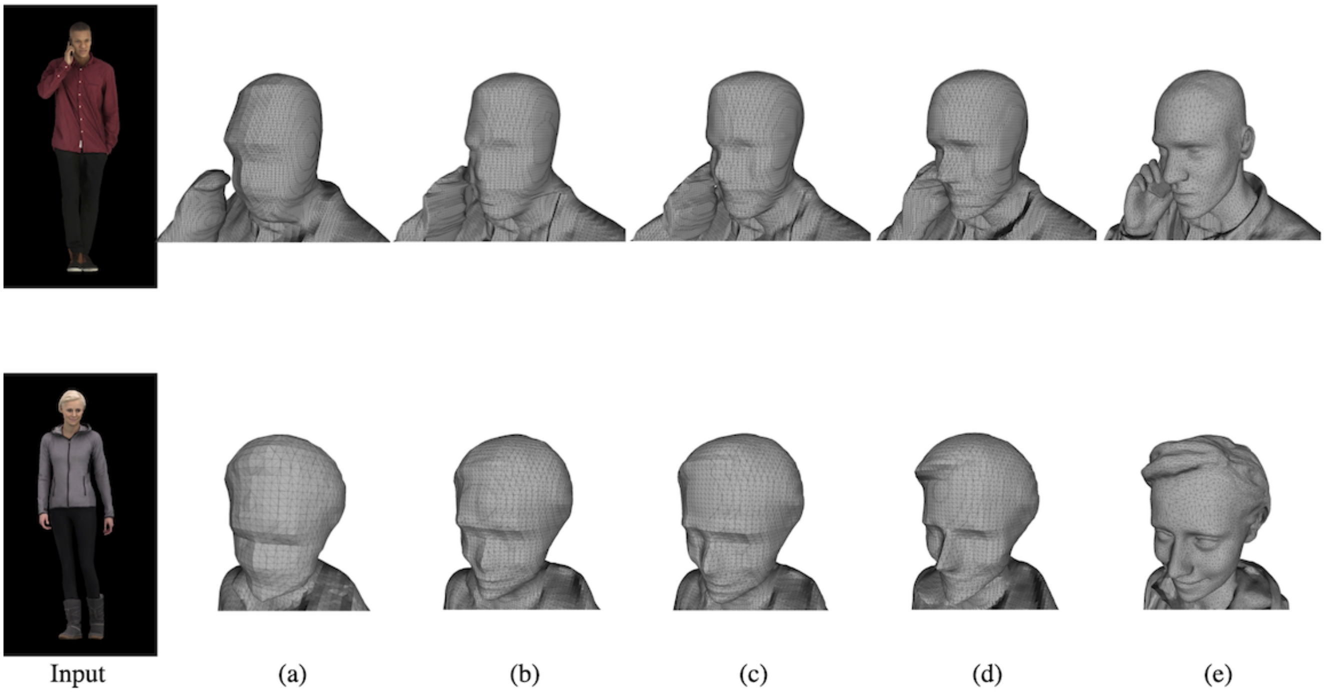


Figure S11. (a) 5000 full-body samples; (b) 5000 full-body samples + 700 face-region samples (c) 5000 samples with 3D face prior; (d) 5000 samples + 700 face-region samples with 3D face prior; (e) Ground-truth mesh

3.3. Effects of different 3D face priors

DECA [2] is a 3D full head model based on FLAME model. It implements differentiable rendering, and applies displacement map learned from 2D image to improve the details.

Figure S12 shows that DECA could also serve as the 3D face prior in JIFF, which results in much better reconstruction than PIFu, indicating the generalizability of our approach to different 3D priors. We observe that 3DMM can give slightly better results and is simpler than DECA. Hence, we use 3DMM as our default choice.

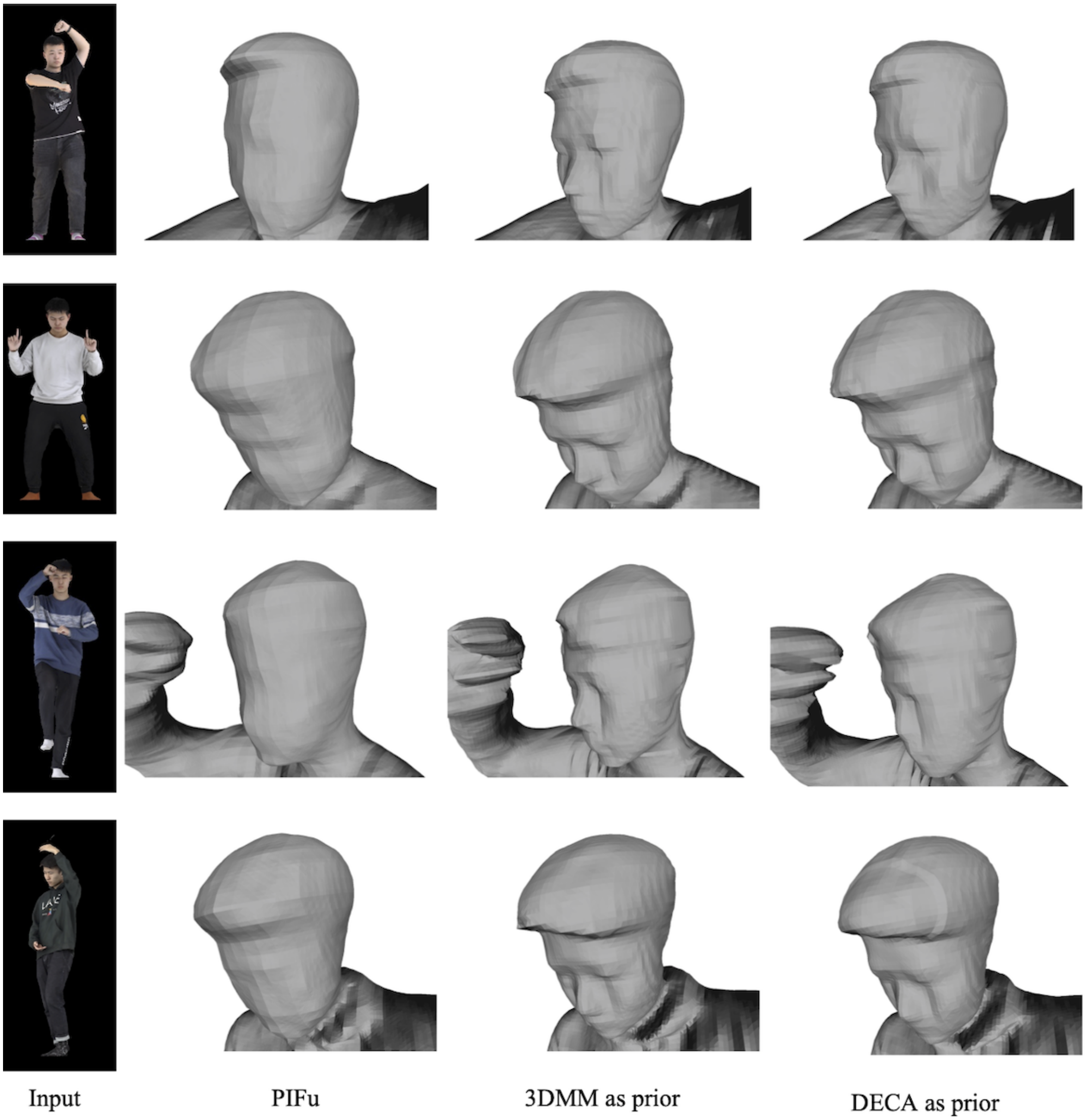


Figure S12. Comparisons between PIFu, our method with 3DMM, and our method with DECA.

4. Discussion on 3D head implicit representations

In this work, we apply 3DMM as a 3D face prior to enhance the pixel-aligned implicit function. Note that there are existing methods (*e.g.*, [3, 6]) focusing on reconstructing the head region. However, we are solving a different problem, as we target at reconstructing full body shapes with fine facial details. Despite the differences in the addressed problems, we discuss the technical differences between our method and the head reconstruction methods [3, 6].

i3DMM[6] The testing input to i3DMM is a watertight 3D scan. During the testing time, it first fit a learned model to optimize the latent vector with the input 3d scan. It then samples points to be the network input to get the prediction. However, the input in our problem is a single image, and it is hard to obtain an accurate watertight 3D head model from the image.

H3D-Net [3] H3D-Net is a multi-view method requiring more than 3 images as input to predict the 3D head model. It has two stages: it first trains a geometry prior with 10000 3D heads, which results in the distribution of SDF representing the 3D heads. Instead of extracting 3D features from that prior, it starts from this prior at the second stage. In the meantime, it uses the losses of SDF, color, and silhouette, and supervises in the image domain with the sphere tracing. It is non-trivial to integrate H3D-Net with pixel-aligned function for full body reconstruction.

With more and more advances in 3D head or face modeling with implicit representation, it would be an interesting direction to study implicit 3D face models a better 3D face prior for full body reconstruction with fine facial details, which we considers as the future work of this paper.

5. Broader impact

Our proposed method JIFF, benefiting from the feature of 3D Morphable Model, improves the quality of face details in the full-body human reconstruction, which could largely facilitate AR and VR applications in the real world. Users can create high-quality 3D models even simply from a single image, and treat them as their digital proxies in the virtual world. However, this should be carefully employed with specific policy so that it will not compromise personal privacy.

References

[1] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[2] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 13

[3] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *IEEE International Conference on Computer Vision*, pages 5620–5629, 2021. 14

[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*, October 2019. 3, 7

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

[6] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 14

[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2

[8] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 7