ICON: Implicit Clothed humans Obtained from Normals
Supplementary Material

Yuliang Xiu Jinlong Yang Dimitrios Tzionas Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{yuliang.xiu, jinlong.yang, dtzionas, black}@tuebingen.mpg.de

We provide more details for the method and experiments, as well as more quantitative and qualitative results, as an extension of Sec. 3, Sec. 4 and Sec. 5 of the main paper.

1. Method & Experiment Details

1.1. Dataset (Sec. 4.2)

Dataset size. We evaluate the performance of ICON and SOTA methods for a varying training-dataset size (Fig. 6 and Tab. 9). For this, we first combine AGORA [12] ($3,109$ scans) and THuman [16] (600 scans) to get $3,709$ scans in total. This new dataset is 8x times larger than the 450 Renderpeople ("450-Rp") scans used in [13, 14]. Then, we sample this "8x dataset" to create smaller variations, for $1/8$x, $1/4$x, $1/2$x, 1x, and 8x the size of "450-Rp".

Dataset splits. For the "8x dataset", we split the $3,109$ AGORA scans into a new training set ($3,034$ scans), validation set (25 scans) and test set (50 scans). Among these, $1,847$ come from Renderpeople [3] (see Fig. 9a), 622 from AXYZ [4], 242 from Humanalloy [2], 398 from 3DPeople [1], and we sample only 600 scans from THuman (see Fig. 9b), due to its high pose repeatability and limited identity variants (see Tab. 1), with the "select-cluster" scheme described below. These scans, as well as their SMPL-X fits, are rendered after every 10 degrees rotation around the yaw axis, to totally generate $(3109\ \texttt{AGORA} + 600\ \texttt{THuman} + 150\ \texttt{CAPE}) \times 36 = 138,924$ samples.

Dataset distribution via "select-cluster" scheme. To create a training set with a rich pose distribution, we need to select scans from various datasets with poses different from AGORA. Following SMPLify [6], we first fit a Gaussian Mixture Model (GMM) with 8 components to all AGORA poses, and **select** 2K THuman scans with low likelihood. Then, we apply M-Medoids ($\texttt{n\_cluster} = 50$) on these selections for **clustering**, and randomly pick 12 scans per cluster, collecting $50 \times 12 = 600$ THuman scans in total; see Fig. 9b. This is also used to split CAPE into "CAPE-FP" (Fig. 9c) and "CAPE-NFP" (Fig. 9d), corresponding to scans with poses similar (in-distribution poses) and dissimilar (out-of-distribution poses) to AGORA ones, respectively.

Perturbed SMPL. To perturb SMPL's pose and shape parameters, random noise is added to θ and β by:

$$\begin{aligned} \theta &\mathrel{+}= s_\theta * \mu, \\ \beta &\mathrel{+}= s_\beta * \mu, \end{aligned} \tag{7}$$

where $\mu \in [-1, 1]$, $s_\theta = 0.15$ and $s_\beta = 0.5$. These are set empirically to mimic the misalignment error typically caused by off-the-shell HPS during testing.

Discussion on simulated data. The wide and loose clothing in CLOTH3D++ [5, 10] demonstrates strong dynamics, which would complement commonly used datasets of commercial scans. Yet, the domain gap between CLOTH3D++ and real images is still large. Moreover, it is unclear how to train an implicit function from multi-layer non-watertight meshes. Consequently, we leave it for future research.

1.2. Refining SMPL (Sec. 3.1)

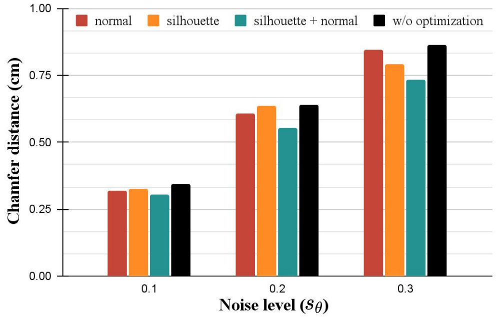


Figure 8. SMPL refinement error (y-axis) with different losses (see colors) and noise levels, s_θ, of pose parameters (x-axis).

To statistically analyze the necessity of $\mathcal{L}_{\text{N\_diff}}$ and $\mathcal{L}_{\text{S\_diff}}$ in Eq. (4), we do a sanity check on AGORA's validation set. Initialized with different pose noise, s_θ (Eq. (7)), we optimize the $\{\theta, \beta, t\}$ parameters of the perturbed SMPL by minimizing the difference between rendered SMPL-body normal maps and ground-truth clothed-body normal maps for 2K iterations. As Fig. 8 shows, $\mathcal{L}_{\text{N\_diff}} + \mathcal{L}_{\text{S\_diff}}$ always leads to the smallest error under any noise level, measured by the Chamfer distance between the optimized perturbed SMPL mesh and the ground-truth SMPL mesh.

	Methods	SMPL-X	AGORA-50			CAPE-FP			CAPE-NFP			CAPE		
		condition.	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓
	ICON	✓	1.583	1.987	0.079	**1.364**	**1.403**	0.080	**1.444**	**1.453**	0.083	**1.417**	**1.436**	0.082
D	SMPL-X perturbed	✓	1.984	2.471	0.098	1.488	1.531	0.095	1.493	1.534	0.098	1.491	1.533	0.097
	$\text{ICON}_{\text{enc(I,N)}}$	✓	1.569	**1.784**	**0.073**	1.379	1.498	**0.070**	1.600	1.580	**0.078**	1.526	1.553	**0.075**
	$\text{ICON}_{\text{enc(N)}}$	✓	**1.564**	1.854	0.074	1.368	1.484	0.071	1.526	1.524	**0.078**	1.473	1.511	0.076
	$\text{ICON}_{\text{N}^\dagger}$	✓	1.575	2.016	0.077	1.376	1.496	0.076	1.458	1.569	0.080	1.431	1.545	0.079

Table 4. Quantitative errors (cm) for several ICON variants conditioned on perturbed SMPL-X fits ($s_\theta = 0.15, s_\beta = 0.5$).

1.3. Perceptual study (Tab. 3)

Reconstruction on in-the-wild images. We perform a perceptual study to evaluate the perceived realism of the reconstructed clothed 3D humans from in-the-wild images. ICON is compared against 3 methods, PIFu [13], PIFuHD [14], and PaMIR [15]. We create a benchmark of 200 unseen images downloaded from the internet, and apply all the methods on this test set. All the reconstruction results are evaluated on Amazon Mechanical Turk (AMT), where each participant is shown pairs of reconstructions from ICON and one of the baselines, see Fig. 10. Each reconstruction result is rendered in four views: front, right, back and left. Participants are asked to choose the reconstructed 3D shape that better represents the human in the given color image. Each participant is given 100 samples to evaluate. To teach participants, and to filter out the ones that do not understand the task, we set up 1 tutorial sample, followed by 10 warm-up samples, and then the evaluation samples along with catch trial samples inserted every 10 evaluation samples. Each catch trial sample shows a color image along with either (1) the reconstruction of a baseline method for this image and the ground-truth scan that was rendered to create this image, or (2) the reconstruction of a baseline method for this image and the reconstruction for a different image (false positive), see Fig. 10c. Only participants that pass 70% out of 10 catch trials are considered. This leads to 28 valid participants out of 36 ones. Results are reported in Tab. 3.

Normal map prediction. To evaluate the effect of the body prior for normal map prediction on in-the-wild images, we conduct a perceptual study against prediction without the body prior. We use AMT, and show participants a color image along with a pair of predicted normal maps from two methods. Participants are asked to pick the normal map that better represents the human in the image. Front- and back-side normal maps are evaluated separately. See Fig. 11 for some samples. We set up 2 tutorial samples, 10 warm-up samples, 100 evaluation samples and 10 catch trials for each subject. The catch trials lead to 20 valid subjects out of 24 participants. We report the statistical results in Tab. 5. A chi-squared test is performed with a null hypothesis that the body prior does not have any influence. We show some results in Fig. 12, where all participants unanimously prefer one method over the other. While results of both methods look generally similar on front-side normal maps, using the body prior usually leads to better back-side normal maps.

	w/ SMPL prior	w/o SMPL prior	P-value
Preference (front)	47.3%	52.7%	8.77e-2
Preference (back)	52.9%	47.1%	6.66e-2

Table 5. Perceptual study on normal prediction.

	w/ global encoder	pixel dims	point dims	total dims
PIFu*	✓	12	1	13
PaMIR*	✓	6	7	13
$\text{ICON}_{\text{enc(I,N)}}$	✓	6	7	13
$\text{ICON}_{\text{enc(N)}}$	✓	6	7	13
ICON	✗	0	7	7

Table 6. Feature dimensions for various approaches. "pixel dims" and "point dims" denote the feature dimensions encoded from pixels (image/normal maps) and 3D body prior, respectively.

1.4. Implementation details (Sec. 4.1)

Network architecture. Our body-guided normal prediction network uses the same architecture as PIFuHD [14], originally proposed in [8], and consisting of residual blocks with 4 down-sampling layers. The image encoder for PIFu*, PaMIR*, and ICON_{enc} is a stacked hourglass [11] with 2 stacks, modified according to [7]. Tab. 6 lists feature dimensions for various methods; "total dims" is the neuron number for the first MLP layer (input). The number of neurons in each MLP layer is: 13 (7 for ICON), 512, 256, 128, and 1, with skip connections at the 3rd, 4th, and 5th layers.

Training details. For training $\mathcal{G}^{\text{N}}$ we do not use THuman due to its low-quality texture (see Tab. 1). On the contrary, $\mathcal{IF}$ is trained on both AGORA and THuman. The front-side and back-side normal prediction networks are trained individually with batch size of 12 under the objective function defined in Eq. (3), where we set $\lambda_{\text{VGG}} = 5.0$. We use the ADAM optimizer with a learning rate of 1.0×10^{-4} until convergence at 80 epochs.

Test-time details. During inference, to iteratively refine SMPL and the predicted clothed-body normal maps, we perform 50 iterations (each iteration takes $\sim$ 460 ms on a Quadro RTX 5000 GPU) and set $\lambda_{\text{N}} = 2.0$ in Eq. (4). We conduct an experiment to show the influence of the number of iterations (#iterations) on accuracy, see Tab. 7.

The resolution of the queried occupancy space is 256^3. We use `rembg`[1] to segment the humans in in-the-wild images, and use `Kaolin`[2] to compute per-point the signed distance, $\mathcal{F}_{\text{s}}$, and barycentric surface normal, $\mathcal{F}_{\text{n}}^{\text{b}}$.

# iters (460ms/it)	0	10	50
Chamfer ↓	1.417	1.413	**1.339**
P2S ↓	1.436	1.515	**1.378**
Normal ↓	0.082	0.077	**0.074**

Table 7. ICON errors w.r.t. iterations

Discussion on receptive field size. As Tab. 8 shows, simply reducing the size of receptive field of PaMIR does not lead to better performance. This shows that our informative 3D features as in Eq. (6) and normal maps $\widehat{\mathcal{N}}^{\text{c}}$ also play important roles for robust reconstruction. A more sophisticated design of smaller receptive field may lead to better performance and we would leave it for future research.

Receptive field	139	271	403
Chamfer ↓	1.418	1.478	**1.366**
P2S ↓	1.236	1.320	**1.214**
Normal ↓	0.083	0.084	**0.078**

Table 8. PaMIR's receptive field

2. More Quantitative Results (Sec. 4.3)

Table 4 compares several ICON variants conditioned on perturbed SMPL-X meshes. For the plot of Fig. 6 of the main paper (reconstruction error w.r.t. training-data size), extended quantitative results are shown in Tab. 9.

Training set scale		1/8x	1/4x	1/2x	1x	8x
PIFu*	Chamfer ↓	3.339	2.968	2.932	2.682	1.760
	P2S ↓	3.280	2.859	2.812	2.658	1.547
PaMIR*	Chamfer ↓	2.024	1.780	1.479	1.350	1.095
	P2S ↓	1.791	1.778	1.662	1.283	1.131
ICON	Chamfer ↓	1.336	1.266	1.219	1.142	**1.036**
	P2S ↓	1.286	1.235	1.184	1.065	**1.063**

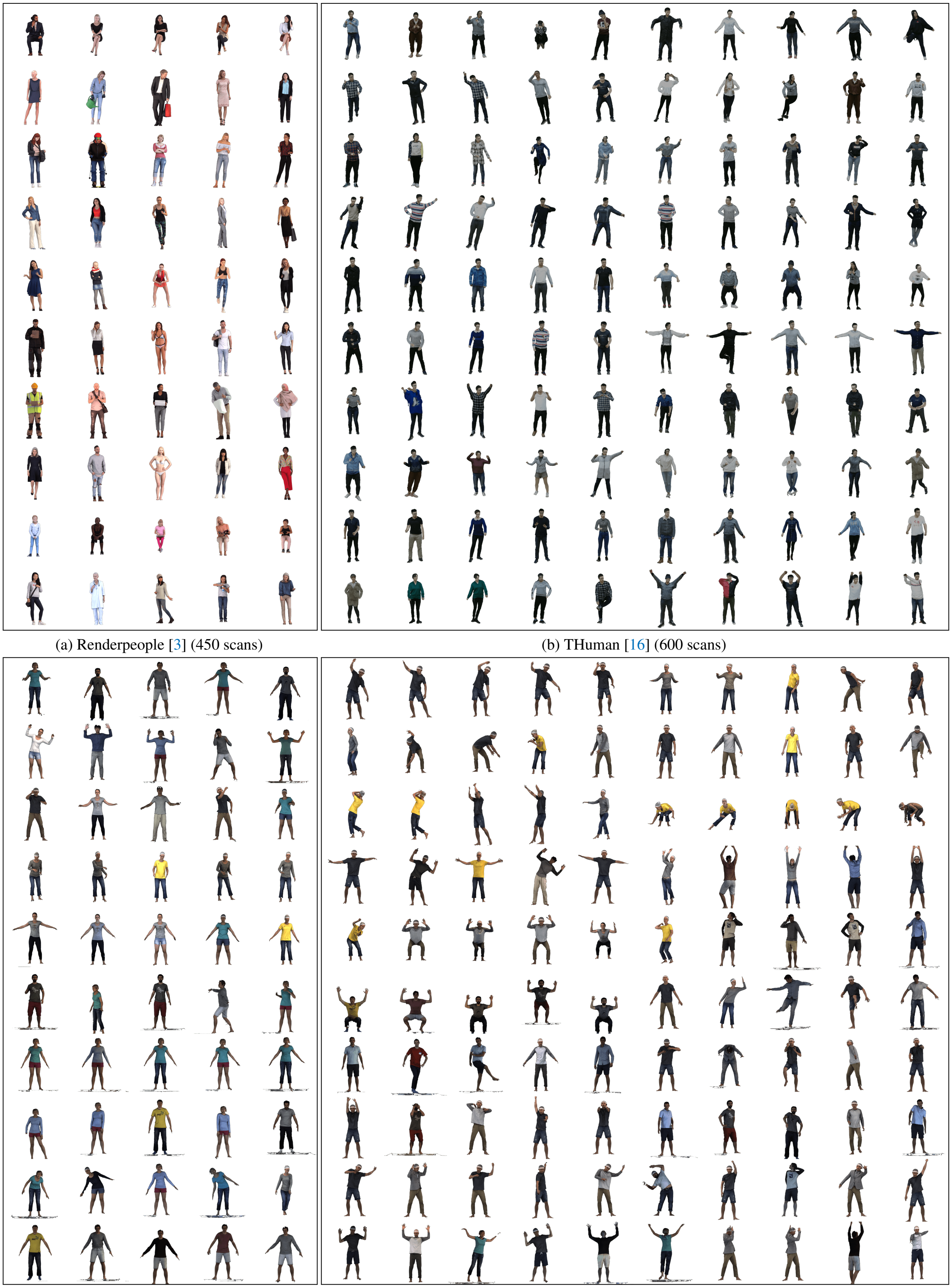
Table 9. Reconstruction error (cm) w.r.t. training-data size. "Training set scale" is defined as the ratio w.r.t. the 450 scans used in [13, 14]. The "8x" setting is all $3,709$ scans of AGORA [12] and THuman [16]. Results outperform ground-truth SMPL-X, which has 1.158 cm and 1.125 cm for Chamfer and P2S in Tab. 2.

3. More Qualitative Results (Sec. 5)

Figures 13 to 15 show reconstructions for in-the-wild images, rendered from four different view points; normals are color coded. Figure 16 shows reconstructions for images with out-of-frame cropping. Figure 17 shows additional representative failures. The video on our website shows animation examples created with ICON and SCANimate.

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

[2] `https://github.com/NVIDIAGameWorks/kaolin`



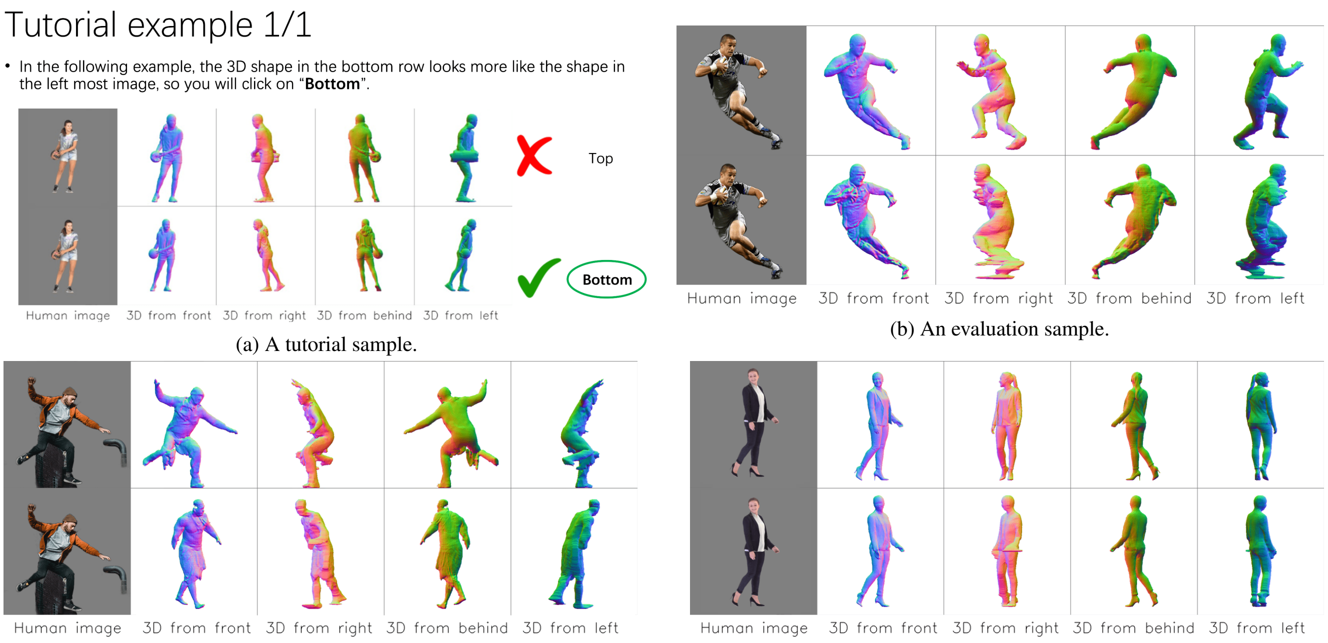
(a) Renderpeople [3] (450 scans)

(b) THuman [16] (600 scans)

(c) "CAPE-FP" [9] (fashion poses, 50 scans)

(d) "CAPE-NFP" [9] (non fashion poses, 100 scans)

Figure 9. Representative poses for different datasets.



Tutorial example 1/1

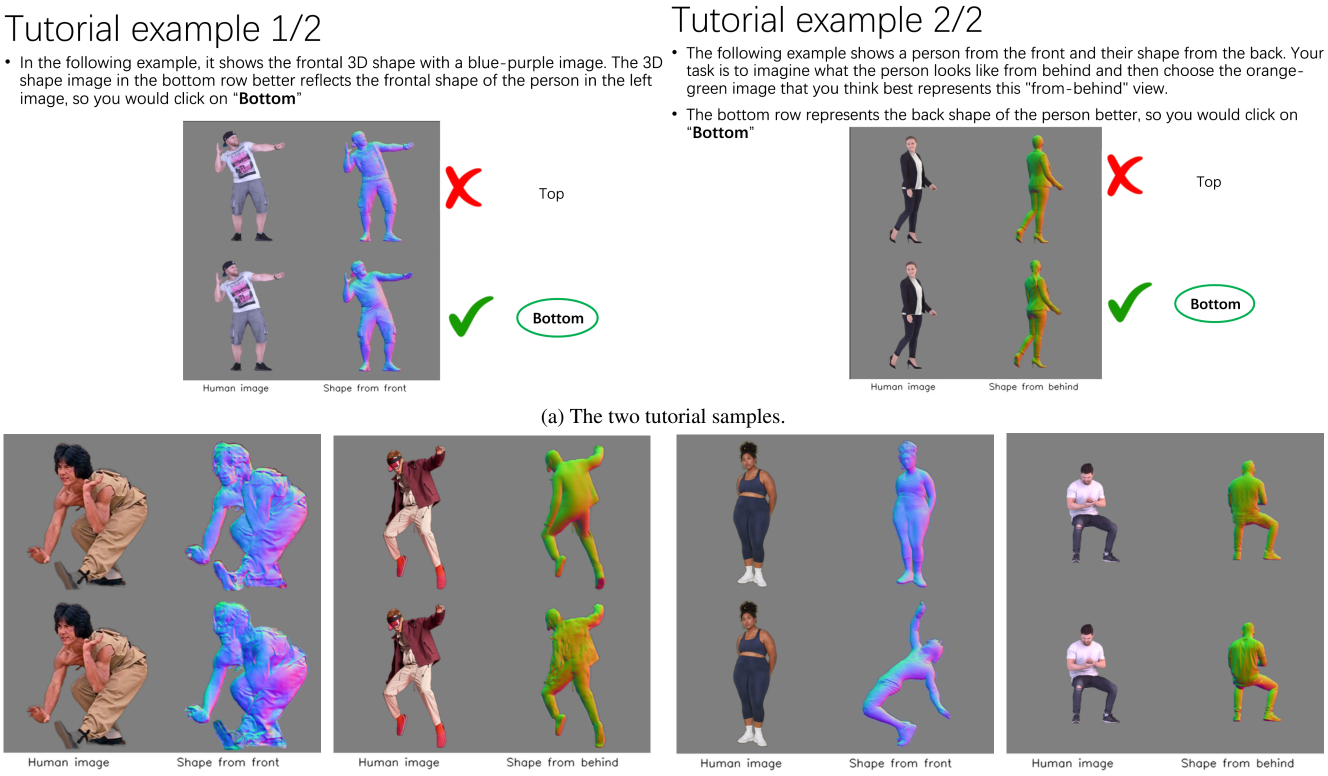
- In the following example, the 3D shape in the bottom row looks more like the shape in the left most image, so you will click on "**Bottom**".

(a) A tutorial sample.

(b) An evaluation sample.

(c) Two samples of catch trials. Left: result from this image (top) vs from another image (bottom). Right: ground-truth (top) vs reconstruction mesh (bottom).

Figure 10. Some samples in the perceptual study to evaluate **reconstructions** on in-the-wild images.



Tutorial example 1/2

- In the following example, it shows the frontal 3D shape with a blue-purple image. The 3D shape image in the bottom row better reflects the frontal shape of the person in the left image, so you would click on "**Bottom**"

Tutorial example 2/2

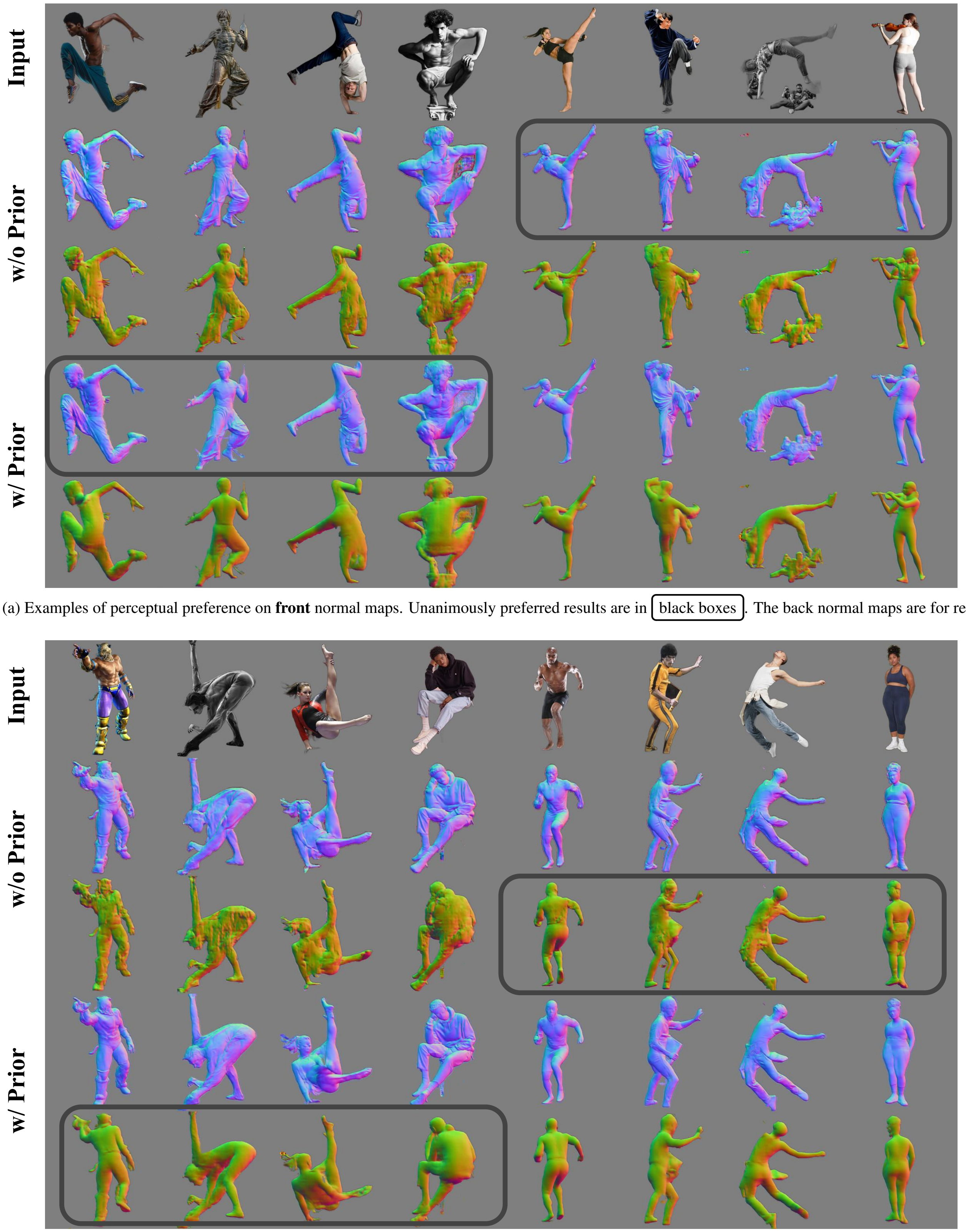
- The following example shows a person from the front and their shape from the back. Your task is to imagine what the person looks like from behind and then choose the orange-green image that you think best represents this "from-behind" view.
- The bottom row represents the back shape of the person better, so you would click on "**Bottom**"

(a) The two tutorial samples.

(b) Two evaluation samples.

(c) Two catch trial samples.

Figure 11. Some samples in the perceptual study to evaluate the effect of the **body prior** for **normal prediction** on in-the-wild images.



(a) Examples of perceptual preference on **front** normal maps. Unanimously preferred results are in black boxes. The back normal maps are for reference.

(b) Examples of perceptual preference on **back** normal maps. Unanimously preferred results are in black boxes. The front normal maps are for reference.

Figure 12. Qualitative results to evaluate the effect of body prior for normal prediction on in-the-wild images.

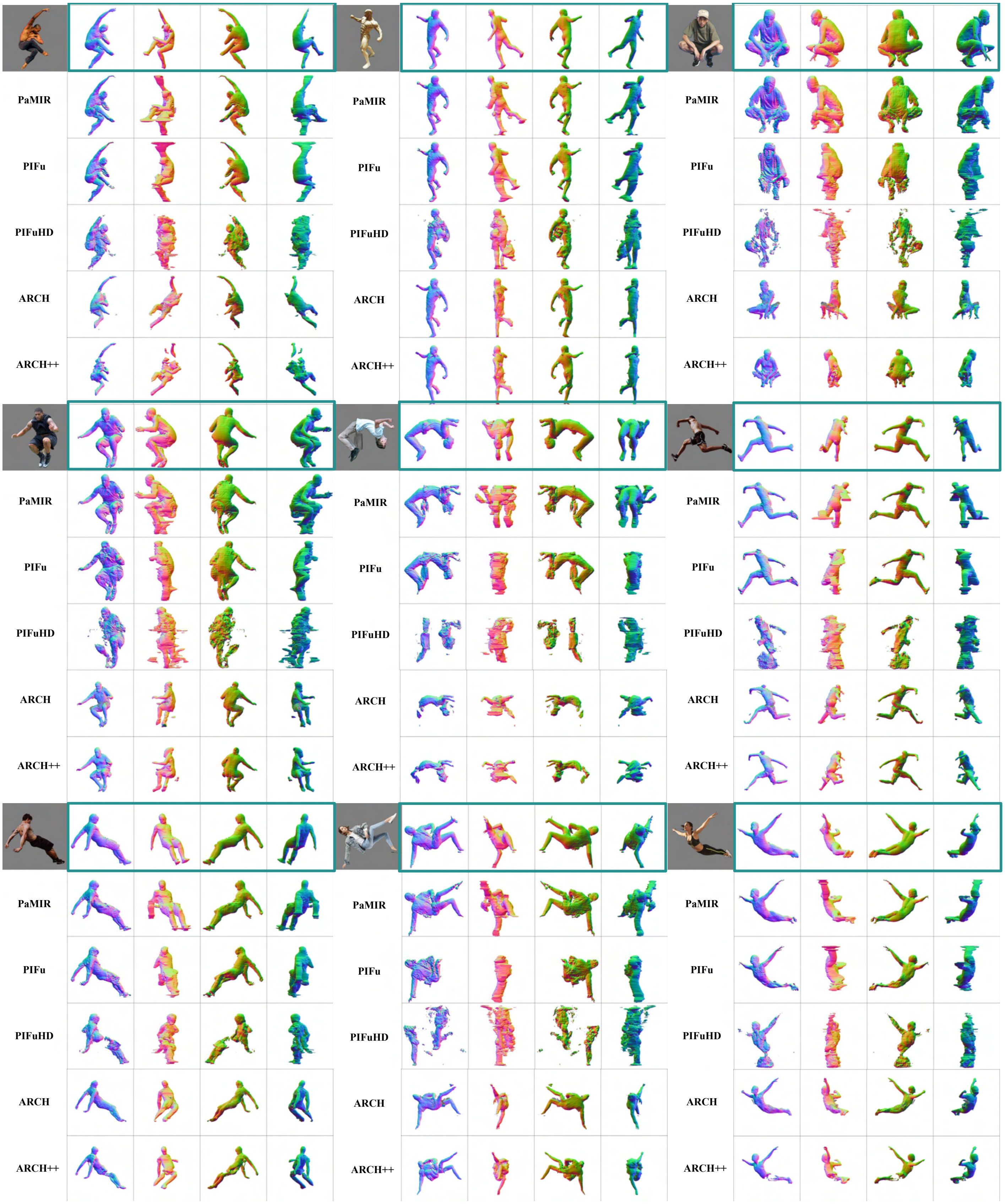


Figure 13. Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

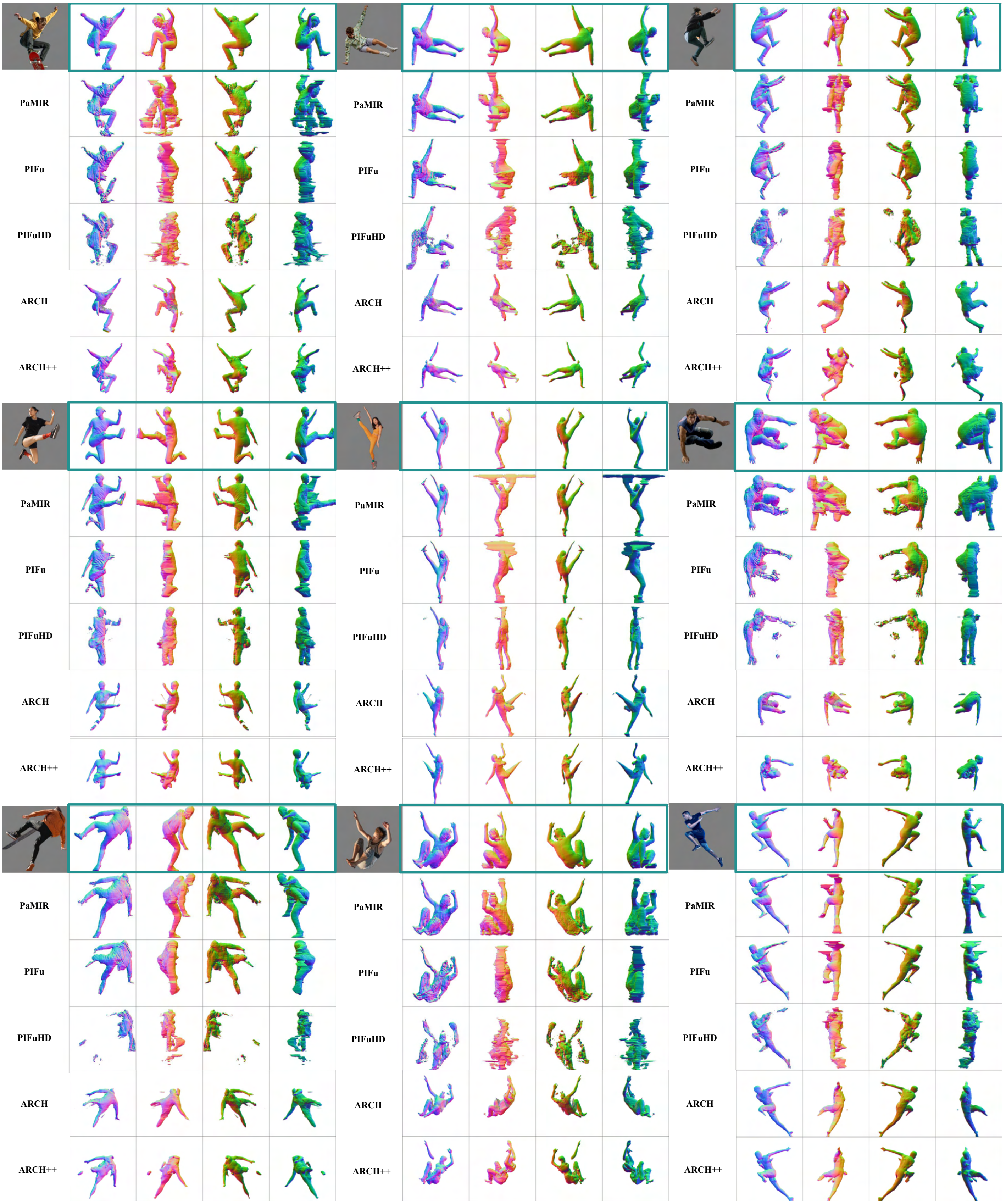


Figure 14. Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

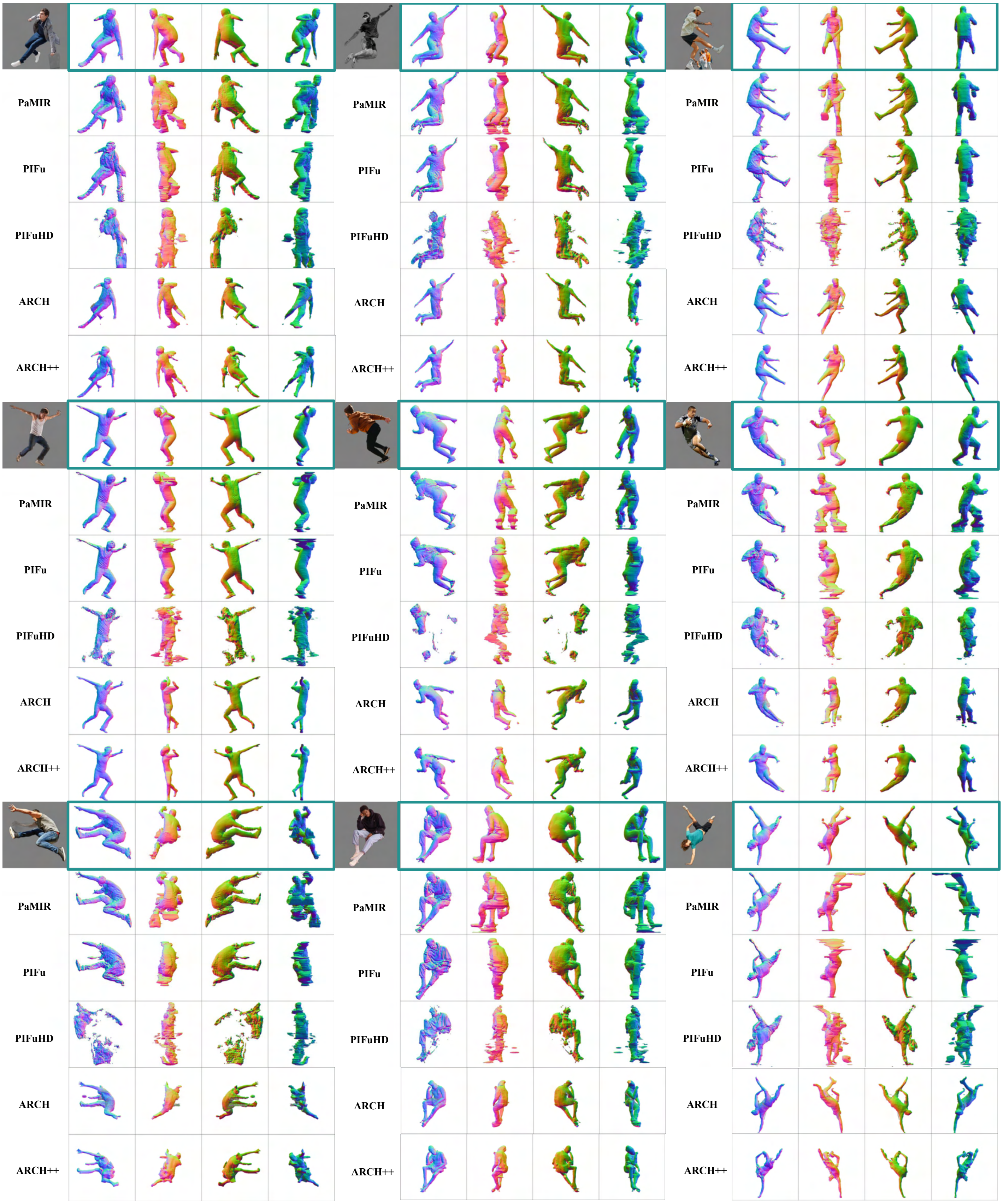


Figure 15. Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

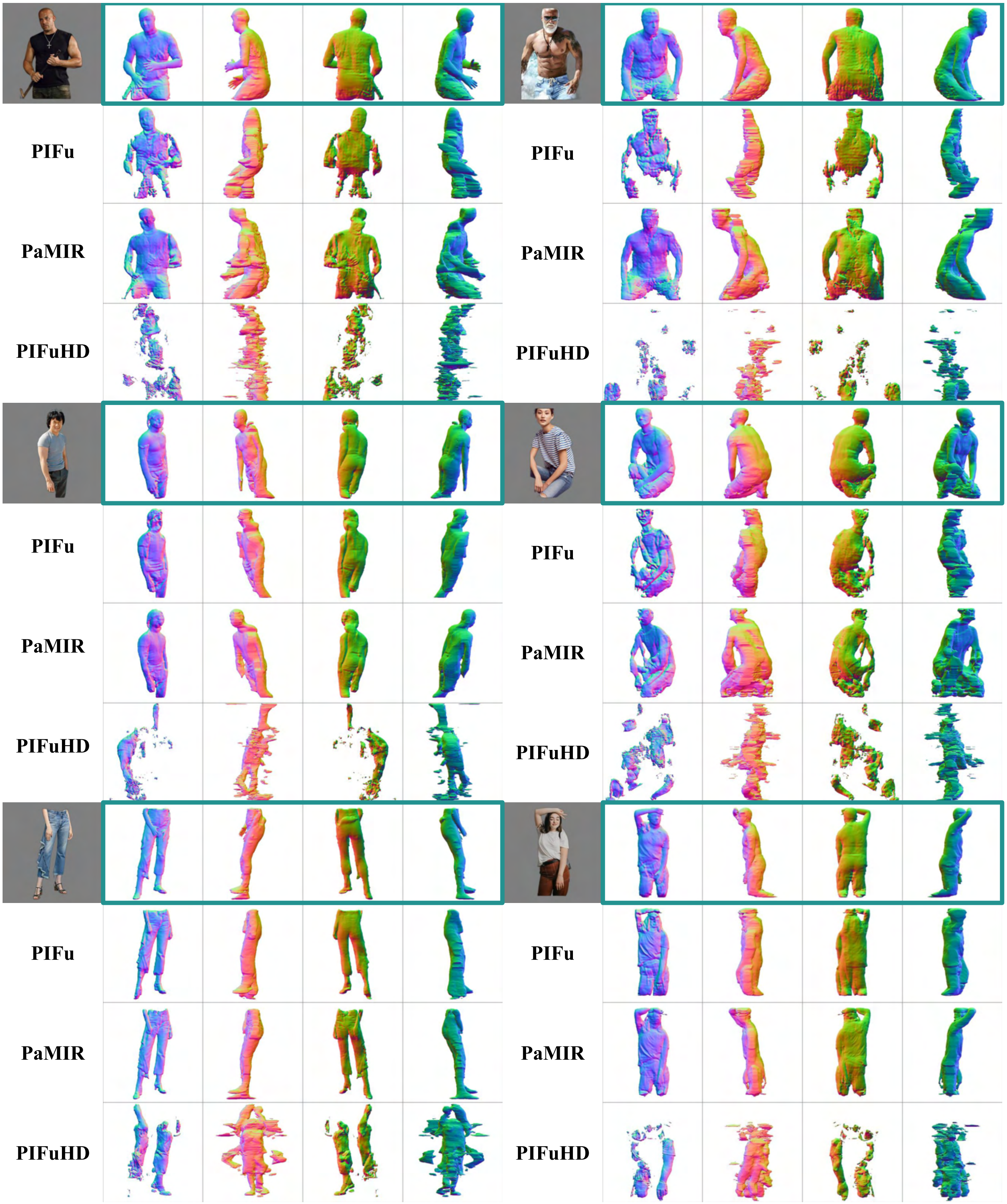
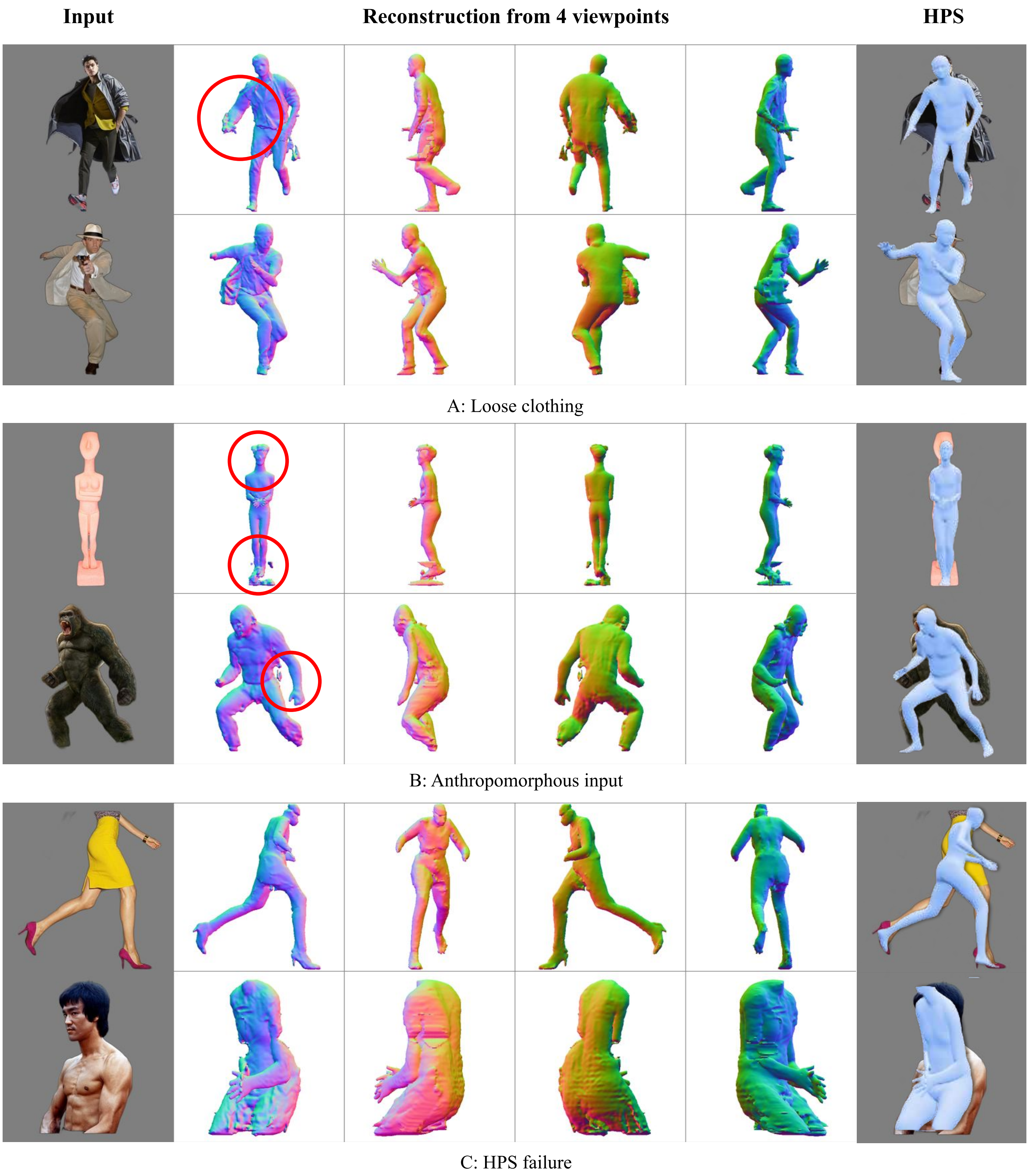


Figure 16. Qualitative comparison (ICON vs SOTA) on images with out-of-frame cropping.



A: Loose clothing

B: Anthropomorphous input

C: HPS failure

Figure 17. More failure cases of ICON.

References

[1] 3DPeople. 3dpeople.com, 2018.

[2] HumanAlloy. humanalloy.com, 2018.

[3] RenderPeople. renderpeople.com, 2018.

[4] AXYZ. secure.axyz-design.com, 2018.

[5] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: Clothed 3D humans. In *European Conference on Computer Vision (ECCV)*, volume 12365, pages 344–359, 2020.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, volume 9909, pages 561–578, 2016.

[7] Aaron S. Jackson, Chris Manafas, and Stefan Roth Georgios Tzimiropoulos. 3D human body reconstruction from a single image via volumetric regression. In *European Conference on Computer Vision Workshops (ECCVw)*, volume 11132, pages 64–77, 2018.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, volume 9906, pages 694–711, 2016.

[9] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477, 2020.

[10] Meysam Madadi, Hugo Bertiche, Wafa Bouzouita, Isabelle Guyon, and Sergio Escalera. Learning cloth dynamics: 3D + texture garment reconstruction benchmark. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR*, volume 133, pages 57–76, 2021.

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 483–499, 2016.

[12] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021.

[13] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[14] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020.

[15] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[16] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019.