VS: Reconstructing Clothed 3D Human from Single Image via Vertex Shift

Supplementary Material

7. Details of Networks

In this section, we describe the details of the ShiftEstNet ($\mathcal{G}_s$) used in the StrechVS module and the DeformNet ($\mathcal{G}_m$) used in the RefineVS module. Please recall these two networks in Fig. 2.

- **ShiftEstNet**

The structure of ShiftEstNet is illustrated in Fig 9. ShiftEstNet consists of 5 convolutional layers, each followed by a LeakyReLU and a BatchNorm. ShiftEstNet takes the body normal maps ($\mathcal{R}$) and the clothing normal maps ($\mathcal{N}$) as input and outputs the shift fields ($\mathcal{S}$). In general, larger resolutions typically require deeper networks and longer runtime. Table 5 demonstrates the reconstruction quality of VS on three datasets when ShiftEstNet is employed with different input resolutions. It can be observed that when utilizing a 5-layer CNN, the reconstruction quality of VS decreases with larger input resolutions for ShiftEstNet. Therefore, the input normal maps are resized to 128×128 in our implementation. With this input resolution, the output shift fields' size is 86×86. To compute the losses in Equ 9, we resize the output shift fields to 128×128 using a bilinear interpolation.

Training Details. As described in Sec. 3.2, ShiftEstNet is trained in an optimized manner (or referred to as "one-instance-one-training"). On the one hand, this training manner does not require a large number of training samples; on the other hand, it allows obtaining the optimal network parameters customized for each instance. We use Adam as an optimizer, with a fixed learning rate of 0.0002. The weights for the loss functions in Equ 9 are set as follows: $\lambda_{\mathrm{w}} = 0.3$, $\lambda_{\mathrm{c}} = 1$, and $\lambda_{\mathrm{s}} = 1$. For each instance, we train it for 500 iterations.

- **DeformNet**

The structure of DefromNet is illustrated in Fig 10. DeformNet consists of 11 mesh convolutional layers, each followed by a LeakyReLU and an InstanceNorm. As mentioned in Sec. 3.3, the input of DeformNet is a 2-manifold triangular mesh, which guarantees that each edge in the mesh is exactly adjacent to four other edges. We refer to the sub-mesh consisting of an edge and its four adjacent edges as a fragment (see Fig. 10). The mesh convolution operation is conducted on each of the fragments.

Training Details. Adam is adopted as the optimizer for training the DeformNet. The learning rate is initiated as 0.0004, and decayed by 0.00008 every epoch after 2 epochs. We totally train DeformNet for 50 epochs with a batch size of 1. The weights for the loss functions in Equ 19 are set as follows: $\lambda_{\mathrm{d}} = 1$, $\lambda_{\mathrm{n}} = 1$, and $\lambda_{\mathrm{e}} = 10$. The number of vertices on a mesh is increased to 20,000 using the method proposed in [17]. The network is trained on an NVIDIA GeForce 3080. Limited by the GPU's memory, we split a mesh into 8 parts and utilize the PartMesh data structure [13] for training.

8. Details of Perceptual Study

In the perceptual evaluation of ECON [49], ICON [48], and TeCH [18], participants are asked to choose the reconstruction they perceive as more realistic between a baseline method and ECON/ICON/TeCH. However, we found that participants can easily "guess" which result corresponds to which method in this 1-vs-1 strategy, resulting in potential perceptual biases.

In our perceptual study, we randomly mix the results produced by all the methods involved in the comparison together, so that the probability of participants guessing which result is produced by which method is greatly reduced, thus effectively reducing the perceptual bias. We design a perceptual evaluation tool where reconstructions produced by all methods are blindly placed on one page. Moreover, using this tool, participants can use a mouse to freely rotate the models from all directions as well as zoom in or out of the models with the mouse wheel. In this way, participants can meticulously examine the quality of each reconstructed 3D model. A screenshot of the perceptual evaluation tool is illustrated in Fig. 11. During the perceptual evaluation, participants are requested to choose the best reconstruction for each input image after a thorough examination of all reconstruction models. We argue that our strategy and tool can yield more fair perceptual valuation results.

9. More Visual Results and Comparisons

In this section, we provide more visual results and comparisons on both benchmark datasets and in-the-wild images. Fig. 12 demonstrates the results produced by our VS. Video1 shows more reconstructions with a rotating virtual camera, for more results. Fig 13, Fig 14, and Fig 15 present the comparisons with PIFu [36], PIFuHD [37], PaMIR [54], ICON [48], and ECON [49] on the THuman2.0 [50], CAPE [30] and RenderPeople [35] datasets, respectively. Fig 16, Fig 17, Fig 18, and Fig 19 illustrate comparisons on in-the-wild images under scenarios of challenging poses, loose clothing, challenging poses + loose clothing, and fashion clothing, respectively. The results of the separate comparison between VS and the latest state-of-the-art method, ECON [49], are presented in Video2.

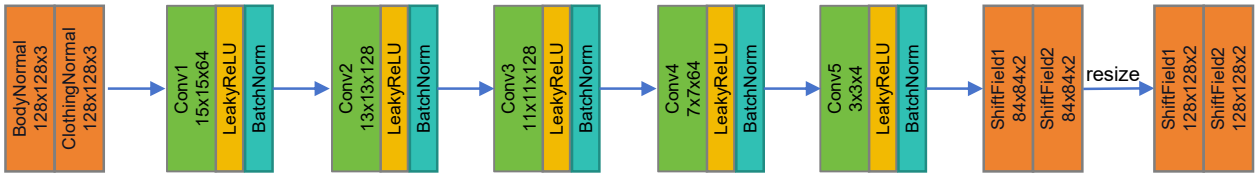

Figure 9. Structure of ShiftEstNet. ShiftEstNet is a 5-layer convolutional neural network that takes the body normal map and the clothing normal map as input and outputs the shift fields.

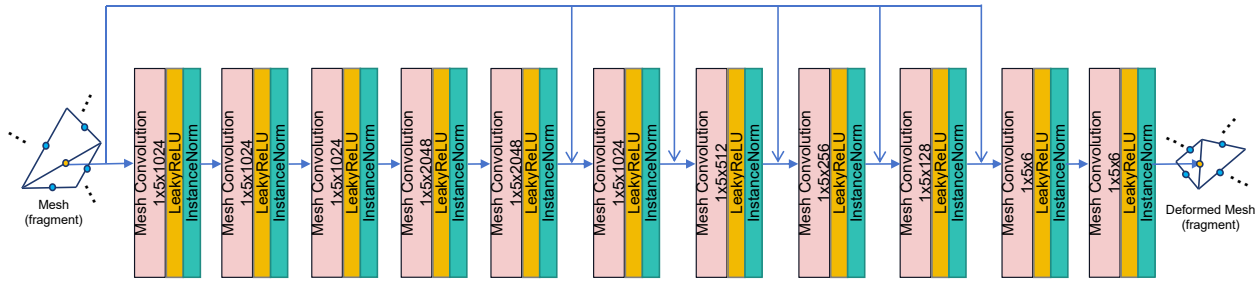

Figure 10. Structure of DeformNet. DeformNet takes a mesh as input and outputs a deformed mesh. We refer to the sub-mesh consisting of an edge and its four adjacent edges as a fragment. The mesh convolution operation is conducted on each of the fragments.

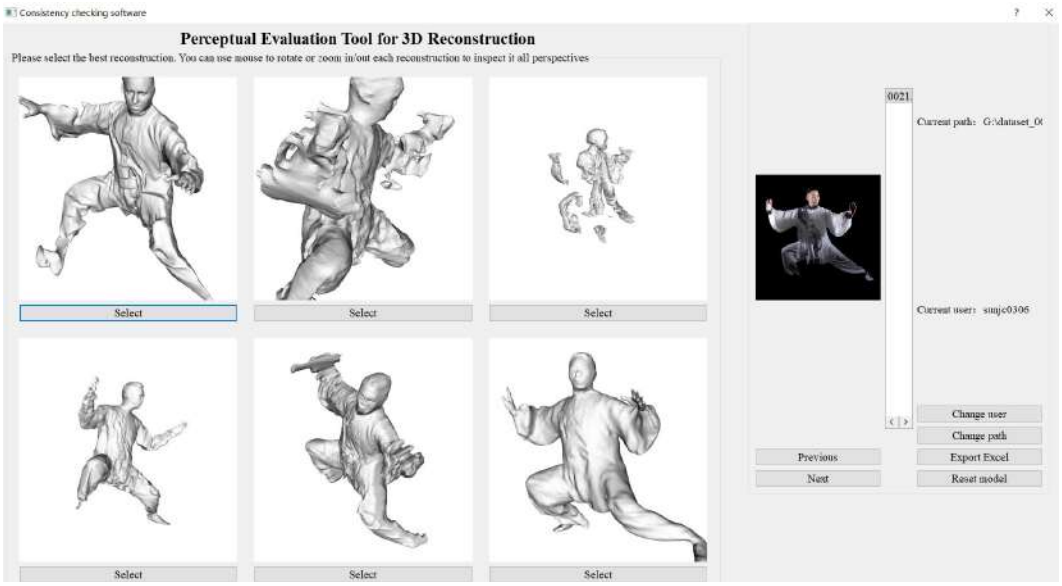

Figure 11. Tool for perceptual evaluation. In this tool, 3D models reconstructed by different methods are blindly placed on one page. Participants can use the mouse to freely rotate the models from the X, Y, and Z directions, as well as zoom in or out of the models using the mouse wheel. In this way, participants can meticulously examine the quality of each reconstructed 3D model. Participants are requested to choose the best reconstruction for each input image after a thorough examination of all reconstruction models.

	THuman 2.0				CAPE				RenderPeople			
Resolution	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$	$\varepsilon_{cd}\downarrow$	$\varepsilon_{p2s}\downarrow$	$\varepsilon_{cos}\downarrow$	$\varepsilon_{l2}\downarrow$
128×128	0.979	0.915	0.0645	0.2469	0.935	0.907	0.0404	0.1743	1.209	1.040	0.0627	0.2443
196×196	1.213	1.176	0.0675	0.2873	1.173	1.112	0.0454	0.2054	1.477	1.231	0.0689	0.2984
196×196	1.213	1.176	0.0675	0.2873	1.174	1.119	0.0455	0.2067	1.489	1.265	0.0713	0.3052
256×256	1.212	1.179	0.0677	0.2886	1.174	1.123	0.0457	0.2074	1.496	1.275	0.0743	0.3102
512×512	1.222	1.197	0.0683	0.2952	1.177	1.135	0.0466	0.2106	1.545	1.334	0.0771	0.3233

Table 5. Reconstruction quality of VS on three datasets when ShiftEstNet is employed with different input resolutions.

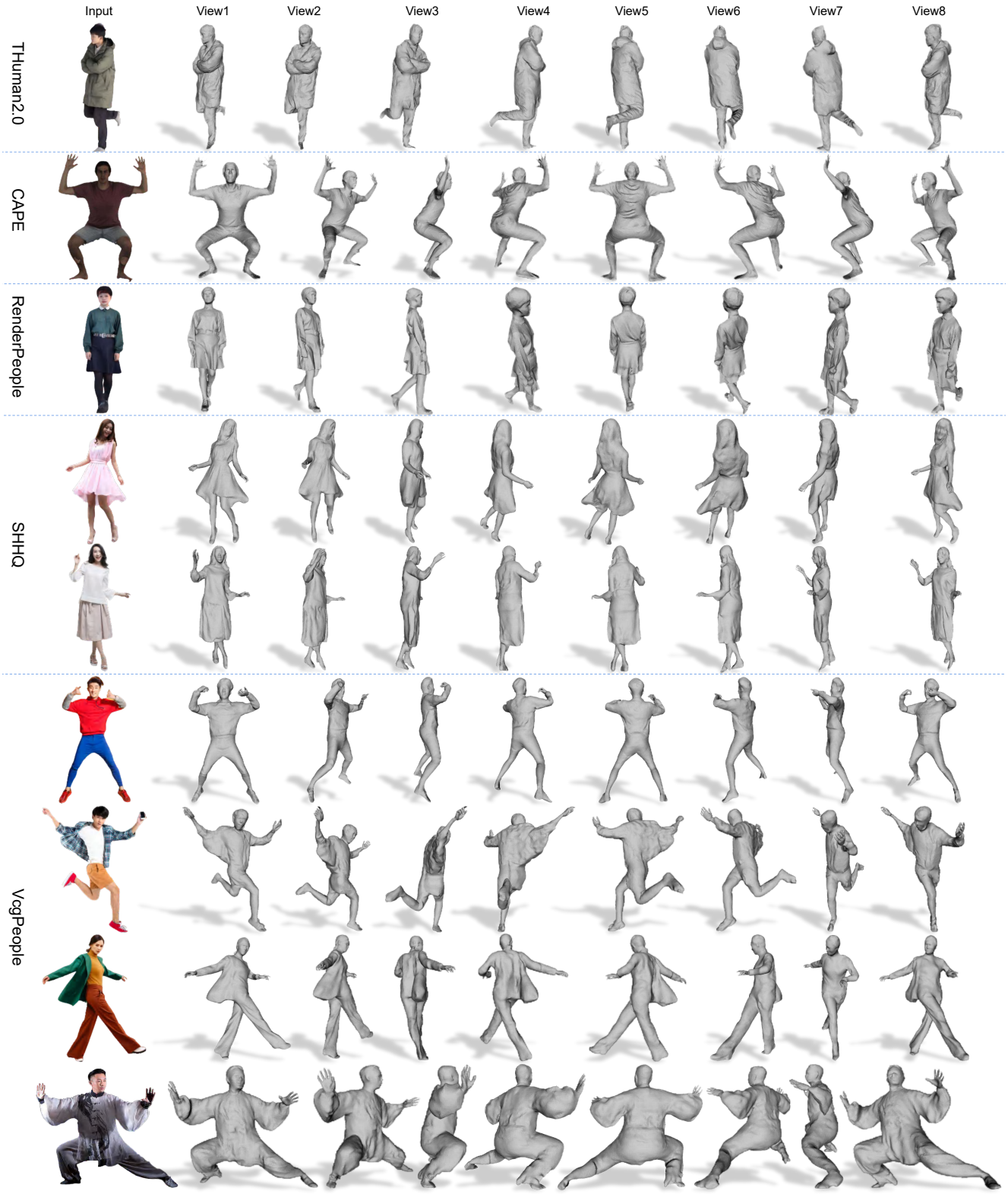

Figure 12. Reconstructions produced by our VS on three benchmarks (i.e., THuman2.0, CAPE, and RenderPoeple) and two in-the-wild datasets (i.e., SHHQ and VcgPoeple). To comprehensively inspect reconstructions from all perspectives, we rotate each reconstruction through a full circle and show it from 8 viewpoints. Please zoom in to see the details.

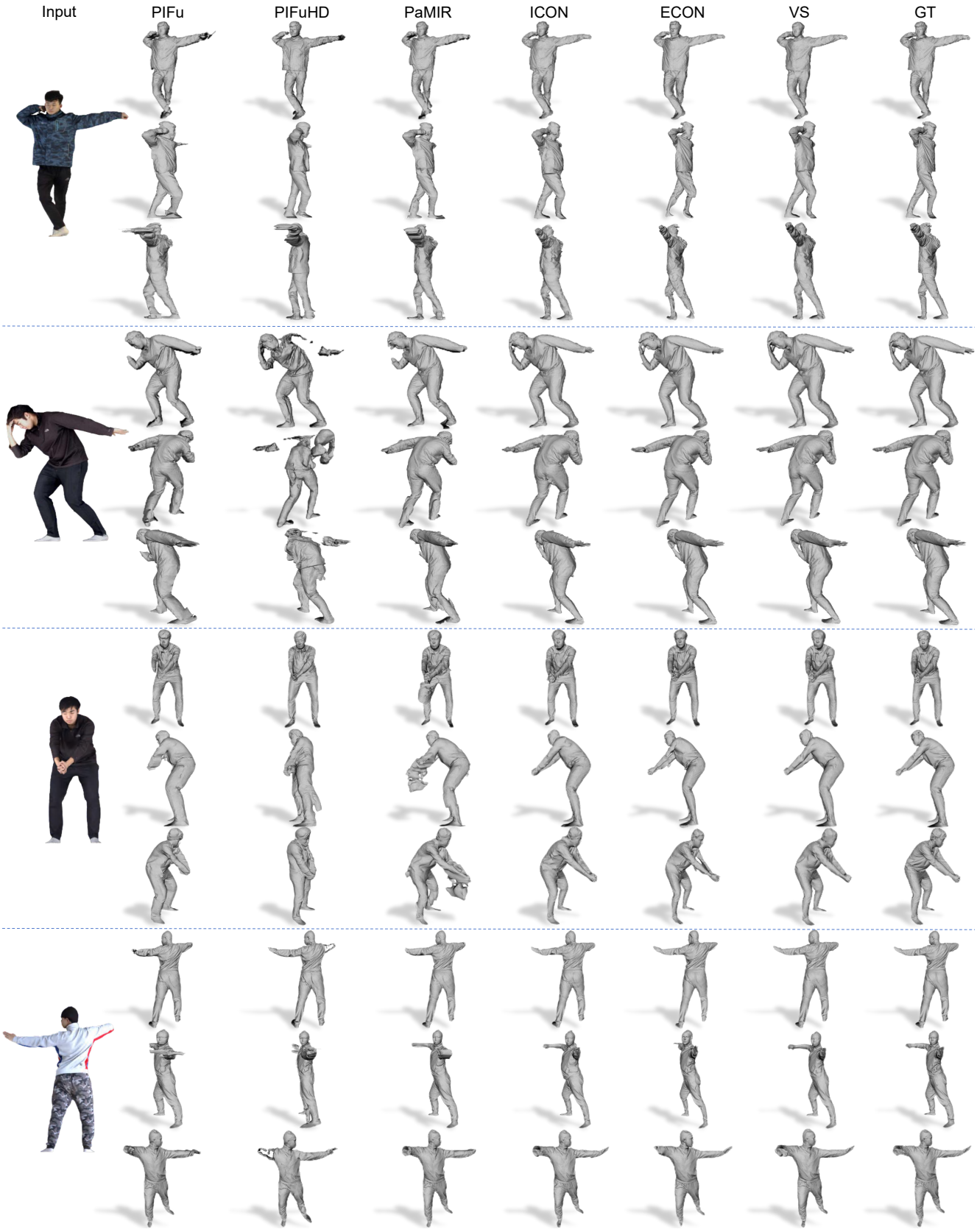

Figure 13. Comparisons with other SOTAs on the THuman2.0 dataset. We show each result from 3 viewpoints. Please zoom in to see the details.

Figure 14. Comparisons with other SOTAs on the CAPE dataset. We show each result from 3 viewpoints. Please zoom in to see the details.

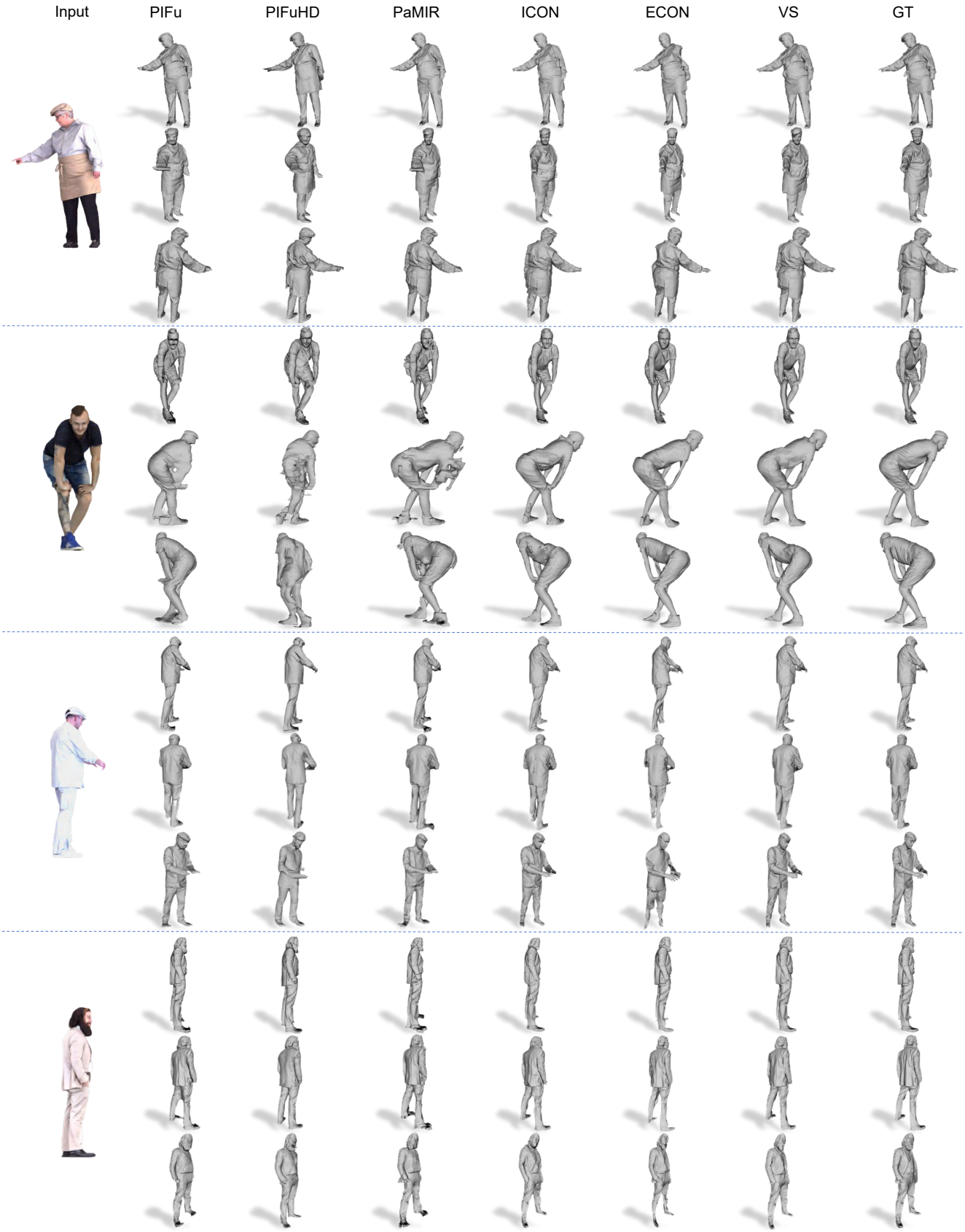

Figure 15. Comparisons with other SOTAs on the RenderPeople dataset. We show each result from 3 viewpoints. Please zoom in to see the details.

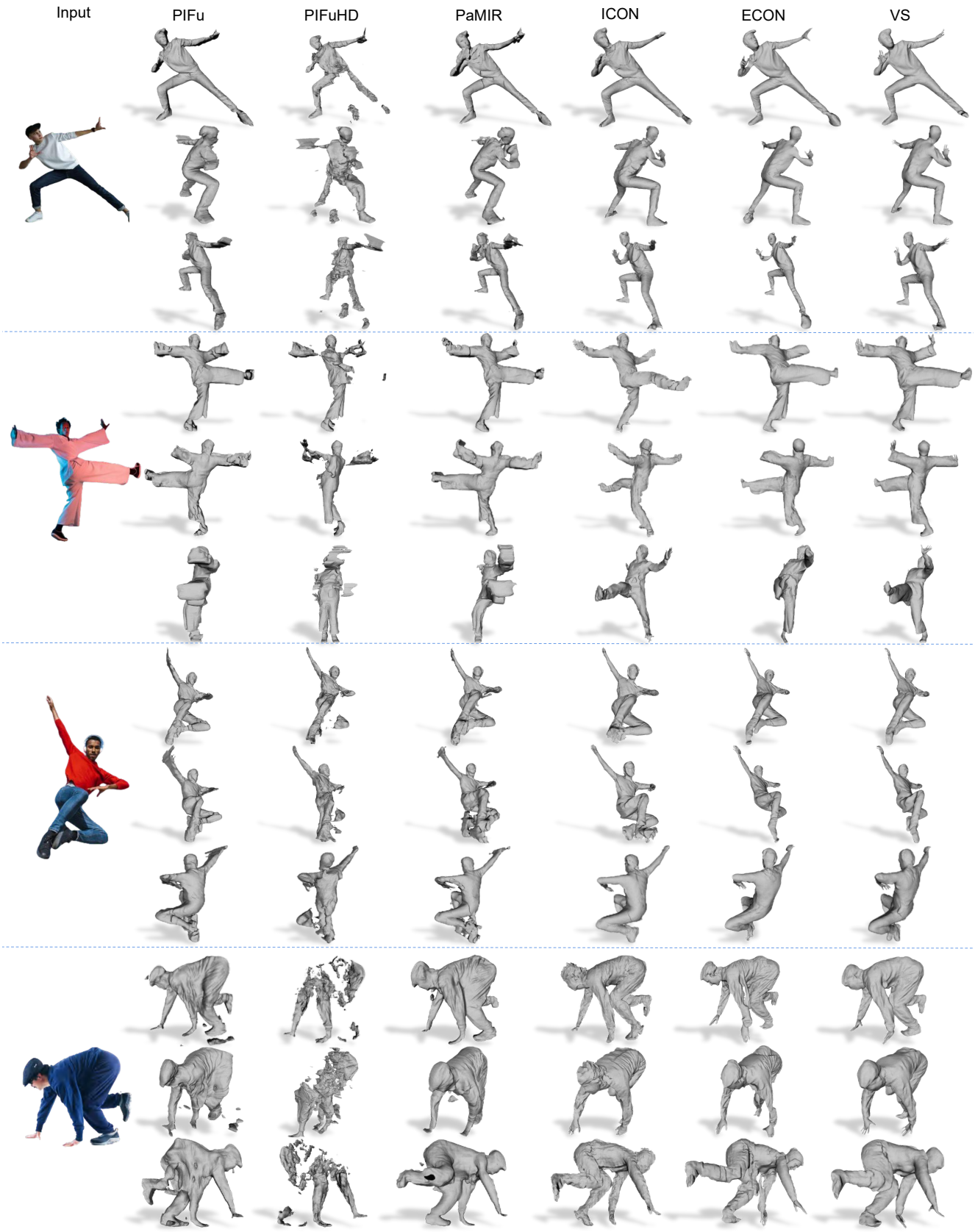

Figure 16. Comparisons against other SOTAs on in-the-wild images with challenging poses. We show each result from 3 viewpoints. Please zoom in to see the details.

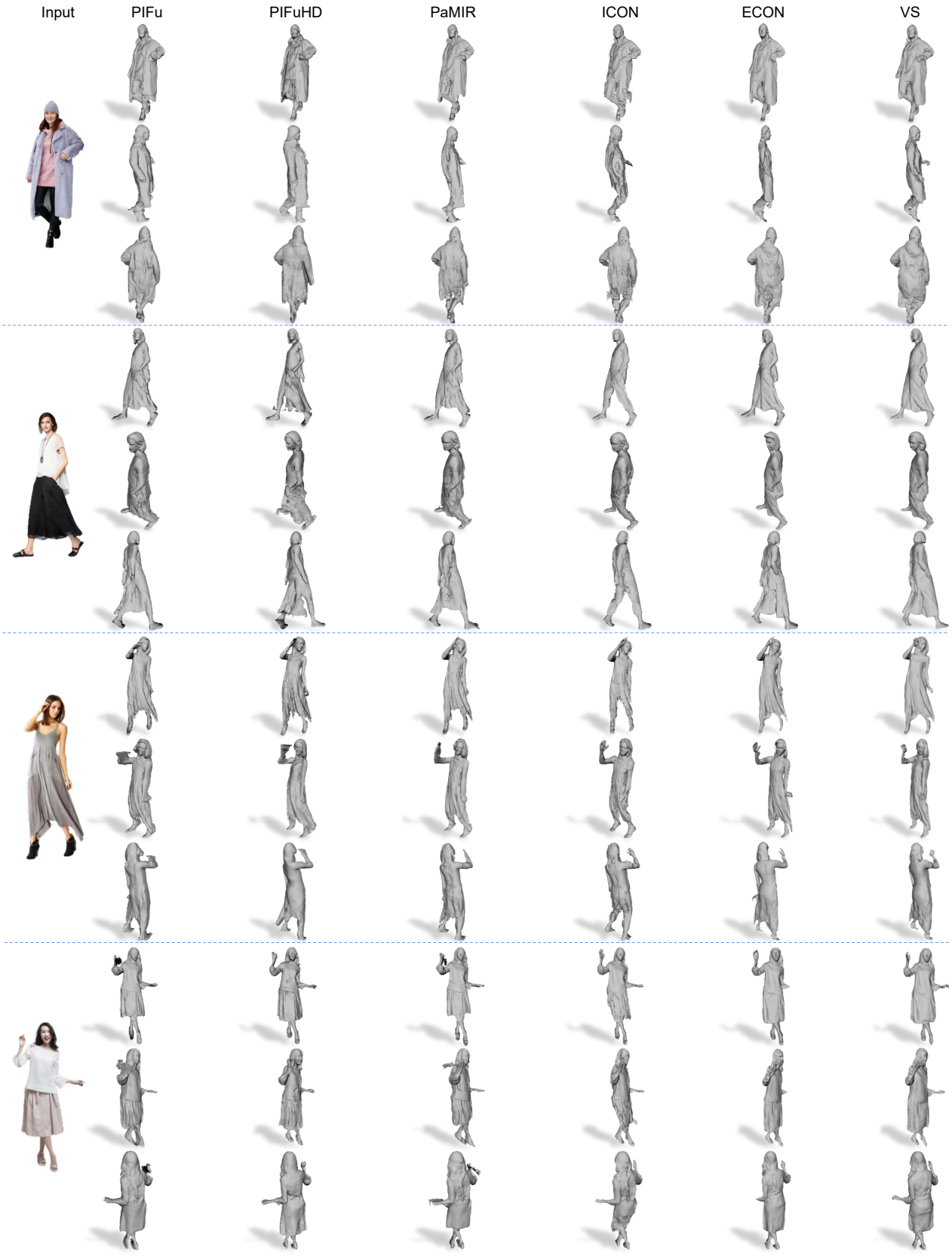

Figure 17. Comparisons against other SOTAs on in-the-wild images with loose clothing. We show each result from 3 viewpoints. Please zoom in to see the details.

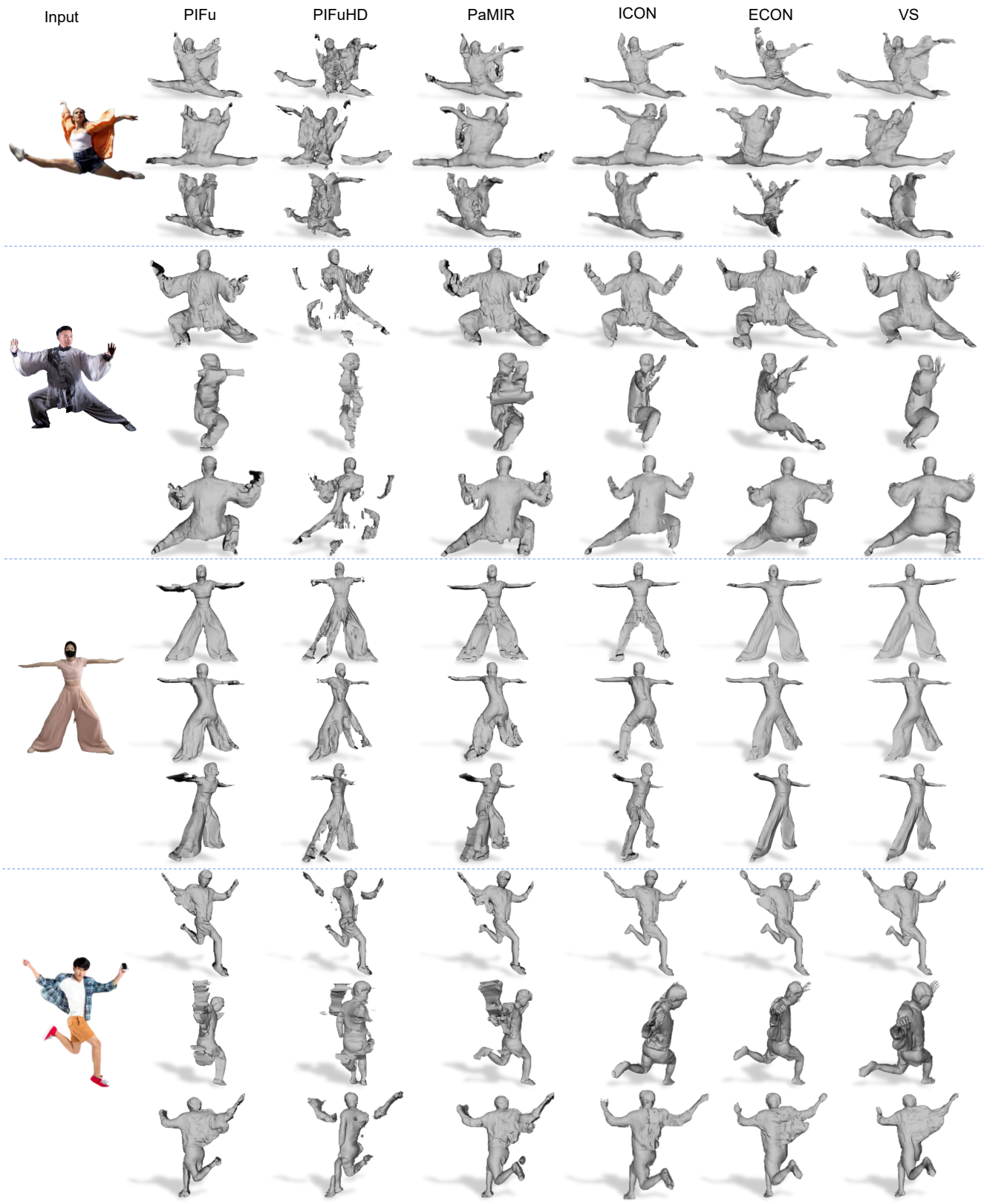

Figure 18. Comparisons against other SOTAs on in-the-wild images with both challenging poses and loose clothing. We show each result from 3 viewpoints. Please zoom in to see the details.

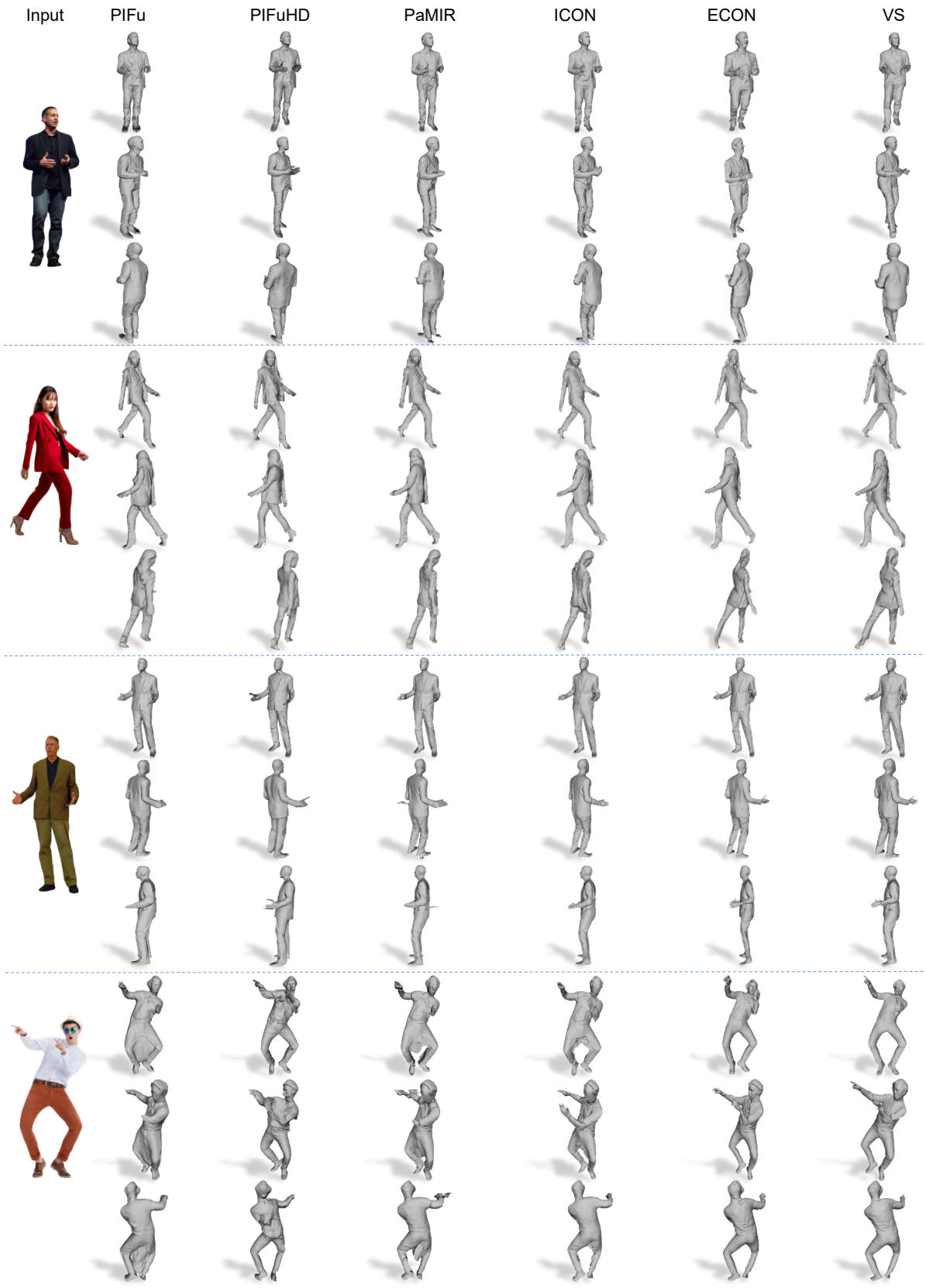

Figure 19. Comparisons against other SOTAs on in-the-wild fashion images. We show each result from 3 viewpoints. Please zoom in to see the details.