Super-resolution 3D Human Shape from a Single Low-Resolution Image - Supplementary Material

Marco Pesavento, Marco Volino, and Adrian Hilton[1]

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey,
Guildford, UK {m.pesavento,m.volino,a.hilton}@surrey.ac.uk

This document presents additional information that supplements the main paper. First, examples of some limitations of the proposed approach are illustrated (Section 1). The architecture of the designed image feature extractor is then explained (Section 2). Different algorithms to create the low-resolution shapes from the high-resolution ones are evaluated (Section 3). The different decimated low-resolution shapes obtained for the 'degradation factor' ablation study of Section 4.2 of the main paper are illustrated (Section 4). Since the related approaches that reconstruct 3D human shape from a single image use higher resolution images than the ones we use in the main paper, we evaluate the approaches using higher resolution images (Section 5). In the main paper, we only show the front of the model since we want to highlight the super-resolution effect of our approach on the shape reconstructed from the low-resolution input image. In this document, we present examples of the back and the sides of the model. Moreover, we show the point-to-surface error maps computed with respect to the ground-truth shapes (Section 6). Finally, more visual results are illustrated (Section 7).

1 Limitations

As explained in the main paper, one of the limitation of SuRS is that it cannot super-resolve parts of the human body that are not present in the input image. No information about the unseen parts has been fed to the network, which estimates a lower resolution representation of those as shown in Fig. 1 a). However, SuRS still achieves higher resolution of the unseen parts compared to other approaches that leverage only RGB images (Section 6). This is related to the fact that SuRS can reproduce fine details that are not clear in the input image thanks to the learning of the map from the low to the high-resolution shape. SuRS tries to increase also the resolution of the geometry of the unseen parts, inferring the difference learnt during training. Since the features of these parts cannot be extracted, the resolution is still lower than that of the seen parts. As a future work, we will try to improve the resolution of the unseen parts.

Like existing approaches, another limitation of SuRS is that it may create unnatural body parts when the features of the input human body significantly differ from the ones present in the training data. Fig. 1 b) shows an example of how SuRS fails to reconstruct sunglasses. Another problem is related to depth ambiguity shown in Fig. 1 c). If for example the input model has naked body parts that are different than faces and hands (like the chest or the belly), SuRS may

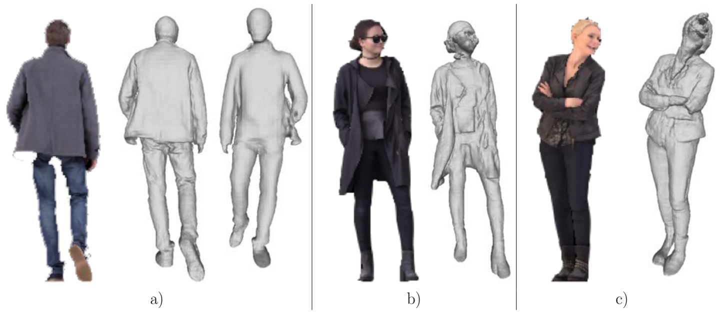

Fig. 1. Examples of limitations of SuRS. a) hidden body parts are not super-resolved; b) fail to reconstruct objects not seen during training; c) depth ambiguity caused by visibility of the chest, which is confused with the face.

mistakenly classify them as part of the face, introducing artefacts in the final shape. This is worsen by the fact that the training data do not contain these features since there are no human models with naked body parts in the training images. These last two problems can be solved by augmenting the training dataset with additional models.

2 Architecture of image feature extractor

We design the image feature extractor architecture as the combination of a novel U-Net [10] architecture and a stacked hourglass network [9]. Given an input image of size $N_I \times N_I$, the novel U-Net part extracts features with the resolution

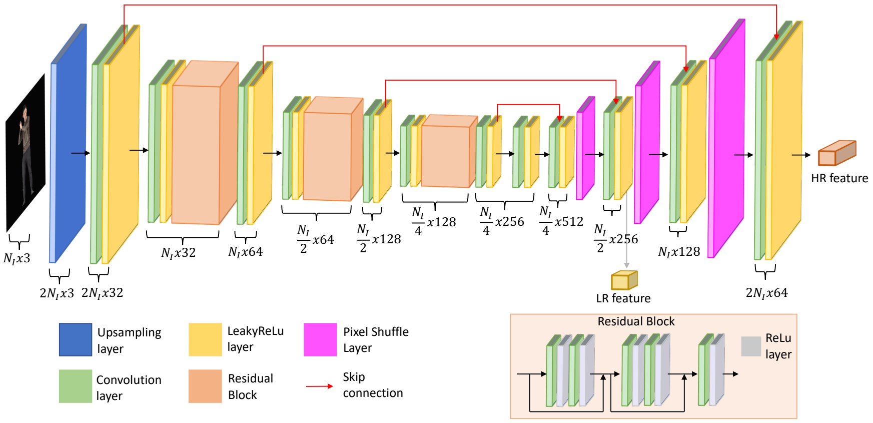

Fig. 2. U-net architecture of the image feature extractor module. The low-resolution output feature are processed by the stacked hourglass part of the feature extractor while the high-resolution output feature are processed by a convolutional layer.

Table 1. Quantitative comparisons between different algorithms applied to retrieve the LR shape from its HR counterpart.

Algorithm	**THuman2.0**			**3D people**		
	CD	Normal	P2S	CD	Normal	P2S
TwoStep [3]	0.959	0.1152	1.207	1.058	0.1232	1.273
Laplacian [13]	0.978	0.1127	1.232	1.101	0.1236	1.332
Subdivision [4]	0.964	0.1099	1.198	1.111	0.1216	1.330
Decimation [5]	**0.931**	**0.1065**	**1.151**	**1.057**	**0.1127**	**1.247**

of $\frac{N_I}{2} \times \frac{N_I}{2}$ to maintain holistic reasoning as well as features with the higher resolution of $2N_I \times 2N_I$. We design the U-Net with convolution layers and residual blocks with skip connections. Fig. 2 shows the architecture of the U-Net with the dimensions of features for each layer. U-Net uses skip connections between matched convolution and deconvolution layers to balance global aspects of the frame with local ones. The local and high frequency details are preserved in the created feature map [17]. The retrieved features are then processed by a stacked hourglass architecture, which has been proved to be efficient for surface reconstruction [11]. We adapt the stacked hourglass network [9] with modifications proposed by [7]. The low-resolution feature is processed by 3 stacks while the high-resolution one is processed by just a convolution layer.

To train SuRS, Adam optimizer is adopted with a learning rate of 1e-4 and the batch size has been set to 8. The training dataset has been augmented by flipping and translating the data.

3 Shape Simplification Algorithms

We evaluate our approach by training with low-resolution shapes obtained by applying different algorithms of simplification of the high-resolution shapes. Namely, we apply TwoStep smoothing algorithm [3], Laplacian smoothing algorithm [13], Butterfly Subdivision algorithm [4] and Quadric Edge Collapse Decimation algorithm [5]. Fig. 3 shows an example of the low-resolution shape obtained by applying the considered algorithm to its high-resolution counterpart. Quantitative (Table 1) and qualitative (Fig. 4) results shows that SuRS achieves the best performance when the Quadric Edge Collapse Decimation algorithm is applied with the highest figures for all the considered metrics. If the other algorithms are applied to create the LR shapes, noise is introduced in the reconstructed meshes. The shapes reconstructed by training SuRS with the LR shapes created by applying the Quadric Edge Collapse Decimation algorithm present more natural details.

4 Factors of Decimation of High-Resolution Shape

In an ablation study of the main paper, we evaluate our approach by creating the low-resolution ground-truth with different factors of decimation. From the high-resolution shape with $\sim 400k$ faces, we applied quadric edge collapse decimation to create different low-resolution surfaces with $100k, 50k, 10k$ and $1k$

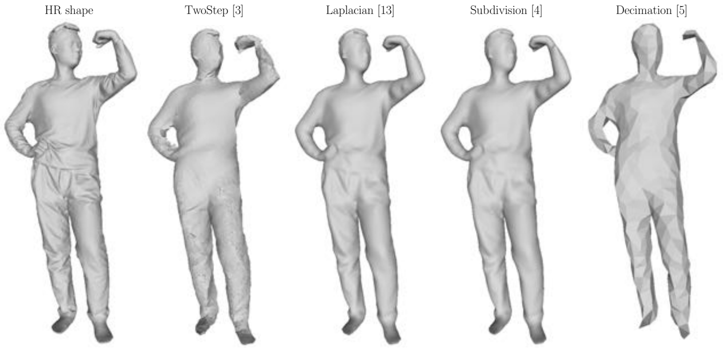

Fig. 3. Example of low-resolution shapes obtained by remeshing the high-resolution shape with different algorithms.

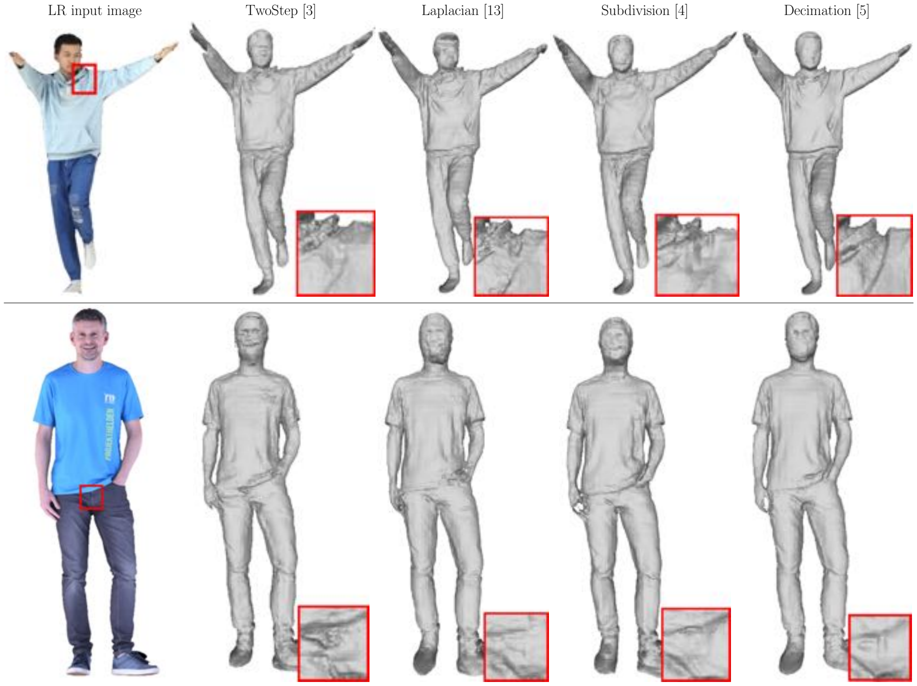

Fig. 4. Visual comparisons between different algorithms applied to retrieve the LR shape from its HR counterpart. The upper model is from THuman2.0, the below one is from 3DPeople.

faces. Examples of these are shown in Fig. 5. As expected, the highest displacement between the high-resolution shape and its low-resolution counterpart is achieved when the number of faces is $1k$.

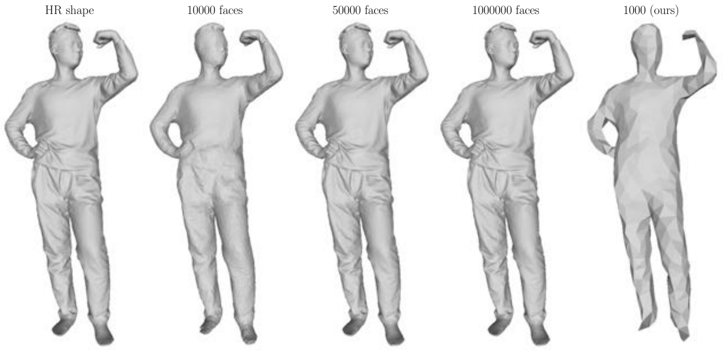

Fig. 5. Example of low-resolution shapes obtained by simplifying the high-resolution shape with different factors of decimation.

5 Higher resolution images examples

Since the studied methods use higher resolution images in their papers, we additionally train and test using images of higher sizes ($N_i = 512, 1024$). We also train and test PIFu with 512×512 images obtained by upscaling 256×256 images by 2× with an image SR network [8] (PIFu$^{\mathrm{up}}$). Similarly, we give as input to PIFuHD a 1024×1024 image obtained by upscaling a 256×256 image by 4× (PIFuHD$^{\mathrm{up}}$, PIFuHD$^{\mathrm{up}}_{\mathrm{no}}$). We evaluate another approach that do not use implicit function but it leverages normal and displacement maps in the reconstruction (Tex2Shape [2]). This is tested only on 1024×1024 images since the training code is not available.

Qualitative evaluation: Fig. 6 illustrates the same shapes of Fig. 8 of the

Table 2. Quantitative comparisons between state-of-the-art approaches with higher sizes ($N_I = 512, 1024$) input image for training and testing. The highest scores are highlighted in red while the second highest scores are blue. The red-bold figures are the highest among all the values.

N_I	Methods	THuman2.0			3D people		
		CD	Normal	P2S	CD	Normal	P2S
512	DeepHuman [16]	1.918	0.1543	2.042	1.697	0.1223	1.709
	PIFu [11]	1.244	0.1299	1.365	1.391	0.1147	1.309
	PIFu$^{\mathrm{up}}$	1.281	0.1395	1.382	1.476	0.1217	1.532
	PIFuHD$_{\mathrm{no}}$	1.086	0.1113	1.114	1.220	0.1145	1.241
	PIFuHD [12]	0.848	0.0975	0.859	0.782	0.0847	0.772
	PaMIR [15]	1.712	0.1314	1.181	1.637	0.1368	1.359
	Geo-PIFu [6]	1.652	0.1413	1.489	1.754	0.1611	1.652
	SuRS (ours)	0.869	0.1106	1.113	0.945	0.1122	1.156
1024	DeepHuman [16]	1.902	0.1450	1.991	1.650	0.1155	1.697
	Tex2Shape [2]	1.780	0.1562	1.798	1.405	0.1490	1.372
	PIFu [11]	1.185	0.1178	1.133	1.329	0.1160	1.238
	PIFuHD$_{\mathrm{no}}$	0.982	0.1063	0.987	1.143	0.1113	1.173
	PIFuHD$^{\mathrm{up}}_{\mathrm{no}}$	1.014	0.1121	1.151	1.275	0.1113	1.284
	PIFuHD [12]	**0.761**	**0.0933**	**0.765**	**0.770**	**0.0806**	**0.769**
	PIFuHD$^{\mathrm{up}}$	0.939	0.1086	0.914	0.807	0.1090	0.784
	PaMIR [15]	1.460	0.1296	1.412	1.219	0.1125	1.210
	Geo-PIFu [6]	1.462	0.1339	1.416	1.691	0.1698	1.619
	SuRS (ours)	0.791	0.1053	0.930	0.802	0.1089	1.034

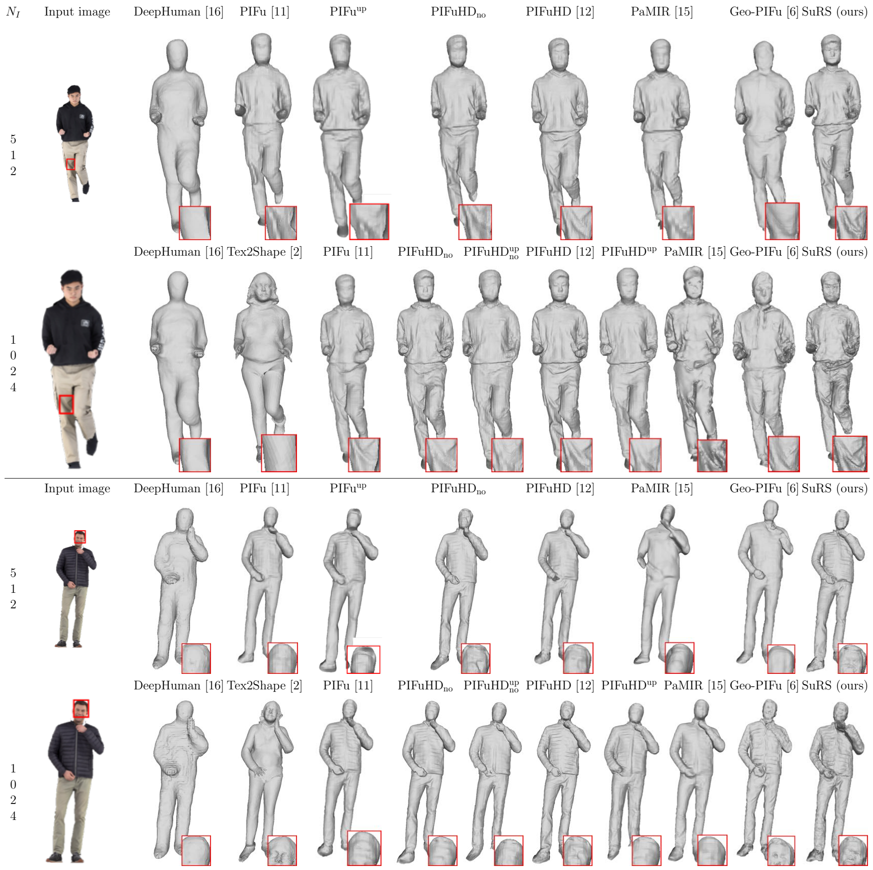

Fig. 6. Visual comparisons using higher sizes of the input image for training and testing. The upper model is from THuman2.0, the below one is from 3DPeople.

main paper reconstructed from higher size images. When the input image is 512 × 512, the shapes reconstructed by SuRS contain the highest level of fine detail. When the related works are trained and tested with HR 1024 × 1024 images, our approach with LR 256 × 256 and 512 × 512 images can reconstruct surfaces with similar level of details as PIFuHD even if the resolution of the image is 4× lower and no auxiliary data is leveraged. PIFuHD and Geo-PIFu are the only related approaches that can achieve similar resolution as SuRS in the reconstructed shape with 1024 × 1024 input images. When the input image is upsampled from 256 to 1024 and processed by PIFuHD, its reconstructed shape contains blurrier details than SuRS.

Quantitative evaluation: In the case of higher resolution images, SuRS outperforms all the methods that leverage only RGB images in training and testing (Table 2). If all the approaches are considered, our method is second to only PIFuHD with normal maps, which obtain the highest figures when is trained and tested with 1024 × 1024 images and normal maps. This is expected since it uses back normal maps to leverage information of unseen parts of the shape.

6 Back of the models and error maps

In the main paper, we present only the part of the 3D human body that is depicted in the input image in order to highlight the super-resolution effect of SuRS. We now show the back of the 3D models and the point-to-surface error maps between the reconstructed shapes and the ground-truth for all the evaluation studies (both ablation and comparisons) presented in the paper. Fig. 7 shows the back of the 3D models obtained by changing the training configuration of SuRS while Fig. 8 illustrates the point-to-surface error maps. Fig. 9 depicts the back of the 3D models obtained by changing the decimation factor. Fig. 10 shows the point-to-surface error maps for this study. The back of the human shapes obtained by changing the architecture of SuRS are shown in Fig. 11 while their point-to-surface error maps are in Fig. 12. Fig. 13 illustrates examples of the back of the 3D models presented in the comparisons section of the main paper and in Section 5 of the supplementary while Fig. 14 shows the point-to-surface error maps for all the tested approaches. As explained in 1, SuRS does not super-resolve unseen parts of the human body but it is still able to retrieve a coarse representation without introducing significant artefacts. The only approaches that reproduce detail in the unseen parts are PIFuHD [12] and Geo-PIFu [6]: the former leverages normal maps of the back and the front of the human subject while the latter uses parametric models to reconstruct unseen parts.
When the colour of the error maps is blue, the Euclidean distance between a sampled point of the reconstructed surface and the ground-truth shape is close to 0 while the red points are the ones with the highest distance.
Fig. 15 illustrates the back of the shapes obtained with real data. In this case the error map cannot be computed because the ground-truth is missing.
Fig. 16 shows the sides of the 3D models reconstructed from synthetic images by SuRS and presented in the main paper while Fig. 17 illustrates the sides of the 3D models of Fig. 9 of the main paper reconstructed from real images.

7 Additional visual results

We present additional visual results of the ‘comparisons’ study with examples taken from THuman2.0 [14] (Fig. 18) and 3Dpeople [1] (Fig. 19) datasets. These examples demonstrate the efficiency of SuRS with different poses of the human subject as well as with various clothes. We finally train and test SuRS with 128x128 images to show that as the LR input image resolution decreases, the super-resolved shape detail from SuRS will decrease due to the absence of visible details (Fig. 20).

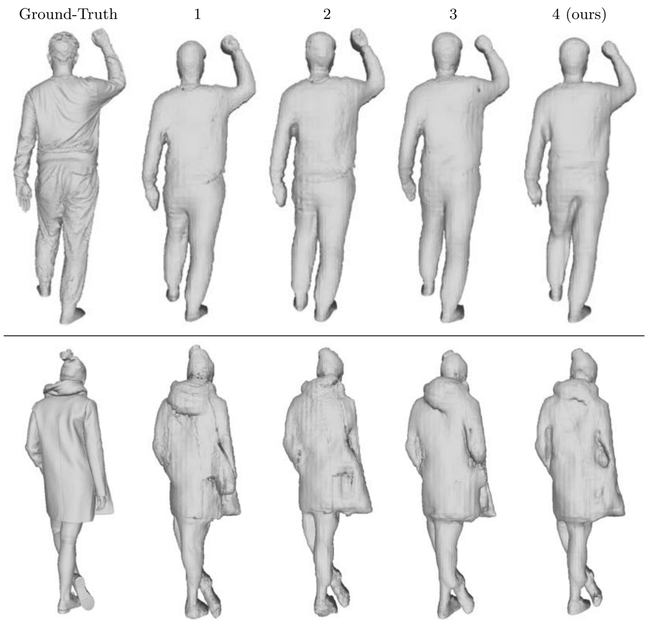

Fig. 7. Back of the models of Fig. 5 of the main paper obtained by changing the configuration of training.

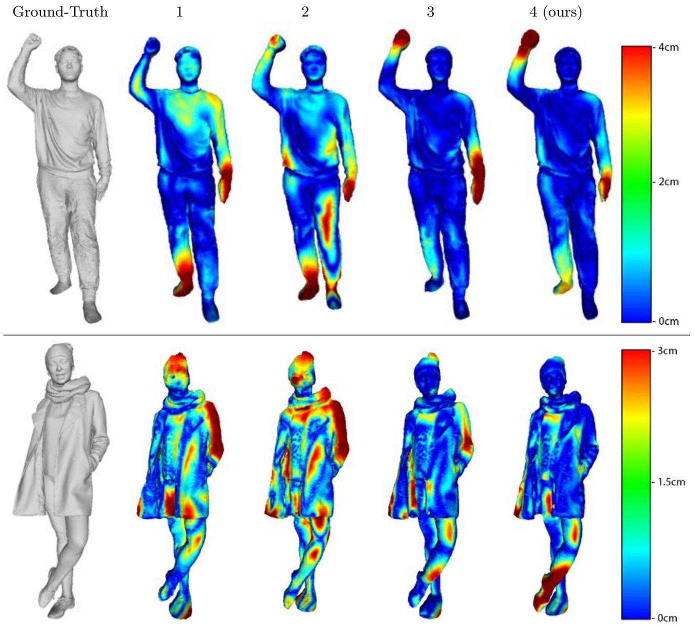

Fig. 8. Point-to-surface error maps of the models of Fig. 5 of the main paper obtained by changing the configuration of training.

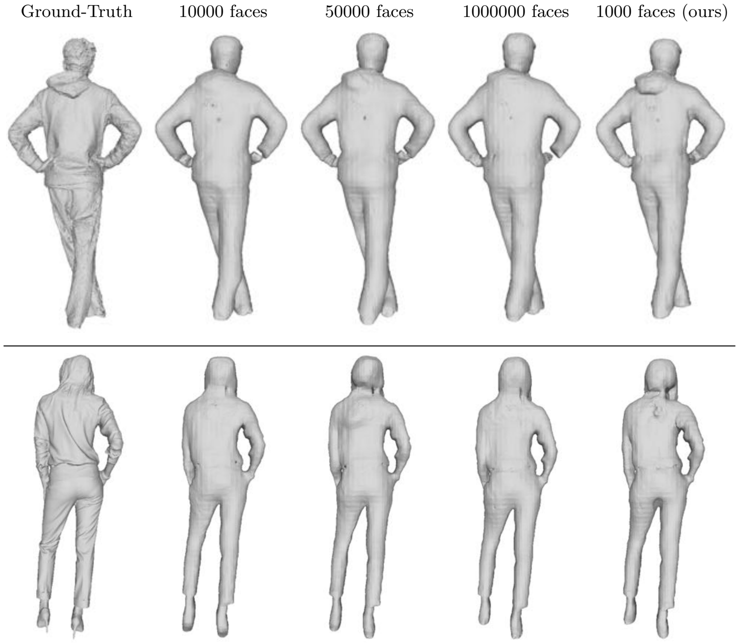

Fig. 9. Back of the models of Fig. 6 of the main paper obtained by using different decimation factors to create the LR ground-truth shape.

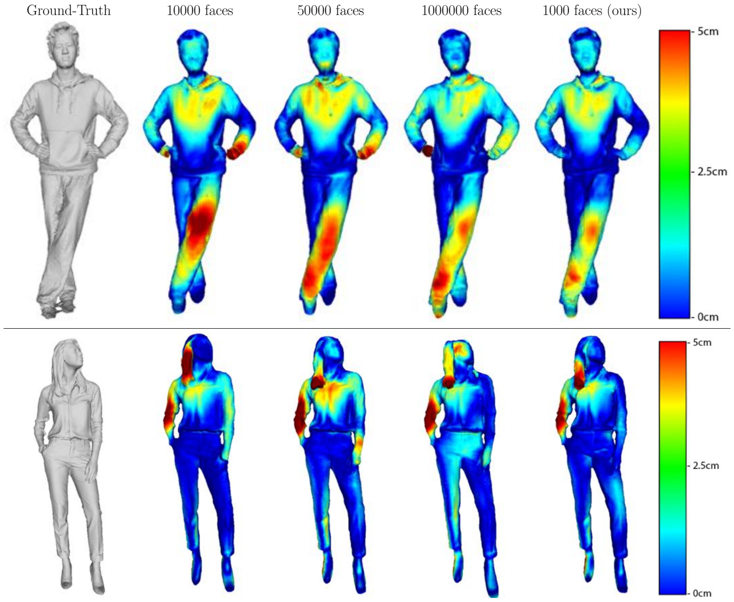

Fig. 10. Point-to-surface error map of the models of Fig. 6 of the main paper obtained by using different decimation factors to create the LR ground-truth shape.

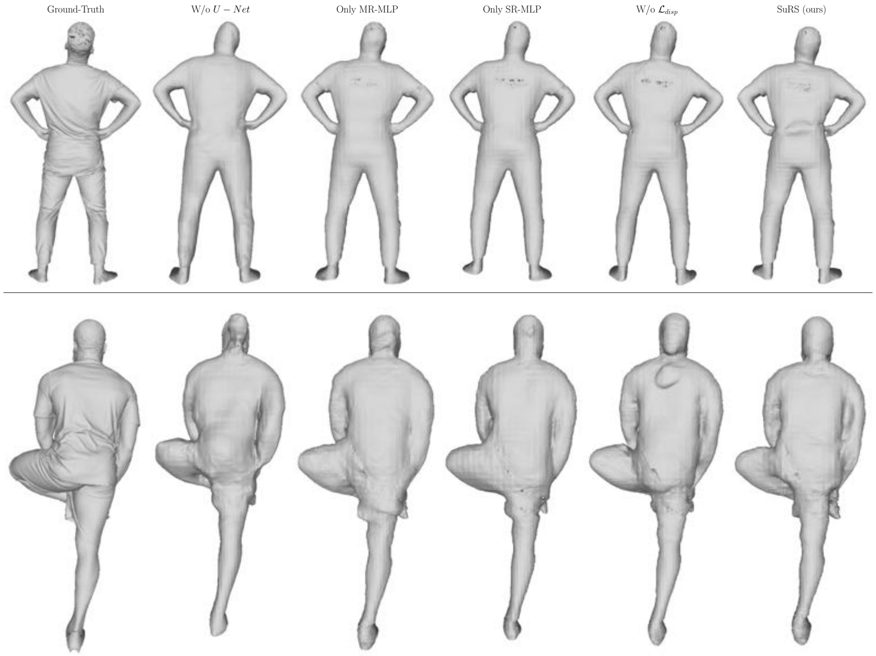

Fig. 11. Back of the models of Fig. 7 of the main paper obtained by changing the architecture of our approach.

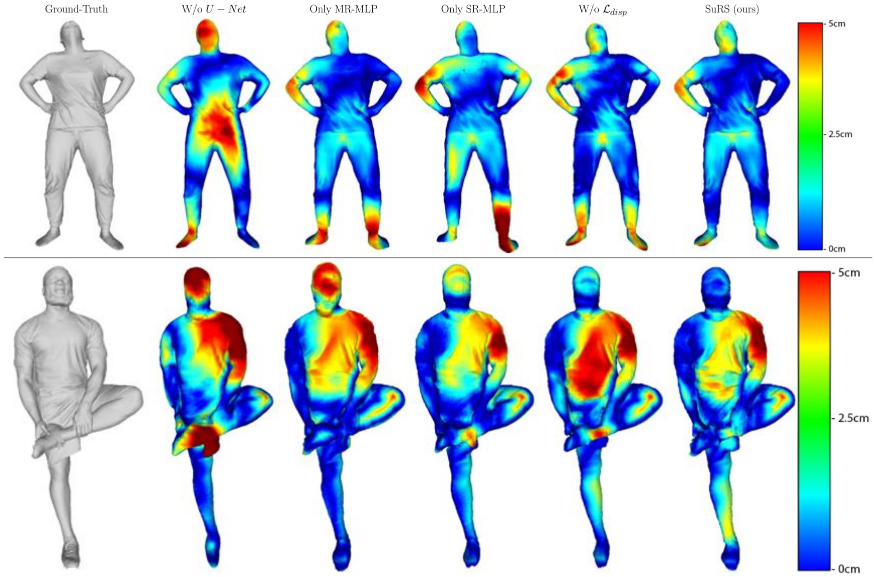

Fig. 12. Point-to-surface error maps of the models of Fig. 7 of the main paper obtained by changing the architecture of our approach.

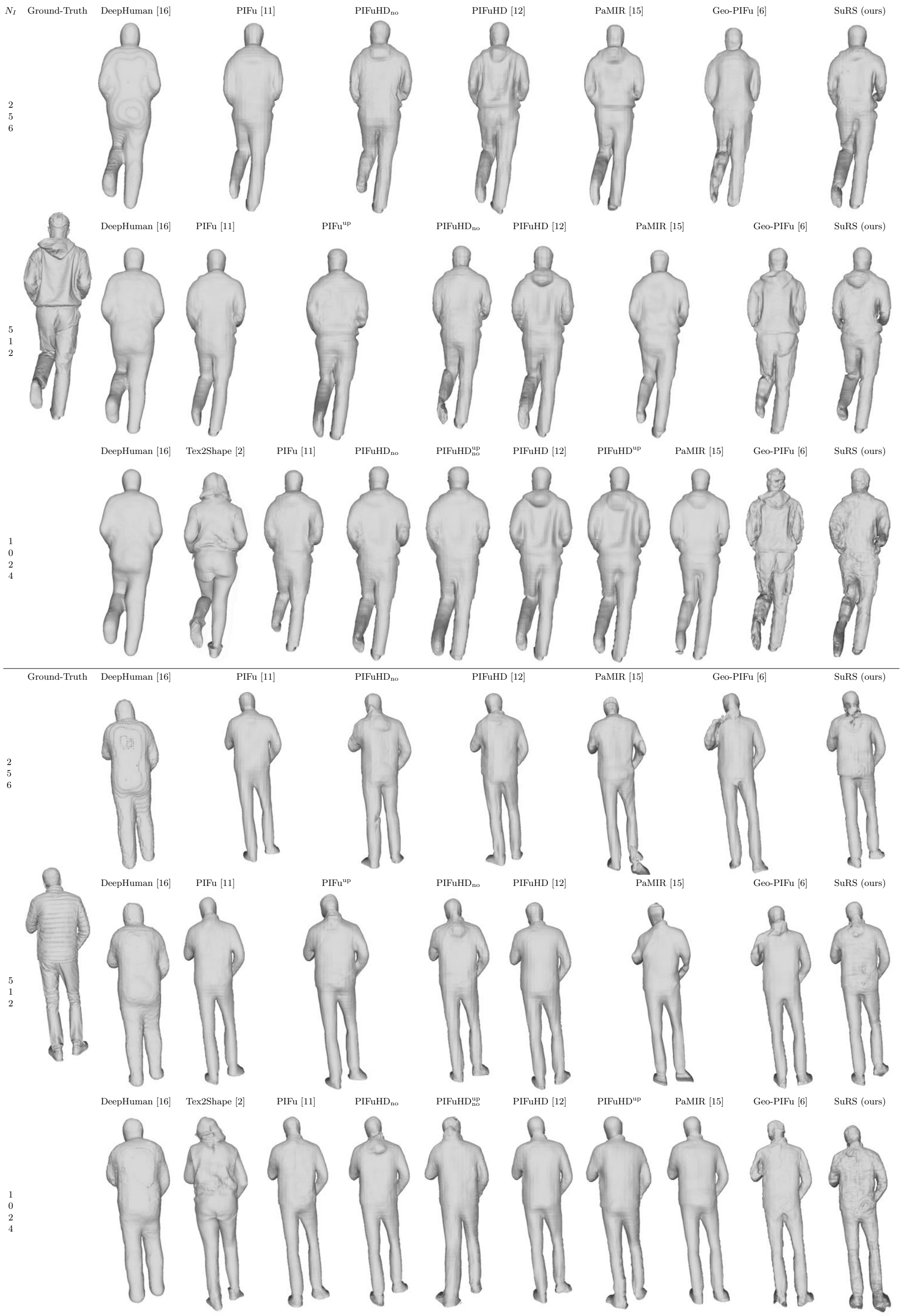

Fig. 13. Back of the models of Fig. 8 of the main paper and of Fig. 6 of the supplementary. The 'no' subscript means that normal maps are not used. The 'up' superscript means that the input image is upscaled from 256 to 1024.

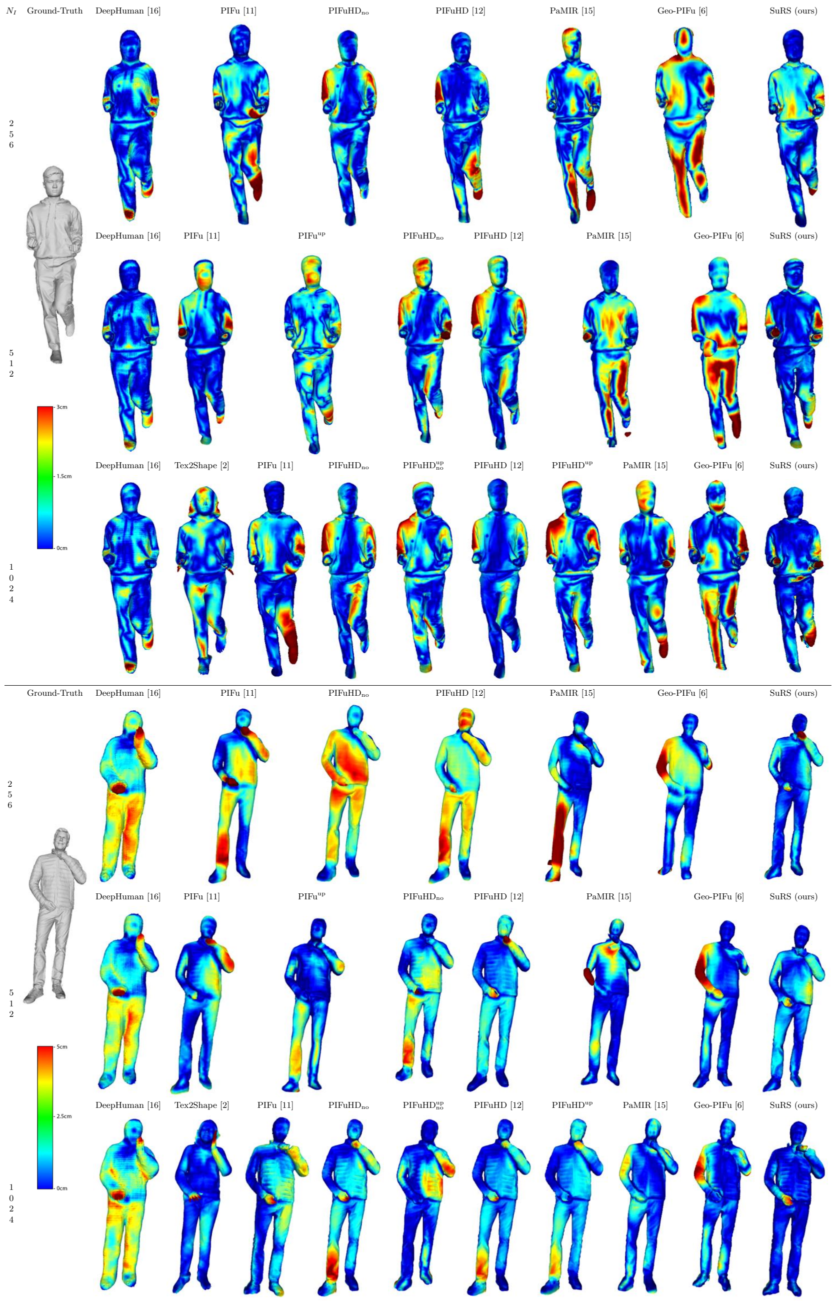

Fig. 14. Point-to-surface error maps between the models of Fig. 8 of the main paper and of Fig. 6 of the supplementary and the ground-truth shape. The 'no' subscript means that normal maps are not used. The 'up' superscript means that the input image is upscaled from 256 to 1024.

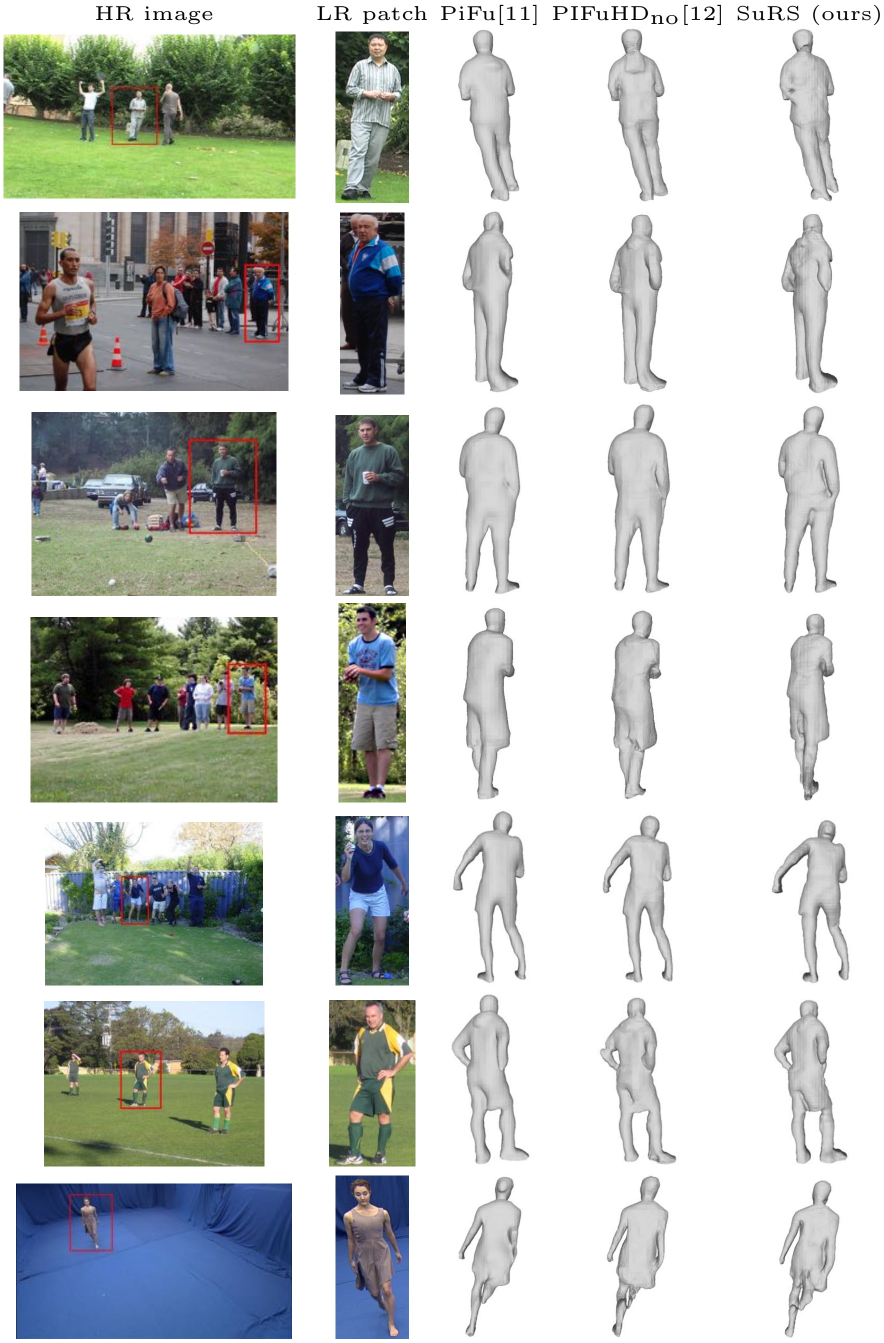

Fig. 15. Back of the models of Fig. 9 of the main paper.

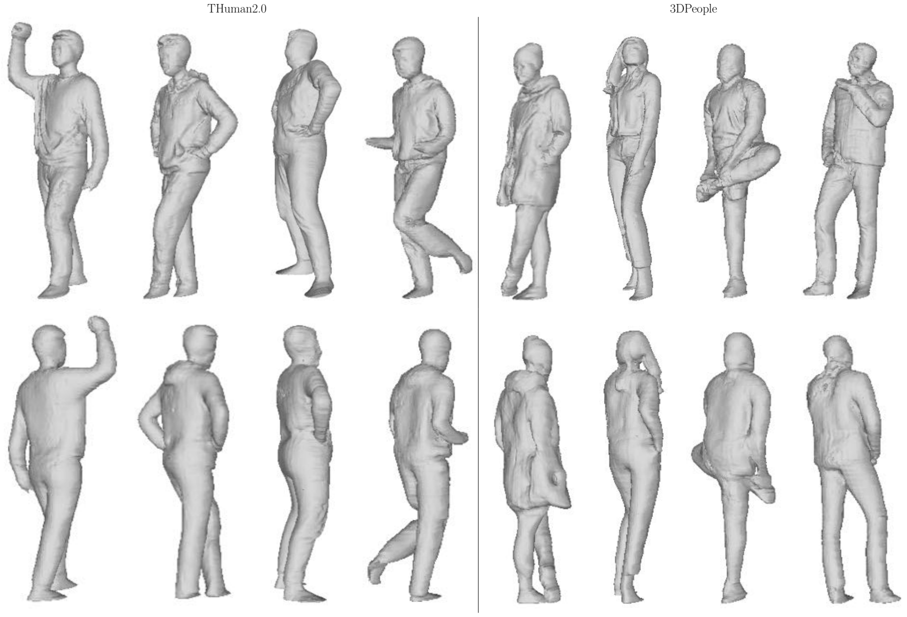

Fig. 16. Sides of the 3D models reconstructed from synthetic images presented in the main paper and reconstructed by SuRS.

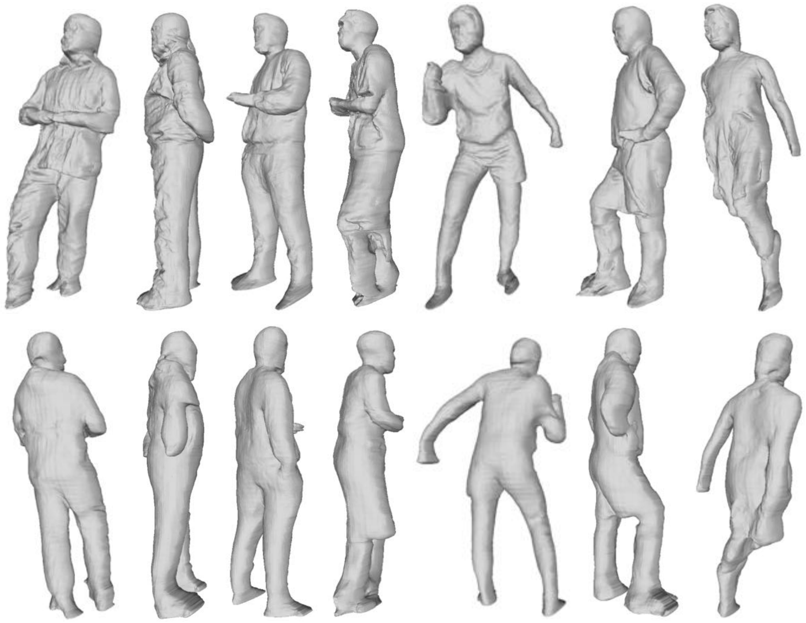

Fig. 17. Sides of the 3D models reconstructed from real images presented in the main paper and reconstructed by SuRS.

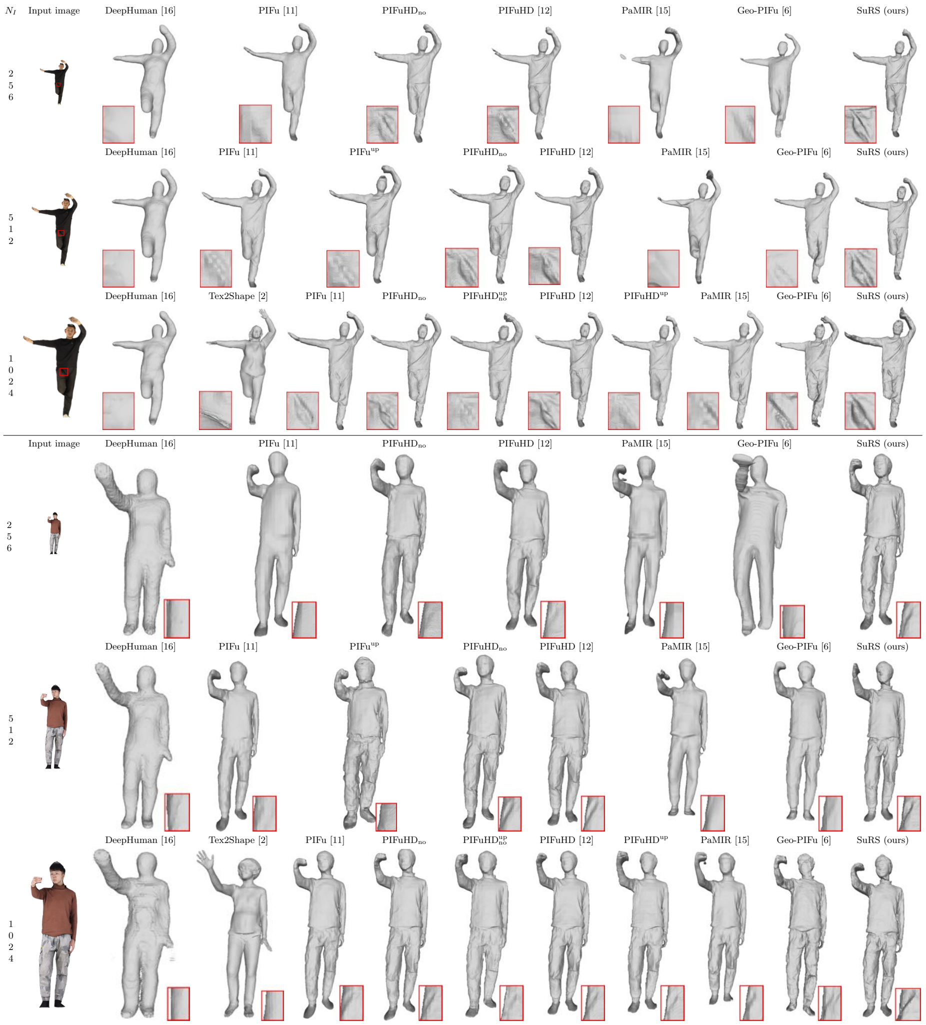

Fig. 18. Visual comparisons using different sizes of the input image for training and testing from THuman2.0 dataset. The 'no' subscript means that normal maps are not used. The 'up' superscript means that the input image is upscaled from 256 to 1024.

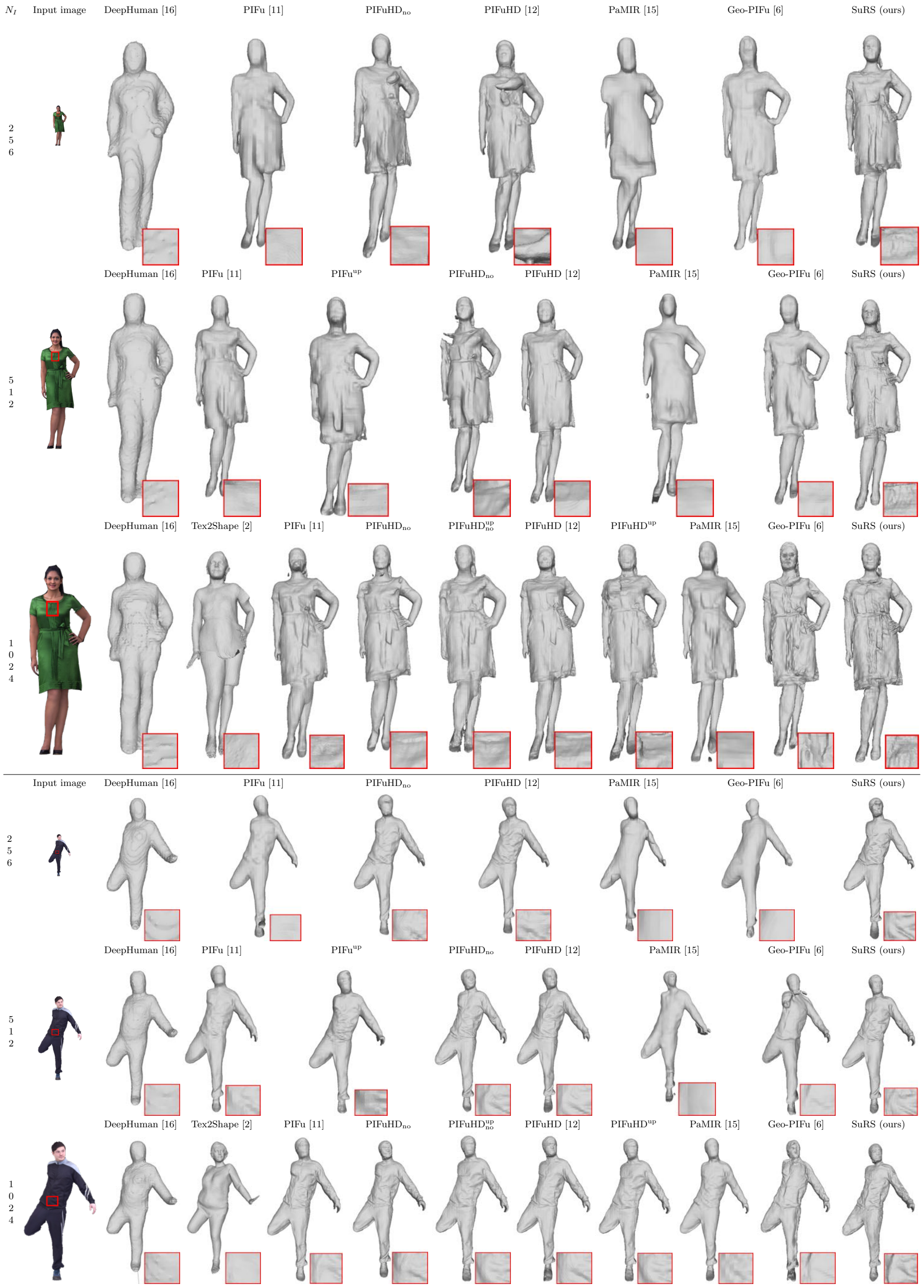

Fig. 19. Visual comparisons using different sizes of the input image for training and testing from 3DPeople dataset. The ‘no’ subscript means that normal maps are not used. The ‘up’ superscript means that the input image is upscaled from 256 to 1024.

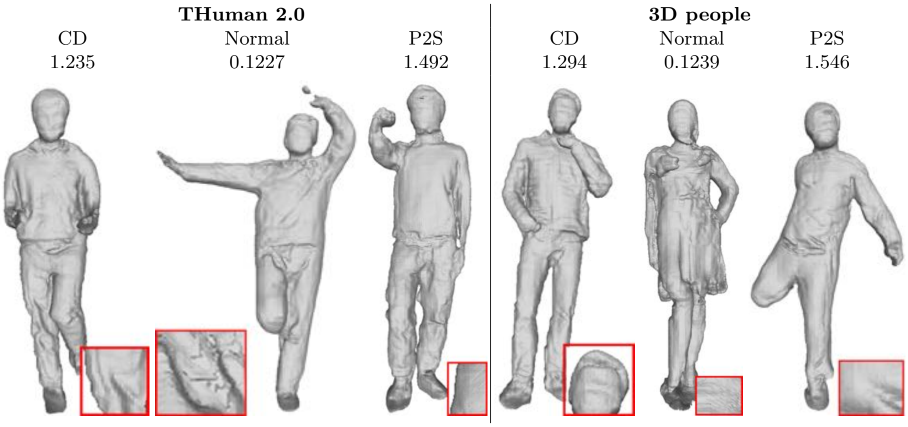

Fig. 20. SuRS trained and tested with 128x128 images.

References

1. 3d people. https://3dpeople.com/en/, accessed: 2021-10-16
2. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2293–2303 (2019)
3. Belyaev, A., Ohtake, Y.: A comparison of mesh smoothing methods. In: Israel-Korea Bi-national conference on geometric modeling and computer graphics. vol. 2. Citeseer (2003)
4. Dyn, N., Levine, D., Gregory, J.A.: A butterfly subdivision scheme for surface interpolation with tension control. ACM transactions on Graphics (TOG) **9**(2), 160–169 (1990)
5. Garland, M., Heckbert, P.S.: Simplifying surfaces with color and texture using quadric error metrics. In: Proceedings Visualization'98 (Cat. No. 98CB36276). pp. 263–269. IEEE (1998)
6. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. arXiv preprint arXiv:2006.08072 (2020)
7. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3d human body reconstruction from a single image via volumetric regression. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
8. Luo, Z., Huang, Y., Li, S., Wang, L., Tan, T.: Unfolding the alternating optimization for blind super resolution. Advances in Neural Information Processing Systems (NeurIPS) **33** (2020)
9. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
11. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)

12. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
13. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175–184 (2004)
14. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)
15. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
16. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7739–7749 (2019)
17. Zhou, B., Franco, J.S., Bogo, F., Boyer, E.: Spatio-temporal human shape completion with implicit function networks. In: 2021 International Conference on 3D Vision (3DV). pp. 669–678. IEEE (2021)