ANIM: Accurate Neural Implicit Model for Human Reconstruction from a single RGB-D image

Supplementary Material

Marco Pesavento[1,3*] Yuanlu Xu[3] Nikolaos Sarafianos[3] Robert Maier[3] Ziyan Wang[3]
Chun-Han Yao[2] Marco Volino[1] Edmond Boyer[3] Adrian Hilton[1] Tony Tung[3]
[1]University of Surrey, CVSSP, UK [2]UC Merced [3]Meta Reality Labs

7. Overview

In this supplementary material we provide:

1. Additional details about the implementation of ANIM
2. Details about the architectural design of VFE (Voxel Feature Extractor)
3. Further details on ANIM-Real
4. Limitations of the proposed approach
5. Additional results obtained by applying ANIM on real noisy data captured with Azure Kinect
6. Qualitative results for the ablations studies presented in the main paper
7. Additional qualitative results, to further demonstrate that ANIM and the new technical contributions we propose clearly outperform prior works on reconstruction quality.

8. Implementation details

In our proposed network architecture, the normals and the RGB images are concatenated and processed by the two hourglass architectures with four stacks each: the HR-FE outputs an embedding of resolution $256 \times 256 \times 256$ while the resolution of the features obtained from LR-FE is $256 \times 128 \times 128$. The former are bi-linearly interpolated with the ground-truth points projected on the input image to align the point and the feature in the 2D space. The latter are given as input to the VFE along with a voxel created from the input depth map. 3D points from the depth map are obtained by transforming 2D image coordinates to 3D world coordinates using the camera parameters, prior to normalization. The voxel is created from these 3D points. The LR features are aligned with the voxel, which is created with as many voxels as the number of channels of the LR feature (256). The VFE is a novel SparseConvNet U-Net style architecture, based on SparseConvNet [19] that has shown to be efficient for the task of 3D object detection when the input is sparse. Following [44], for any point in 3D space, we tri-linearly interpolate the latent codes from multi-scale code volumes with the ground truth point. The VFE and the HR-FE features are concatenated and finally classified by the MLP with a number of neurons equal to (369, 512, 256, 128, 1). The same features extracted from the VFE and the HR-FE are then interpolated with the point cloud for the depth-supervision. We implement our proposed framework using PyTorch and run training and testing with NVIDIA Tesla V100 GPUs. We train the neural networks with Adam optimizer and a learning rate $lr = 1e-4$ and $\delta = 1.25$. Inference time for one image, without code optimization, is in the order of the second.

For the comparisons in Sec. 5.3 IF-Net, PaMIR, ICON, SuRS, OcPlans, and (6) PIFu and IF-Net variants are retrained with the same dataset and configuration as ANIM. PIFuHD, ECON, PHORHUM and NormalGAN are not retrained due to unavailability of training code. We used their checkpoints for evaluation. All methods are tested on the same datasets (RenderPeople [1], THuman2.0 [60]).

Ethical concerns. ANIM was trained on public datasets that do not reveal the identity of subjects. ANIM aims at faithfully capturing full-body humans without alteration and body distortion, avoiding potential misuse or misrepresentation.

9. VFE Architecture

We report in Tab. 4 the detailed architecture of VFE, which consists of a SparseConvNet U-net that we designed for ANIM. The SparseConvNet implements spatially sparse convolutional networks [19]. The VFE architecture is implemented using sub-manifold sparse convolution operations. The table gives the sizes of the different layers and of the receptive fields. We experimented with various variants and report the ones that returned the best results in our experiments.

[*] Work performed during an internship at Meta Reality Labs, Sausalito

Layer	Layer Description	Output Dimension
	Input volume	$D \times H \times W \times 256$
1-3	(3×3×3 conv, 16 features, stride 1) ×2	D×H×W×16
4	(3×3×3 conv, 32 features, stride 2)	1/2D×1/2H×1/2W×32
5-6	(3×3×3 conv, 32 features, stride 1) ×2	1/2D×1/2H×1/2W×32
7	(3×3×3 conv, 64 features, stride 2)	1/4D×1/4H×1/4W×64
8-10	(3×3×3 conv, 64 features, stride 1) × 3	1/4D×1/4H×1/4W×64
11	(3×3×3 conv, 128 features, stride 2)	1/8D×1/8H×1/8W×128
12-15	(3×3×3 conv, 128 features, stride 1)×4	1/8D×1/8H×1/8W×128
16	3×3×3 invConv, 64 features, stride 1	1/4D×1/4H×1/4W×64
-	concat output 16/10	1/4D×1/4H×1/4W×128
17	3×3×3 conv, 32 features, stride 1	1/4D×1/4H×1/4W×64
18-20	(3×3×3 conv, 32 features, stride 1) × 3	1/4D×1/4H×1/4W×64
21	3×3×3 invConv, 32 features, stride 1	1/2D×1/2H×1/2W×32
-	concat output 21/6	1/2D×1/2H×1/2W×64
22	3×3×3 conv, 32 features, stride 1	1/2D×1/2H×1/2W×32
23-24	(3×3×3 conv, 32 features, stride 1) × 2	1/2D×1/2H×1/2W×32
25	3×3×3 invConv, 16 features, stride 1	D×H×W×16
-	concat output 25/3	D×H×W×32
26	3×3×3 conv, 16 features, stride 1	D×H×W×16
27-28	(3×3×3 conv, 16 features, stride 1) × 2	D×H×W×16

Table 4. VFE SparseConvNet U-net Architecture.

10. ANIM-Real dataset details

As explained in Sec. 4 of the main paper, the performance of neural implicit models significantly deteriorates when tested with raw data from consumer-grade sensors due to the severe input noise. To address this problem, we curated a new dataset (ANIM-Real) consisting of RGB-D noisy data captured with Azure Kinect and high-quality 3D ground-truth meshes reconstructed using a high-resolution camera system that employs active stereo and multi-view cameras [24]. We fine-tune ANIM with this dataset to reconstruct accurate and high-quality 3D human shapes from real-world data, mitigating the impact of the sensor noise. This section provides further details on the system used for data capture and presents examples of data of ANIM-Real. The capture system comprises two subsystems, with 1 Azure Kinect camera and 32 multi-view stereo cameras from [24]. To acquire the data, we calibrate the two systems in order to align the 3D ground-truth meshes with the RGB-D data. The collected dataset consists of 31 subjects, with 16 women and 15 men captured, each subject performing a set of scripted animations (*e.g.*, standing, walking, turning, jogging, stretching, putting on/taking off clothes). Some examples of data are shown in Fig. 10.

Datasets that integrate high-resolution 3D ground-truth shapes with raw RGB-D data are currently unavailable. The introduction of ANIM-Real is a valuable contribution to the research community in the context of neural implicit 3D human reconstruction. This dataset helps to mitigate domain gaps, providing researchers with a resource that facilitates the development of effective techniques in this domain.

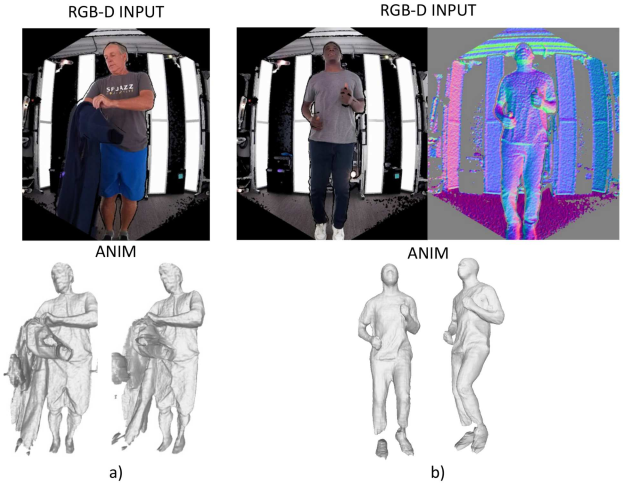

Figure 8. Limitations of ANIM. Accuracy is reduced in challenging scenes (a). Noise still affects the final reconstruction is some body parts of the shape (b).

11. Limitations

Failure cases can arise from challenging scenes that include arbitrary objects or complex motions (*e.g.* taking of clothes) as shown in Fig. 8a.

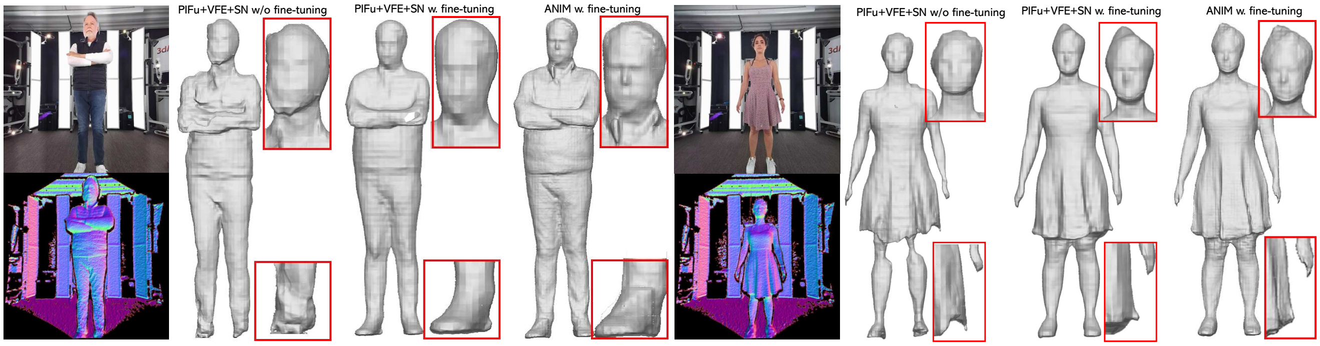
PIFu+VFE+SN w/o fine-tuning
PIFu+VFE+SN w. fine-tuning
ANIM w. fine-tuning
PIFu+VFE+SN w/o fine-tuning
PIFu+VFE+SN w. fine-tuning
ANIM w. fine-tuning

Figure 9. Qualitative comparison with PIFu+VFE+SN using real-world data.

The accuracy of ANIM applied to real-world data is slightly lower than the one achieved with synthetic data since ANIM is still influenced by the noise of the input raw data, which can affect the reconstruction as shown in Fig. 8b where the ankle of the model is not reconstructed.
The model could be further fine-tuned to learn specific sensor noise and mitigate domain gaps.

12. Additional results on real-world data

We test ANIM on real-world data obtained with an Azure Kinect after fine-tuning with additional 16k frames, consisting of around 800 frames in average from a single view of 21 subjects. Fig. 10 shows examples of ANIM reconstruction from real single RGB-D images captured with a Kinect Azure. Our approach can retrieve high-quality details on the final mesh even if the input normals and depth are noisy. ANIM can eliminate the noise of the consumer-grade sensor, significantly improving the reconstruction with accurate and high-quality 3D human shapes. We present further qualitative comparisons among different methods using real-world data. Given the inherent challenges associated with the real-world dataset, we show results from one of the most competitive methods, PIFu+VFE+SN, both before and after finetuning. As illustrated in Figure 9, finetuning PIFu+VFE+SN on ANIM-Real yeilds qualitative improvements, yet not on par with ANIM.

13. Ablation Studies

We illustrate qualitative comparisons for the ablation studies presented in Sec. 5.2 of the main paper. The labels used in the figures are consistent with the ablation study conducted in Tab. 1 and Tab. 2 in the main paper. Fig. 11 illustrates the role that each module of ANIM plays in representing high-quality details in the final reconstruction, with the highest-quality shapes obtained when all the modules are exploited. More specifically, fewer details are represented in the face and hands of the model when spatial-aware sampling is not applied. The importance of normals and HR feature can also be noticed by the reduced amount of details in the final reconstruction. Less accurate shapes are then obtained if LR feature is not used. The introduction of depth supervision further increases the accuracy and the details in the reconstructed shapes. Fig 12 demonstrates the effectiveness of the architecture of ANIM. Each key component was tested one-by-one and it is proved that the complete model outperforms the others with more accurate and highly-detailed 3D shapes.

14. Additional qualitative Results

Additional qualitative comparisons for approaches that reconstruct the 3D shape from an input different than RGB-D are presented in Fig. 13 while Fig. 14 shows additional results obtained by reconstructing 3D shapes from RGB-D data. ANIM consistently generates high-fidelity reconstructions, with cloth wrinkles and high-quality faces and hands in accordance with the input RGB images thanks to our depth-supervision strategy. Depth ambiguity issues are also solved by leveraging the depth channel of the input data. Moreover, it is shown how the contributions we propose, such as using the VFE and the multi-resolution features of HR-FE and LR-FE, can be used to improve other approaches, but only our complete ANIM model design returns the best results.
Fig. 15 shows results of reconstructing 3D shapes from input different than RGB-D for other related methods that are not shown in the paper.
Fig. 16 show the the side-view reconstruction of the results showed in Fig. 7 and Fig. 14.

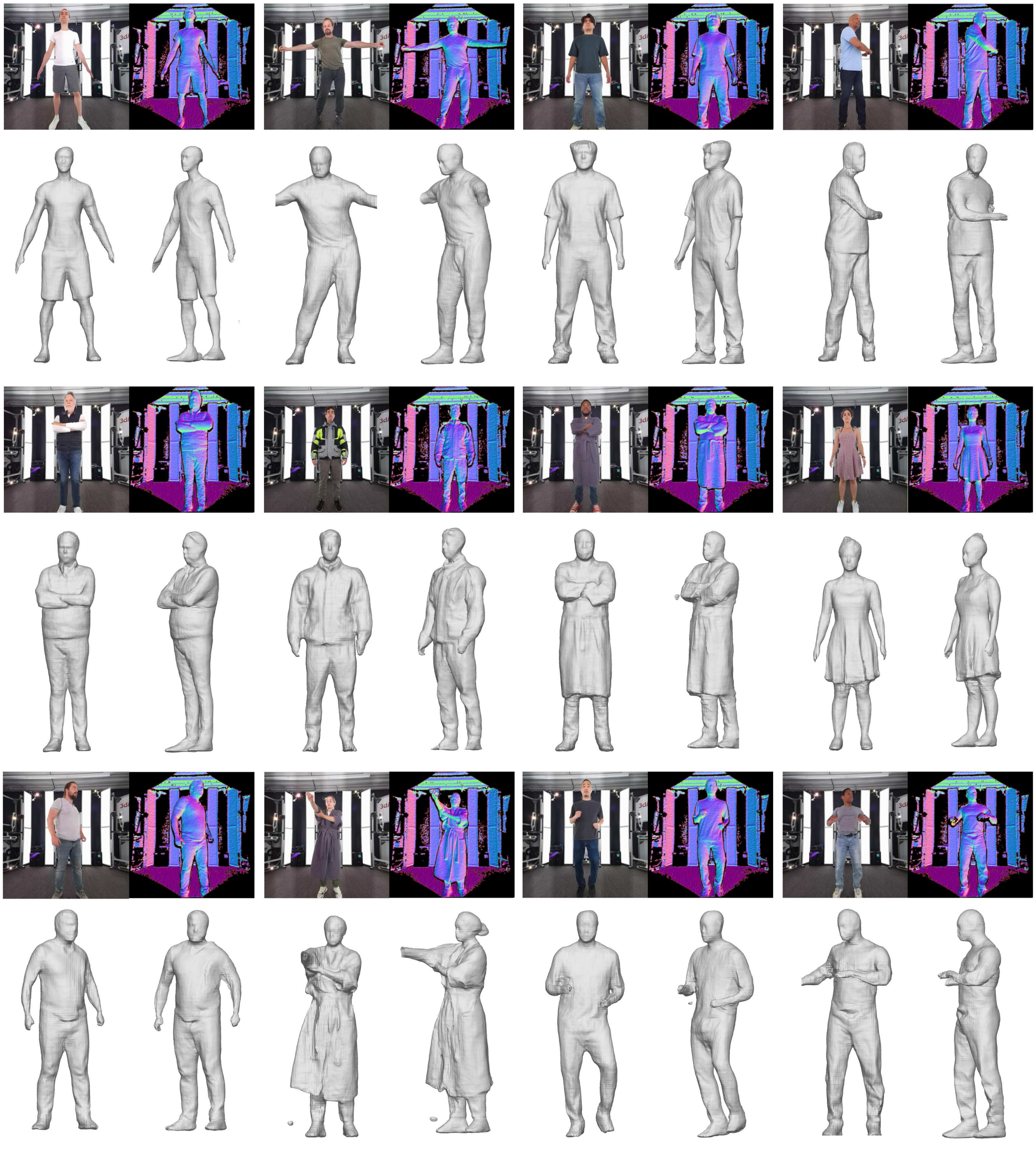

Figure 10. More reconstruction results by ANIM using a consumer-grade RGB-D camera (Azure Kinect) as an input. ANIM is capable of handling various human body and cloth typologies ranging from a skirt to a bath robe and is agnostic to diverse human poses.

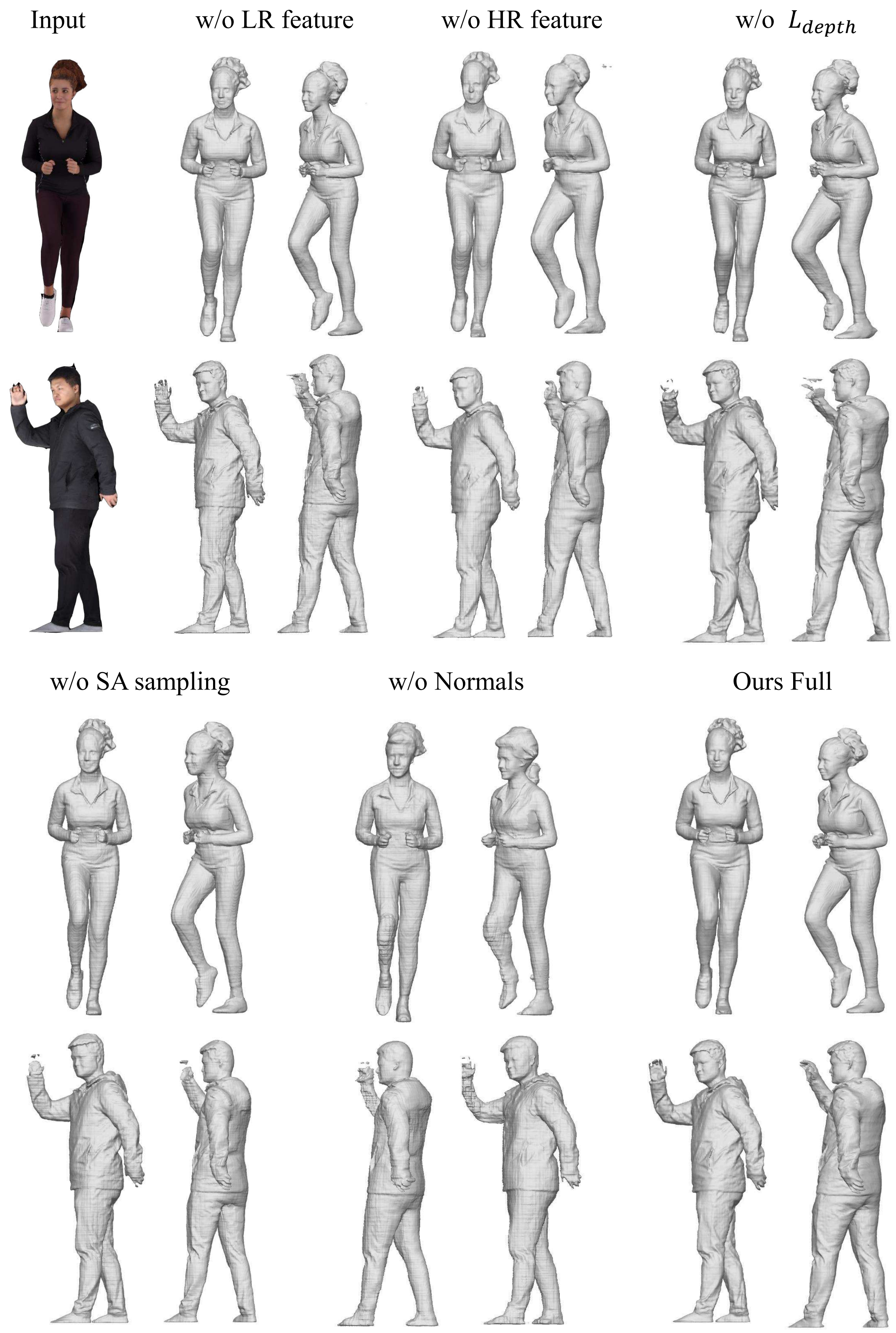

Figure 11. We conducted an ablation study on the components of ANIM that influence the reconstruction quality. We show reconstructions of 2 subjects (one from THuman2.0 [60] and the other from RenderPeople [1]) captured by a single-view RGBD image (*i.e*. partial view), from frontal and 45-deg side views. Our full ANIM model provides high-quality reconstructions with facial expressions, hands, and cloth wrinkles with fine-level details, without shape distortion along the camera view. Please zoom in the figure to better see details.

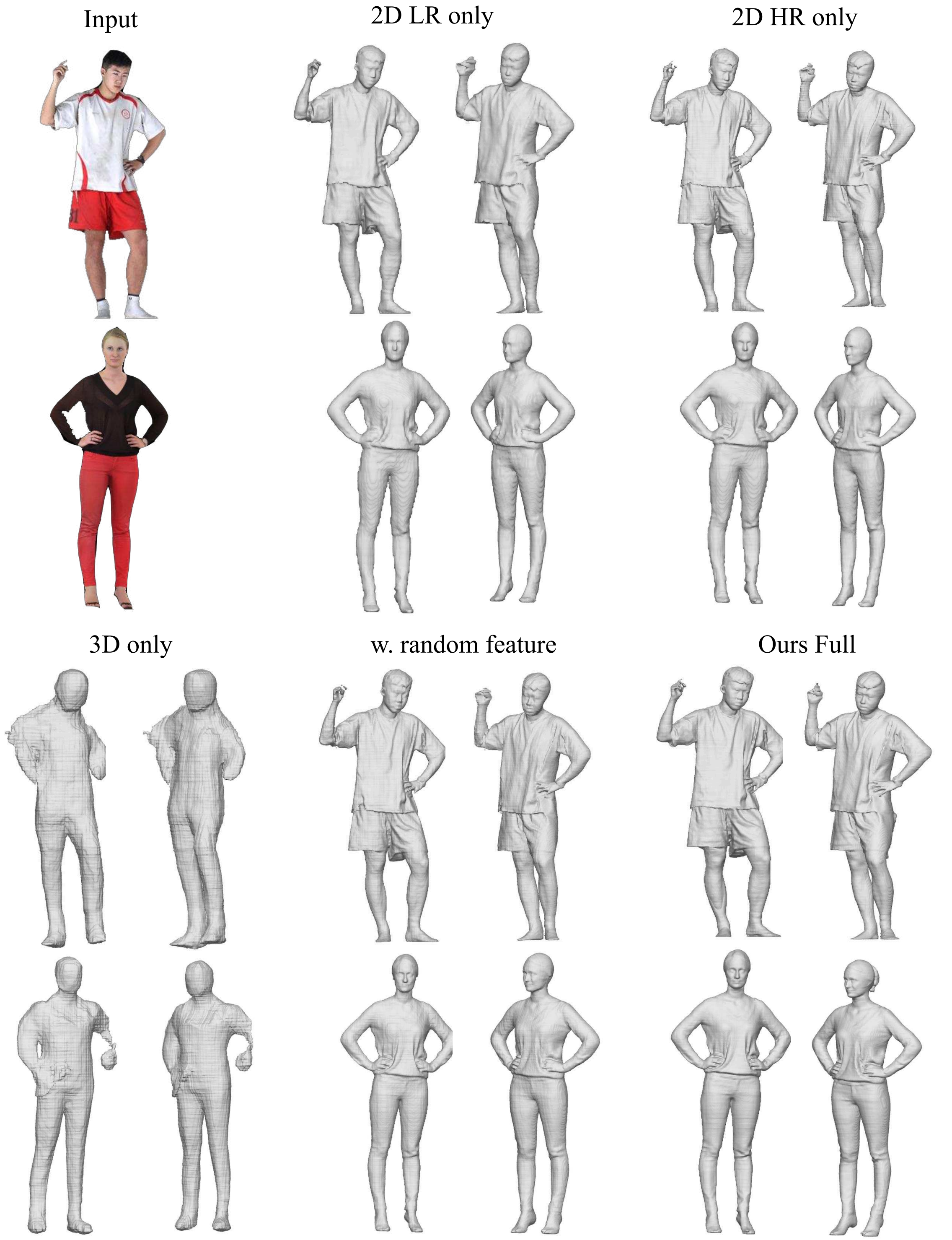

Figure 12. We conducted an ablation study where components of ANIM were removed one-by-one to prove the superiority of the proposed architecture. We show reconstructions of 2 subjects (one from THuman2.0 [60] and the other from RenderPeople [1]) captured by a single-view RGBD image (*i.e.* partial view), from frontal and 45-deg side views, with colored normals. Our full ANIM model provides more accurate results. Please zoom in the figure to better see details.

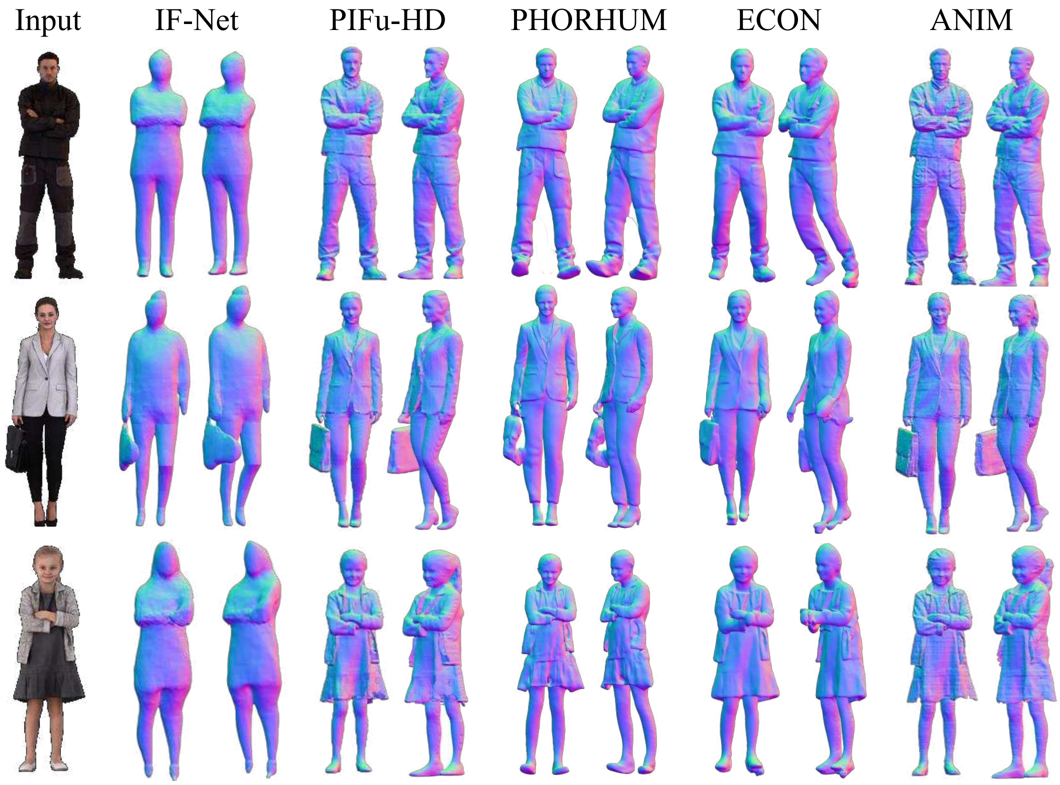

Figure 13. Additional comparisons with approaches that use single RGB image or partial point clouds as input. Data from RenderPeople [1]. ANIM reconstructs full-body models with high accuracy, with cloth wrinkles, face and hand details, and without depth ambiguity (*i.e.* distortion along camera view).

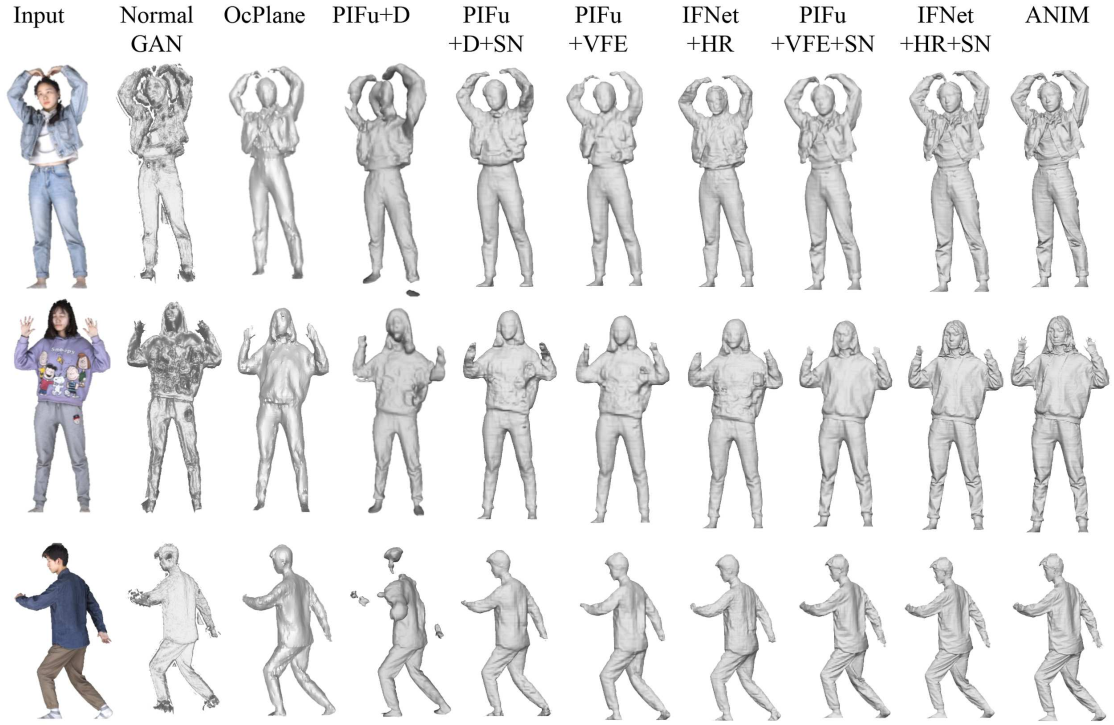

Figure 14. Additional comparisons with methods that use a single RGB-D image as input. Our core contributions can leverage state-of-the-art models, but only our complete ANIM model design returns the best results. We show reconstruction from the front view. Data from THuman2.0 [60].

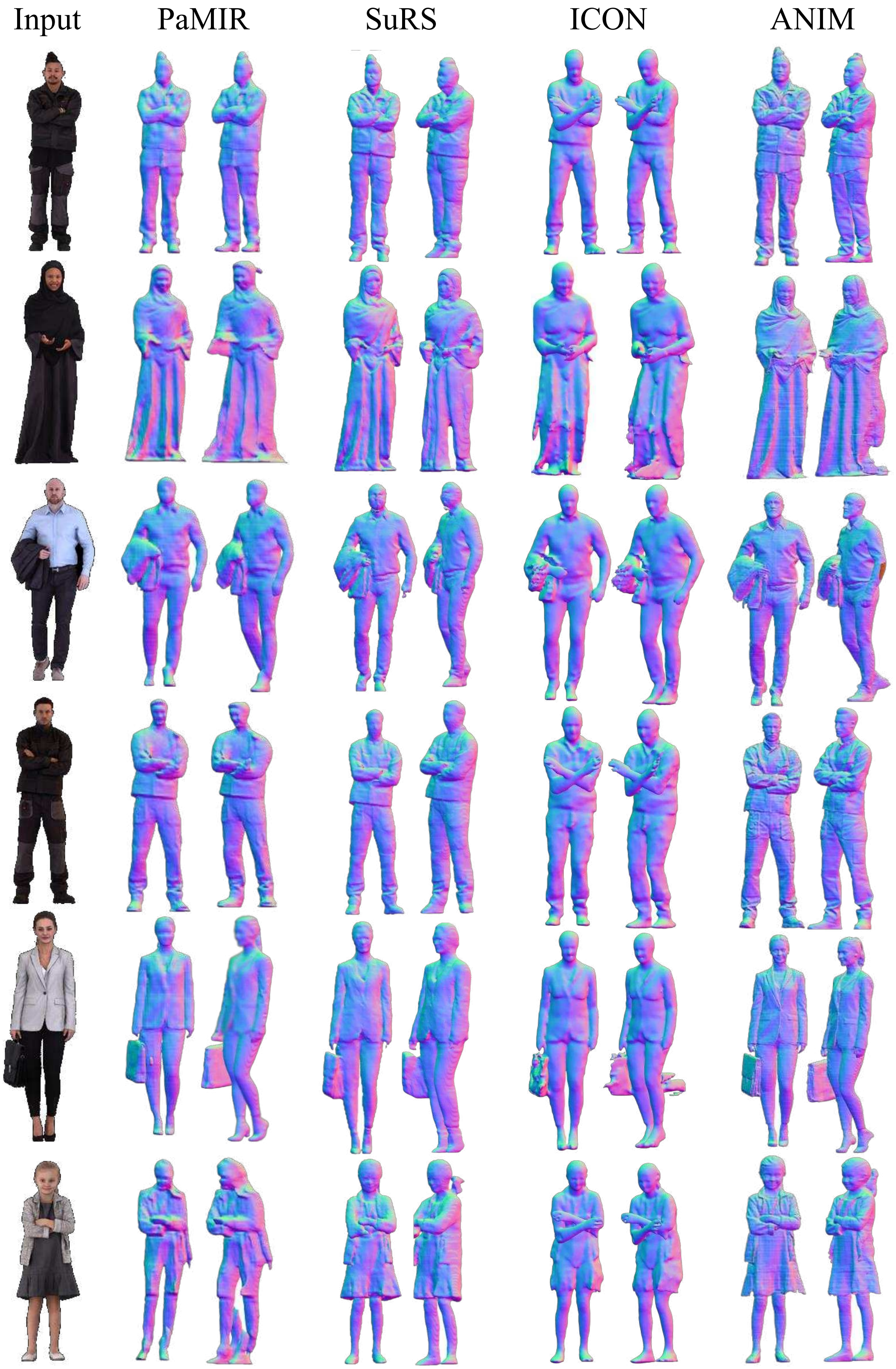

Figure 15. Qualitative comparisons with approaches not illustrated in the main paper that use a single RGB image or partial point clouds as input. Data from RenderPeople [1].

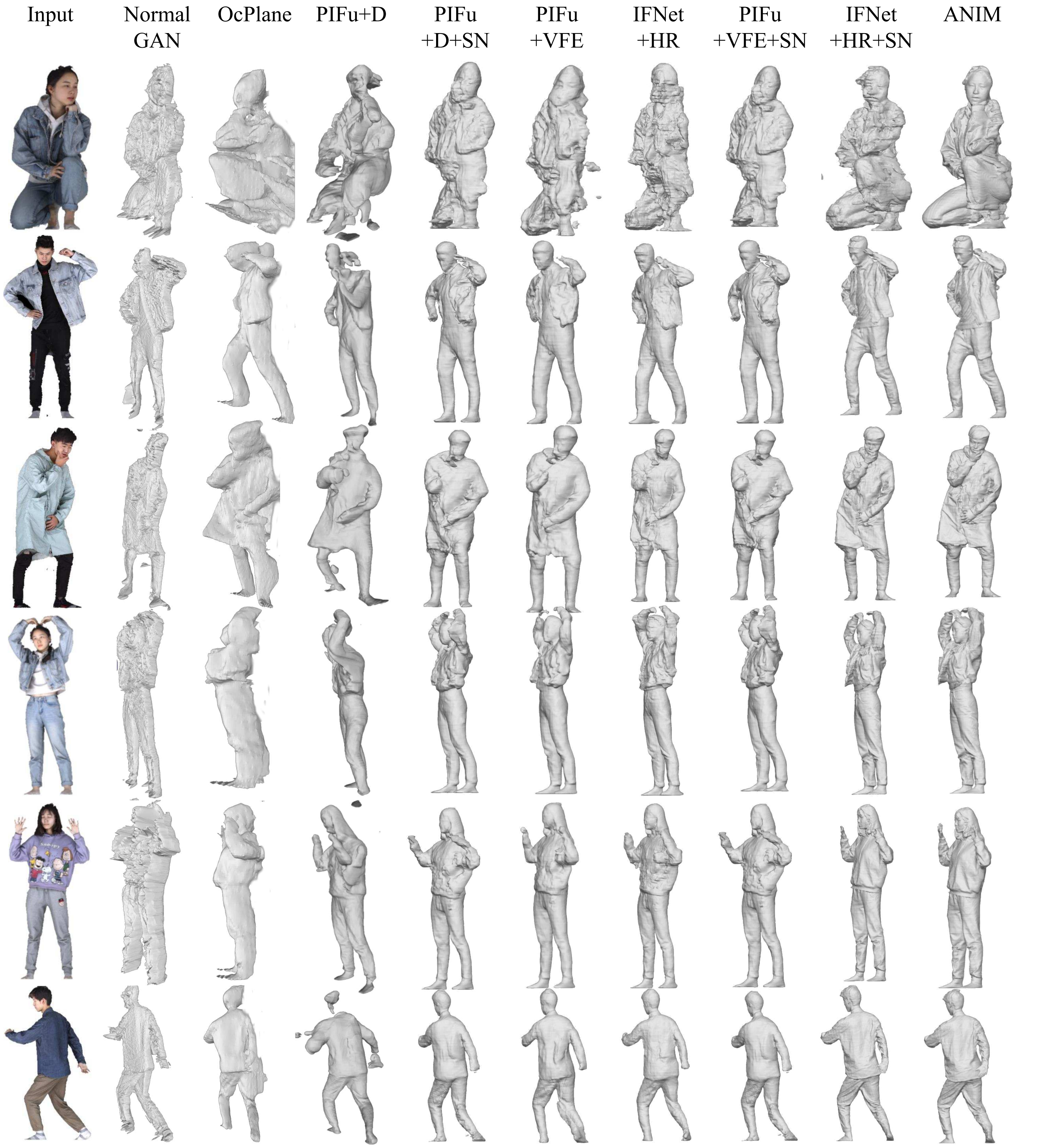

Figure 16. Side-views of the 3D shapes reconstructed from an input RGB-D data showed in Fig. 7 and Fig. 14.

References

[1] Renderpeople. https://renderpeople.com/. Accessed: 2020-07-26. 5, 6,

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision*, 2019. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[6] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *European Conference on Computer Vision Workshops*, 2012. 2

[7] Gavin Barill, Neil Dickson, Ryan Schmidt, David I.W. Levin, and Alec Jacobson. Fast winding numbers for soups and clouds. *ACM Transactions on Graphics*, 2018. 5

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, 2020. 1

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, 2020. 2

[10] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter 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*, 2016. 2

[11] A. Burov, M. Niesner, and J. Thies. Dynamic surface function networks for clothed human bodies. In *IEEE International Conference on Computer Vision*, 2021. 2

[12] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 6, 7

[13] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 2

[14] C. Coppola, D. Faria, U. Nunes, and N. Bellotto. Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016. 2

[15] Zheng Dong, Ke Xu, Ziheng Duan, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Geometry-aware two-scale pifu representation for human reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 1

[16] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE International Conference on Computer Vision*, 2019. 2

[18] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45(5):586–597, 2014. 2

[19] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3,

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[21] Azmi Haider and Hagit Hel-Or. What can we learn from depth camera sensor noise? *Sensors*, 22(14):5448, 2022. 5

[22] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Annual Conference on Neural Information Processing Systems*, 2020. 2, 4

[23] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *IEEE International Conference on Computer Vision*, 2021. 2

[24] https://3dmd.com/. 3dmd 4d scanner. 4, 5,

[25] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[26] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. Bodymap: Learning full-body dense correspondence map. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 5

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[28] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. *European Conference of Computer Vision Workshops*, 2018. 2

[29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[31] Dongping Li, Tianjia Shao, Hongzhi Wu, and Kun Zhou. Shape completion from a single rgbd image. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1809–1822, 2017. 2

[32] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*, 2020. 2

[33] Xing Li, Yangyu Fan, Di Xu, Wenqing He, Guoyun Lv, and Shiya Liu. Sfnet: Clothed human 3d reconstruction via single side-to-front view rgb-d image. In *International Conference on Virtual Reality*, 2022. 1, 2

[34] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6): 248, 2015. 2

[36] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987. 4

[37] Yang Lu, Han Yu, Wei Ni, and Liang Song. 3d real-time human reconstruction with a single rgbd camera. *Applied Intelligence*, pages 1–11, 2022. 1

[38] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *IEEE International Conference on Computer Vision*, 2021. 2

[39] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision*, 2022. 2

[40] Aihua Mao, Hong Zhang, Yuxin Liu, Yinglong Zheng, Guiqing Li, and Guoqiang Han. Easy and fast reconstruction of a 3d avatar with an rgb-d sensor. *Sensors*, 17(5), 2017. 2

[41] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. *Person Re-Identification*, pages 161–181, 2014. 2

[42] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[44] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3, 6,

[45] Marco Pesavento, Marco Volino, , and Adrian Hilton. Super-resolution 3d human shape from a single low-resolution image. In *European Conference on Computer Vision*, 2022. 1, 2, 6, 7

[46] 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. 2, 3, 5, 6

[47] 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*, 2020. 1, 2, 5, 6, 7

[48] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *IEEE International Conference on Computer Vision*, 2019. 2

[49] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[50] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[51] Chris Sweeney, Greg Izatt, and Russ Tedrake. A supervised approach to predicting noise in depth images. In *International Conference on Robotics and Automation*, 2019. 5

[52] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *IEEE International Conference on Computer Vision*, 2021. 1

[53] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, 2018. 2

[54] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *European Conference on Computer Vision*, 2020. 1, 2, 6, 7

[55] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014. 2

[56] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 2

[57] 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. 1, 2, 6, 7

[58] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6, 7

[59] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4490–4504, 2021. 2

[60] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5, 6,

[61] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[62] Xiaoming Zhao, Yuan-Ting Hu, Zhongzheng Ren, and Alexander G Schwing. Occupancy planes for single-view rgb-d human reconstruction. *arXiv preprint arXiv:2208.02817*, 2022. 1, 2, 6, 7

[63] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE International Conference on Computer Vision*, 2019. 2

[64] 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*, 44(6): 3170–3184, 2021. 1, 2, 3, 4, 5, 6, 7