Supplementary Material
Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition

Chen Guo[1] Tianjian Jiang[1] Xu Chen[1,2] Jie Song[†1] Otmar Hilliges[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen

In this **supplementary document**, we provide additional materials to supplement our main submission. In the **supplementary video**, we show more reconstruction results of our method on monocular in-the-wild videos.

Contents

1. Implementation Details

1.1. Network Architecture

Human. The canonical human shape network f^H_{sdf} (Eq. 1 in the main manuscript) includes 8 blocks, each of which consists of a fully connected layer, a weight normalization layer [17] and a softplus activation layer [5]. Each fully connected layer contains 256 neurons. The pose condition $\boldsymbol{\theta}$ is obtained by concatenating all axis angles represented in radians. We apply positional encoding with 6 frequency components to the input points to better model high-frequency details [16]. The canonical human texture network f^H_{rgb} (Eq. 5 in the main manuscript) includes 4 blocks with the same architecture as the human shape network except using the Sigmoid activation function for the last layer and using a ReLU activation function for the rest layers.

Background. The background network f^B (Eq. 7 in the main manuscript) also consists of two parts: the background density network and the texture network. The density network has the same architecture as the canonical human shape network with 10 frequency components to the input points. And the texture network only includes 1 block of a fully connected layer with 128 neurons, a weight normalization layer, and a ReLU activation layer, ending up with a Sigmoid activation layer.

1.2. Training Details

We train our networks using the Adam optimizer [12], with an initial learning rate of $l = 5e^{-4}$ which will decay in half after each scheduled milestone. In our implementation, the milestones are set to be 200 and 500 epochs respectively. The other Adam hyper-parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A model is trained for 36 to 48 hours on a single NVIDIA RTX 3090 GPU.

1.3. Composited Volume Rendering

As introduced in the main manuscript, we extend the inverted sphere parametrization of NeRF++ [25] to represent the scene: an outer volume (i.e., the background) covers the complement of a spherical inner volume with radius $R = 3$ (i.e., the space assumed to be occupied by the human).

Foreground Rendering. For rendering the foreground component, we combine the implicit neural avatar repre-

[†] Corresponding author

sentation (Sec. 3.1 in the main manuscript) with SDF-based volume rendering [24] which allows us to convert the SDF to a density value σ. Similar to [24], we first uniform sample 128 points along the ray $\mathbf{r}$ in the inner volume and then perform the inverse Cumulative Distribution Function (CDF) sampling returning 64 sampled points. For more details, we refer to [24].

Background Rendering. To obtain the background component color value, we follow [25] and sample 32 points outside the inner sphere. This is achieved by uniformly sampling $\frac{1}{r}$ in the range [0, $\frac{1}{R}$]. Given the sampled $\frac{1}{r}$, we calculate the corresponding background point $\mathbf{x}'_b$ using the geometric relationship derived in [25].

1.4. Initialization

We first utilize the body pose regressor [18] to estimate the initial SMPL [14] parameters of the human in the videos. Similar to [7], we refine the SMPL estimates by simply minimizing the 2D distance between 2D joint predictions from OpenPose [3] and the 2D projection of 3D SMPL joints along with a temporal pose stability loss which penalizes temporal pose jittering. The total objective $\mathcal{L}_{\text{SMPL}}$ for the SMPL refinement optimization is:

$$\mathcal{L}_{\text{SMPL}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{joint}}(\boldsymbol{\theta}) + \lambda_{\text{stab}}\mathcal{L}_{\text{stab}}(\boldsymbol{\theta}) \tag{18}$$

with

$$\mathcal{L}_{\text{joint}}(\boldsymbol{\theta}) = \sum_{i=1}^{N_{\text{joint}}} w_i \rho \left(\Pi\left(J(\boldsymbol{\theta})_i\right) - J_{2D,\text{est},i}\right), \tag{19}$$

$$\mathcal{L}_{\text{stab}}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{N_{\text{joint}}} \|J(\boldsymbol{\theta})_i - J'_i\|_2^2, \tag{20}$$

where $J(\boldsymbol{\theta})$ is the 3D SMPL joints given the SMPL parameters $\boldsymbol{\theta}$. We sum up the distances for each joint i overall counter joints N_{joint}. We denote Π as the 3D to 2D projection of joints with camera parameters. To account for detection noise, the error terms are weighted by the corresponding detection confidence w_i. To down-weight outlier 2D detections, a robust Geman-McClure error function ρ [6] is applied. J'_i represents the 3D SMPL joints of the last frame.

We pretrain the shape network in canonical space based on SMPL mesh deformed into the canonical pose for the purpose of accelerating the training process.

1.5. Data Preprocessing

Given the estimated SMPL parameters for all frames, we relocate the individual SMPL meshes at the space origin $\mathbf{O}$ by subtracting their center of gravity. We further apply a global scale of $\frac{3}{R*1.1}$ to ensure all estimated camera centers are inside the inner spherical volume, similar to [24].

1.6. Opacity Sparseness Regularization

We deploy the opacity sparseness regularization loss term to encourage the global sparsity of the ray opacity. This is achieved by leveraging the dynamically updated human shape in canonical space. In particular, we warp the sampled points into canonical space via inverse warping and calculate the signed distance to the human shape in canonical space. Instead of simply using the queried SDF values from the canonical shape network for these points, we extract the canonical human shape explicitly and then calculate the signed point-to-surface distances in the mesh space. This is because the SDF distribution for the space that is far away from the canonical human shape could be irregular due to the lack of observations. For the rays whose nearest point to canonical human shape has a distance larger than a pre-defined threshold value 5 cm, we classify these rays as non-intersecting rays. Note that, we progressively update the canonical human shape during the whole training process, and thus, the threshold value does not tightly bound the canonical human shape but is a regularizer for a conservative update.

1.7. In-shape Stabilization Loss

Even though the canonical human shape is initialized as an SMPL mesh deformed into the canonical pose, the training process could fail in case the background model dominates the representation of the entire scene. To further stabilize the training progress, we can encourage the intersecting rays (distance $\leq$ 0) to be fully opaque (the opacities equal to 1). We gradually decay the weight of this loss to zero as it is optional and only needed in the early stage of the training.

2. Evaluation Details

2.1. 2D Segmentation Comparisons

Dataset. Following MonoPerfCap [23], we use the MonoPerfCap dataset to compare our method with other off-the-shelf 2D segmentation approaches to validate the scene decomposition quality of our method. The evaluation sequence Helge_outdoor contains 131 frames of in-the-wild human performance with ground-truth masks.

2.2. View Synthesis Comparisons

We use the NeuMan dataset [11] for the evaluation of novel view synthesis. This dataset collects 6 videos of 10 to 20 seconds long captured by a mobile phone. For more details and training/testing split of this dataset, we refer to [11]. The original dataset does not contain ground-truth human masks and thus is not suitable for the evaluation of the rendering quality of **humans** under test views. To this end, we manually segment the human from images in the test set for the quantitative comparison. We use HumanNeRF [21] and NeuMan [11] as our baselines. By default, NeuMan uses [9] to segment humans and we run RVM [13] for HumanNeRF. We simply use the pre-trained checkpoints of NeuMan for evaluation (7 days required for training) and train HumanNeRF with the same human poses as ours.

2.3. Reconstruction Comparisons

We use 3DPW [20] and SynWild datasets (see also Sec. 2.4) for reconstruction comparisons. Specifically, we evaluate the outdoors_fencing_01 sequence in 3DPW which contains 942 frames. ICON [22] and SelfRecon [10] are used as baseline approaches. By default, ICON applies [2] to obtain the human masks and we run RVM [13] for SelfRecon.

2.4. SynWild Dataset

We propose a new dataset called **SynWild** to evaluate the monocular human surface reconstruction task. Dynamic human subjects are captured in a dense multi-view system and reconstructed with detailed surface geometry and realistic textures via commercial software [4]. More specifically, the capturing setup is equipped with 106 synchronized cameras (53 RGB and 53 IR cameras) and human motion sequences are filmed at 30 FPS. Then we place the textured 4D scans into realistic 3D scenes/HDRI panoramas and render monocular videos from virtual cameras with a 35mm focal length and 1920x1080 image resolution, leveraging a high-quality game engine [1]. In total, this dataset includes 6 video sequences (1091 frames) with different motions, human subjects, and backgrounds. This is the first dataset that allows for quantitative comparison of monocular human reconstruction in a realistic setting via semi-synthetic data. We show the rendered images and their corresponding ground-truth meshes in Fig. 8.

3. More Results

3.1. Additional Ablation Studies

Jointly Pose Optimization: We compare our full model to a variant version without jointly optimizing human poses. The qualitative result is shown in Fig. 9, in which we can see that our method can refine the poses and achieve better reconstruction quality.

Modeling of Background: One key component of our method is to model the background along with the human. To show the importance of background modeling, we compare our full model to a variant without background modeling. Instead, we run RVM [13] to segment the human from images. **Results:** Tab. 5 and Fig. 10 indicate that modeling human and background jointly is crucial for human reconstruction in the wild.

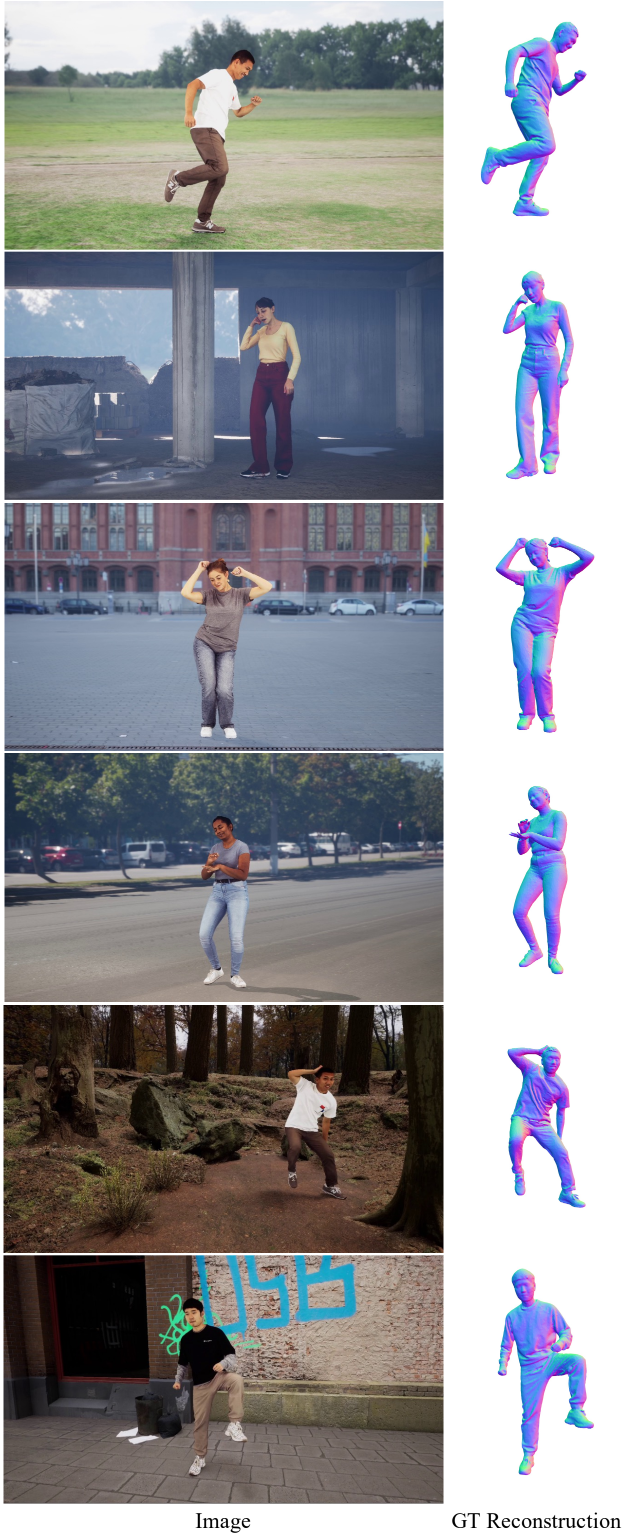

Figure 8. **SynWild Dataset.** We show sample images and their corresponding ground-truth meshes from the SynWild dataset.

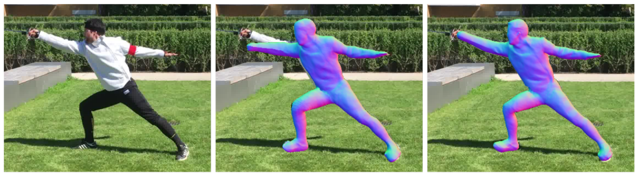

Figure 9. **Importance of jointly pose optimization.** Jointly pose optimization corrects initial pose estimates and achieves better reconstruction quality.

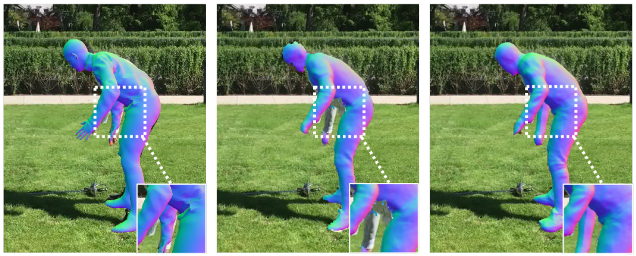

Figure 10. **Importance of modeling background.** Without modeling the background, the decoupling between the human and the background is upper-bounded by the off-the-shelf segmentation tool and only yields worse reconstruction results.

Method	IoU ↑	$\mathbf{C}-\ell_2$ ↓	NC ↑
w/o BG Modeling.	0.811	3.13	0.728
Ours	**0.818**	**2.66**	**0.753**

Table 5. **Importance of background modeling.** Without background modeling, our method cannot recover the complete human body and the reconstruction quality is upper-bounded by the 2D segmentation module.

3.2. Comparison with Template-based Methods

A direct *quantitative* comparison to state-of-the-art template-based approaches is not feasible. This is because 1) their code is not publicly available and 2) no accurate 3D ground-truth human scans for their in-the-wild testing sequences are available. Hence, we conduct *qualitative* comparisons to state-of-the-art template-based approaches: MonoPerfCap [23] and DeepCap [8].

We show in Fig. 11 and Fig. 12 the comparison of our method with MonoPerfCap and DeepCap respectively. We can see that our method achieves better human reconstruction results in terms of the global pose alignment and the surface detail recovery. To be noted, our method does not require any cumbersome (pre-scanning and manual rigging) pre-processing.

3.3. Background Component Rendering

At the core of our method lies the idea to jointly learn the dynamic foreground and the background from images, leading to self-supervised scene decomposition. As shown in Fig. 16, our method is able to model the complete background even under occlusions caused by human motions by leveraging the temporal information in the whole video sequence via the proposed global optimization formulation. In other words, these results also reflect a clean and robust decoupling of the human and background in the scene.

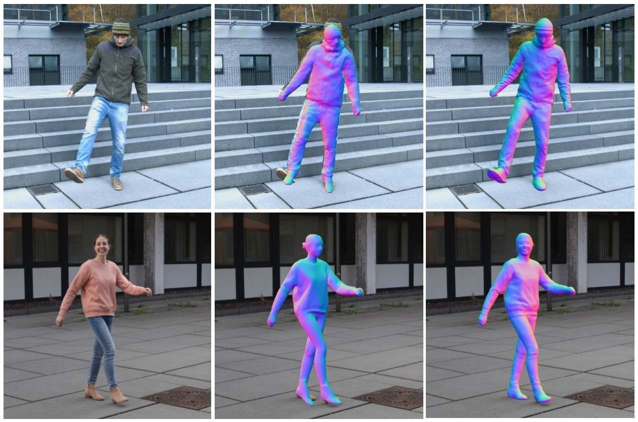

Figure 11. **Qualitative comparison with MonoPerfCap.** Our method recovers better dynamic surface details (e.g., cloth wrinkles) and realistic facial features.

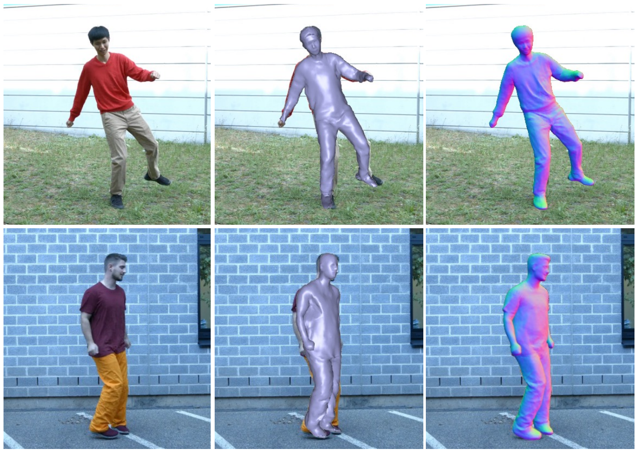

Figure 12. **Qualitative comparison with DeepCap.** Our method recovers better dynamic surface details (e.g., cloth wrinkles) and realistic facial features.

3.4. Additional Qualitative Results

We provide an additional qualitative novel view synthesis comparison in Fig. 13. We also show the front and back view of our reconstructed 3D avatars in canonical space in Fig. 17.

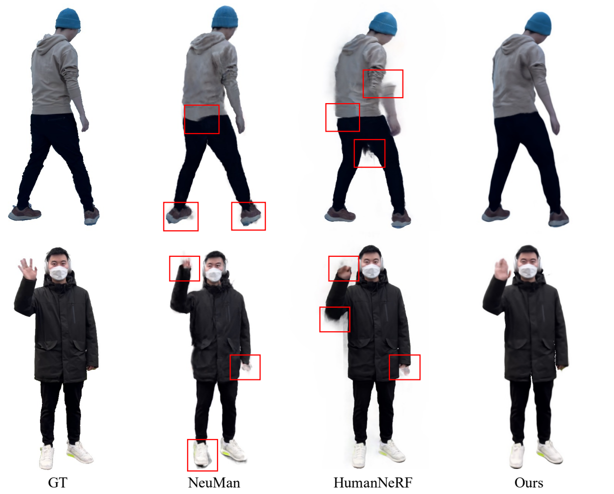

Figure 13. **Additional qualitative view synthesis comparison.** Our method achieves comparable and even better novel view synthesis results compared to NeRF-based methods.

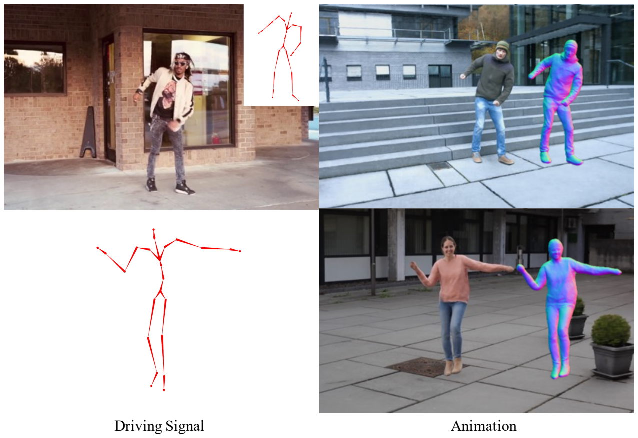

Figure 14. **Animation results.** Given driving signals from different sources, our reconstructed 3D avatars can be animated to novel new poses and can be rendered along with the original background.

4. Animation

Apart from capturing the dynamic human performance over the entire video sequence, the reconstructed 3D avatar can also be animated to novel poses. The animation results are shown in Fig. 14. The 3D avatars can be animated using the motion from another training video or the poses from the off-the-shelf large corpus of motion capture data [15, 19]. Note that, since we also learn the background, we can put the posed avatar into the original scene with high-fidelity rendering results as demonstrated in the second column of Fig. 14.

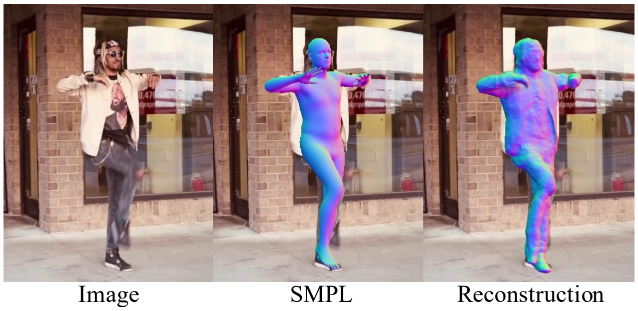

Figure 15. **Failure case.** Unreasonable pose initialization leads to the incorrect reconstruction, especially when the RGB information is not enough for the pose correction.

5. Limitations and Societal Impact Discussion

Although readily available, Vid2Avatar still relies on reasonable pose estimates as inputs. A poor pose initialization may lead to artifacts on that particular frame, especially in the case of motion blurs where the photometric information is missing to correct the poses as shown in Fig. 15. And if regions of the clothed human are not or barely visible in the entire video input, the 3D surface in these regions tends to be relatively smooth without high-frequency details. Furthermore, our method performs well for garments that are topologically similar to the body. Loose clothing such as skirts or free-flowing garments poses challenges due to their fast dynamics.

Vid2Avatar enables the digitization of humans from a single RGB video, which has many potential applications in movies, AR/VR, and telepresence applications. As the result of our method is a human avatar, that can be animated with unseen poses, there is a risk that it might be misused for purposes such as deep-fakes. Such concerns must be addressed first before deploying digital human avatars in products. Clearly, our goal with this work is to enable uses of the technology that are beneficial for society. Unfortunately, we cannot prevent nefarious uses of such technology, but we argue that studying these methods in a maximally transparent way, including discussion of technical details in the paper, and release of code and data, should be preferred over undisclosed research, as this will help to build counter measures to mitigate the potential for dubious uses.

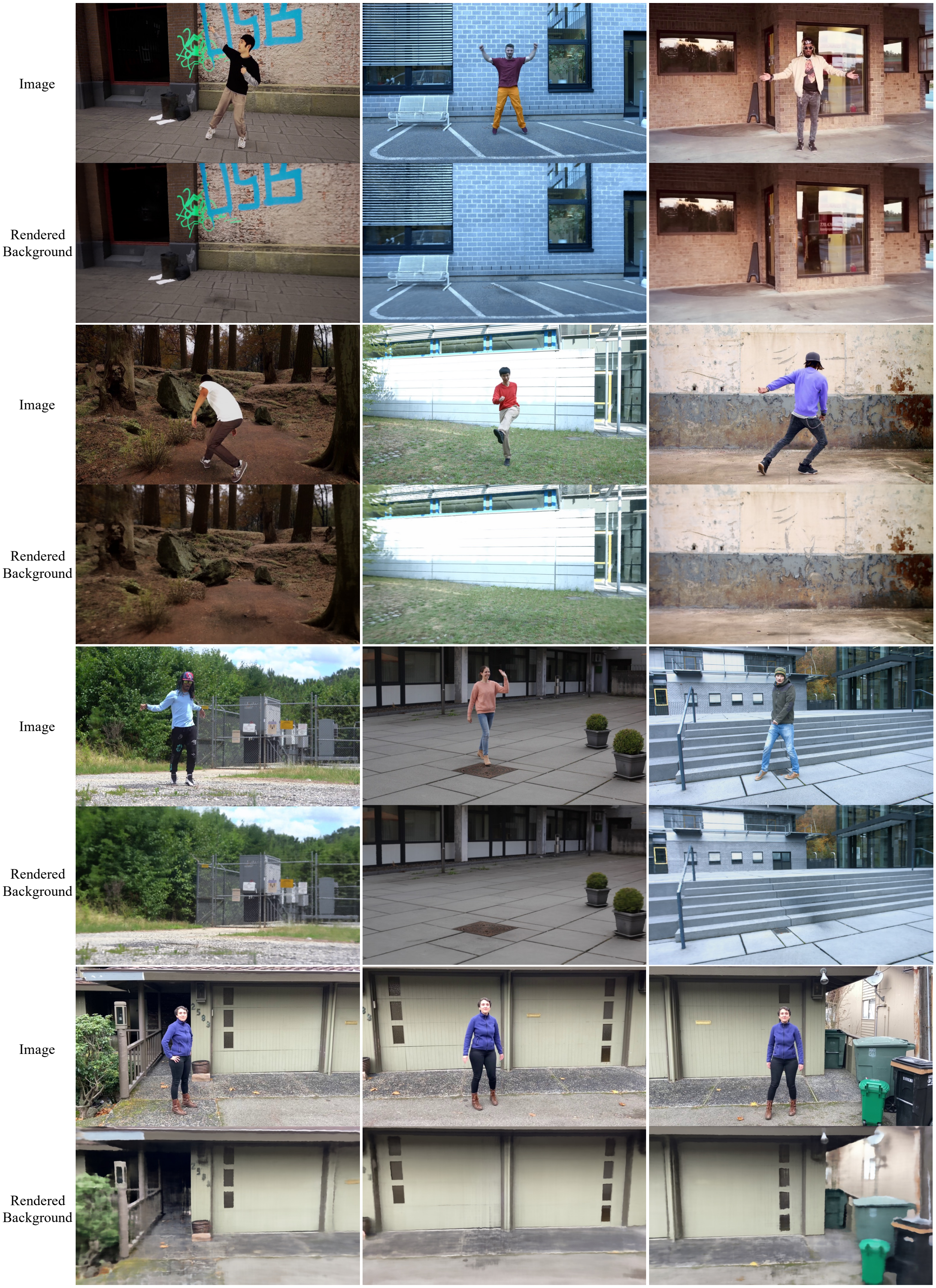

Figure 16. **Qualitative results of background rendering.** We show qualitative results of our modeled backgrounds.

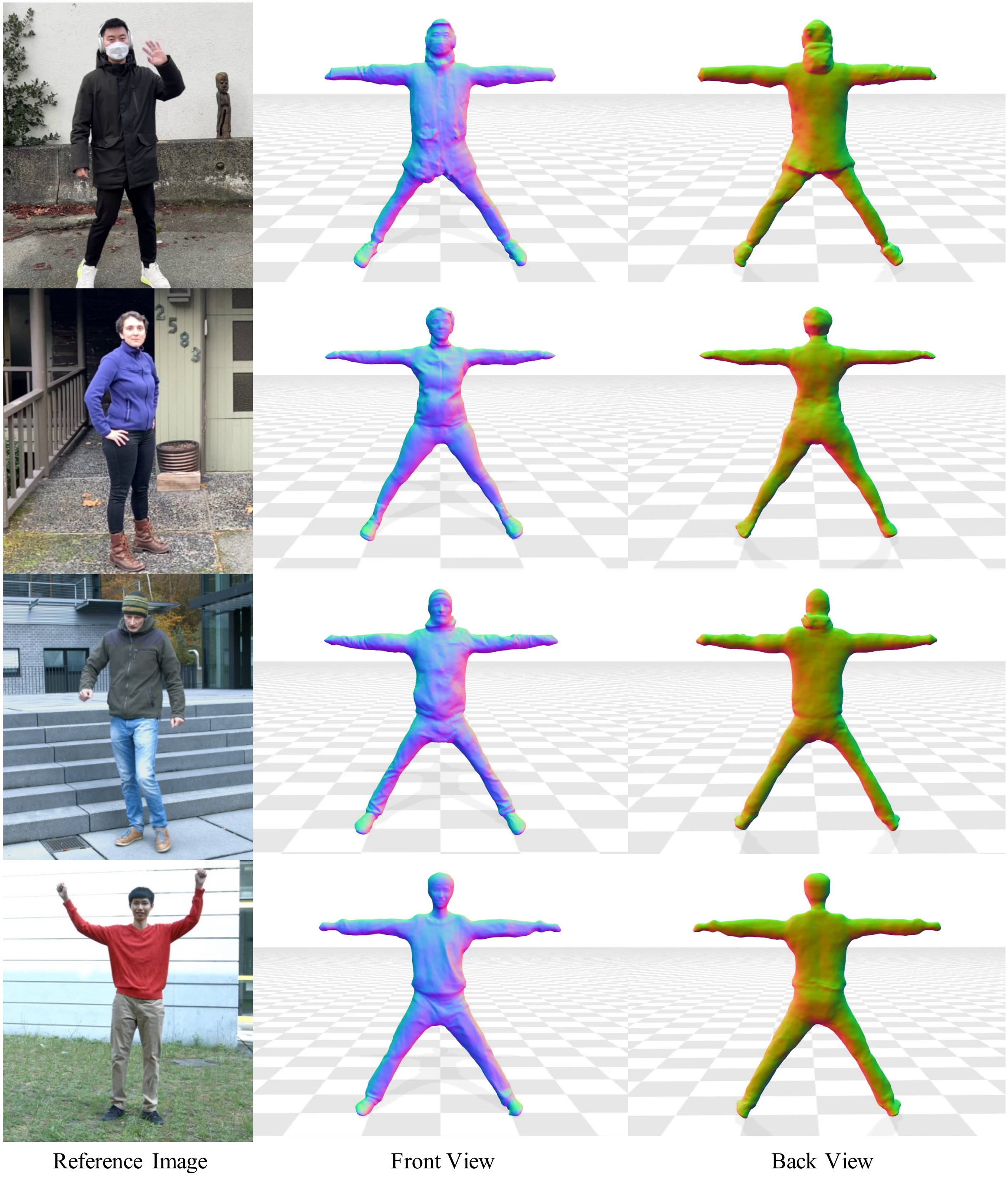

Figure 17. **Visualization of front and back view of reconstructed 3D avatars in canonical space.**

References

[1] *Unreal*, 2020. https://www.unrealengine.com. 3

[2] *Rembg*, 2022. https://github.com/danielgatis/rembg. 3

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), jul 2015. 3

[5] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. 1

[6] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. 1987. 2

[7] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2

[8] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 4

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[10] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[11] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 1

[13] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 3

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[15] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 5

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1

[18] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 2

[19] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 5

[20] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 3

[21] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 2

[22] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 3

[23] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. 37(2):27:1–27:15, May 2018. 2, 4

[24] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[25] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 1, 2