Supplementary Material
MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild

Zeren Jiang*[1] Chen Guo*[1] Manuel Kaufmann[1] Tianjian Jiang[1]
Julien Valentin[2] Otmar Hilliges[1] Jie Song[1]
[1]ETH Zürich [2]Microsoft

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

Contents

1. Implementation Details

1.1. Neural Human Model

Parameterization Details. We simplify the human representation with one neural network f^p in our manuscript. In practice, we model the geometry and texture field using two separate neural networks similar to [3, 18]. Our SDF network f_s^p takes the point and the human pose parameters $\boldsymbol{\theta}^p$ as input and outputs the signed distance value s^p along with global geometry features $\boldsymbol{z}$ of dimension 256. Our texture network f_c^p takes the point, the human pose parameters $\boldsymbol{\theta}^p$, points' normals $\boldsymbol{n}_d$ in deformed space, and the extracted 256-dimensional global geometry feature vectors $\boldsymbol{z}$ from the SDF network as input and predicts the radiance value $\boldsymbol{c}^p$. Specifically, the points' normals $\boldsymbol{n}_d$ are calculated by the spatial gradient of the signed distance field f_s^p w.r.t. the 3D position in deformed space, following [3, 21]. This facilitates better disentanglement of human geometry and appearance reconstruction.

Deformation Module. We define the SMPL-based mapping $T_{\text{smpl}}(\cdot)$ from canonical space to deformed space and its inverse mapping $T_{\text{smpl}}^{-1}(\cdot)$ as follows:

$$T_{\text{smpl}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{n_b} w_c^i(\boldsymbol{x})\, \mathbf{B}_i(\boldsymbol{\theta})\boldsymbol{x}, \tag{15}$$

$$T_{\text{smpl}}^{-1}(\boldsymbol{x}, \boldsymbol{\theta}) = (\sum_{i=1}^{n_b} w_d^i(\boldsymbol{x})\, \mathbf{B}_i(\boldsymbol{\theta}))^{-1}\boldsymbol{x}, \tag{16}$$

where $\mathbf{B}_i(\boldsymbol{\theta})$ derived from the body pose parameters $\boldsymbol{\theta}$ represents the bone transformation matrix for joint $i \in \{1, ..., n_b\}$. Here n_b is equal to the total number of SMPL bones. And $\boldsymbol{w}_{(\cdot)}(\boldsymbol{x}) = \{w_{(\cdot)}^1(\boldsymbol{x}), ..., w_{(\cdot)}^{n_b}(\boldsymbol{x})\}$ denotes the skinning weights for $\boldsymbol{x}_{(\cdot)}$. Canonical points $\boldsymbol{x}_c^p$ are associated with the nearest SMPL vertex' skinning weights and analogously for points $\boldsymbol{x}_d^p$ in deformed space. Hence, a canonical point $\boldsymbol{x}_c^p$ is mapped to the point $\boldsymbol{x}_d^p$ in deformed space via LBS: $\boldsymbol{x}_d^p = T_{\text{smpl}}(\boldsymbol{x}_c^p, \boldsymbol{\theta}^p)$. Similarly, the canonical correspondence $\boldsymbol{x}_c^p$ for point $\boldsymbol{x}_d^p$ can be found through $\boldsymbol{x}_c^p = T_{\text{smpl}}^{-1}(\boldsymbol{x}_d^p, \boldsymbol{\theta}^p)$.

Network Architecture. The canonical human shape network f_s^p for the p-th person is modeled as an MLP with 8 fully connected layers, each of which consists of a weight normalization layer [14] and a softplus activation layer. Each fully connected layer contains 256 neurons. Given the input point, we apply positional encoding with 6 frequency components to better model high-frequency details [11]. The canonical human texture network f_c^p is modeled

*These authors contributed equally to this work

as an MLP with 4 fully connected layers, each of which has the same architecture as the human shape network layers. Except, it uses the Sigmoid activation function for the last layer and uses a ReLU [12] activation function for the rest layers. All subjects in the scene are modeled individually and share the same network architectures.

1.2. Background Modeling and Scene Composition

Quadruple Reparameterization We follow the inverted sphere parameterization of NeRF++ [20] to represent the background. Our human models are defined to be within a spherical inner volume with a radius equal to 3 and the background covers the complementary space. Specifically, each 3D background point $\boldsymbol{x}_b = (x_b, y_b, z_b)$ is reparametrized by the quadruple $\boldsymbol{x}'_b = (x'_b, y'_b, z'_b, \frac{1}{r})$, where $\|(x'_b, y'_b, z'_b)\| = 1$ and $(x_b, y_b, z_b) = r \cdot (x'_b, y'_b, z'_b)$. Here r denotes the magnitude of the vector from the camera origin to $\boldsymbol{x}_b$. This reparameterization of the background points helps to improve the numerical stability and weight farther away points with lower resolution. To obtain the background component RGB value, we follow NeRF++ and sample 32 background points. This is done by uniformly sampling $\frac{1}{r}$ in the range [0, $\frac{1}{3}$], where 3 corresponds to the pre-defined inner volume radius. Given the sampled $\frac{1}{r}$, we calculate the corresponding background point $\boldsymbol{x}'_b$ using the geometric relationship derived in [20].

Scene Composition To obtain the final rendered pixel value, we raycast the human layers and the background volume separately and composite the rendered color of humans $\hat{C}^H$ with the one of the background $\hat{C}^B$. The final pixel color value is calculated by:

$$C = C^H + (1 - \hat{O}^H)\, C^B, \tag{17}$$

where $\hat{O}^H = \sum_{i=1}^{N} \sum_{p=1}^{P} \left[o_i^p \prod_{q=1}^{P} \prod_{j \in \mathcal{Z}_i^{q,p}} \left(1 - o_j^q\right) \right]$ is the total opacity for all the person in the scene, and we follow the same notations as our manuscript.

Network Architecture The background network f^b consists of two parts: the 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 background points. The texture network only includes 1 block of a fully connected layer with 128 neurons, a weight normalization layer, a ReLU activation layer, and a Sigmoid activation layer at the end. Both the density network and the texture network take the quadruple parameterization of the sampled background point, view direction, and per-frame learnable time encoding as input and output the density and the view-dependent radiance value. The per-frame time encoding helps to compensate for dynamic changes in the environment.

1.3. Preprocessing

To obtain pose initialization, we first estimate the SMPL [9] parameters and the tracking ID for each person by the pre-trained TRACE model [15]. Then we extract the bounding box of the estimated SMPL model and feed it to the ViTPose-H to obtain the corresponding 2D keypoints for each subject. Non-maximum suppression (NMS) is applied for each bone to filter out the potential duplicated estimation. Finally, we employ the 2D reprojection loss L_{2d} to optimize SMPL parameter based on the estimated 2D joints:

$$L_{2d}^p = \sum_{i=1}^{K} s_i^2 \rho(\Pi(\mathcal{J}_i(\boldsymbol{\theta}^p, \boldsymbol{\beta}^p)) - J_i), \tag{18}$$

where K denotes the number of 2D joints. J_i and s_i represents the i-th estimated 2D keypoint from ViTPose and its confidence score respectively. $\mathcal{J}(\boldsymbol{\theta}^p, \boldsymbol{\beta}^p)$ are the corresponding 3D SMPL keypoints given the pose $\boldsymbol{\theta}^p$ and body shape $\boldsymbol{\beta}^p$ parameters for subject p, and Π is the camera projection function, $\rho(\cdot)$ denote the robust Geman-McClure error function [4]. To alleviate the jittering of the pose estimates, we also deploy a temporary consistency loss:

$$L_{\text{temp}}^p = \sum_{i=1}^{N_{\text{frame}}} \left\| \mathcal{J}^i(\boldsymbol{\theta}^p, \boldsymbol{\beta}^p) - \mathcal{J}^{i-1}(\boldsymbol{\theta}^p, \boldsymbol{\beta}^p) \right\|_2^2, \tag{19}$$

where $\mathcal{J}^i(\boldsymbol{\theta}^p, \boldsymbol{\beta}^p)$ denotes all 3D joints for i-th frame. Then, we calculate the total loss for the preprocessing stage:

$$L_{\text{pre}} = \sum_{p=1}^{P} (L_{2d}^p + \lambda_{\text{temp}} L_{\text{temp}}^p), \tag{20}$$

where P is the number of subjects and λ_{temp} is the hyper-parameter to balance the weight of the temporal loss term.

1.4. Final Training Objectives

The final objective L is defined by:

$$\begin{aligned} L = {} & L_{\text{rgb}} + \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{inter}} L_{\text{inter}} + \lambda_e L_e \\ & + \lambda_{\text{bce}} L_{\text{bce}} + \lambda_{\text{stab}} L_{\text{stab}}, \end{aligned} \tag{21}$$

where L_{rgb}, L_{mask}, L_{depth}, L_{inter}, and L_e are introduced in the manuscript exhaustively (Eq. 9 - Eq. 14). Following [3], a self-supervised ray classification loss L_{bce} is applied to delineate dynamic foreground and background. Moreover, in the early training stage, a stabilization loss L_{stab} is enforced to supervise the opacity $\hat{O}^p$ of p-th person to be close to 1 if it is inside the corresponding SMPL surface. The formula of L_{bce} and L_{stab} are shown as follows:

$$L_{\text{bce}} = -\hat{O}^H \log(\hat{O}^H) + (1 - \hat{O}^H) \log(1 - \hat{O}^H), \tag{22}$$

Figure 7. **MMM Dataset.** We show the captured image, ground-truth meshes, and ground-truth camera trajectories of our MMM dataset.

$$L_{\text{stab}} = \sum_{p=1}^{P} \sum_{\mathbf{r} \in \mathcal{R}^{p}_{\text{smpl}}} \left\| \hat{O}^{p}(\mathbf{r}) - 1 \right\|, \tag{23}$$

where $\mathcal{R}^{p}_{\text{smpl}}$ denotes the set of rays that intersect with the SMPL of person p.

1.5. Training Details

It is essential to note that we initialize our canonical human shape network with a generic SMPL body shape by using a subset of motion sequences released in AMASS [10]. We optimize our neural networks and pose parameters using the Adam optimizer [6]. The learning rate for training our neural networks is set to $l = 5e^{-4}$ and the learning rate for optimizing the pose parameters is set at one-tenth of l initially. We decay the learning rates in half after 400 and 800 epochs respectively. The other Adam hyper-parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A model is trained on a single NVIDIA A100 (80GB) with batch size 2048 for 2 to 5 days, depending on the number of subjects in the scene.

2. Evaluation Details

2.1. Hi4D Dataset

For all the experiments, we use 4 sequences (in total 515 frames), *i.e. pair15-fight15-view4*, *pair16-jump16-view4*, *pair17-dance17-view28*, and *pair19-piggyback19-view4*, in Hi4D [19] dataset to evaluate our method. We downsample all the frames by factor 2 for training efficiency with 470×640 image resolution.

2.2. MMM Dataset

We collect a new dataset called Monocular Multi-huMan (MMM), which is captured by a monocular dynamic camera, to evaluate the monocular multi-person human reconstruction task. This dataset contains six sequences with two to four subjects in each sequence. Half of the sequences (569 frames) are captured in the stage with ground-truth meshes and camera trajectory for quantitative evaluation and the others are captured in the wild for qualitative evaluation. The ground-truth meshes are reconstructed by 106 synchronized cameras (53 RGB and 53 IR cameras) via commercial software [2], while the hand-held camera trajectories are tracked by AprilTag [16]. Specifically, we first stick the printed AprilTag to the backside of the smartphone. Then we locate the AprilTag in the stage coordinate using our dense multi-view camera system. However, there is still an offset between the AprilTag and the camera of the smartphone. To get the ground-truth camera trajectories, we utilize the silhouette loss between the projected ground-truth mesh mask and human segmentation mask estimated by RVM [8] to optimize the offset. The captured image, the ground-truth meshes, and the ground-truth camera trajectories are shown in Fig. 7.

2.3. Reconstruction Comparisons

ECON [17] is a state-of-the-art regression-based model for reconstructing 3D humans from individual frames and Vid2Avatar (V2A) [3] is a self-supervised method that reconstructs a single performer from monocular videos without using 3D human data. For a fair comparison, we extend V2A to multi-person scenarios by learning a distinct human model for each subject in the scene individually. Specifically, we simply use the pre-trained checkpoints of ECON and infer the 3D human shapes from our video frames. We feed V2A with the same pose and mask initialization as ours without modifying the original framework, objectives, and hyperparameters. In order to visualize the multi-person reconstructions of V2A in overlapped format, we generate the depth map for each subject and compose the rendered normal maps based on the depth value that represents the distance between the point and the camera origin, *i.e.*, we select the pixel value of the subject that is closer to the camera origin.

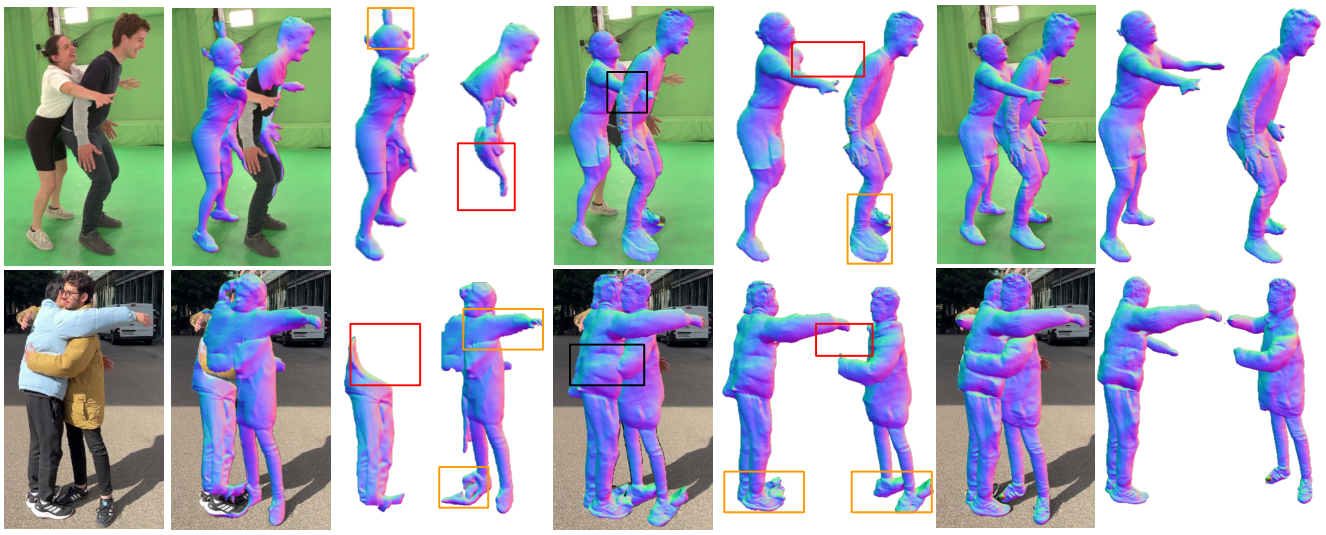

Figure 8. **Additional qualitative reconstruction comparison.** We show both the overlaid and separated reconstruction results for each method. **Red** bounding boxes: the incomplete reconstruction of the occluded part. **Orange** bounding boxes: incorrect instance segmentation results caused by the surrounding visual complexities. **Black** bounding boxes: inaccurate spatial arrangement due to pose error.

Figure 9. **Additional qualitative rendering comparison.** Our method achieves more plausible renderings with sharp boundaries.

Figure 10. **Addtional instance segmentation comparisons.** Our method produces more accurate instance segmentation masks while SCHP fails to associate the pixels to the correct subject when people closely interact.

We follow the evaluation protocol of Hi4D for quantitative comparisons, *i.e.*, we first perform the ICP between each reconstructed mesh and the corresponding ground-truth instance mesh and then we calculate the reconstruction metrics. However, since ECON doesn't include human tracking inherently, we additionally apply the Hungarian matching based on the cost matrix of the ICP between each reconstructed mesh and the ground-truth instance mesh after the convergence of ICP.

3. Additional Experimental Results

3.1. Reconstruction Comparisons

We provide additional qualitative reconstruction comparisons with ECON [17] and Vid2Avatar [3] in Fig. 8. Com-

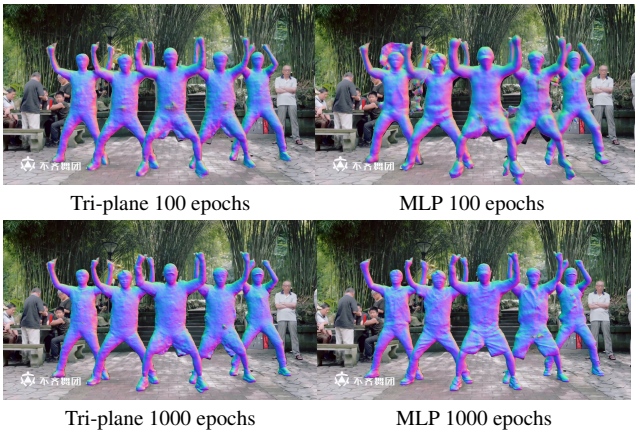

Figure 11. **Comparison between Tri-plane and MLP representation.** Tri-plane representation with a shared decoder converges faster than MLP-based human representation. However, Tri-plane representation produces less fine-grained surface details.

pared to state-of-the-art human reconstruction methods in both categories (learning-based and optimization-based), our method outperforms them by a substantial margin both quantitatively (*cf*. Tab. 2 in manuscript) and qualitatively (*cf*. Fig. 8). Our framework recovers the fine-grained surface details and maintains complete human bodies.

3.2. Novel View Synthesis Comparisons

We show additional qualitative novel view synthesis comparisons in Fig. 9. Our method achieves better disentanglement between multiple people, yielding more plausible novel view renderings with clearly sharp boundaries.

3.3. Instance Segmentation Comparisons

Our primary goal is not to provide a standalone method for instance segmentation, but accurate segmentation of people is an essential component to achieve high-quality reconstruction. We complement the qualitative comparisons with SCHP [7] in Fig. 10. Our instance segmentation is most effective when people are in contact. Our method produces more accurate instance segmentation masks while SCHP fails to associate the pixels to the correct subject when people closely interact. To provide further evidence, we recalculate the metrics on frames labeled as "contact" by Hi4D, which correlates with occlusion. Tab. 6 reveals a more pronounced difference between MultiPly, SCHP, and the initialization than Tab. 4 in the main paper did (for instance 95.1% IoU vs 91.2% (Init.) and 90.5% (SCHP)).

Method	IoU ↑	F1 ↑	Recall ↑	Precision ↑
SCHP	0.905	0.972	0.973	0.972
Ours (Init.)	0.912	0.975	0.992	0.959
Ours (Progressive)	**0.951**	**0.987**	**0.994**	**0.979**

Table 6. **Instance segmentation** evaluation on "contact" Hi4D.

3.4. Pose Estimation Comparisons

We supplement the qualitative comparison with state-of-the-art multi-person pose estimation approaches in Fig. 13. Our framework performs effectively in correcting implausible human poses and spatial arrangement.

3.5. Neural Tri-Plane Human Representation

We demonstrate in this section that our strategies to obtain robust instance segmentation and plausible 3D human poses generalize well to different neural human representations. Similar to our neural human representation in the manuscript, we proposed another neural human model that is designed based on the tri-plane representation [1] with a commonly shared feature-conditioned MLP decoder. Specifically, each human model is represented as a separate tri-plane in canonical space which allows us to query points' high-frequency feature vectors efficiently from the feature plane parallel to xy, yz, xz. We conduct the ablation studies under the same evaluation protocol as the main paper and show the quantitative results in Tab. 7. It demonstrates our key algorithmic components (layer-wise volume rendering, progressive prompting strategy, and confidence-guided alternating optimization) consistently improve the reconstruction quality, showing the superiority of our proposed learning schemes.

Furthermore, the proposed neural tri-plane human representation is able to accelerate the training procedure by a considerable margin (halved the convergence time compared to MLP-based human representation). The reasons are twofold: 1) tripled feature planes provide more expressive high-frequency latent features and querying points' features directly from tri-plane perform more efficiently compared to using fully connected layers with positional encoding, 2) our shared decoder can be trained to decode the common features among all subjects in the scene which is particularly beneficial since the people in the same scene tend to have a similar dressing style. However, this representation produces less fine-grained surface details in turn, as shown in Fig. 11.

4. Visualization

As shown in Fig. 12, our method MultiPly generalizes to various people with different human shapes and miscellaneous clothing styles and performs robustly against different levels of occlusions, close human interaction, and environmental visual complexities.

5. Limitations and Societal Impact Discussion

Our method struggles with fast human movement that potentially results in strong motion blurs in image observations. This would negatively impact our pose initialization and obstruct photorealistic reconstructions. Our ap-

Figure 12. **Additional qualitative results.** Our method MultiPly generalizes to various people with different human shapes and miscellaneous clothing styles and performs robustly against different levels of occlusions, close human interaction, and environmental visual complexities.

Metrics	Pose Estimation				Human Reconstruction			
	MPJPE ↓	MVE ↓	CD ↓	PCDR ↑	V-IoU ↑	C $-\ell_2$ ↓	P2S ↓	NC ↑
Initial pose	75.3	90.8	235.6	0.566	-	-	-	-
Layer-wise volume rendering	70.8	85.1	248.6	0.605	0.746	3.85	3.69	0.719
+ Progressive SAM	70.7	85.0	247.7	0.609	0.793	2.72	2.44	0.777
+ Confidence-guided OPT	**69.7**	**83.6**	**217.0**	**0.689**	**0.803**	**2.62**	**2.39**	**0.785**

Table 7. **Quantitative ablation studies on Hi4D based on tri-plane representation.** We demonstrate the importance of the proposed progressive prompt for SAM and confidence-guided alternating optimization. Both key components effectively contribute to the final reconstruction quality. This improvement generalize to different neural human representations.

Figure 13. **Qualitative pose estimation comparisons.** CLIFF, TRACE, and our pose initialization all fail to estimate the correct spatial arrangement of the close interacted pairs on the Hi4D dataset. Simply optimizing both pose and shape jointly during training doesn't help to refine inaccurate pose estimates. In contrast, our confidence-guided alternating optimization performs effectively in correcting implausible human poses and spatial arrangement (e.g. the interpenetration between the arm and the body, and the wrong depth order of legs).

proach demonstrates its superior performance with garments that exhibit topological similarity to the human body. Loose clothing, such as skirts or free-flowing garments, still presents challenges owing to their rapid dynamics. Based on the current formulation and human modeling, the computational complexity of our framework increases linearly with the number of involved persons, making it inefficient for a crowd of people in the scene. Our neural tri-plane human representation with shared decoder architecture presented in the supplementary material makes a step towards a more efficient formulation. However, this is far from sufficient. Future work could incorporate recent advances in fast and memory-efficient neural representations [5]. Furthermore, our method does not explicitly model hands and we believe the integration of an expressive human model [13] is a promising future direction.

MultiPly, for the first time, enables high-fidelity digitization of multiple people under natural interaction from a single monocular video, which has the potential to make a broad range of downstream applications in movie and gaming industries, and AR/VR. The final outcome of our framework is multiple realistic human avatars, that can be animated with driven signals. This could raise risks of privacy leaks and misuse of human avatars, for example, deepfakes. Primary attention should be given to addressing these concerns before incorporating digital human avatars into products. Openly, the objective of this work is to facilitate the application of the technology in ways that are beneficial for society. Regrettably, the prevention of malevolent applications of such technology remains unattainable. Nevertheless, we contend that a thorough examination of these methodologies with maximum transparency, including the discussion of technical intricacies in the paper, along with the release of code and data, should be prioritized over undisclosed research. This approach is crucial for developing effective countermeasures to mitigate the potential for unscrupulous applications.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 5

[2] 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), 2015. 3

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4

[4] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, pages 337–346. Wiley Online Library, 2009. 2

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 7

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 3

[7] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

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

[9] 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

[10] 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, 2019. 3

[11] 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

[12] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, page 807–814, Madison, WI, USA, 2010. Omnipress. 2

[13] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 7

[14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 1

[15] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[16] John Wang and Edwin Olson. AprilTag 2: Efficient and robust fiducial detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016. 3

[17] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[18] 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. 1

[19] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

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

[21] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1