GUAVA: Generalizable Upper Body 3D Gaussian Avatar

Supplementary Material

Overview

This supplementary material presents more details and additional results not included in the main paper due to page limitation. The list of items included is:

- Video demo at Project page with a brief description in Appendix A.
- More model implementation details in Appendix B.
- More results and EHM tracking in Appendix C.
- Further discussion and ethical considerations in Appendix D.2.

A. Video Demo

We highly recommend readers watch the video demo in the supplementary materials. The video showcases GUAVA's self-reenactment and cross-reenactment animation results, as well as novel view synthesis. Additionally, we compare GUAVA with both 2D-based [5, 6, 11] and 3D-based [2, 18, 19] methods under self-reenactment. We also compare it with MimicMotion [18], Champ [19], and MagicPose [2] under cross-reenactment. Finally, we present the visual results of ablation studies. These results demonstrate that our method enables more detailed and expressive facial and hand motion while maintaining ID consistency with the source image across various poses.

B. More Implementation Details

B.1. Training details

We train the model for a total of 200,000 iterations with a fixed learning rate of 1e-4. The learning rate for certain MLPs gradually decreases linearly to 1e-5 over the training process, while the weights of DINOv2 [12] remain frozen. Initially, we set the LPIPS [17] loss weight λ_{lpips} to 0.025 and increase it to 0.05 after 10,000 iterations. Other loss weights are set as follows: $\lambda_f = 0.25$, $\lambda_h = 0.1$, $\lambda_1 = 1.0$, $\lambda_p = 0.01$, $\lambda_s = 1.0$. The hyperparameters ϵ_{pos}, ϵ_{sca} are set to 3.0 and 0.6, respectively.

B.2. Model details

In the reconstruction model, the appearance feature map F_a output by the image encoder is passed through convolutional layers to transform its dimensions to 32 and 128, which are then used for the UV Gaussians prediction and template Gaussians prediction branches, respectively.

In the UV branch, the appearance feature map is concatenated with the original image, and inverse texture mapping is applied to map the features to the UV space, resulting in $F_{uv} \in \mathbb{R}^{H\times W\times 35}$. This is passed into the UV decoder's StyleUnet, which outputs a 96-dimensional feature map. The feature map is then further processed by a convolutional module to decode the Gaussian attributes for each pixel. Additionally, the ID embedding f_{id} is injected into StyleUnet via an MLP.

In the template Gaussian prediction branch, the projection feature f_p and the base vertex feature f_b are both set to a dimension of 128. The ID embedding f_{id} is mapped to 256 dimensions via an MLP.

For Gaussian representation, we discard the spherical harmonic coefficients and use a latent feature c with a dimension of 32 to model the Gaussian appearance. Through splatting, we obtain a rough feature map with a dimension of 32. To help the refiner decode finer images from the rough feature map, we use a loss function to ensure that the first three channels of the latent feature represent RGB. Some details of the model are not shown in Fig. 2 of the main paper, for clarity.

B.3. Evaluation details

For self-reenactment evaluation, MagicPose struggles with synthesizing black backgrounds. To avoid the background color influencing the evaluation metrics, we use the ground truth mask to remove the background. Similarly, for Champ, since it uses SMPL-rendered maps [9] (*e.g.*, normal and depth) as input, the generated images may include legs. To ensure accurate metric calculations, we also apply the ground truth mask to filter out the irrelevant parts.

B.4. Inverse texture mapping

Here, we explain inverse texture mapping with added details for clarity and ease of understanding. Given the tracked mesh and its corresponding information, including the vertices of each triangle and the three UV coordinates for each triangle, we can locate the area covered by each triangle on the UV map. Then, we identify which triangle each pixel belongs to and calculate its barycentric coordinates. For each pixel, we use its barycentric coordinates to interpolate the triangle vertices and calculate the corresponding position t on the mesh. Next, we project each pixel onto screen space based on its position t:

$$x_{uv}^j = \mathcal{P}(t^j, RT_s),\ j \in [0, H \cdot W], \tag{1}$$

where RT_s is the viewing matrix of the source image and $\mathcal{P}$ denotes projection. Finally, we perform linear sampling on the appearance feature map, reshape it to $H \times W \times 35$, and obtain F_{uv}, completing the inverse texture mapping of the appearance feature map to UV space. To filter out features

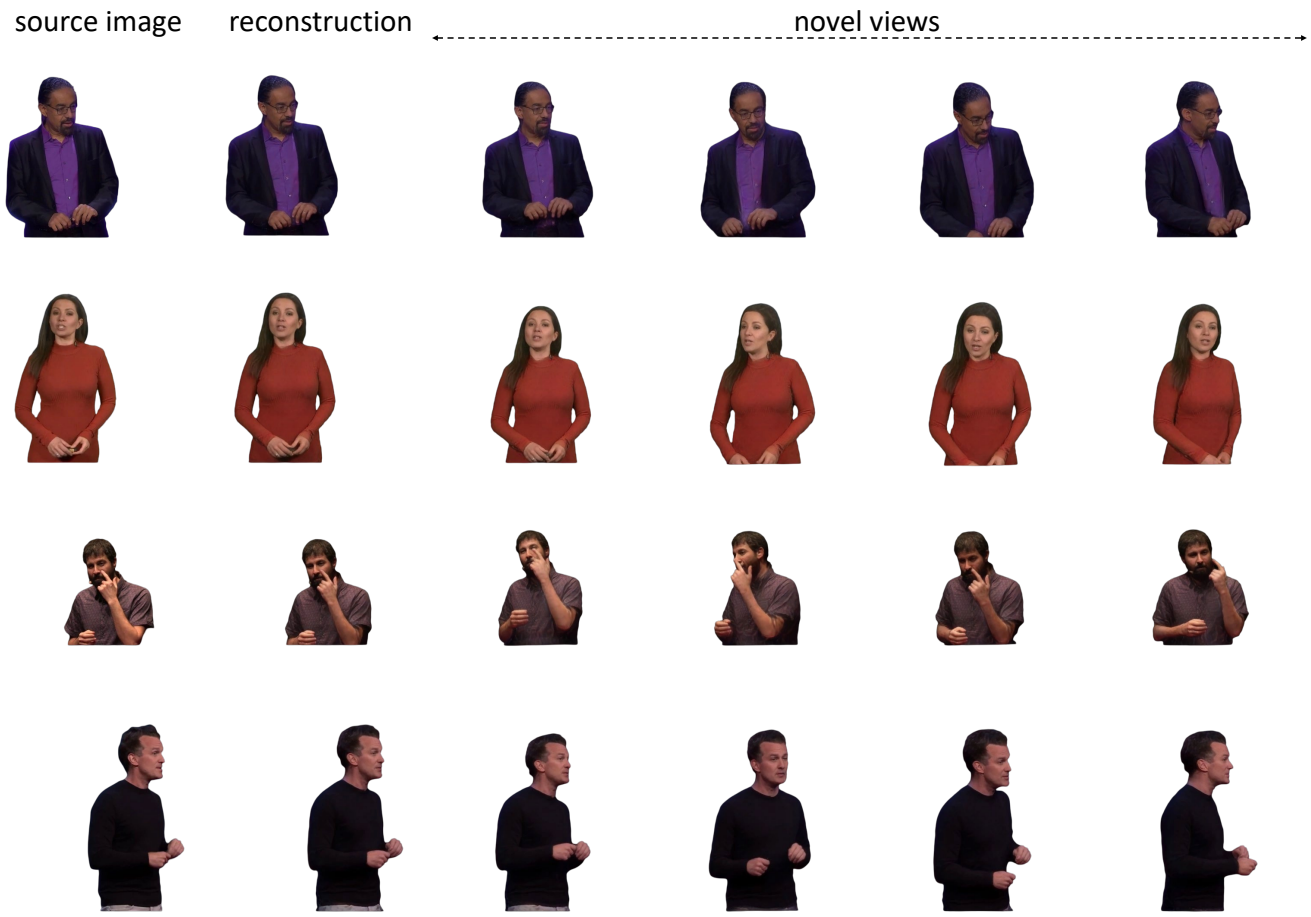

Figure S1. Visual results on novel view synthesis. Our method effectively generates reasonable 3D information while ensuring strong multi-view consistency and preserving the details of the source image.

from invisible regions, we use the Pytorch3D rasterizer to render the tracked mesh and acquire the visible triangles. For pixels corresponding to invisible triangles, their features are set to 0.

B.5. EHM tracking

In the main paper Sec. 3.1, we briefly introduce the tracking of the template model's parameters using keypoint loss, omitting some details for clarity. However, the actual process is more complex. Here, we provide a more detailed description of our tracking method. Given images of a human's upper body, we first estimate the keypoints K_b using DWPose [16]. Based on these detected keypoints, we crop the head and hand regions. During the cropping process, we also record the affine transformation matrices A_f and A_h. For the cropped hand image, we use HaMeR [14] to estimate the parameters $Z_h = (\beta_h, \theta_h)$ for both hands. For the upper body, we estimate the SMPLX [13] parameters $Z_b = (\beta_b,\ \theta_b)$ using PIXIE [4].

Face Tracking. Based on the cropped head image, we estimate three sets of keypoints K_f^1, K_f^2 and K_f^3 using FaceAlignment [1], MediaPipe [10], InsightFace [3], where we only use the mouth keypoints from K_f^3. We also perform a rough estimation of the FLAME [7] parameters $Z_f = (\beta_f,\ \psi_f,\ \theta_f)$ using Teaser [8], where θ_f includes the jaw pose θ_{jaw}, eye pose θ_{eye}, neck pose θ_{neck} and a global pose θ_{fg}. Then, we optimize the rotation R and translation T of the camera parameters, as well as Z_f, for 1000 iterations, with the loss function defined as follows:

$$\begin{aligned}\mathcal{L}_{face-track} = \sum_{i}^{3} \lambda_{k,i}^{f} \mathcal{L}_1(K_f^i, \hat{K_f^i}) + \\ \lambda_{smo}^{f} \mathcal{L}_{smo}(Z_f, R, T) + \lambda_{reg}^{f} \mathcal{L}_{reg}(Z_f).\end{aligned} \tag{2}$$

Here, $\mathcal{L}_{smo}$ and $\mathcal{L}_{reg}$ represent the smoothness loss between adjacent frames and the regularization loss (constraining parameters toward zero), respectively. Next, we optimize the eye pose θ_{eye} for 500 iterations using keypoint loss and smoothness loss, focusing on the eye keypoints.

Additionally, since the FLAME model does not include

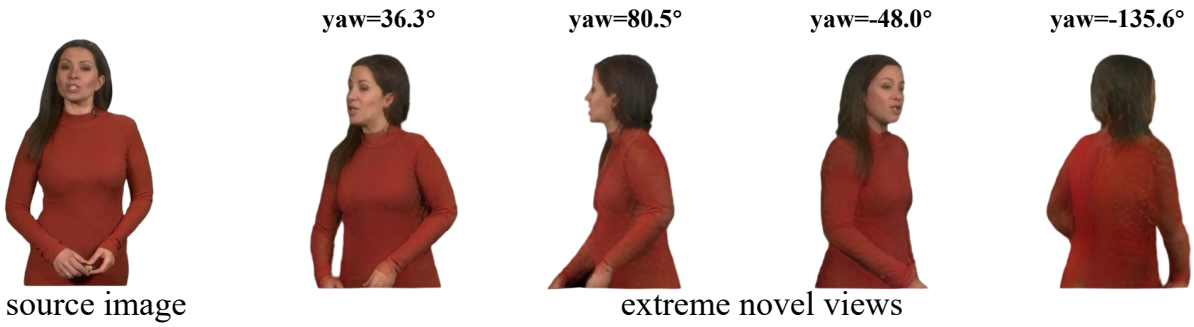

Figure S2. Visual results on extrapolated novel view synthesis. Renderings of back-facing regions are slightly lower quality due to the lack of backside data in our training set.

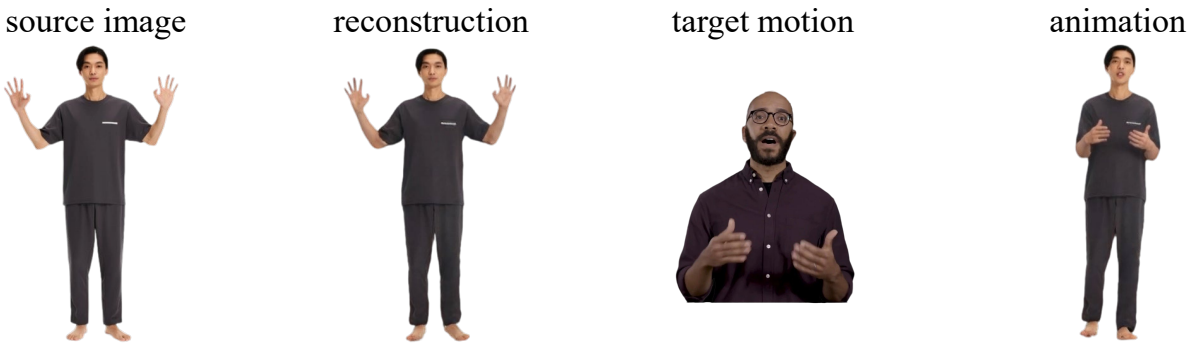

Figure S3. Full-Body reconstruction results. Our method successfully performs full-body reconstruction. Acquiring additional full-body data is expected to improve the results in such cases.

a mouth interior mesh, we follow [15] to incorporate teeth into FLAME. The upper and lower teeth meshes are initialized accordingly, with their poses driven by the neck and jaw joints, respectively.

Body Tracking. After head tracking, we replace the head part of SMPLX with the neutral-pose expressive FLAME model $M_f(\beta_f, \psi_f, \theta_{jaw}, \theta_{eye})$ (with zero global and neck pose) to obtain EHM, as described in main paper Eq. 1. We then optimize the body parameters using not only 2D keypoint loss but also a 3D guidance loss from the tracked FLAME model $M_f(Z_f)$ and the tracked MANO model $M_h(Z_h)$. Since these tracked models align with their respective cropped image regions, they serve as accurate guidance. By applying the recorded affine transformation, we convert their vertices form local to global space and compute the error between EHM's head and hand vertices and these references, enhancing pose optimization and alignment accuracy. The following loss function is used to optimize Z_b, β_f, θ_h as well as camera parameters R and T:

$$\begin{aligned}\mathcal{L}_{body-track} = \lambda^b_k \mathcal{L}_1(K_b, \hat{K}_b) + \lambda^b_{reg}\mathcal{L}_{reg} + \\ \lambda^b_{smo}\mathcal{L}_{smo}(Z_b, \theta_h, R, T)(Z_b, \theta_h) + \\ \lambda^b_{3d}\mathcal{L}_{3d}(M_{ehm}, A_f^{-1}M_f(Z_f), A_h^{-1}M_h(Z_h)) + \\ \lambda^b_{prior}\mathcal{L}_{prior}(Z_b),\end{aligned} \tag{3}$$

where, $\mathcal{L}_{prior}$ represents the constraint loss applied to the pose parameters using VPoser [13], enforcing a prior distribution.

C. More Results

C.1. Novel views synthesis

Reconstructing a 3D upper-body avatar from a single image is an ill-posed problem. However, GUAVA learns the animation of diverse subjects from different viewpoints, enabling it to generalize and infer certain 3D information. As a result, the reconstructed avatar not only supports animation but also enables novel view synthesis. Fig. S1 presents our results, demonstrating high multi-view consistency while preserving the subject's identity. Additionally,

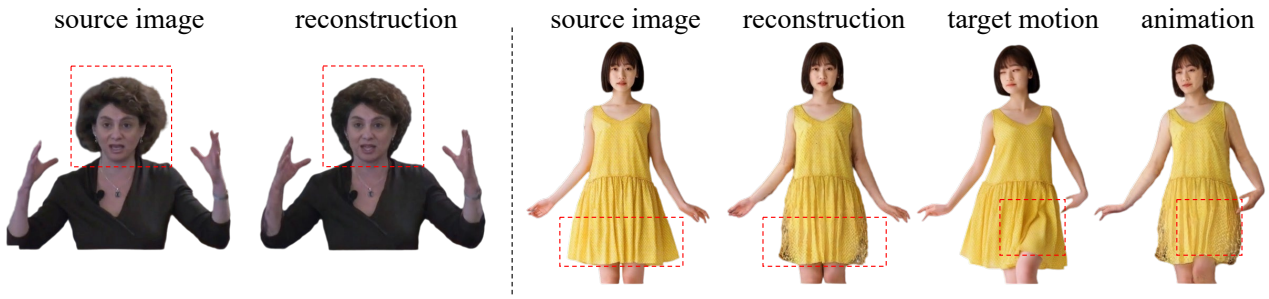

Figure S4. Visualization of failure cases. Our method exhibits limitations when handling fluffy hair, loose clothing, and flowing folds.

Figure S5. Visual results of our EHM tracking method. Without EHM, the model can only capture basic mouth-opening and -closing movements, whereas our method accurately tracks subtle facial expressions. Additionally, our approach successfully captures complex and detailed hand gestures.

synthesized novel views exhibit high-quality rendering with fine texture details.

We also demonstrate novel view synthesis from extreme angles, as shown in Fig. S2. Since our training dataset includes only front views, back rendering is suboptimal. Training on 360° datasets, as IDOL [20] could help.

C.2. Full-body

Our method is also applicable to full-body settings as Fig. S3. However, due to few full-body data in our dataset, adding more data could improve generalization.

C.3. Failure cases

As shown in Fig. S4, our method has limitations with fluffy hair, loose clothing, and flowing folds — all calling for future improvements.

C.4. EHM tracking results

Although we have demonstrated the improvement in model performance with the EHM model from both qualitative and quantitative perspectives in the main paper Sec. 4.3, to provide a more intuitive comparison, we further present the visual tracking results in Fig. S5. It is important to note that "w/o EHM" refers to tracking with SMPLX, which still uses our designed tracking framework, but without the FLAME integration step. From the results, it is clear that SMPLX can only roughly capture mouth movements, while EHM captures detailed facial expressions. Furthermore, our designed tracking framework not only captures accurate facial expressions but also tracks fine gestures, including finger movements, with high precision.

D. More discussion

D.1. EHM vs SMPLX

As discussed in the main paper Sec. 2.1, although SMPLX [13] integrates SMPL [9] and FLAME [7], its expression space is newly trained on full-body scans, which may overlook fine facial details. This results in SMPLX having less expressive facial expressiveness compared to FLAME, a limitation also noted in ExAvatar (Sec. 3.1) [11]. EHM's

main contribution is to improve this facial expressiveness.

D.2. Ethical considerations

The generalization of 3D human avatar reconstruction technology raises several potential ethical concerns. First, unauthorized data collection and processing could lead to privacy violations, particularly with sensitive personal information like facial features and body shape. Second, this technology could be misused to create deepfake content, leading to identity theft, fraud, and other illegal activities. Additionally, the rapid reconstruction and real-time animation could be exploited to spread misinformation or engage in online harassment. Therefore, strict adherence to data protection regulations, ensuring informed consent, and taking measures to prevent misuse are essential. Transparency and traceability of technology should also be prioritized to build public trust and minimize potential risks.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 2

[2] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 1

[3] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018. 2

[4] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 2

[5] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[6] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19876–19887, 2024. 1

[7] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 4

[8] Yunfei Liu, Lei Zhu, Lijian Lin, Ye Zhu, Ailing Zhang, and Yu Li. Teaser: Token enhanced spatial modeling for expressions reconstruction. *arXiv preprint arXiv:2502.10982*, 2025. 2

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1, 4

[10] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[11] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024. 1, 4

[12] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[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. 2, 3, 4

[14] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2

[15] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 3

[16] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 2

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[18] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 1

[19] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[20] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26308–26319, 2025. 4