Supplementary Material
Vid2Avatar-Pro: Authentic Avatar from Videos in the Wild via Universal Prior

Chen Guo[*1,2] Junxuan Li[*1] Yash Kant[1,3] Yaser Sheikh[1] Shunsuke Saito[†1] Chen Cao[†1]
[1]Codec Avatars Lab, Meta [2]ETH Zürich [3]University of Toronto

In this **supplementary document**, we provide additional materials to supplement our manuscript. In Sec. 1, we provide further implementation details of our proposed method Vid2Avatar-Pro. Sec. 2 explains details of our experiment, including dataset descriptions and the evaluation protocol. Furthermore, in Sec. 3, we show additional qualitative comparisons to show our superior performance over prior art and additional ablation studies of Vid2Avatar-Pro. Sec. 4 includes more qualitative results on the avatar creation from in-the-wild videos. Finally, we discuss our limitations and potential negative societal impacts in Sec. 5. The **supplementary video** includes additional animation results.

1. Implementation Details

1.1. Normalized Conditioning Data Acquisition

Canonical Pose Definition Our canonical pose $\boldsymbol{\theta}_{\text{cano}}$ characterized by a 60-degree angle between the legs and an additional 90 degrees of palm rotation compared to the standard T-pose of SMPL-X [19] as shown in the Fig. 1 of the manuscript. This configuration, with increased space between the legs, facilitates the refinement of poses when initial estimates around the legs or feet are inaccurate. Additionally, the rotated palms result in palm-facing canonical maps, which enhance the reconstruction of hands, particularly in capturing individual finger details.

SDF-based Canonical Template. To acquire high-quality conditioning data for universal model training, we first reconstruct the canonical templates for all training subjects. This static reconstruction is based on a single automatically selected keyframe in which the human pose $\boldsymbol{\theta}$ exhibits the greatest similarity to our pre-defined canonical pose $\boldsymbol{\theta}_{\text{cano}}$. Note that the human pose in the selected keyframe is not identical to our canonical pose. Thus, we consider the space in the keyframe as the deformed (posed) space.

Specifically, we represent the 3D shape of the clothed human using an implicit signed-distance field (SDF) and capture the appearance with a texture field within a predefined canonical space with pose $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{cano}}$. Both the SDF and texture field are modeled with a neural network f_s and f_c respectively, similar to [2, 29]. Our SDF network f_s, which models the geometry, takes the canonical point $\mathbf{x}_c$ as input and outputs the signed distance value s along with global geometry features $\mathbf{z}$ of dimension 512. The texture network f_c, receives the canonical point $\mathbf{x}_c$, the points' normals $\mathbf{n}_d$, and the extracted 512-dimensional global geometry feature vectors $\mathbf{z}$ extracted from the SDF network as input, and predicts the radiance value $\mathbf{r}$. In particular, the points' normals $\boldsymbol{n}_d$ are calculated in the deformed space as the spatial gradient of the signed distance field f_s w.r.t. the 3D position in deformed space, following [2, 29, 32]. Note that unlike [2] which aims to reconstruct dynamically moving humans, our goal is to obtain a high-quality canonical textured template that is consistent across different camera views and frames. Thus, we do not inject human pose information into the networks to keep consistency.

We use an inverse mapping approach, similar to [2], to unwarp the ray samples into canonical space. This process enables us to extract the signed distance values and radiance values, which are then used to perform volume integration to obtain the per-pixel color. This allows us to formulate a training loop by comparing the volume-rendered pixel color and the groundtruth image color, thereby updating the weights of f_s and f_c.

We run MISE [14] to extract the canonical template meshes from f_s. Compared to templates reconstructed in Li *et al*. [13], we achieve higher-quality templates with more geometric details which enhance the performance of our universal prior model.

SDF Network Architecture. The canonical shape network f_s is implemented as an MLP comprising 8 fully connected layers. Each layer includes a weight normalization layer [23] and a Softplus activation function. Each fully connected layer consists of 256 neurons. For the input point, we apply positional encoding with 6 frequency components to better model high-frequency details [15]. The canonical texture network f_c is modeled as an MLP with 4 fully connected layers, each of which has the same architecture as the geometry network layers, except that it uses the Sigmoid activation

[*]Equal contribution [†]Equal advisory

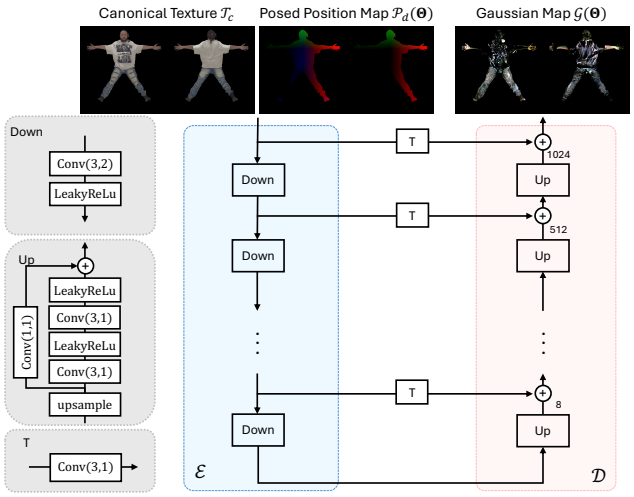

Figure 9. **Universal Prior Model Architecture.**

function for the last layer and ReLU [17] for the rest of the layers. We initialize the shape network f_s with a generic SMPL-X body [19] by directly supervising f_s with signed distance loss that is calculated based on the 3D sample points around the canonical SMPL-X surface.

Canonical Texture Unwrapping. We inherently already obtain per-vertex colors of the canonical template mesh through the texture field. However, we empirically find that these per-vertex colors are limited by the template mesh resolution and can often miss high-frequency texture details. In order to maintain continuous textures with minimal texture information loss, we reformulate this problem as a canonical texture unwrapping task. Specifically, we regard the canonical texture map (with resolution 1024×2048) as a variant of UV parametrization with invalid pixels outside of the projected template mask $\mathcal{M}_c$. We then conduct traditional UV texture unwrapping to obtain the colors on the canonical texture map. Combined with the template mesh and skeleton-based normalization strategy (*cf*. Sec. 3,1 in manuscript), we obtain the normalized identity conditioning data for all training subjects.

1.2. Universal Prior Model Architecture

Our universal prior model backbone is a U-Net, following [1, 12, 22]. The network architecture is depicted in Fig. 9, which includes the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ branches. The "Down" block consists of a convolutional layer with a kernel size of 3 and stride of 2, followed by a leaky ReLU activation function. The "Up" block is composed of a upsampling operation, two duplicated convolutional layers with a kernel size of 3 and stride of 1 followed by leaky ReLU activation functions. The input to the upsampling block is upsampled by a factor of 2 using bilinear interpolation. It then passes through the convolutional layers. We also add a skip connection for each upsampling block. The encoder branch takes the concatenation of the normalized identity conditioning data, *i.e*. $\mathcal{T}_c$, and the posed position map as input, and generates multi-scale feature maps encoded with identity and pose information. The generated feature maps are then added to the corresponding layer of the decoder branch to output the pose-dependent Gaussian maps. The decoder is composed of 8 "Up" blocks, which take a 4×4 map as input and output Gaussian maps at a resolution of 1024×1024. Specifically, the input texture and position map are at a resolution of 1024×1024 with 12 channels in total (incl. front and back maps). In practice, we employ three separate U-Nets to predict three Gaussian maps, which are stored with color offsets (3 channels), position offsets (3 channels), and other Gaussian attributes (8 channels), respectively, similar to [13].

1.3. Diffusion-based Texture Inpainting

Our model of choice is the Diffusion Transformer (DiT [20]) featuring 1.3B parameters and pre-trained on nearly 3B human images. We fine-tune the pre-trained model on a dataset of 1000 unwrapped canonical texture maps (front and back). To augment the training dataset and minimize the domain gap between studio data and in-the-wild videos, we create 200 inpainting masks (visibility masks) for each subject through rasterization of the pre-acquired studio canonical templates using $2-4$ randomly positioned sparse cameras. To tailor the DiT for the inpainting task, we incorporate a ControlNet-like module [30] to inject spatial information from the inpainting mask. We also integrate the MoVQ [31] autoencoder, which achieves an 8x spatial downsampling with a 4-channel latent space. Given the limited dataset, we find the placement of ControlNet modulation within the DiT crucial to avoid disrupting the original network weights. During pre-training, we provide image-only embeddings via CLIP [21] and DinoV2 [18] with ViT-L/14 architecture to provide weak supervision and alignment towards target generations. The training process unfolds in two stages: initially at a low resolution of 128×128 for 50K iterations, followed by a high resolution 1024×1024 per view (front and back) for 30K iterations. We use a learning rate of $1e$-4 in all stages and use the AdamW optimizer with beta values of $[0.9, 0.98]$ and an epsilon value of $1e-6$. We also use linear warmup over the first 1000 learning steps starting from $1e-6$.

1.4. Preprocessing for In-the-Wild Personalization

Given the monocular in-the-wild video, we estimate the shape and per-frame pose parameters by using an off-the-shelf SMPL-X estimator [24]. This initial estimation is often inaccurate with a wrong camera assumption (very large focal length). Thus, we employ state-of-the-art 2D keypoint predictor Sapiens [8] to estimate the 2D joint positions. We formulate an offline optimization to refine the human shape

Figure 10. **Additional comparisons with ExAvatar.** Compared to ExAvatar, our method creates higher-quality 3D human avatars with finer-grained appearance details (*e.g.*, clothing wrinkles and facial features), and generalizes better to out-of-distribution driving signals.

and pose estimates by minimizing the 2D keypoint projection error. These 2D keypoints also serve as point prompts for SAM-HQ [7] to extract the human segmentation masks.

Given the human shape/pose and foreground masks of the monocular video, we apply a similar method as stated in Sec. 1.1 to obtain the canonical template. Different from reconstructing the canonical templates for the studio data, we also optimize the estimated human shape/pose parameters $\mathbf{\Theta}$ jointly with f_s and f_c. The extracted template is then normalized based on an average human skeleton scale to attain the spatially aligned identity conditioning data. For the preprocessing of in-the-wild videos, we try to follow a similar strategy as done for studio data that is used for training the universal prior model. This can largely help to mitigate the inherent domain gap in conditioning data between in-the-wild sequences and the high-quality multi-view training data of the prior model. The preprocessing for in-the-wild videos takes approximately 6 hours in total, which is on a similar level as ExAvatar [16] ($\sim$ 6-7 hours).

1.5. Training Details and Efficiency

For the training of our universal prior model, we use Adam optimizer [9] with a learning rate of $5e^{-4}$. We train the model with a batch size of 64 for 500k iterations, using 64 NVIDIA A100 GPUs. It takes about 5 days to converge. For the personalized avatar fine-tuning stage, we use the same optimizer and learning rate, and fine-tune the pre-trained universal prior model using only 1 NVIDIA A100 GPU for 2k iterations. The fine-tuning stage takes about 10 minutes. During inference, the rendering speed is about 20 fps for 960×540 resolution image rendering. Note that the current implementation is research-only, and both the training and testing efficiency can be further improved with codebase optimization.

2. Evaluation Details

2.1. NeuMan Dataset

For the interpolation view synthesis comparisons on NeMan dataset [6], we use the same estimated human shape, poses, and segmentation masks as ExAvatar [16] to run all baseline comparisons. We follow the official training and testing splits to train/fine-tune on the monocular observations. The quantitative results of HumanNeRF [27], InstantAvatar [5] and GaussianAvatar [4] in Tab. 1 of the manuscript are sourced from [4, 16].

2.2. MonoPerfCap Dataset

For the extrapolation view synthesis comparisons on MonoPerfCap dataset [28], we curate 4 sequence clips *i.e.*, *Helge_outdoor*, *Nadia_outdoor*, *Natalia_outdoor*, and *Weipeng_outdoor*, where in total 1490 frames (first 80%)

Figure 11. **Number of training IDs.** The rendering quality consistently increases when the universal prior model is trained on more identities/data. Especially, more appearance details can be recovered, *e.g.*, the eyes and the pocket of the jeans, and our full model trained on 1000 subjects do not suffer from the problem caused by inaccurate opacity predictions, *cf*. the collar of the hoodie.

Table 4. **Importance of inpainting.** Our diffusion-based inpainting module can effectively complete the textures that are missing from the monocular observations (*cf*. Fig. 12).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Ours w/o Inpainting	30.17	0.977	2.22
Ours	30.22	0.977	2.18

are used for training/fine-tuning and 373 frames (remaining 20%) are used for testing.

2.3. Self-Captured Videos

In addition to the publicly available datasets, we also capture monocular in-the-wild videos using an iPhone to demonstrate the superior performance of our method on in-the-wild videos. In total, 18 participants have freely volunteered to participate in this data collection with a signed consent form. They have been duly informed about the intended use and are recorded with daily motions. All captures are conducted outdoors with both static and moving cameras. The avatars created using the iPhone captures are already shown in Fig. 1 and Fig. 7 of the manuscript. The original captured videos can be found in the supplementary video.

2.4. Evaluation Protocol

Following previous works [5, 16], for the evaluation on testing frames of both NeuMan and MonoPerfCap datasets, we fit SMPL-X parameters [19] of testing frames while freezing all other parameters with the loss stated in the Eq. 7 of the main paper. For baseline methods (incl. NeuMan [6], Vid2Avatar [2], and ExAvatar [16]) that jointly model the human and the background, we only compare the foreground (*i.e.*, human) rendering quality. This evaluation protocol is slightly different from the one that was applied in ExAvatar [16], where the estimated segmentation masks from SAM [10] are first used to mask out the background in the rendered images which include both the foreground and the scene. That explains the mismatched quantitative results of

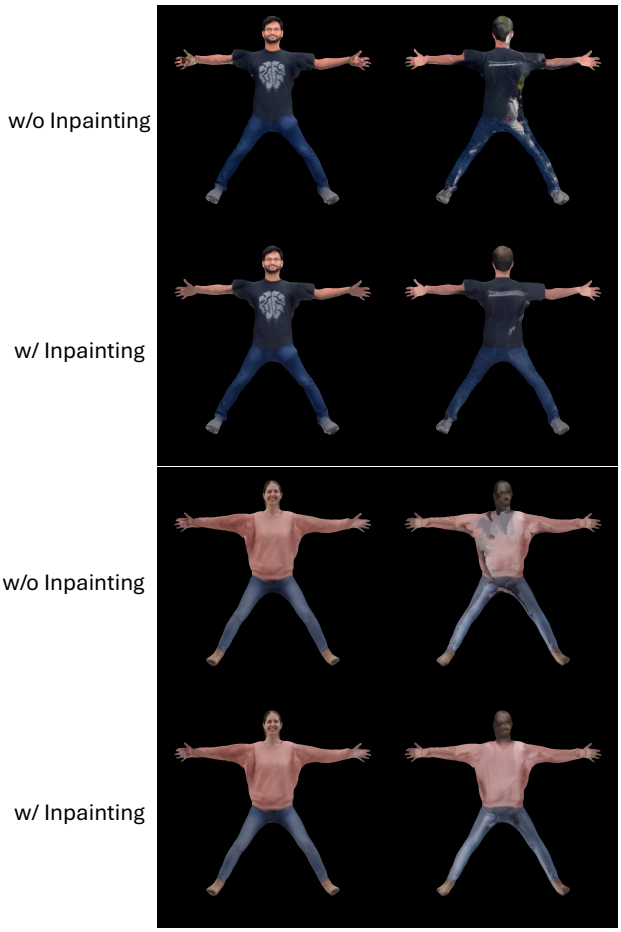

Figure 12. **Canonical texture inpainting.** Our diffusion-based inpainting module can effectively complete the textures that are missing from the monocular observations.

Figure 13. **Additional Visual animation results of avatars created from monocular in-the-wild videos.** The created 3D avatars can be animated using novel human poses and demonstrate highly detailed appearance from arbitrary view points.

ExAvatar in our comparisons and its original report.

3. Additional Experimental Results

3.1. Animation Comparisons with ExAvatar

To further illustrate the superior 3D human avatars rendering quality of our method, we present additional qualitative comparison results with ExAvatar [16] on out-of-distribution human poses. ExAvatar is the state-of-the-art approach to reconstructing animatable human avatars from monocular videos and it is the method achieved highest quality among all our baseline methods. As shown in Fig. 10, our method creates higher-quality 3D human avatars with finer-grained appearance details (*e.g.*, clothing wrinkles and facial features), and generalizes better to unseen novel motions.

3.2. Training Data

We show the qualitative ablation study on the number of training identities in Fig. 11. We observe that the final rendering quality consistently improves when the universal prior model is trained on more identities/data. Especially, more appearance details such as the eyes and the pocket can be recovered. Our full model trained on 1000 subjects does not suffer from the problem caused by inaccurate opacity predictions, *cf*. the collar of the hoodie.

3.3. Diffusion-based Texture Inpainting

We provide more qualitative ablation studies on the effectiveness of in Fig. 12. We also quantitatively measure the improvement of our diffusion-based inpainting module, presented in Tab. 4. Specifically, we select the *Nadia_outdoor* sequence from MonoPerfCap dataset [28] in which the test split contains most body regions that are invisible from the training split. Fig. 12 and Tab. 4 show that our diffusion-based canonical inpainting module improves our final results both qualitatively and quantitatively.

4. Visualization

Fig. 13 presents additional animation results of avatars created from monocular videos captured in under-controlled environments. Vid2Avatar-Pro demonstrates its ability to generalize across diverse identities and garment styles, yielding highly realistic renderings of novel human poses and view points.

5. Limitations and Societal Impact Discussion

Our method Vid2Avatar-Pro still relies on reasonable initial human shape and pose estimates, and segmentation masks as inputs. The robustness against imperfect segmentation masks can be improved by incorporating the background

Figure 14. **Results on loose outfits.** Our method generates plausible renderings for less challenging driving signals but fail to output promising results for challenging human poses.

Figure 15. **Results on extreme lighting.** The brightness of the created human avatars is in its imperfection in case of a dark capture environment.

modeling, similar to [11, 16].

The current training dataset for our universal prior model lacks the capture of diverse facial expressions and dynamic capture data of human subjects wearing loose garments. Therefore, our method currently does not support animatable faces or highly realistic animations of human avatars dressed in loose outfits. We show two examples from [3, 25] in Fig. 14. Our method can generate plausible renderings for less challenging driving signals but fails to output promising results for challenging human poses. Future work could incorporate more training data with rich human facial expressions and performance capture data of human subjects dressed in free-flowing garments to realize more authentic 3D human avatars.

Additionally, Vid2Avatar-Pro assumes standard lighting conditions and may not perform optimally in environments with extreme lighting variations. For example, as shown in Fig. 15, when the environment is dark, the brightness of the created human avatars is also in its imperfection. We believe training a universal relightable prior model for clothed humans is a promising future direction to address this issue.

The current efficiency bottleneck for the in-the-wild personalization lies in the preprocessing stage. The acceleration strategies can be borrowed from [5, 26].

Vid2Avatar-Pro enables high-fidelity 3D digitization of humans from monocular videos captured in uncontrolled environments. This capability holds significant potential to enhance a variety of downstream applications, including those in the film and gaming industries, as well as virtual communication within augmented and virtual reality (AR/VR) environments. The ultimate output of Vid2Avatar-Pro consists of photorealistic 3D human avatars, which can be animated into novel poses based on corresponding driving signals. However, this capability raises potential concerns related to privacy breaches and the misuse of digital assets. Specifically, there is a risk of creating digital avatars of individuals without their consent, followed by the possible misappropriation of these avatars for unethical or dubious purposes. When developing methods for avatar creation, whether for research purposes or commercial products, it is imperative to prioritize addressing these concerns. Our goal is to facilitate the use of such technology in ways that benefit society. However, it is important to acknowledge that it is not possible to fully guarantee the prevention of malicious applications. We advocate for a transparent and comprehensive approach to developing these methodologies, emphasizing the importance of openly discussing technical details in research papers and making code and data accessible. This strategy is essential for fostering the development of effective countermeasures that can mitigate the potential risks associated with unethical applications, rather than pursuing undisclosed research endeavors.

References

[1] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 2

[2] 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, 4

[3] 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, 2020. 6

[4] 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. 3

[5] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4, 6

[6] 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. 3, 4

[7] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 3

[8] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *Computer Vision – ECCV 2024*, pages 206–228, Cham, 2025. Springer Nature Switzerland. 2

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

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 4

[11] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[12] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2

[13] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[14] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 1

[15] 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

[16] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 3, 4, 5, 6

[17] 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

[18] 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. 2

[19] 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. 1, 2, 4

[20] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Infor-*

mation Processing Systems. Curran Associates, Inc., 2016. 1

[24] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1834–1843, 2024. 2

[25] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[26] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6

[27] 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, 2022. 3

[28] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *SIGGRAPH*, 37(2):27:1–27:15, 2018. 3, 5

[29] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, 2021. 1

[30] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2

[31] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: modulating quantized vectors for high-fidelity image generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[32] 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