Text-driven Human Avatar Generation by Neural Re-parameterized Texture Optimization

Kim Youwang[1] Tae-Hyun Oh[1,2,3]

[1]Dept. of Electrical Engineering and [2]Grad. School of Artificial Intelligence, POSTECH
[3]Institute for Convergence Research and Education in Advanced Technology, Yonsei University

Abstract

We present TexAvatar, a text-driven human texture generation system for creative human avatar synthesis. Despite the huge progress in text-driven human avatar generation methods, modeling high-quality, efficient human appearance remains challenging. With our proposed neural re-parameterized texture optimization, TexAvatar generates a high-quality UV texture in 30 minutes, given only a text description. The generated UV texture can be easily superimposed on animatable human meshes without further processing. This is distinctive in that prior works generate volumetric textured avatars that require cumbersome rigging processes to animate. We demonstrate that TexAvatar produces human avatars with favorable quality, with faster speed, compared to recent competing methods.

1. Introduction

Virtual human avatar is a key component in multimedia industrial fields such as movies, games, and AR/VR. Professional graphic designers strive to embody realistic or creative virtual human characters. Still, the hand-designed generation of animatable and textured 4D human avatars requires cumbersome and time-consuming efforts with intensive labor and the pain of creation. To reduce such burdens, approaches for capturing or generating natural human motions have been extensively studied [4, 24, 20, 9, 18], but studies for capturing or generating realistic human textures remain challenging.

Several attempts tried to capture human textures from visual observations [2, 15, 10, 1, 23, 5]. While those methods may faithfully reconstruct human textures, they require carefully captured multi-view images or expensive texture datasets. More importantly, they cannot generate human texture, which limits the scope of their application. To build an affordable and easy-to-use human texture generation system for real-world applications, we focus on the task of generating high-quality human textures from text descriptions, *i.e.*, text-driven human texture generation.

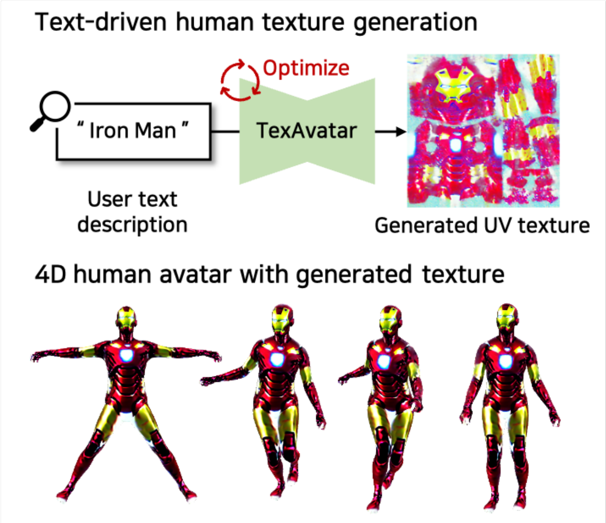

Text-driven human texture generation
Optimize
" Iron Man "
TexAvatar
User text description
Generated UV texture
4D human avatar with generated texture

Figure 1. Given the user's text description, TexAvatar generates high-resolution UV texture by synthesis-through-optimization. Generated textures can be rasterized onto any SMPL-X motion sequence to synthesize 4D human avatars.

In this work, we present *TexAvatar*, an efficient text-driven human texture generation system. Given a text description about the avatar's appearance, TexAvatar generates a high-quality UV texture image that conforms to the input prompt via synthesis-through-optimization (see Fig. 1). Due to the absence of high-resolution UV texture map datasets paired with the text descriptions, TexAvatar leverages the powerful text-to-image diffusion model for its supervision during optimization [17, 19]. The core of TexAvatar is the neural re-parameterized optimization for UV texture generation. Instead of directly optimizing the pixel values of the UV texture map, we propose to re-parameterize the initial UV texture map into a neural network and optimize its neural parameters.

We highlight that TexAvatar is preferable than recent text-driven human avatar generation works [25, 6, 8, 3, 7] in that

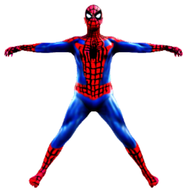
	AvatarCLIP (SIGGARPH 2022)	DreamAvatar (arxiv 2023)	AvatarCraft (ICCV 2023)	DreamWaltz (arxiv 2023)	TexAvatar (Ours)
					
3D representation	NeuS (SDF)	NeRF	NeuS (SDF)	NeRF	Mesh
Optimization time	~ 6 hrs	~ 2 hrs	~ 3 hrs	~ 2 hrs	< 0.5 hrs
Texture	3D point color	3D point color	3D point color	3D point color	UV Texture map
Post-processing for animation	Rigging + learned motion module	Per-avatar learned transform	Per-vertex transform	Learned 3D point transform	Not needed

Figure 2. **Comparison w/ competing methods**. TexAvatar generates a human UV texture map with synthesis-through-optimization. The optimization takes about 0.5 hours, and no post-processing is required for superimposing the texture onto the meshes and animating meshes.

it optimizes the UV texture map, which can be efficiently rasterized with a parametric human mesh model, *i.e.*, SMPL-X [16] while achieving 4∼10× faster optimization speed with favorable human appearance quality.

We summarize our main contributions as follows:

- *TexAvatar*, a text-driven generation of high-quality UV texture maps for human avatars.
- Neural re-parameterized texture optimization that utilizes a neural network prior to generating high-resolution and locally-smooth texture.
- Significantly faster texture optimization for enabling real-world applications and seamless rasterization of generated texture maps with animatable human meshes.

2. Related work

Our task is related to text-driven human texture generation. Recently, a few impressive works tried to ease such a process with an avatar generation method that only requires simple text prompts [13, 25, 6, 8, 3, 7]. These methods can be categorized into (1) Volumetric avatar-based and (2) Mesh-based texture generation methods.

Volumetric avatar-based texture generation. With the huge success of neural radiance fields, *i.e.*, NeRF [14], and volume rendering, most recent methods [6, 8, 3, 7] model human texture as per-point color of 3D volumetric spaces. As a seminal work, AvatarCLIP [6] implemented a 3D human avatar as a volumetric representation called NeuS [22], which is initialized with the template SMPL [11] mesh. AvatarCLIP optimizes the geometry and per-point color of 3D points sampled from volumetric spaces. Many follow-up works [8, 3, 7] showed improved results using the text-to-image diffusion model [19] as their supervision or modifying volumetric representations. However, most volumetric avatar-based texture generation methods suffer from significant drawbacks, where they need cumbersome post-processing to animate their generated 3D avatar. More importantly, such methods require 2 to 6 hours to synthesize a static avatar, making the real-world application impractical.

Our proposed TexAvatar leverages the efficient rasterization of high-resolution UV texture maps and achieves faster generation, about 30 minutes. As a favorable by-product, TexAvatar does not need any post-processing to animate the textured avatars, showing its potential for real-world graphics applications (Fig. 2).

Mesh-based texture generation. Another line of work is a mesh-based texture generation method. Seminal works, Text2Mesh [13] and CLIP-Actor [25] have shown plausible visual qualities and faster texture generation speed. However, their generation qualities are limited because they generate per-vertex features for representing human textures. Such texture representations generate diffuse color values for discrete and coarse mesh vertices and interpolate the vertex colors to express mesh face colors, resulting in blurry textures. Moreover, such per-vertex texture cannot guarantee a locally-smooth texture. In other words, it is hard to regularize adjacent vertices to have similar colors, which is mandatory for realistic human textures.

In this work, TexAvatar aims to generate high-resolution and human UV texture map images, which can express locally-smooth and high-fidelity texture while efficiently texturizing animated meshes.

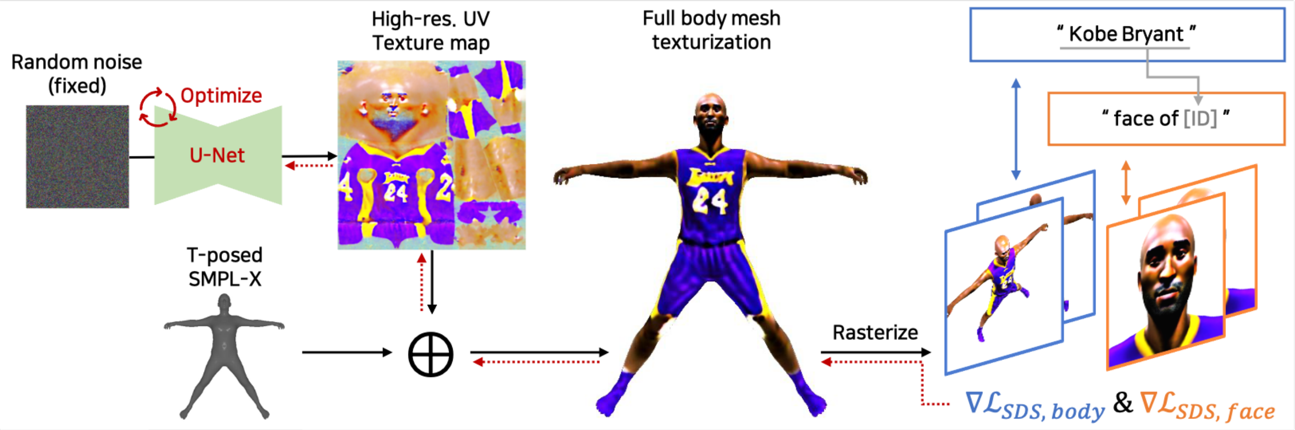


Figure 3. **TexAvatar: Overall system**. TexAvatar generates a high-resolution UV texture map via synthesis-through-optimization. Starting from a randomly sampled noise, TexAvatar generates a UV texture map. Then, the generated texture map is superimposed on the T-posed SMPL-X mesh, generating a full-body textured mesh. Given a user's text description of the appearance of the human avatar, TexAvatar differentiably rasterizes the textured mesh and computes the loss to obtain the update gradient. Note that TexAvatar re-parameterizes a UV texture map with U-Net, and updates U-Net with the obtained gradient rather than directly optimizing the pixel values of the texture map.

3. Text-driven Neural Re-parameterized Human Texture Optimization

In this section, we first explain the TexAvatar optimization. Then, we elaborate more on the benefits of neural re-parameterized texture optimization.

TexAvatar optimization. TexAvatar consists of two main parts: a U-Net-based texture map generation and a pre-trained diffusion model-based texture update (see Fig. 3). TexAvatar generates a high-resolution UV texture map by optimizing U-Net G_θ. Note that an arbitrary input is fixed during the optimization. Since TexAvatar synthesizes a texture map through optimization, it will produce a random texture map at the very early stage of its optimization. The generated UV texture map is superimposed on the canonical, *i.e.*, T-posed, SMPL-X mesh. We implement this by leveraging pre-defined texture mapping of the SMPL-X mesh model. Then, TexAvatar rasterizes the textured human mesh into multi-view full-body and face-focused images.

Given multi-view images of the avatar's body and face images $\mathbf{I}$, Score-Distillation Sampling (SDS) [17] is applied to compute the text-conditioned texture update gradients. Specifically, we employ a denoising function of the pre-trained text-to-image latent diffusion model $\mathcal{D}_\phi$, *i.e.*, Stable-Diffusion [19]. We then add random noise for t steps, obtaining noisy image $\mathbf{I}_t$. Given the text prompt y, the denoising function estimates the injected noise $\hat{\epsilon}_\phi(\mathbf{I}_t; y, t)$, where its error measured between the actual injected noise ϵ becomes the update direction $\nabla_\theta\mathcal{L}_{\text{SDS}}$ for TexAvatar, *i.e.*,

$$\nabla_\theta\mathcal{L}_{\text{SDS}}(\phi, \mathbf{I}) = \mathbb{E}_{t,\epsilon}\left[m(t)(\hat{\epsilon}_\phi(\mathbf{I}_t; y, t) - \epsilon)\frac{\partial \mathbf{I}}{\partial \theta}\right], \quad (1)$$

where $m(t)$ denotes a weighting function conditioned on the diffusion noise timestep, and θ denotes the neural parameters of the texture generator U-Net. Iterative update of U-Net G_θ with $\nabla_\theta\mathcal{L}_{\text{SDS}}$ finally yields a high-quality UV texture map that conforms to the user's text description.

Neural re-parameterized texture optimization. Before trying our neural re-parameterization, one might try a naïve baseline method to optimize a UV texture map, *i.e.*, direct pixel optimization of UV texture, with the same supervision.

Figure 4 compares direct texture map pixel optimization and our proposed neural re-parameterized texture map optimization. Our proposed optimization's generated UV texture and final textured avatars achieve a much smoother and high-fidelity texture than the baseline method. For the baseline, the supervision $\nabla_\theta\mathcal{L}_{\text{SDS}}$ is strong enough to make coarse texture look plausible, but the lack of pixel-wise dependency leads the locally non-smooth texture, showing substantial texture jitters. On the other hand, our neural re-parameterized optimization inherits the prior [21] induced from the architecture of U-Net G_θ, where it consists of a diverse scale of convolution kernels. Such convolution kernels allow texture optimization to consider locality in texture map space, generating a locally-smooth texture map. Furthermore, since human textures are likely to have similar or repeated patterns across the body, the convolution kernels can help reduce the difficulty of optimization.

4. Results

In this section, we show TexAvatar's qualitative results. In Fig. 5, we visualize the generated human avatars in the canonical pose and animated poses. We utilize the text-to-motion retrieval system proposed by CLIP-Actor [25] to animate the generated human avatars in the canonical pose.

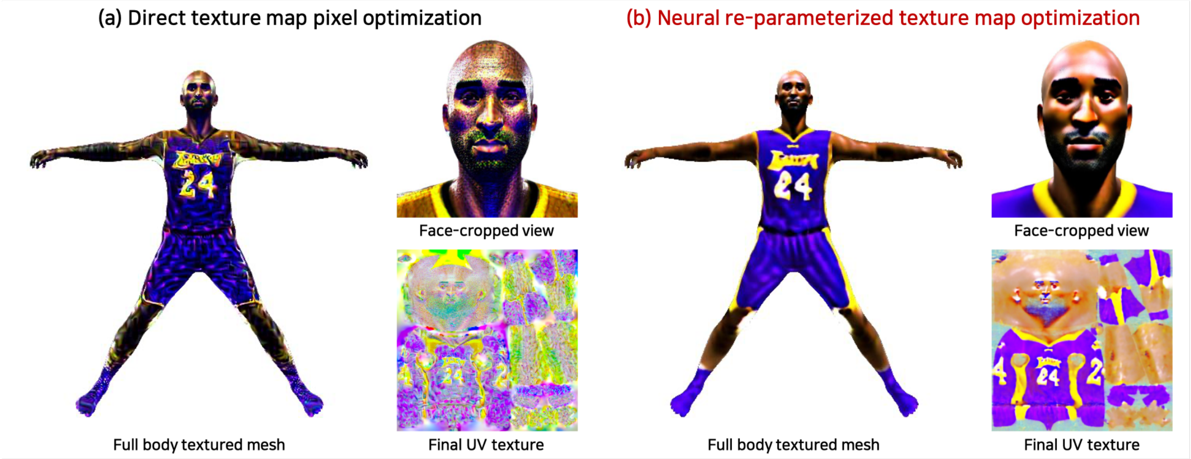


Figure 4. **Benefits of neural re-parameterized texture optimization**. By updating the convolutional kernels in texture generator U-Net, G_θ, TexAvatar yields high-quality and locally-smooth UV textures compared to the naïve baseline of direct texture map pixel optimization.

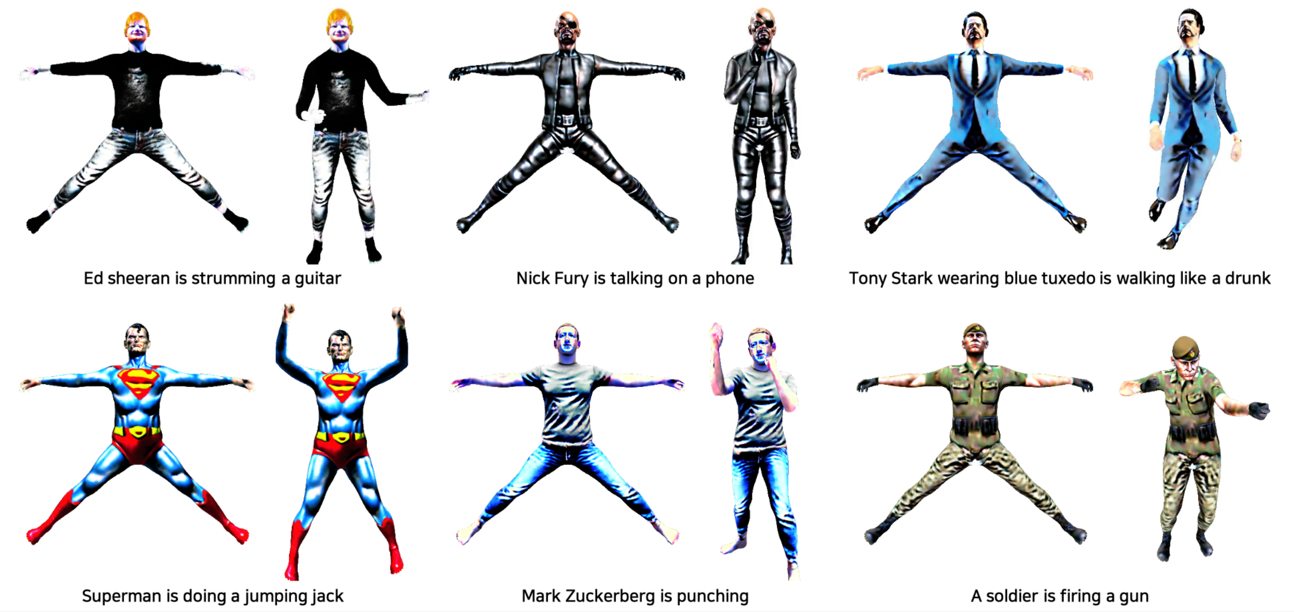

Ed sheeran is strumming a guitar
Nick Fury is talking on a phone
Tony Stark wearing blue tuxedo is walking like a drunk
Superman is doing a jumping jack
Mark Zuckerberg is punching
A soldier is firing a gun

Figure 5. **Qualitative results**. Generated UV texture maps can be superimposed on any animated SMPL-X mesh sequences without post-processing. MoCap datasets [12], text-to-motion retrieval, or generation [25, 20, 9] can be used for textured human avatar animation.

The qualitative results demonstrate the plausible and high-quality human texture generation of TexAvatar.

5. Conclusion

We present *TexAvatar*, a text-driven human UV texture generation method. A synthesis-through-optimization system with neural re-parameterized texture and pre-trained text-to-image diffusion model as a supervision yield significantly improved quality of human texture and faster optimization.

The limitation of TexAvatar is its absence of geometric details. Since this work is in progress, we plan to extend TexAvatar to support surface geometry displacements for more details and more vivid texture in our future work.

Acknowledgment. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.2022-00164860, Development of Human Digital Twin Technology Based on Dynamic Behavior Modeling and Human-Object-Space Interaction and No.2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation)

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[2] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[3] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models, 2023. 1, 2

[4] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[5] A. Grigorev, K. Iskakov, A. Ianina, R. Bashirov, I. Zakharkin, A. Vakhitov, and V. Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[6] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4):1–19, 2022. 1, 2

[7] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars, 2023. 1, 2

[8] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[9] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 1, 4

[10] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, 2019. 1

[11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[12] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[13] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[14] 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 (ECCV)*, 2020. 2

[15] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[16] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 3

[18] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 4

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[22] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[23] Xiangyu Xu and Chen Change Loy. 3D human texture estimation from a single image with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[24] Kim Youwang, Kim Ji-Yeon, Kyungdon Joo, and Tae-Hyun Oh. Unified 3d mesh recovery of humans and animals by learning animal exercise. In *British Machine Vision Conference (BMVC)*, 2021. 1

[25] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4