SIAVATAR: ANIMATABLE 3D GAUSSIAN AVATAR FROM A SINGLE IMAGE

Anonymous

ABSTRACT

Despite the progress of 3D Gaussian Splatting (3DGS), reconstructing one-shot animatable 3D human avatars from a single image remains a challenging task. Existing 3DGS-based methods primarily rely on appearance observation and motion cues from monocular videos to reconstruct animatable 3D avatars. However, when applied to single-image setups, these methods struggle to extract accurate 3D features from a 2D image, limiting their ability to capture fine-grained appearance details and dynamic deformation, especially from challenging viewpoints. In this work, we propose SIAvatar, a novel single-image human reconstruction method that integrates diffusion-based appearance prior and parametric human model geometry prior within a 3DGS framework. Secondly, a vertex-based adaptive Gaussian densification scheme is introduced to effectively represent human geometry while mitigating artifacts. Extensive experiments demonstrate that SIAvatar generates realistic 3D avatars with plausible appearance details and novel pose animation from a single input image. Video demo: https://sigport.org/documents/siavatar.

***Index Terms*—** single-image human reconstruction, 3D Gaussian Splatting, novel pose animation

1. INTRODUCTION

Customized avatar reconstruction is a significant yet challenging task in computer vision, with broad applications in gaming, virtual reality, and telepresence. 3D human reconstruction methods can generally be categorized into explicit and implicit approaches. Explicit methods [1, 2] optimize mesh parameters based on parametric body models [3, 4] to align with observed images but struggle with detailed clothing and complex deformation. Implicit methods [5, 6, 7], which rely on continuous functions such as signed distance functions [8] and neural radiance fields [9], offer greater flexibility in handling topology but suffer from high computational costs and inefficiencies.

Recently, the emergence of 3D Gaussian Splatting [10] provides a promising framework for 3D reconstruction and novel view synthesis. Existing 3DGS-based approaches [11, 12, 13] leverage abundant appearance observation and motion cues from monocular videos to reconstruct animatable 3D avatars. However, video data often requires a controlled capture environment and motion sequences of a subject performing specific actions, which may not always be feasible, especially in casual environments. In contrast, single-image data offers a more flexible and convenient solution to quick avatar customization without complex capture processes. Nevertheless, 3DGS-based methods typically lack the ability to infer sufficient 3D content from a single 2D image, resulting in difficulties in fine-grained appearance reconstruction and novel pose animation.

To infer rich 3D content from a single image, significantly reduce data requirements, and achieve high-quality rendering, we propose SIAvatar for reconstructing animatable 3D Gaussian avatars from a single image. Our key insight lies in integrating appearance prior from a diffusion model and geometry prior from a parametric human model into a 3DGS framework. The diffusion-based prior enables SIAvatar to generate the plausible appearance details by lifting 2D observation to 3D content, while the geometry prior provides accurate human body structures and a robust foundation for capturing complex poses. Specifically, we utilize a pretrained 3D-aware diffusion model [14] to hallucinate invisible regions of human appearance for multi-view supervision. Then, we estimate a canonical SMPL-X [4] human mesh, which serves as the geometry prior for 3D Gaussian initialization and animation. A point-to-point correspondence is established between 3D Gaussians and the vertices of the human mesh, enabling the application of our vertex-based adaptive Gaussian densification scheme to capture fine human geometry while reducing artifacts. We animate the 3D avatars by performing linear blend skinning (LBS) to the 3D Gaussians using SMPL-X parameters and render them as 2D images in novel views.

Our contributions are summarized as follows:

- We present SIAvatar for animatable 3D Gaussian avatar reconstruction, a single-image method that incorporates diffusion-based appearance prior and parametric human model geometry prior into a 3DGS framework.
- We introduce a vertex-based adaptive Gaussian densification scheme to produce a reasonably compact and precise representation of human geometry and mitigate artifacts.
- Extensive experiments demonstrate the effectiveness of SIAvatar in generating high-quality 3D human avatars with plausible appearance and novel pose animation from a single image.

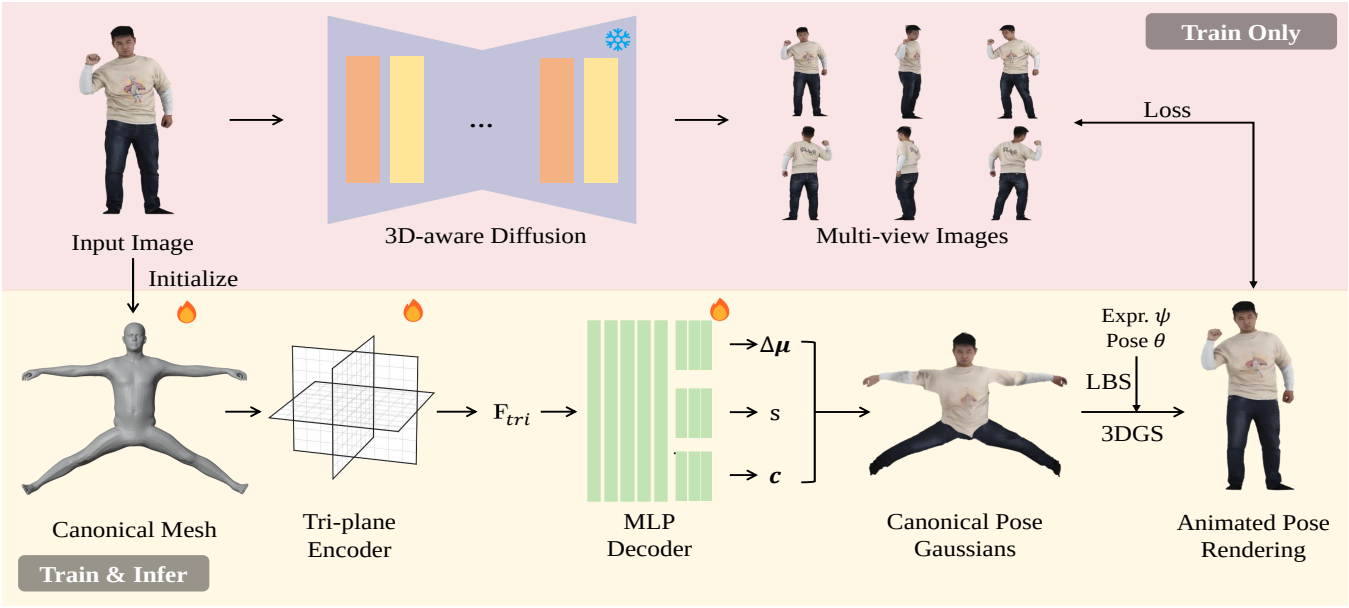

Fig. 1. Overview of SIAvatar. Given a single image input, SIAvatar leverages diffusion-based appearance prior and SMPL-X geometry prior to regress the 3DGS representation of a 3D human avatar. During the inference stage, SIAvatar animates the 3D avatar by applying LBS to the 3D Gaussians using SMPL-X facial expression and pose parameters, and then renders the avatar as 2D images in arbitrary views.

2. METHOD

2.1. Preliminary

3D Gaussian Splatting [10] represents 3D content as a set of 3D Gaussians $G = \{\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{q}_i, \mathbf{c}_i, \boldsymbol{\sigma}_i\}_{i=1}^{N}$, where $\boldsymbol{\mu}_i$ denotes the 3D position, $\mathbf{s}_i$ the scale, $\mathbf{q}_i$ the rotation, $\mathbf{c}_i$ the color, $\boldsymbol{\sigma}_i$ the opacity, and N is the number of 3D Gaussians. This representation preserves properties of volumetric rendering for optimization, while enabling real-time differentiable rasterization for 2D image rendering.

SMPL-X [4] is a unified parametric human body model with learned blend shapes based on LBS. Given pose parameters θ, shape parameters β, and facial expression parameters ψ, the SMPL-X model deforms and generates the human body mesh $\mathcal{M}(\theta, \beta, \psi) \in \mathbb{R}^{N\times 3}$ comprising N = 10475 vertices.

2.2. Overview

Given a whole-body image I_0 as input, our goal is to create a 3D human avatar that can be rendered from novel views and animated with novel poses. As illustrated in Fig. 1, SIAvatar leverages a diffusion model as multi-view appearance prior and SMPL-X as human geometry prior to regress the 3DGS representation of a human avatar. First, given an input human image, we utilize a pretrained 3D-aware diffusion model [14] to produce multi-view pseudo ground truth I_{pGT} for appearance supervision. Next, we estimate SMPL-X parameters from the input image [15] to initialize a canonical human mesh serving as geometry guidance. Each vertex of the canonical mesh is fed into a tri-plane [16] encoder E_{tri} to aggregate 3D latent features $\mathbf{F}_{tri}$. A lightweight MLP decoder D_{mlp} then interprets the tri-plane features as 3D Gaussian attributes. These 3D Gaussians can be posed with SMPL-X facial expression ψ and pose θ, and then be rendered differentiably into 2D images from novel views. To accurately model human geometry details and mitigate artifacts, we introduce a vertex-based adaptive Gaussian densification scheme for the density control of 3D Gaussians.

2.3. 3D Gaussian Regression

We initialize isotropic 3D Gaussians [11, 12] from the canonical human mesh $\mathcal{M} \in \mathbb{R}^{N\times 3}$. The neutral position $\boldsymbol{\mu}_0 \in \mathbb{R}^{N\times 3}$ of 3D Gaussians are initialized by 3D location of the vertices. The scale dimension is set to 1. The rotation $\mathbf{q}_i$ and opacity σ_i are fixed to $[1, 0, 0, 0]$ and 1 for all 3D Gaussians, respectively. Inspired by EG3D [16], we adopt a tri-plane architecture as 3D human geometry enconder for efficient and expressive representation, coupled with a lightweight MLP as the 3D Gaussian attributes decoder. Specifically, we query each vertex of the canonical mesh by projecting it onto each of the three axis-aligned orthogonal feature planes $E_{tri} \in \mathbb{R}^{3\times H\times W\times C}$ to retrieve the feature components, where H, W and C denote the height, width, and number of channels of each feature plane, respectively. These components are concatenated as 3D latent features $\mathbf{F}_{tri} = (\mathbf{F}_{xy}, \mathbf{F}_{xz}, \mathbf{F}_{yz})$, which are fed into the lightweight MLP D_{mlp} to predict the

position offset $\Delta\boldsymbol{\mu} \in \mathbb{R}^{N\times 3}$, scale $\mathbf{s} \in \mathbb{R}^{N\times 1}$, and color $\mathbf{c} \in \mathbb{R}^{N\times 3}$ for all 3D Guassians.

To animate the 3D Gaussian avatar, we transform the canonical 3D Gaussians using SMPL-X facial expression ψ and pose θ. The facial expression parameters ψ are mapped to the 3D expression offset $\Delta\boldsymbol{\mu}_{expr} \in \mathbb{R}^{N\times 3}$ based on expression blendshapes. We then add all offsets to the neutral position $\boldsymbol{\mu}_0$ and perform LBS to the 3D Gaussians with the transformation matrix computed from pose θ. This process can be formulated as:

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu} + \Delta\boldsymbol{\mu}_{expr}, \tag{1}$$

$$\boldsymbol{\mu}_{pose} = LBS(\boldsymbol{\mu}, \theta, \mathbf{w}), \tag{2}$$

where $\mathbf{w}$ denotes the skinning weight. We then follow 3DGS rasterization [10] to obtain the rendering image I:

$$I = R(\boldsymbol{\mu}_{pose}, \mathbf{s}, \mathbf{q}, \mathbf{c}, \boldsymbol{\sigma}; \mathbf{K}, \mathbf{E}), \tag{3}$$

where $R(\cdot)$ is rasterization function, $\mathbf{K}$ represents camera intrinsic parameters, and $\mathbf{E}$ denotes camera extrinsic parameters.

2.4. Vertex-based Adaptive Gaussian Densification

As shown in Fig. 2, we begin with the initial set of 3D Gaussians derived from the SMPL-X vertices and then apply our vertex-based adaptive Gaussian densification scheme to achieve a more accurate and detailed representation of human geometry. Thanks to the point-to-point correspondence between 3D Gaussians and the human mesh vertices, our method just need to focus on under-reconstruction regions because over-reconstruction regions, where one Gaussian cover large areas, are naturally resolved during the optimization of Gaussian attributes.

The densification is guided by the observation that regions requiring fine-grained geometry detail modeling often exhibit large view-space position gradients of 3D Gaussians. Specifically, for each Gaussian whose magnitude of the view-space position gradient exceeds a predefined threshold τ, we identify its corresponding vertex and the associated triangular faces. Each of these faces is subdivided into four smaller triangles by introducing midpoints along its three edges. These midpoints are then used to initialize new Gaussians, effectively increasing the density of 3D Gaussian representation in under-reconstruction regions. By leveraging the mesh structure and directly associating Gaussians with vertices, our 3D Gaussian representation integrates seamlessly with the inherent geometry of the human body. The adaptive densification scheme ensures that denser Gaussians are densified in the regions that require fine-grained geometry details modeling like body joints, while maintaining a compact representation in the coarse-grained regions.

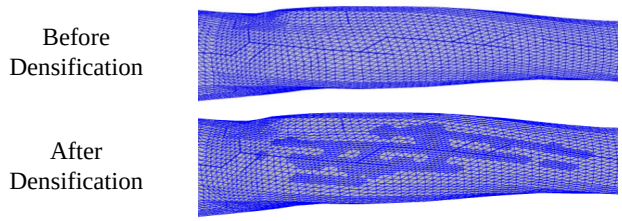

Fig. 2. Example of vertex-based adaptive Gaussian densification in the arm region.

2.5. Loss Function

During the training, we optimize the canonical mesh $\mathcal{M}$, tri-plane encoder E_{tri}, and MLP decoder D_{mlp}. To minimize image reconstruction error between pseudo ground truth I_{pGT} and the rendered images I from the corresponding views, we employ a combination of $L1$ loss $\mathcal{L}_{rgb}$, SSIM [17] loss $\mathcal{L}_{ssim}$, and LPIPS [18] loss $\mathcal{L}_{lpips}$. The reconstruction loss $\mathcal{L}_{rec}$ is formulated as:

$$\mathcal{L}_{rec} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{lpips}\mathcal{L}_{lpips}, \tag{4}$$

where λ_{rgb} = 0.8, λ_{ssim} = 0.2, and λ_{lpips} = 0.2.

Without fine-tuning on the input image I_0 of the specific subject, the pretrained diffusion model may introduce incorrect identity information in the generated multi-view images. To ensure identity consistency, we apply ArcFace [19] loss L_{arc} and the aforementioned reconstruction loss $\mathcal{L}^f_{rec}$ to the face region between input I_0 and the rendered image I. The identity loss $\mathcal{L}_{id}$ is defined as:

$$\mathcal{L}_{id} = \lambda_{arc}\mathcal{L}_{arc} + \mathcal{L}^f_{rec}, \tag{5}$$

where λ_{arc} = 0.1.

To prevent floating 3D Gaussians, we apply $L2$ regularizer $\mathcal{L}_{l_2}$ and Laplacian regularizer $\mathcal{L}_{lap}$ to the position offsets and scales, The regularization loss $\mathcal{L}_{reg}$ is formulated as:

$$\mathcal{L}_{reg} = \lambda_{l_2}\mathcal{L}_{l_2} + \lambda_{lap}\mathcal{L}_{lap}, \tag{6}$$

where λ_{l_2} = 10, and λ_{lap} = 1.

The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{id} + \mathcal{L}_{reg}. \tag{7}$$

3. EXPERIMENTS

3.1. Implementation Details

We configure the tri-plane encoder with H = 128, W = 128, and C = 32. For each specific identity, we take a single 1024×1024 front-view image as input and train the model for 36,000 steps with AdamW [22] optimizer. The adaptive Gaussian densification scheme starts at 16,000 steps and ends at 26,000 steps, with a densification interval of 2,000 steps

Fig. 3. Qualitative comparison on THuman2.0 [20] and X-Humans [21]. All methods are trained on a front-view image for each subject. In the first two rows, we render each avatar from different views with the same pose. In the last three rows, we animate each avatar with novel poses and render them from novel views. We highly recommend readers view our supplementary video for intuitive comparisons.

Table 1. Quantitative comparison on THuman2.0 [20]. * indicates that this method is trained on multi-view images.

Methods	T0034			T0103			T0139		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
InstantAvatar [5]	14.93	0.880	0.181	17.16	0.903	0.140	13.81	0.860	0.185
GaussianAvatar [11]	18.20	0.909	0.112	19.12	0.922	0.082	19.04	0.904	0.111
GaussianAvatar* [11]	20.27	0.946	0.062	21.99	0.950	0.042	20.42	0.933	0.063
ExAvatar [12]	19.48	0.913	0.106	22.69	0.943	0.064	20.88	0.918	0.092
SIAvatar (Ours)	**26.00**	**0.966**	**0.037**	**26.45**	**0.968**	**0.034**	**26.46**	**0.957**	**0.044**

Table 2. Quantitative comparison on X-Humans [21]. * indicates that this method is trained on a monocular video.

Methods	X00028			X00034			X00088		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
InstantAvatar [5]	17.14	0.921	0.114	18.42	0.907	0.115	15.48	0.890	0.138
GaussianAvatar [11]	19.14	0.937	0.071	21.91	0.931	0.060	19.84	0.926	0.066
GaussianAvatar* [11]	21.20	0.959	0.043	23.41	0.946	**0.037**	21.49	0.947	0.041
ExAvatar [12]	21.11	0.950	0.056	23.38	0.944	0.049	21.64	0.935	0.057
SIAvatar (Ours)	**23.53**	**0.965**	**0.035**	**25.41**	**0.954**	**0.037**	**24.83**	**0.961**	**0.031**

and a threshold $\tau = 0.0003$. The model is trained on a single NVIDIA RTX 4090 GPU, taking approximately three hours to complete.

3.2. Comparison

We take a single front-view image as input for SIAvatar and state-of-the-art methods [5, 11, 12] to train the avatar for each subject. Quantitative and qualitative comparisons are conducted on THuman2.0 [20] and X-Humans [21] datasets.

Quantitative Comparison. As shown in Tab. 1 and Tab. 2, SIAvatar consistently surpasses state-of-the-art methods in PSNR, SSIM and LPIPS across all subjects, demonstrating its superior performance in reconstructing high-fidelity 3D avatars from a single image. By leveraging diffusion-based appearance prior and SMPL-X geometry prior, SIAvatar can infer plausible and detailed appearance directly from the single image and provide a robust structural foundation for human shape and deformation, thus achieving better or comparable quantitative results compared to GaussianAvatar [11] trained on multi-view images or a monocular video.

Qualitative Comparison. As illustrated in Fig. 3, SIAvatar generates more photorealistic and detailed 3D avatars in novel views and poses than state-of-the-art method. Specifically, InstantAvatar [5], which utilizes NeRF [9] as implicit representations, exhibits ghosting artifacts when rendering avatars in novel views. GaussianAvatar [11] and ExAvatar [12] struggle to generate high-quality avatars from challenging angles due to their limited ability to extract meaningful 3D features from a single image. In contrast, SIAvatar leverages diffusion-based appearance prior and SMPL-X geometry prior, enabling the faithful reconstruction and animation of high-fidelity 3D human avatars across diverse views and poses.

3.3. Ablation Study

As shown in Tab. 3 and Fig. 4, we conduct an ablation study to verify the effect of the identity loss and the vertex-based adaptive Gaussian densification scheme. Experimental results demonstrate that incorporating the identity loss significantly improves identity consistency, while the vertex-based adaptive Gaussian densification scheme effectively mitigates artifacts and enhances the modeling of human geometry.

Table 3. Ablation study of the identity loss and the vertex-based adaptive Gaussian densification scheme.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o Identity Loss	22.06	0.943	**0.066**
w/o Densification	22.06	0.942	0.075
SIAvatar (Ours)	**22.12**	**0.944**	**0.066**

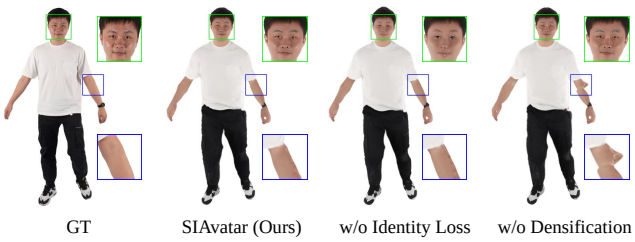

Fig. 4. Ablation study of the identity loss and the vertex-based adaptive Gaussian densification scheme. Please zoom in for a detailed view.

4. CONCLUSION

We present SIAvatar, a one-shot animatable 3D Gaussian avatar reconstruction method that integrates diffusion-based appearance prior and parametric human model geometry prior within a 3DGS framework. A vertex-based adaptive Gaussian densification scheme is designed for the density control of 3D Gaussians, achieving accurate 3D Gaussian representation of human geometry details. Comprehensive experiments demonstrate that SIAvatar excels in realistic 3D avatar reconstruction and robust novel pose animation from a single image. Future work will explore relighting and dynamic clothing modeling for 3D Gaussian avatars.

5. REFERENCES

[1] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou, "Reconstructing 3d human pose by watching humans in the mirror," in *CVPR*, 2021, pp. 12814–12823.

[2] Yang Lu, Han Yu, Wei Ni, and Liang Song, "3d real-time human reconstruction with a single rgbd camera," *Applied Intelligence*, vol. 53, no. 8, pp. 8735–8745, 2023.

[3] 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)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[4] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019, pp. 10975–10985.

[5] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges, "Instantavatar: Learning avatars from monocular video in 60 seconds," in *CVPR*, 2023, pp. 16922–16932.

[6] 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 *CVPR*, 2022, pp. 16210–16220.

[7] 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 *CVPR*, 2023, pp. 12858–12868.

[8] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *CVPR*, 2019, pp. 165–174.

[9] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[11] 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 *CVPR*, 2024, pp. 634–644.

[12] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito, "Expressive whole-body 3D gaussian avatar," in *ECCV*, 2024.

[13] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang, "3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting," in *CVPR*, 2024, pp. 5020–5030.

[14] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang, "Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement," *arXiv preprint arXiv:2408.14211*, 2024.

[15] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu, "Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery," *arXiv preprint arXiv:2304.05690*, 2023.

[16] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al., "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022, pp. 16123–16133.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] 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, pp. 586–595.

[19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.

[20] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in *CVPR*, 2021, pp. 5746–5756.

[21] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges, "X-avatar: Expressive human avatars," in *CVPR*, 2023, pp. 16911–16921.

[22] I Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.