OmniAvatar: Geometry-Guided Controllable 3D Head Synthesis (Supplementary)

Hongyi Xu[1] Guoxian Song[1] Zihang Jiang[1,2] Jianfeng Zhang[1,2] Yichun Shi[1]
Jing Liu[1] Wanchun Ma[1] Jiashi Feng[1] Linjie Luo[1]
[1]ByteDance Inc [2]National University of Singapore
{hongyixu, guoxian.song, zihang.jiang, jianfeng.zhang, yichun.shi,
jing.liu, wanchun.ma, jshfeng, linjie.luo}@bytedance.com

In this supplementary material, we provide additional implementation details in Section A, additional baseline comparisons, ablations and more visual results in Section B. We discuss limitations and future work in Section C.

A. Implementation Details

Implicit Semenatic Signed Distance Function Training. We train the parametric semantic SDF W using 150K FLAME instances with Gaussian sampled shapes and expressions, and jaw and neck poses within their corresponding joint limits. For SDF computation with watertight geometry, we close the mesh with consistent hole filling on the mouth cavity and below the neck. Before the training of W, we pretrain a canonical SDF. As such, the signed distance value of a spatial point $\mathbf{x}$ with respect to the FLAME mesh $\mathbf{S}(\mathbf{p})$ can be obtained by querying the canonical SDF with its canonical correspondence point $\bar{\mathbf{x}}$. We train W with 150 epochs using a decaying learning rate of 0.0002, on a single Tesla V100 GPU. We set the loss weights $1., 0.5, 0.1$ for L_{iso}, L_{eik} and L_{sem} respectively.

Controllable 3D GAN Training. Our training largely follows the the scheme of EG3D [2], although we only train at a 64×64 neural rendering resolution. Better view consistency potentially could be achieved by fine tuning the network on 128×128 neural rendering resolution while is not applied in our training given it is not the primary focus of this work. We empirically set the training weights $0.1, 1.$ for our geometric prior loss L_{prior} and control loss L_{enc}, whereas our conditioning expression and joint pose for modeling dynamic details is perturbed with a Gaussian noise of magnitude 0.1. Following EG3D, we mirror each training image from FFHQ [8] and rebalance the dataset by replicating large-pose images. Additionally, for training images with jaw opening larger than $10\circ$, we duplicate them 4 times in our dataset. With a batch size of 32 on 8 Tesla V100 GPUs, the network is trained with iterations of $160K$ images. Due to the additional computational overhead in correspondence mapping, our training is about $1.43\times$ slower than the original EG3D.

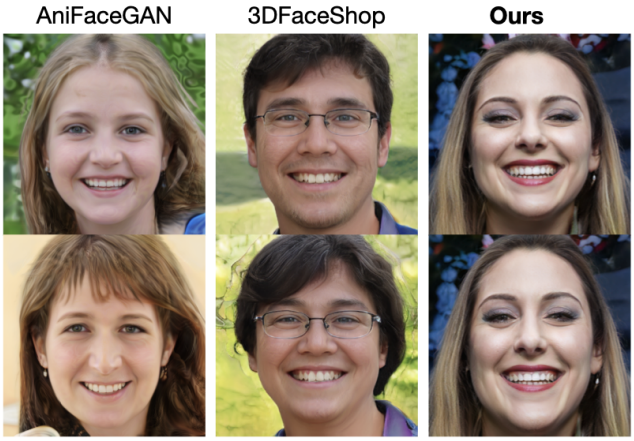

Figure S1. Qualitative comparison on shape editing (under the same expression in each column). Our method strictly preserves facial identity when shape varies whereas significant appearance variations are observed for AniFaceGAN [15] and 3DFaceShop [14].

FLAME Parameters Fitting. We associate each training image with a 16-dimensional camera extrinsics, an empirical 9-dimensional camera intrisics, and a 206-dimensional FLAME parameter $\mathbf{p}$ (shape $\boldsymbol{\alpha} \in R^{100}$, expression $\boldsymbol{\beta} \in R^{100}$, jaw pose $\boldsymbol{\theta}_{jaw} \in R^3$ and neck pose $\boldsymbol{\theta}_{neck} \in R^3$). We fit the FLAME parameters $\mathbf{p}$ with a nonlinear optimization,

$$\min_{\mathbf{p},\mathbf{R},\mathbf{t}} \|\Pi\mathbf{S}(\mathbf{p},\mathbf{R},\mathbf{t}) - L_{2d}\|_2 + \|\mathbf{p} - \mathbf{p}_0\|_2 \qquad (1)$$

where Π is a 3D landmarks extractor from FLAME mesh $\mathbf{S}$ concatenated with a default camera projection, L_{2d} the detected image 2D landmarks and $\mathbf{p}_0$ the initialization of FLAME parameters with DECA [5]. We fix $\boldsymbol{\theta}_{neck}$ to be 0

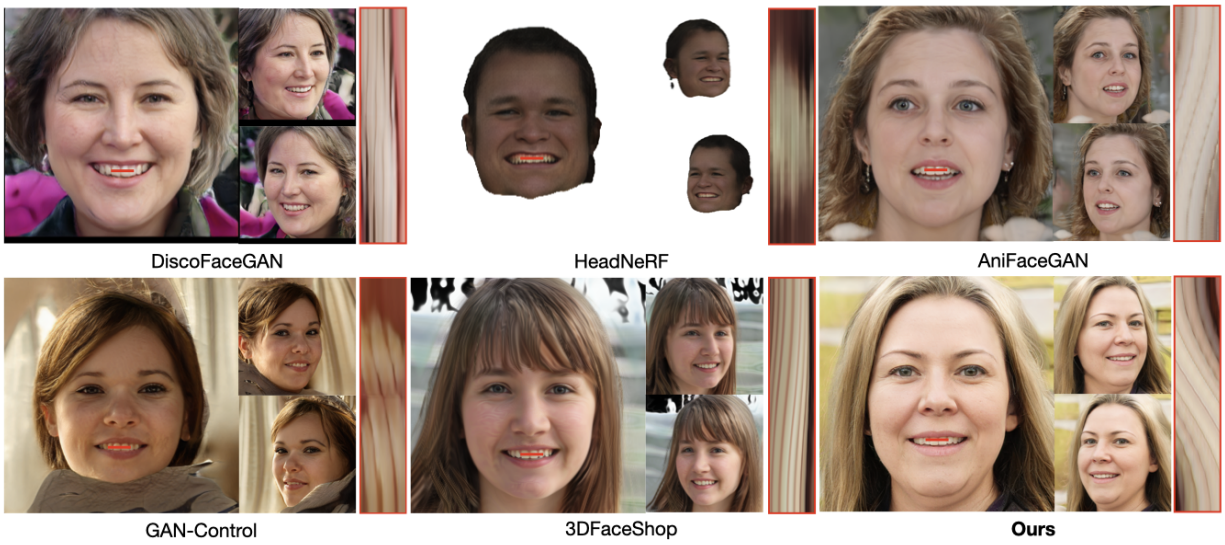

Figure S2. Qualitative comparison on multiview consistency. Our method demonstrates smooth pattern transition as camera rotates, comparable with other controllable 3D GANs (AniFaceGAN [15], 3DFaceShop [14]). Noisy feature transitions occur for 2D GANs [4, 13] and HeadNeRF [6].

during optimization 1 and co-optimize the root rotation $\mathbf{R}$ and translation $\mathbf{t}$ with $\mathbf{p}$. After the optimization, we transform the ($\mathbf{R}$, $\mathbf{t}$) into the camera extrinsics while locating the optimized FLAME meshes centered at the world origin facing towards Z axis.

Inversion. For single-view portrait manipulation and animation, we need to first find the corresponding latent embedding of the input image $\bar{I}$. Following the aforementioned fitting procedure, we estimate the rendering camera $\mathbf{c}$ and FLAME parameter $\mathbf{p}$ from the input image. With that, we perform a nonlinear optimization in the latent $\mathbf{Z}^+$ space as

$$\min_{\mathbf{z}^+} \|\bar{I} - I_{RGB}(\mathbf{z}^+|\mathbf{c},\mathbf{p})\|_1 + \\ +L_{lp}(\bar{I}, I_{RGB}(\mathbf{z}^+|\mathbf{c},\mathbf{p})) + \|\text{Div}(\mathbf{z}^+)\|_2, \quad (2)$$

where the first term is a pixel-wise L_1 loss between input and synthesized image I_{RGB}, L_{lp} is the image perceptual loss with LPIPS [17] and the last term evaluates the divergence of $\mathbf{Z}^+$. In contrast of modulating each layer of StyleGAN synthesis [8] with a duplicated $\mathbf{w}$ code mapped from $\mathbf{z}$, $\mathbf{Z}^+$ is more expressive with a different mapped $\mathbf{w}$ code for each layer. We prevent far drifting from the original high-quality $\mathbf{Z}$ latent space with the divergence term, which otherwise leads to image quality degeneration.

For better visual reconstruction of the input portrait, we follow the optimization with pivot tuning [11] that fine-tunes the parameters of the triplane synthesis network with a fixed latent code. We do not alter the parameters of the neural decoder and super-resolution module to prevent the degeneration of multi-view consistency from overfitting. After the fine-tuning, one can synthesize a new image using the fine-tuned generator at novel views and expressions by modifying the $\mathbf{c}$ and $\mathbf{p}$ accordingly. We note that optimizing in the less regularized $\mathbf{Z}^+$ domain followed with Pivot Tuning is necessary for identity preservation but might sacrifice visual quality and 3D view consistency.

Foreground and Background Decomposition. While not being the focus of the paper, our 3D-aware generator also decomposes the 3D-aware foreground synthesis from the background. Different from the original EG3D, our feature image $I^*(\mathbf{z})$ is composed with 3D-aware human foreground integrated from the neural radiance field and a 2D background image $I^+(\mathbf{z})$ synthesized with 2D convolution kernels, as

$$I^*(\mathbf{z}) = I^*(\mathbf{z}) + (1 - M^*) \odot I^+(\mathbf{z}), \quad (3)$$

where M^* is the accumulated foreground density along camera rays. Similar to [12], we associate each training image with a 2D foreground mask, and augment our discriminator D with a dual convolution branch for pairs of (I, M) where M is up-sampled from M^*. To this end, our generator G is able to synthesize view-consistent images while keeping the background static.

B. Experimental Analysis

Shape Manipulation Comparison. We qualitatively compare the capability on shape editing against AniFaceGAN [15] and 3DFaceShop [14] in Figure. S1, whereas

Figure S3. Our approach enables realistic synthesis of dynamic details, adapted to the appearance of the subject.

2D GANs (DiscoFaceGAN [4],GAN-Control [13]) do not support shape manipulation. Our approach well preserves the facial identity feature even with significant shape variation, indicating a clean disentanglement of geometric variation from appearance generation. Noticeable identity variation is observed when changing the shape by AniFaceGAN [15] and 3DFaceShop [14]. This clearly shows the benefits of our design where the neural scene generation is not directly conditioned on the shape or expression code. Thus by nature our appearance generation is minimally affected with a modified shape and expression code, whereas explicit regularizations are required to disentangle the control from neural appearance generation in AniFaceGAN and 3DFaceShop.

Multiview Consistency Comparison. In Figure. S2, we visually compare the view consistency using Epipolar Line Images (EPI) similar to [16]. The 2D generative models shows inferior view consistency compared to NeRF-based 3D-aware generative models. The appearance transition of HeadNeRF [6] is more noisy, largely due to the lack in fine details in such areas as teeth. Our method shows natural and continuous pattern transition when smoothly changing the views, comparable with 3DFaceShop [14] and AniFaceGAN [15].

Ablations on Correspondence Field Training. In addition to the ablation study in our main text, we further ablate the training of the signed distance correspondence field W. In our pipeline, we pretrain W from FLAME meshes and freeze the weights during the 3D GAN training. Such a two-stage training scheme decouples the 3D learning from unsupervised image-based adversarial training. It also avoids the heavy computation and overloaded memory in obtaining the SDF gradient for Eikonal loss L_{eik}. Nevertheless, we experiment with fine-tuning the weights of W together with image-based adversarial training. The network does not converge (FID=90.2), indicating more regularization terms are required if we co-learn W with image synthesis. In practice, we find our current strategy well preserves the rich 3d prior knowledge from the 3D statistical head model, and the synthesis network can compensate the imperfections in our pretrained correspondence function.

Dynamic Details. In Figure S3, we show a zoomed-in view for the dynamic details synthesis. For both subjects, our method substantially enhances the temporal realism with dynamic details, such as dimples and wrinkles, when transiting from a neutral expression to smiling. Moreover, our dynamic details are adapted to the subject's appearance, where we observe much less eye wrinkles on a smiling face of a younger subject. Such capability is sourced from our novel design, where we decode the neural radiance field from both triplane features (appearance) and expressions.

Control Accuracy. In Figure. S4, We visualize more synthesized identities with the same shape and expression code, compared with our ablated framework without our geometric prior loss L_{prior} and control loss L_{enc}. We overlay the images with the projected landmarks of the controlling FLAME. Our full pipeline shows more consistent shape and expression as specified by the input control, whereas more variations are observed in shapes without L_{prior} and in expressions without L_{enc}.

Geometry. In Figure. S5, we show the visualization of iso-surface geometry extracted from the prior FLAME SDF

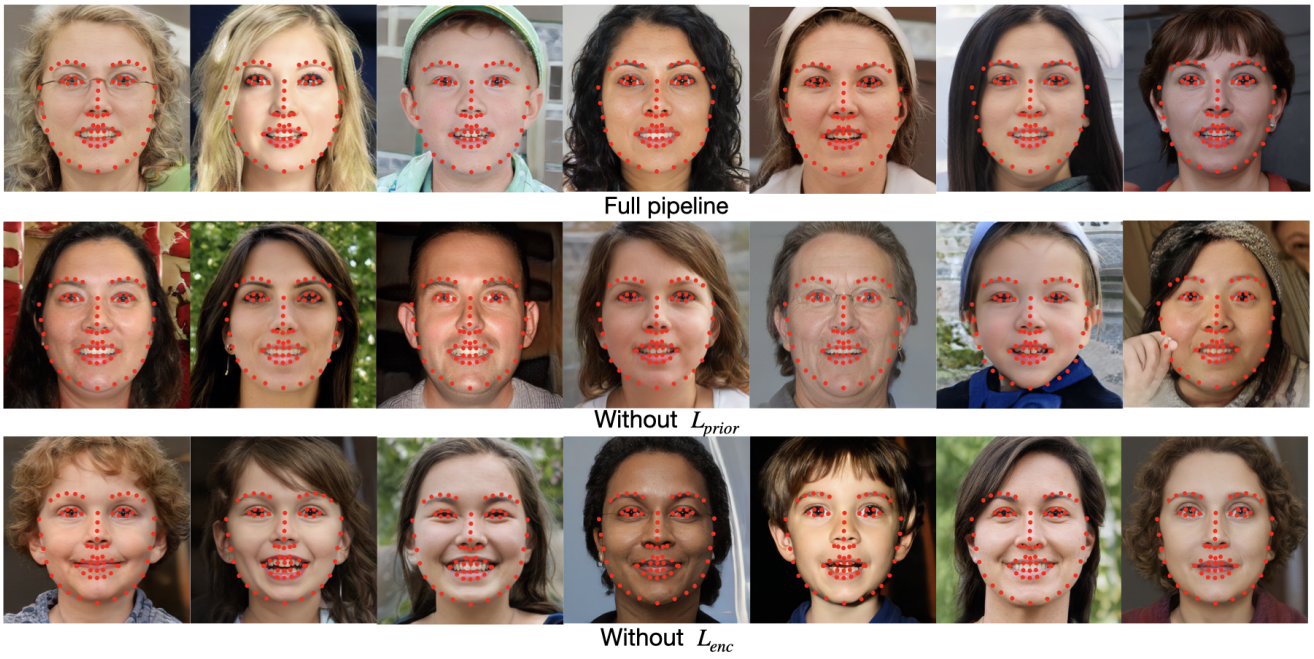

Figure S4. Identities synthesized under the same shape and expression code with our full and ablated pipelines. Red dots indicate the 68 projected 2D landmarks of the control FLAME mesh. Our full pipeline generates images more consistent with the input control, whereas degeneration in shape and expression control accuracy is observed after removing geometric prior loss L_{prior} and control loss L_{enc} respectively.

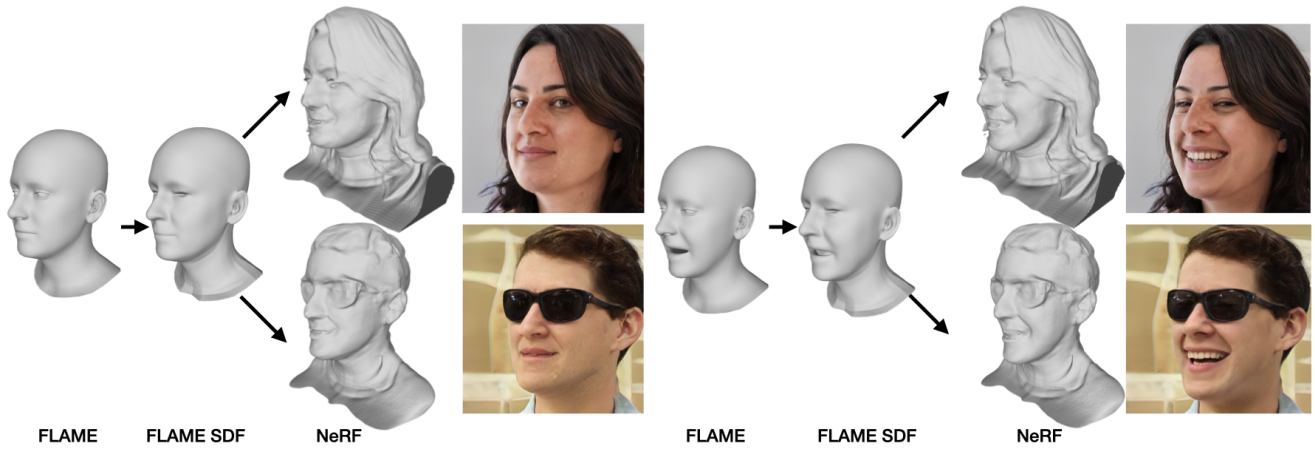

Figure S5. We visualize the geometry guidance flow. Our signed distance correspondence field generates a SDF from the input FLAME mesh and guides the generation of the neural density field. Our dynamic details are reflected in the geometric level as well.

and the density of the neural radiance field using Marching Cubes algorithm [10]. Our approach preserves the detailed geometry synthesis capability from EG3D [2], supporting multiview consistency in image generation. Our dynamic details can also be observed in the synthesized neural density field. Moreover, the FLAME SDF also acts as a reasonable geometry proxy for regions with little image supervisions, and enable building a complete full head shape. We further provides quantitative evaluation on our FLAME SDF reconstruction accuracy on a test dataset of 1000 randomly sampled FLAME instances using bi-directional Chamfer L_2 Distance↓ (7.4×10^{-5}), Normal Consistency↑(0.933) and Volumetric Intersection of Union (IoU↑, 0.968), following the geometric metrics in imGHUM [1].

More Talking Head Animations. In Figure. S6, we showcase more talking head animation examples in differ-

Figure S6. Synthesized multiview-consistent talking head animations. Each column is generated under the same expression and camera poses.

ent views, indicating the robustness of our approach in synthesizing diverse identities while maintaining the control accuracy.

More 3D-Aware Face Reenactment Results. In Figure. S7, we show more multiview-consistent face reenactment videos by inverting a single-view portrait image and manipulating the expressions by following temporal FLAME reconstructions from a reference video. All the reference photos and video are downloaded with Creative Commons licences.

C. Limitations and Future Work.

Expressiveness. Our method demonstrates superior expressive controllability than prior controllable 3D GANs and is able to synthesize images with subtle expressions, such as eye blinks. However, the expressiveness is still largely constrained to the FLAME model and it is diffi-

Figure S7. Our approach enables multi-view consistent face video reenactment from a single-view portrait (shown on the leftmost column).

cult to synthesize expressions that are under-represented in the parametric expression space of FLAME, such as mouth puckering or blowing cheeks. There is no gaze control as well. Additionally for the application of face reenactment, the quality is also limited by the accuracy of FLAME tracking over the reference video. In the future, we would like to augment the control in more expressive representations such as dense landmarks or even image features.

Visual Artifacts. Even with the rich 3D geometric guidance from the statistical head model, there is fundamental ambiguity in observation-to-canonical one-to-one point correspondences. For example, when the mouth is closed, it is undetermined for its canonical correspondence, if a point is touching both the upper and lower lip in the observation space. SNARF [3] provides a differential optimization-based solution for this ambiguity issue but suffers from slow computational performance. Therefore modeling of the mouth cavity is still the most challenging part due to this correspondence ambiguity. Moreover, there is also less visual supervision in the training images with clear observations inside the mouth cavity. Even though we are able to alleviate this issue by rebalancing the image dataset with replicated open-mouth images, noticeable artifacts still exist when the mouth is opened widely as depicted in Figure S8. We would like to explore more advanced differential correspondence functions and augment our training dataset with

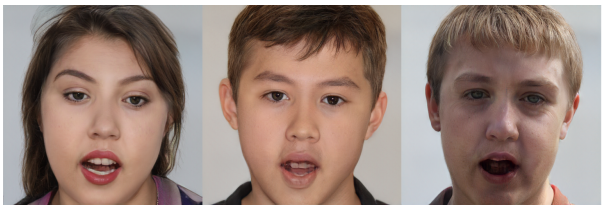

Figure S8. Mouth opening artifacts: blurriness in the lower (left) or upper teeth (middle), and even missing teeth (right).

more diverse and extreme expressions.

Temporal coherence. Inherited from the StyleGAN2 [9] synthesis network, noticeable high-frequency noise can be observed in our synthesized animations. By switching to StyleGAN3 [7], we expect better temporal coherence but leave it as a future work. Additionally, with our expression-conditioned dynamic details modeling, we occasionally observe unnatural shading variation with expression changes. This is due to the limitation of the MLP-decoder in generalizing to novel poses, which we consider it as an interesting future direction to explore.

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, pages 5461–5470, 2021. 4

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. *CVPR*, 2022. 1, 4

[3] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 6

[4] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *cvpr*, pages 5154–5163, 2020. 2, 3

[5] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics*, volume 40, 2021. 1

[6] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *cvpr*, pages 20374–20384, 2022. 2, 3

[7] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 7

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 7

[10] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 4

[11] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2

[12] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, 2022. 2

[13] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14083–14093, 2021. 2, 3

[14] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434*, 2022. 1, 2, 3

[15] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *arXiv preprint arXiv:2210.06465*, 2022. 1, 2, 3

[16] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 3

[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*, pages 586–595, 2018. 2