AgileGAN3D: Few-Shot 3D Portrait Stylization by Augmented Transfer Learning

Guoxian Song[1]
[1]ByteDance Inc

Abstract

While substantial progresses have been made in automated 2D portrait stylization, admirable 3D portrait stylization from a single user photo remains to be an unresolved challenge. One primary obstacle here is the lack of high quality stylized 3D training data. In this paper, we propose a novel framework AgileGAN3D *that can produce 3D artistically appealing and personalized portraits with detailed geometry. New stylization can be obtained with just a few (around 20) unpaired 2D exemplars. We achieve this by first leveraging existing 2D stylization capabilities,* style prior creation*, to produce a large amount of augmented 2D style exemplars. These augmented exemplars are generated with accurate camera pose labels, as well as paired real face images, which prove to be critical for the downstream 3D stylization task. Capitalizing on the recent advancement of 3D-aware GAN models, we perform* guided transfer learning *on a pretrained 3D GAN generator to produce multi-view-consistent stylized renderings. In order to achieve 3D GAN inversion that can preserve subject's identity well, we incorporate* multi-view consistency loss *in the training of our encoder. Our pipeline demonstrates strong capability in turning user photos into a diverse range of 3D artistic portraits. Both qualitative results and quantitative evaluations have been conducted to show the superior performance of our method. Code and pretrained models will be released for reproduction purpose.*

1. Introduction

Portrait painting as an art form goes back to prehistoric times, and has been primarily serving the rich and powerful. Fast forward with the technology advancement, people nowadays can enjoy a high fidelity digital portrait within seconds, and even capturing one's detailed facial 3D geometry[6, 55]. Driven by the creative nature of human beings, people are no longer satisfied with simply a faithful depiction of their appearance. Portraiture has evolved into more expressive interpretations with a plethora of styles, such as abstract art, cubism and cartoon. However, most previous works are limited to stylized portraits in 2D image space[9, 38, 47, 56]. Automatically creating 3D stylized portrait with detailed geometry using just a single selfie as input is still an open problem. To the best of our knowledge, we are the first work that can automatically create 3D stylized portrait with detailed geometry, using just a single selfie as input. The result 3D portrait can be adapted into a wide range of artistic styles, as long as a few 2D style exemplars are provided (see Fig.1). Such new format enables a lot of applications like 3D printed postcards, dynamic profile pictures (by changing viewing angles, or lighting directions), as well as personalized 3D contents in augmented and virtual reality worlds.

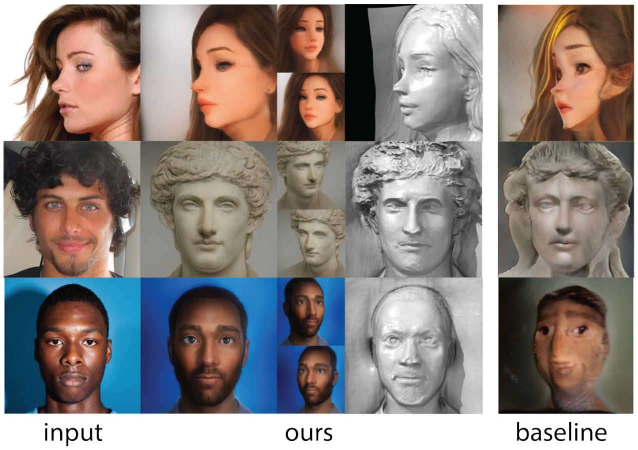

Figure 1. Our AgileGAN3D enables 3D stylized portraits creation from a single input image. A new 3D style can be obtained with only a few unpaired 2D style exemplars (∼ 20). Compared to the baseline method (directly fine-tune the 3D GAN model [8]), our approach produces high-quality, multi-view-consistent renderings of the portrait, with detailed in-style geometry.

The core challenge that prevents us from creating visually appealing, personalized 3D portraits is rooted in the shortage of high quality 3D data. A traditional approach to

[*] This work authors are Guoxian Song, Hongyi Xu, Jing Liu, Tiancheng Zhi, Yichun Shi, Jianfeng Zhang, Zihang Jiang, Jiashi Feng, Shen Sang and Linjie Luo.

create customized 3D portraits for users is by assembling a 3D avatar system with tons of graphics assets (e.g. Zepeto[1], ReadyPlayer[2], and [44]). However, it's almost impossible to capture all the diversities in real world appearances given only a few hundred assets and a base morphable 3D face model. Therefore such approach usually generates less personalized results.

The latest generative models such as StyleGAN2 [25], latent diffusion models [43] are very powerful architectures in producing highly diversified imageries, largely credited to the huge data sources that these models have been trained on. Fine-tuning a generative model to produce highly personalized portraits thus becomes possible, though still primarily in 2D image space[9, 38, 47, 56]. Lifting arbitrary stylized portraits from 2D into 3D space remains to be an unsolved problem, partly due to the lack of 3D prior knowledge for target artistic pictures. Promising early results have emerged for general 3D objects [20, 40], but none of them can produce desirable quality for 3D portraits yet.

Recent rapid progress in geometry-aware GANs [7, 8, 11, 36, 37, 46, 59] inspired our work. Particularly EG3D [8] demonstrated strikingly realistic 3D face synthesis capability by using only unstructured 2D photos. Its tri-plane structure serves as an efficient representation for 3D content generation, combined with a neural volumetric renderer, making multi-view-consistent, photo-realistic image synthesis possible. However, even with such a powerful 3D generator, there are still a few obstacles ahead before we can have a usable 3D stylized portrait generator. First, even in 2D image format, it is nontrivial to collect a large number of diverse portraits in a consistent style, let alone obtaining robust camera pose estimations from artistic portraits, which is a critical step in training 3D GANs. Secondly, in order to personalize user photos into 3D stylized portraits, a reliable 3D GAN inversion method is required. The inversion module has to work in a way that balances the reconstruction fidelity and the stylization quality.

To tackle the aforementioned challenges, we propose *AgileGAN3D*, a novel *augmented transfer learning* framework for generating high quality stylized 3D portraits, using only a few 2D style exemplars (∼20). With extensive experiments, we observe the successful transfer learning for 3D GANs relies on adequate style visual supervisions with well estimated camera labels. To address the shortage of stylized training data issue, we started with *style prior creation*, that leverages the existing 2D portrait stylization capabilities. Specifically, we trained a 2D portrait stylization module following AgileGAN[47], to first obtain a large number of stylized portraits using real face photos as inputs. Some extra benefits of this way of generating 2D style exemplars are that we naturally obtain: 1) pair data between stylized faces and real faces; 2) fairly accurate head pose estimations of generated stylized portraits (by reusing the poses estimated from the corresponding real faces). Both of these benefits are incorporated into our *guided transfer learning* step with a reconstruction loss, that helps improve the 3D stylization for out-of-domain samples. Equipped with a transfer-learned style generator, we further introduce an 3D GAN encoder that embeds a real image into an enlarged latent space for better identity-preserved 3D stylization. A cycle consistency loss is proposed in the 3D GAN encoder training to further improve the multi-view reconstruction fidelity. To best of our knowledge, we are the first paper to propose generative NeRF based 3D stylized portrait creation using only a limited number of 2D style exemplars.

To summarize the contributions of our work:

- A novel pipeline for creating 3D stylized portraits with detailed geometry, given only a single user photo as input. New stylization can be achieved with only a few unpaired 2D style exemplars (around 20).
- A simple yet efficient way to fine-tune 3D GAN, first with *style prior creation* to improve data diversity, combined with *guided transfer learning* to increase the stylization domain coverage;
- A 3D GAN encoder that inverts real face images into corresponding latent space, trained with cycle consistent loss to improve identity preservation, while achieving high stylization quality.

2. Related Work

Face Stylization Stylizing facial images in an artistic manner has been explored in the context of nonphotorealistic rendering. Early approaches relied on low level histogram matching using linear filters [18]. Neural style transfer [14], by matching feature statistics in convolutional layers, led to early exciting results via deep learning. However, they usually fail on styles involving significant geometric deformation of facial features, such as cartoonization. For more general cross-domain stylization, Toonify [38] proposed a GAN interpolation framework for controllable cross-domain image synthesis for cartoonization. A following method AgileGAN [47] proposed VAE inversion to enhance distribution consistency in the latent space, leading to fewer artifacts and better results for real input images. Besides, Huang *et al*. [19] achieves multidomain stylization via a layer swapping technique. Recent exemplar-based approaches [9, 27, 56] enable one-shot portrait stylization given a single style exemplar. There are also several 2D stylization works for video generations [4, 57]. In contrast, our proposed *AgileGAN3D* produces highly detailed 3D stylized portraits using the same amount of input as used in 2D stylization methods.

[1] https://zepeto.me/

[2] https://readyplayer.me/

Geometry-Aware GANs Generative adversarial networks [15, 21, 24] have been used to synthesize images ideally matching the training dataset distribution via adversarial learning. Built on the success of 2D GANs, recent works have extended to multiview consistent image synthesis with unsupervised learning from unstructured single-view images. The key idea is to combine differential rendering with 3D representations, such as meshes [13, 29, 48], point clouds [28], voxels [34, 35, 53], and recently implicit neural representation [7, 8, 11, 36, 37, 46, 59]. Especially the Neural Radiance Fields (NeRF)[16, 37] representations, which have proven to generate high-fidelity results in novel view synthesis, are introduced to 3D-aware generative models. They typically use StyleGAN2 [25] as the backbone to generate intermediate features for MLP to query and perform neural volumetric rendering for image synthesis. Recently, EG3D[8] uses an efficient triplane-based neural radiance field, combined with CNN-based upsampling and a pose-aware dual discriminator to improve synthesis quality and multi-view consistency. Our work further extends the success of existing 3D GAN models to generate stylized results with only a few unpaired 2D stylized exemplars.

GAN Inversion Given an input image, GAN inversion addresses the complementary problem of finding the most accurate latent code to reconstruct that image. Existing approaches roughly fall into three categories: optimization-based, learning-based and hybrid GAN inversion. Optimization-based approaches [1, 22, 49] directly optimize the latent code to minimize the pixel-wise reconstruction loss for a single input instance. Learning-based approaches [61] train a deterministic model by minimizing the difference between the input and synthesized images. Some works combine these ideas, e.g. learning an encoder that produces a good initialization for subsequent optimization [5]. In addition to image reconstruction, some methods also use inversion when undertaking image manipulation. For example, Zhu *et al*. [60] introduced a hybrid method to encode images into a semantic manipulable domain for image editing. Richardson *et al*. [41] presented the generic Pixel2Style2Pixel (pSp) encoder to embed image into StyleGAN W^+ space, based on a dedicated identity loss. To balance reconstruction and editing ability for inversion, E4E encoder [50] later uses a progressive training strategy to stimulate W space for W^+ space. ReStyle [2] trains a residual based encoder with iterative refinement. Besides training encoder, Pivotal Tuning Inversion[42] also fine-tunes the generator parameters each time for recovering image details that cannot be encoded in the latent space to improve reconstruction. Wang *et al*. [52] proposed an approach to achieve high fidelity inversion without inference time optimization. Recent works [3, 12] employ hyper networks [17] to improve StyleGAN inversion. However, directly adopting 2D GAN encoders for 3D GAN models won't work well, as it is not taking account of the multi-view generation aspect of a 3D GAN model. In our work, we introduce a *multi-view cycle consistent loss* to improve our 3D GAN encoder's capability to preserve subject identities while balancing the stylization quality.

Diffusion-based Multi-view A recent noteworthy study, Zero-1-to-3[32, 33], employs a stable diffusion model to leverage geometric priors extracted from a comprehensive synthetic dataset, leading to the production of high-quality predictions. Furthermore, Consistent123[30], adopts a case-aware strategy that utilizes Zero-1-to-3's 3D priors for constructing an initial structural representation, subsequently enhancing texture fidelity. However, it's important to highlight that these methodologies predominantly focus on general objects. This focus results in a compromise in quality when these techniques are applied to the synthesis of portraits.

3. Method

As shown in Fig. 2, we first use *style prior creation* to augment the limited 2D style exemplars, in order to supply downstream 3D GAN transfer learning with sufficient training data with well-estimated camera labels (Section 3.1). Then we train an encoder to map input images into 3D GAN latent space (W^+), which well preserves facial identities with a *multi-view cycle consistent loss* (Section 3.2). To further improve the stylization quality, we add *guided transfer learning* that removes out-of-domain stylization artifacts(Section 3.3).

3D-Aware Image Generation. For multi-view consistent image generation, we build our pipeline on top of a state-of-the-art geometry-aware 3D GAN model named EG3D [8]. To synthesize an image, the 3D generator $\mathcal{G}_\phi$ takes two variables: a latent code z, from a standard Gaussian distribution, that determines the geometry and appearance of a subject; a conditional camera pose label $\hat{p}$ added to the latent code. z and $\hat{p}$ are passed through a multi-layer perceptron (MLP) mapping network to obtain a w code, which is duplicated multiple times to modulate the synthesis convolution layers that produce tri-plane features. These features are sampled into a neural radiance field at the desired camera angle p and accumulated to generate a raw feature image via volumetric rendering. Finally, the raw feature images are up-sampled by a super resolution module to synthesize the final RGB images. A camera-conditioned dual discriminator D is used to examine the image fidelity in adversarial training, while ensuring multi-view consistency.

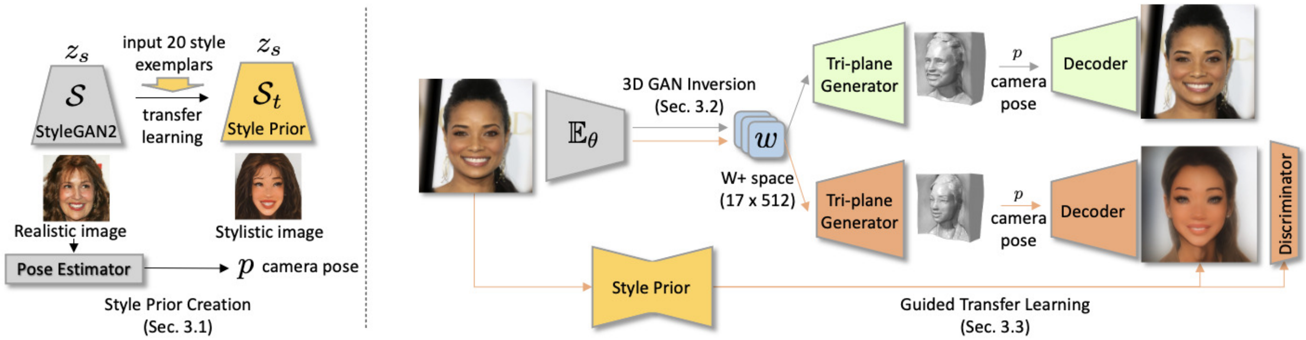

Figure 2. Pipeline overview. Our 3D stylization pipeline consists of three stages, style prior creation, 3D GAN inversion and guided style transfer learning, with different data training flows indicated in different colors. Specifically given a few 2D style exemplars, we create a 2D style prior (left, yellow) that augments stylistic training samples with well-estimated camera labels from real images. We then perform transfer learning of 3D-aware image generator to target styles using augmented labeled style samples (orange), under the paired guidance of 2D stylization. Additionally we train an encoder for 3D GAN image inversion (green) into a corresponding latent code in W^+ space, from which we can turn into an identity-preserved 3D style portraits.

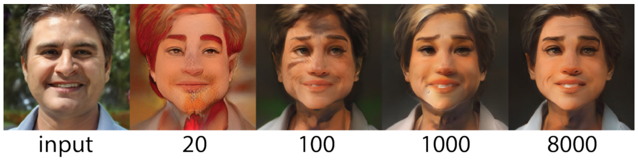

Figure 3. The evolution of transfer learning results using different number of training samples by our 2D style prior. Better visual quality is achieved with more training style exemplars.

3.1. Style Prior Creation

Unlike the 2D GAN-based stylization tasks, the few-shot transfer learning on 3D GAN model is less well studied. A straightforward attempt to 3D portrait stylization will be through fine-tuning a pretrained 3D generator $\mathcal{G}_\phi$ directly with the few shot samples, e.g., 20 stylized exemplars. However, plain transfer learning generates poorly in both perceptual quality and user similarity. We suspect the problem is rooted in two aspects: insufficient style exemplars due to the more complicated nature of a 3D GAN architecture and using inaccurate camera pose estimation from style exemplars.

To mitigate the above problems, instead of directly using given stylized exemplars to fine-tune the 3D generator, we create a style prior based on 2D GAN to guide the transfer learning, which are less complicated and does not need camera pose. Here we leverage the capability of the state-of-the-art 2D stylization methods for *style prior creation*. Inspired by recent 2D stylization works [38, 47], we perform transfer learning on top of the original StyleGAN2 generator $\mathcal{S}$ trained on FFHQ dataset [21], with the few shot style exemplars. We denote the fine-tuned styled generator as $\mathcal{S}_t$. This gives us the capability to turn widely-accessible real portrait images into augmented 2D style samples. Moreover, this augmentation approach naturally offers pairs of stylized and real images, from the latter of which we can obtain accurate camera pose labels with off-the-shelf pose estimator such as [55].

Transfer Learning Loss By sampling from prior latent space, we can get infinite high-quality diverse stylized images for transfer learning on 3D GANs. We use an adversarial loss to fine-tune the pre-trained 3D-aware generator $\mathcal{G}$ with respect to its parameter ϕ as well as its dual discriminator D, that matches the distribution of the translated images to the style prior distribution:

$$\begin{aligned}\mathcal{L}_{prior} =& \mathbb{E}_{z_s\sim N(0,I)}[min(0,-1+D(\mathcal{S}_t(z_s),p))]+ \\ & \mathbb{E}_{z\sim N(0,I)}[min(0,-1-D(\mathcal{G}_\phi(z,p),p))]\end{aligned} \tag{1}$$

where the latent code z_s and z are from StyleGAN2 and EG3D latent space respectively. We also apply regularization terms for stable fine-tuning. For discriminators, we use R_1 path regularization.

Our style prior leads to significant quality improvements, as shown in Fig. 3. We believe this is a requirement imposed by the NeRF module inside the 3D GAN pipeline. Typically, visual observations from a wide camera distribution are necessary for NeRF to reason the underlying 3D scene geometry and its corresponding appearance. Thus fine-tuning the cross-domain 3D generation of neural radiance features requires much more style samples than prior 2D GAN-based approaches.

Another point is the camera pose. For certain artistic styles, direct camera pose estimations from the input style exemplars might not be very accurate, which also affects 3D stylization. Different from 2D StyleGANs that directly up-sample feature maps into images via several convolution layers, 3D GAN synthesizes images by first accumulating

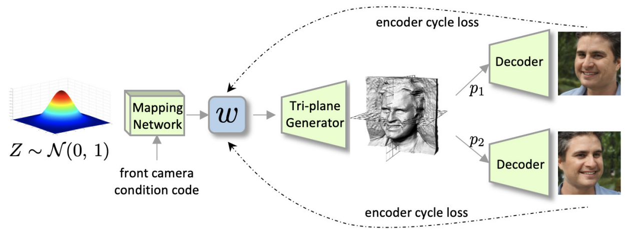

Figure 4. Illustration of multiview cycle consistency loss. Sampling from the Gaussian-distributed latent space, we synthesize facial images at different random poses, from which we minimize the difference of our encoded latent codes with the input.

neural radiance features via volume rendering to a feature map and then rely on super resolution to obtain the final image. Both the volume rendering and dual discriminator require reliable estimation of camera parameters, which are not easy to accurately obtained from non-realistic style examples.

3.2. Encoder Inversion with W^+

Assisted with the style prior creation step, we are able to achieve desirable 3D GAN stylization quality. One remaining challenge to complete the pipeline is the capability to invert a real face image into the latent space for 3D stylization.

Embedding Space and encoding The pre-trained EG3D model [8] is equipped with two latent spaces: the original latent space Z under a Gaussian distribution, and a less entangled W space via a mapping network from z with a conditional camera pose label $\hat{p}$. In spite of the camera swapping strategy, we observe the tri-plane generation is not fully decoupled from the pose, inducing varying geometry and appearance along with the change of $\hat{p}$. Therefore we use W space for our image embedding, and augment it to W^+ space that significantly increases the model expressiveness. In contrast to modulating the convolutional kernels with a same code w, W^+ produces a different w latent code for each layer, allowing for individual attribute control. In our case, a code w^+ in W^+ space has a dimension of 17×512, which can be represented as 17 w_i codes, where the w_0,...,w_{13} codes are for tri-plane generation, and $w_{14}, ..., w_{16}$ are used in super resolution module.

For fast inference, we train an encoder for image inversion, with the expectation of preserving user features to the largest extent. We design our encoder based on the architecture used in StyleGAN2 encoder, E4E [50], but fully exploit the unique proprieties of 3D GAN in generating view-consistent contents. In particular, we utilize the hierarchy of a pyramid network to capture different levels of detail from different layers. The input image resized at 256×256 resolution is passed through a headless pyramid network $\mathcal{E}_\theta$ to produce three levels of feature maps at different sizes. Each level's feature map then goes through a separate sub-encoder block to produce the W^+ style code.

Encoder Training Loss Even though our chosen W^+ space offers a large degree of freedom and expressiveness in representing real human faces, straightforward encoding without regularization can easily lead to out-of-domain issues, where the synthesized images present undesired artifacts like blurriness and noise. To prevent the encoder from over-drifting from the representation domain of $\mathcal{G}_\phi$, we introduce a *multi-view cycle consistent loss*, as shown in Fig 4. The core idea is that the encoder should reproduce the latent code from a synthesized image conditioned on w but rendered from arbitrary views. Specifically, a collection of latent codes randomly sampled under the standard Gaussian distribution, together with a fixed frontal camera pose, are fed into the mapping network and obtain w samples. Note that these in-domain w samples are complied with the original distribution of EG3D and likely to synthesize high-quality images without artifacts, and are a special form in W^+ space as well. Essentially training the encoder with these in-domain samples prevents the output w^+ codes from drifting far-away from the W space. We synthesize the images with N random camera poses $p_1, p_2, ..$ from training dataset camera distribution and supervise the training of $\mathcal{E}_\theta$ with ground-truth w^+ labels.

$$\mathcal{L}_{cyc} = \sum_{i=1}^{N} \mathcal{L}_2(w^+, \mathcal{E}_\theta(\mathcal{G}_\phi(w, p_i))). \tag{2}$$

In addition to the *multi-view cycle consistent loss*, our encoder is at the same time trained with reconstruction losses and regularization loss in a weighted combination manner, while freezing the EG3D generator weights.

Let x be the input image, passed through an encoder and decoder to yield $\hat{x} = \mathcal{G}_\phi(\mathcal{E}_\theta(x))$

$$\mathcal{L}_{rec} = \mathcal{L}_2(x, \hat{x}) + \mathcal{L}_{lpips}(x, \hat{x}) + \mathcal{L}_{arc}(x, \hat{x}) \tag{3}$$

The $\mathcal{L}_2, \mathcal{L}_{lpips}, \mathcal{L}_{arc}$ respectively measure the pixel-level, perceptual-level similarities [58] and facial recognition-level similarity differences. $\mathcal{L}_{arc}$ is based on the cosine similarity between intermediate features extracted from a pre-trained ArcFace recognition network [10], evaluating the identity similarity. A regularization term is further introduced to reduce the divergence of w^+ code to mimic the origin W space for the best of image quality,

$$\mathcal{L}_{reg} = ||\text{div}(\mathcal{E}_\theta(x))||_2. \tag{4}$$

3.3. Guided Transfer Learning

To further improve the 3D stylization quality, especially for cases where the inverted codes might be still not well

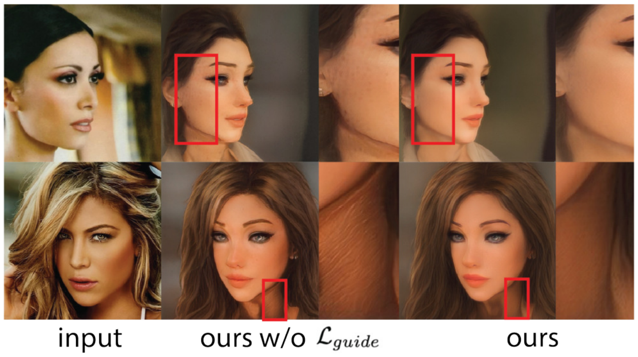

Figure 5. Our guided transfer learning loss helps improve generative quality and resolve fine-level visual artifacts.

aligned with the original distribution, the stylized images might contain artifacts, such as blurriness. By combining fine-tuning and inversion, we propose a guided transfer learning to enlarge the transfer learning space from its Z space to W^+ space with stronger generative stylization capability.

Thanks to the real to stylized face paired data that are produced in *style prior creation* step, we are able to guide the transfer learning of our 3D generator with a reconstruction loss. Given a real image x with estimated camera p and its 2D stylized pair x_s, let $\hat{x_s} = \mathcal{G}_\phi(\mathcal{E}_\theta(x), p)$ we have:

$$\mathcal{L}_{guide} = \mathcal{L}_2(x_s, \hat{x_s}) + \mathcal{L}_{lpips}(x_s, \hat{x_s}) \tag{5}$$

This guidance loss can help stabilize the generation training, and also improve the generative quality and user similarity, as illustrated in Fig. 5. We fine-tune the 3D generator and discriminator with the encoder and style prior frozen.

4. Experimental Analysis

4.1. Implementation Details

Our encoder is trained on the CelebA-HQ dataset [26] containing 30,000 high quality face images, where we use the first 28000 for training and the rest 2000 for testing. For consideration of computational efficiency, the input images are down-sampled to 256×256. The pre-trained EG3D uses the weights from the FFHQ 512-128 model [23]. We empirically set $\lambda_{reg} = 0.001$ and $\lambda_{cyc} = 1$ with 2 random camera poses. We minimize the objective function for 20 epochs using the Rectified Adam solver [31], with a batch size of 2 and a learning rate of 5×10^{-4}.

For transfer learning, we collect the initial 20 2D style exemplars from multi-image asset websites[39, 51] for each style, with which we train the style prior. For 3D GAN transfer learning, we use CelebA-HQ as real images in the guided transfer learning loss. Initialized with a pretrained EG3D model, the weights of the generator and discriminators are fine-tuned at a learning rate of 0.002 with a batch size of 4. We limit the number of iterations around 8K images.

Table 1. Stylization LPIPS ↓ for different stylization methods

	Ours	AgileGAN-EG3D	Toonify-EG3D
Cartoon	**0.195**	0.440	0.481
Oil Painting	**0.212**	0.433	0.525
Comic	**0.218**	0.379	0.486
Sam Yang	**0.20**	0.445	0.506
Sculpture	**0.234**	0.469	0.549

4.2. Comparisons

4.2.1 3D-Aware Stylization

In Fig. 6, we present more 3D portrait stylization results from our method. A diverse range of styles demonstrate that our method can robustly handle input images that represent a variety of genders, face shapes and hair styles under different illumination conditions, creating visually appealing, multi-view consistent stylization.

Since there is no prior few-shot 3D stylization methods that we can compare directly, we fine-tune a 3D generator as used in our method, following 2D stylization methods AgileGAN[47] and Toonify[38]. We name these two hybrid methods as *AgileGAN-EG3D* and *Toonify-EG3D*, and compare to them both quantitatively and qualitatively. For *Toonify-EG3D*, we perform direct transfer learning of the generator with the provided style exemplars, and the inversion is achieved with an optimization. For *AgileGAN-EG3D*, in addition to transfer learning the generator, we follow their setting to train a hierarchical variational encoder for image inversion.

Qualitative Evaluation In Fig. 7, we present visual comparisons against the two baseline methods. In constrast with *AgileGAN-EG3D* and *Toonify-EG3D* results exhibiting noticeable artifacts, our approach demonstrates 3D stylization with superior perceptual quality and identity preservation.

Quantitative Evaluation In Table 1, we also quantitatively measure the visual quality by evaluating a perceptual distance loss between the results of 3D style generator and 2D style prior, which we refer as Stylization LPIPS. The evaluation is performed on CelebA-HQ test images. Given the high quality 2D stylization (without 3D consistent manipulation capability though), we can consider a lower perceptual distance indicating higher visual quality, where our method outperforms the baselines substantially. We also compute a perceptual evaluation with user study, and please refer to our supplemental materials.

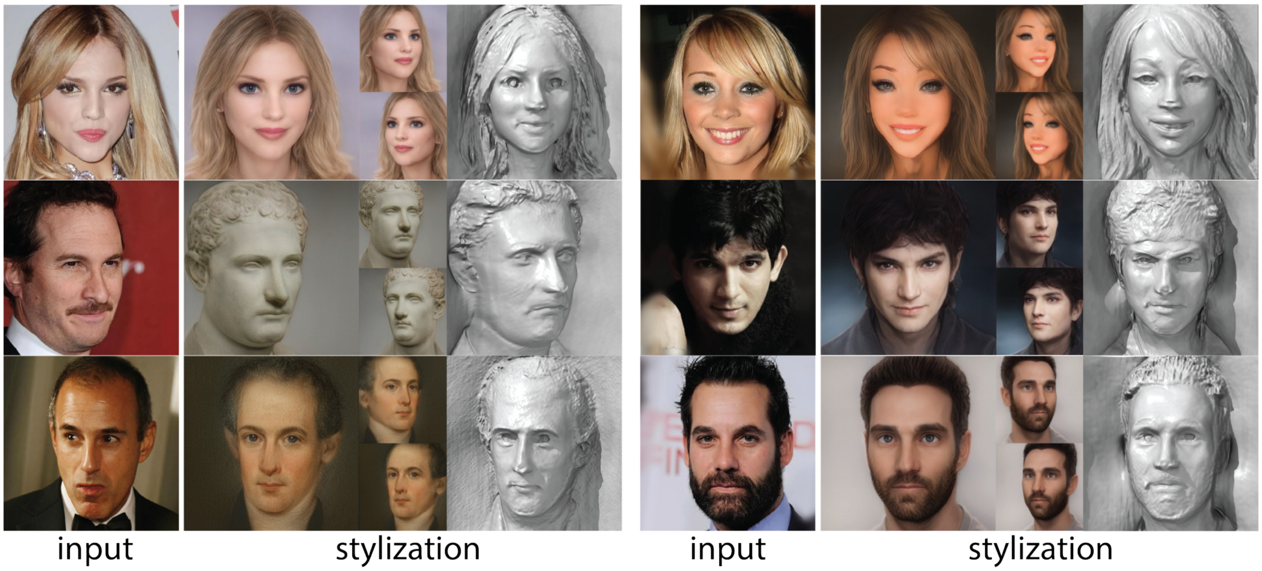

Figure 6. 3D artistic portraits generated from a variety of input images. From left to right, we show the input image, and generated stylistic multi-view images and geometry with our pipeline. Please refer to supplementary materials for higher-resolution qualitative results.

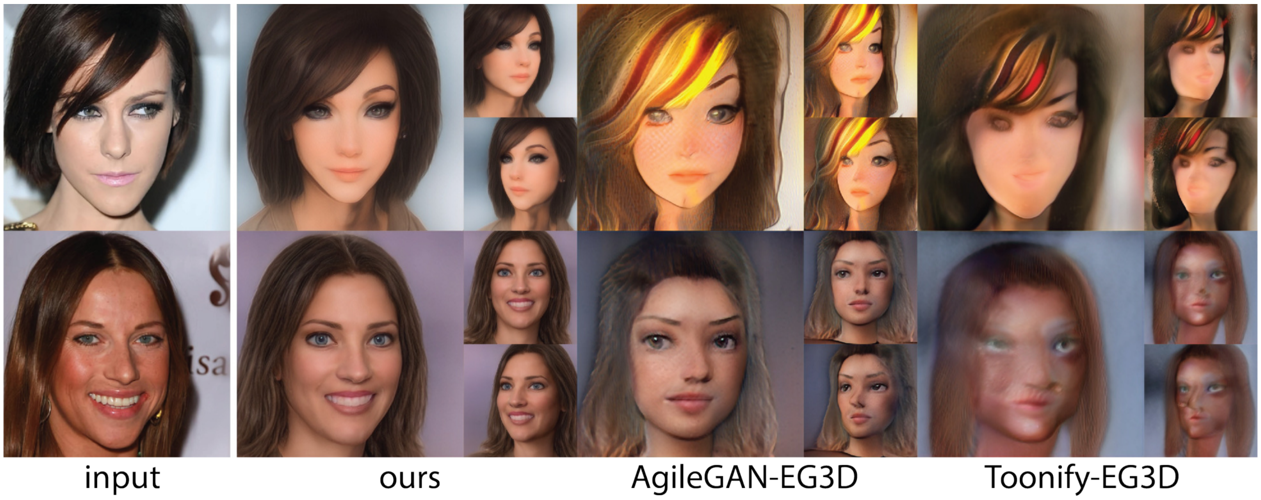

Figure 7. Our method visually outperforms direct transfer learning of EG3D generator following 2D few-shot stylization AgileGAN[47] and Toonify[38]. Our AgileGAN3D depicts identity-preserved 3D style portraits with fine-level details.

4.2.2 Multiview Manipulation Consistency

By leveraging a 3D-aware image generator, our method achieves multiview consistent stylization. 2D stylization approaches like AgileGAN[47] support limited view manipulation via modifying the latent code[45] but exhibits noticeable visual inconsistency. In Fig. 8, we visually compare the view consistency using Epipolar Line Images (EPI) similar to [54], where our method shows smooth and natural pattern transition when rotating the rendering camera. Additionally benefiting from camera-disentangled image synthesis capability, our AgileGAN3D is more robust in large-pose stylization as depicted in Fig. 9.

4.3. Ablation Studies

Inversion Learning In Fig. 10, we evaluate the efficacy of our cycle consistency loss introduced in our encoder for image inversion. With our loss, the encoder presents higher perceptual quality and identity similarity, numerically evidenced with better reconstruction losses in Tab. 2.

Transfer Learning We evaluate the effect of training sample quantity over the stylization quality both numerically in Tab. 3 and qualitatively in Fig. 3. The experiment

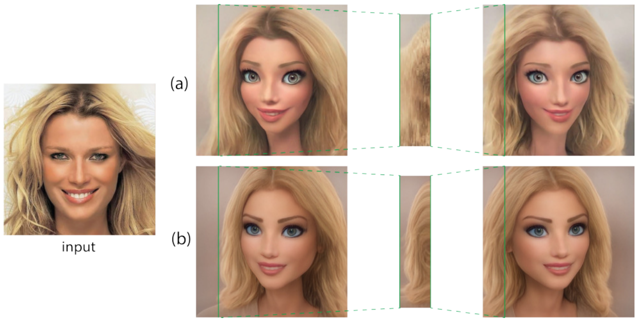

Figure 8. We compare our method (b) against 2D AgileGAN[47] (a) in multiview consistency. We manipulate stylization result from the left to right and horizontally stack a vertical segment of pixels from each generated image (middle). Our method shows a more natural visual transition, indicating better view consistency.

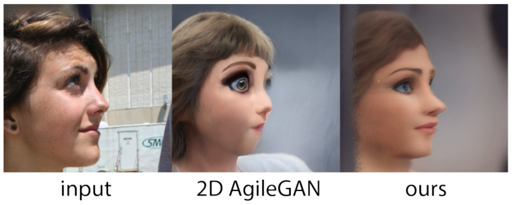

Figure 9. With the prior knowledge of camera extrinsics, our method demonstrates more robust large-pose stylization, compared to 2D AgileGAN[47].

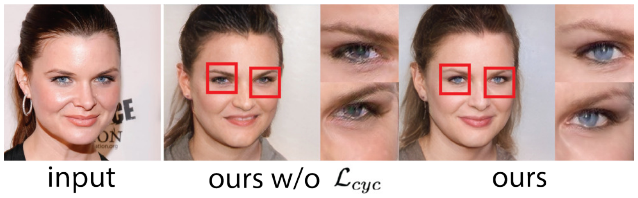

Figure 10. Multiview cycle consistent loss improves image inversion with higher perceptual quality and similarity.

Table 2. Quantitative ablation of cycle consistency loss, evaluated from 2K testing images.

Algorithm	MSE ↓	SSIM ↑	LPIPS ↓
Ours w/o Cycle Loss	0.0203	0.525	0.317
Ours	**0.0200**	**0.543**	**0.194**

Table 3. Stylization Perceptual scores ↓ with different training samples

# Training samples	LPIPS↓
20	0.41
100	0.252
1000	0.236
8000	0.227
8000(with guided loss, without accurate pose)	0.211
8000(with guided loss)	**0.200**

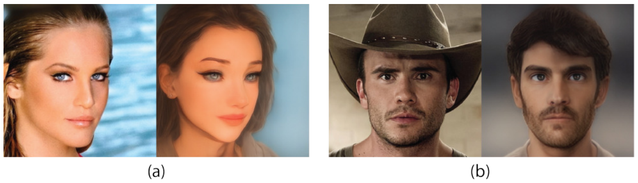

Figure 11. Failure examples. (a) inconsistent gaze directions, (b) unmodeled hat.

is performed over sam yang style. We compare against 2D AgileGAN stylization over test CelebA-HD images, where we observe closer perceptual quality as the number of training exemplars increase, demonstrating the efficacy of our augmented transfer learning. Additionally, our guided transfer learning further improves the perceptual score, as also visually evidenced in Fig. 5. We note that using camera labels estimated from style images leads to degenerated perceptual quality. That being said, our paired camera labels from real images help 3D GAN transfer learning.

5. Conclusion

We presented *AgileGAN3D*, the first few-shot pipeline generating high quality 3D stylistic portraits with detailed in-style geometry from a single user image, which sheds light on many potential applications. Our method only uses a limited number (around 20) of unpaired 2D style exemplars for a new target style. This is achieved via a novel framework incorporating *style prior creation* into *guided transfer learning*, which addresses the inadequate supervision issue of 3D GAN transfer learning with accurate camera labels. We also introduce a 3D GAN inversion module with *multi-view consistency loss* to improve identity preservation while achieving appealing stylization quality. Experimental results show that the algorithm produces high-quality multi-view consistent stylized 3D portraits.

Limitations We presented a variety of compelling 3D portrait stylization results, but there is still space for further improvement in our framework. Fig. 11 shows two example failure cases. (a) In some situations, we found that the generated eye gaze direction is biased towards frontal gaze, which may not be consistent with the input. (b) Occasionally, our approach fail to preserve accessories such as hat and glasses after stylization, as such cases are under-represented in the input datasets. These problems could potentially be mitigated by including more diverse input exemplars.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 4

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 4

[3] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 4

[4] Jihye Back, Seungkwon Kim, and Namhyuk Ahn. Webtoonme: A data-centric approach for full-body portrait stylization. *arXiv preprint arXiv:2210.10335*, 2022. 3

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019. 4

[6] Menglei Chai, Linjie Luo, Kalyan Sunkavalli, Nathan Carr, Sunil Hadap, and Kun Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6), 2015. 2

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3, 4

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3, 4, 6

[9] Min Jin Chong and David Forsyth. Jojogan: One shot face stylization. *arXiv preprint arXiv:2112.11641*, 2021. 2, 3

[10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[11] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *CVPR*, 2022. 3, 4

[12] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. 4

[13] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances In Neural Information Processing Systems*, 2022. 4

[14] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 4

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *CVPR*, 2022. 4

[17] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. 4

[18] David J. Heeger and James R. Bergen. Pyramid-based texture analysis/synthesis. In *ACM Trans. Graph.*, 1995. 3

[19] Jialu Huang, Jing Liao, and Sam Kwong. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia*, 24:1435–1448, 2021. 3

[20] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. *arXiv*, 2021. 3

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4, 5

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4

[23] 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

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 4

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 4

[26] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 7

[27] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, 2021. 3

[28] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *ICCV*, 2019. 4

[29] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *CVPR*, 2020. 4

[30] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors, 2023. 4

[31] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, 2020. 7

[32] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45:

Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 4

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 4

[34] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019. 4

[35] Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *NeurIPS*, 2020. 4

[36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3, 4

[37] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *CVPR*, 2022. 3, 4

[38] Justin N. M. Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. In *NeurIPS Workshop.*, 2020. 2, 3, 5, 7, 8

[39] pinterest. pinterest, 2021. `https://www.pinterest.com/`. 7

[40] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[41] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 4

[42] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 4

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[44] Shen Sang, Tiancheng Zhi, Guoxian Song, Minghao Liu, Chunpong Lai, Jing Liu, Xiang Wen, James Davis, and Linjie Luo. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. In *SIGGRAPH Asia 2022 Conference Papers*, New York, NY, USA, 2022. 3

[45] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *ECCV*, 2020. 8

[46] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 3, 4

[47] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: Stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2021. 2, 3, 5, 7, 8, 9

[48] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv*, 2019. 4

[49] Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. In *ACM Trans. Graph.*, 2020. 4

[50] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 4, 6

[51] turbosquid. turbosquid, 2021. `https://www.turbosquid.com/Search/3D-Models/`. 7

[52] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022. 4

[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NeurIPS*, 2016. 4

[54] 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. 8

[55] Sicheng Xu, Jiaolong Yang, Dong Chen, Fang Wen, Yu Deng, Yunde Jia, and Xin Tong. Deep 3d portrait from a single image, 2020. 2, 5

[56] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *CVPR*, 2022. 2, 3

[57] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 3

[58] 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. 6

[59] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arXiv*, 2021. 3, 4

[60] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 4

[61] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 4