Dr.3D: Adapting 3D GANs to Artistic Drawings

Wonjoon Jin
POSTECH
South Korea
jinwj1996@postech.ac.kr

Nuri Ryu
POSTECH
South Korea
ryunuri@postech.ac.kr

Geonung Kim
POSTECH
South Korea
k2woong92@postech.ac.kr

Seung-Hwan Baek
POSTECH
South Korea
shwbaek@postech.ac.kr

Sunghyun Cho
POSTECH
South Korea
Pebblous
South Korea
s.cho@postech.ac.kr

(a) Input (b) π-GAN (c) StyleNeRF (d) Dr.3D (e) Dr.3D + Editing

Figure 1: GAN inversion and semantic editing examples on a portrait drawing. For comparison, we perform naïve domain adaptation to π-GAN [Chan et al. 2021] and StyleNeRF [Gu et al. 2022] by finetuning them on portrait drawings. Then, we invert the input image in (a) using an off-the-shelf GAN inversion method to a latent code and reconstruct the image and its shape at a different camera pose using each 3D GAN model. The results in (b) and (c) show that naïve adaptations of existing 3D GANs fail to handle the input drawing. On the other hand, our method can successfully reconstruct the input image, and also allow semantic editing as shown in (d) and (e). Image in (a): Portrait of a Member of the Wedigh Family, 1532 by Hans Holbein the Younger, WikiArt [Public Domain] via (https://bit.ly/3KfgKPI)

ABSTRACT

While 3D GANs have recently demonstrated the high-quality synthesis of multi-view consistent images and 3D shapes, they are mainly restricted to photo-realistic human portraits. This paper aims to extend 3D GANs to a different, but meaningful visual form: artistic portrait drawings. However, extending existing 3D GANs to drawings is challenging due to the inevitable geometric ambiguity present in drawings. To tackle this, we present Dr.3D, a novel adaptation approach that adapts an existing 3D GAN to artistic drawings. Dr.3D is equipped with three novel components to handle the geometric ambiguity: a deformation-aware 3D synthesis network, an alternating adaptation of pose estimation and image synthesis, and geometric priors. Experiments show that our approach can successfully adapt 3D GANs to drawings and enable multi-view consistent semantic editing of drawings.

CCS CONCEPTS

• **Computing methodologies** → *Image processing*; **Artificial intelligence**.

KEYWORDS

Generative adversarial networks, domain adaptation, artistic drawings, 3D-aware image synthesis

ACM Reference Format:
Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. Dr.3D: Adapting 3D GANs to Artistic Drawings. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3550469.3555422

1 INTRODUCTION

Generative adversarial networks (GANs) [Goodfellow et al. 2014] have achieved remarkable success in learning to synthesize realistic images, which is crucial for a plethora of applications in computer

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9470-3/22/12…$15.00
https://doi.org/10.1145/3550469.3555422

graphics and vision [Karras et al. 2019, 2020b]. Notably, GANs allow us to explore synthesized images and edit real images in a semantically meaningful way [Shen et al. 2020; Shen and Zhou 2021; Wu et al. 2021]. Among many image categories that GAN methods have dealt with, it is not surprising that the human face is one of the most popular targets in computer graphics and vision. Recently, making GANs aware of 3D geometry has received great attention, opening up an exciting research field of 3D GANs. They tackle the ill-posed problem of learning the 3D-aware distribution of real images by explicitly modeling 3D light transport between a camera and a target object. 3D GANs enable the synthesis and editing of photographs not only in a semantically meaningful way, but also in consideration of 3D scene geometry [Chan et al. 2022, 2021; Gu et al. 2022; Niemeyer and Geiger 2021; Zhou et al. 2021].

To date, 3D GANs have been mainly demonstrated only on real-world photographs, which are the exact recordings of real-world scenes through perspective cameras. In this paper, we aim to extend the capability of 3D GANs to handle a different, but meaningful visual form: *drawing*. Drawing plays a crucial role in human history by depicting both real-world and imaginary subjects with intended and/or unintended variations. Existing 2D GAN methods have been extended to cope with drawings by adapting 2D GANs pretrained on real-world photographs into drawings, so-called domain adaptation [Isola et al. 2017; Karras et al. 2020a; Ojha et al. 2021; Zhu et al. 2017]. The adaptation strategy exploits common features between photographs and drawings, allowing us to bring the synthesis and editing capability of 2D GANs to the drawing domain [Wu et al. 2022]. Unfortunately, extending 3D GANs to the drawing domain turns out to be more challenging as shown in Figure 1.

One fundamental reason for this difficulty is that drawings have intrinsic geometric ambiguity on the subject and camera pose. Artists intentionally or unintentionally assume nondeterministic geometry of subjects from an imaginary viewpoint deviating from the physical one, resulting in drawing with creative ambiguity. This further increases the ill-posedness of learning a 3D-aware image distribution of drawings and hinders the direct application of previous domain adaptation methods used in 2D GANs for 3D GAN methods. Figure 1 shows that the application of state-of-the-art 3D GANs [Chan et al. 2021; Gu et al. 2022] on drawings via domain adaptation fails to synthesize faithful 3D-consistent images.

This paper proposes Dr.3D, a novel 3D GAN domain adaptation method for portrait drawings. Dr.3D effectively handles the fundamental geometric ambiguity of drawings with three remedies. First, we present a deformation-aware 3D synthesis network suitable for learning a large distribution of diverse shapes in drawing. Second, we propose an alternating adaptation scheme for 3D-aware image synthesis and pose estimation, effectively reducing the learning complexity of ambiguous 3D geometries and camera poses in drawings. Third, we impose geometric priors that enable stable domain adaptation from real photographs to drawings. The resulting domain adaptation method, Dr.3D, is the first method that enables stable editing and synthesis of drawing images in a 3D consistent way. We validate the effectiveness of Dr.3D via extensive quantitative and qualitative evaluations.

2 RELATED WORKS

3D-aware GANs. Several recent works have extended 2D GANs to be aware of the 3D structures of subjects and camera poses. Voxel-based 3D GANs [Nguyen-Phuoc et al. 2019] directly represent 3D structures with 3D voxel grids parameterized by 3D convolutional neural networks. Unfortunately, they typically suffer from large memory requirements. Mesh-based GANs [Liao et al. 2020; Szabó et al. 2019] lift the memory problem by using sparse meshes as a geometric representation. However, dealing with such sparse primitives with neural networks is challenging due to their unstructured data types. Recently, implicit 3D GANs [Chan et al. 2022, 2021; Gu et al. 2022; Niemeyer and Geiger 2021; Schwarz et al. 2020] have shown promising performance in terms of image fidelity and 3D consistency. GRAF and π-GAN [Chan et al. 2021; Schwarz et al. 2020] first proposed to learn to generate neural radiance fields (NeRF) [Mildenhall et al. 2020] and synthesize images via differentiable volumetric rendering. Since then, several attempts have been made to further improve the synthesis quality by incorporating feature projection and upsampling with the expense of losing multi-view consistency [Gu et al. 2022; Niemeyer and Geiger 2021]. Most recently, EG3D [Chan et al. 2022] demonstrates the synthesis of high-resolution 3D-aware images based on a tri-plane representation and a StyleGAN generator [Karras et al. 2020b]. Albeit great progress has been made in 3D GANs, directly applying them to drawings fails to learn meaningful 3D structures due to the large domain gap between real photographs and drawings (Figure 1).

Photo-to-Drawing Domain Adaptation. Applying GANs to drawings has often been practiced via domain adaptation in the 2D image space where we first train a GAN model on real photographs, and then finetune the model on a drawing dataset. This domain-adaptation technique has achieved notable success in synthesizing high-fidelity drawing images. Moreover, the adapted models inherit the semantically-meaningful editing capability of previous 2D GANs, thus enabling semantic editing of drawing images. Thus, their applications span the diverse computer graphics and vision fields, resulting in new applications such as image cartoonization [Pinkney and Adler 2020; Yang et al. 2022] and automatic caricature generation [Jang et al. 2021].

However, extending the success of 2D GANs to 3D GANs has been challenging. Drawings have ambiguous and diverse geometric shapes and appearances, resulting in a large domain gap between real photographs and drawings as witnessed by recent works [Gu et al. 2022]. Typical failure examples are flattened 3D shapes, inconsistent multi-view images, and low-fidelity images as shown in Figure 6. We aim to overcome this hurdle by proposing a 3D domain adaptation method designed explicitly for drawings and demonstrates compelling results via our stable photo-to-drawing domain adaptation.

Non-generative 3D-aware Image Editing. Editing an input image considering its 3D structure can be also done without using generative models. For instance, fitting a 3D parametric shape model [Blanz and Vetter 1999; Li et al. 2017; Paysan et al. 2009] to an image allows us to have a geometrically-editable 3D model textured with the image [Deng et al. 2019; Feng et al. 2021]. StyleRig [Tewari et al. 2020] propose combining 3DMM parameters

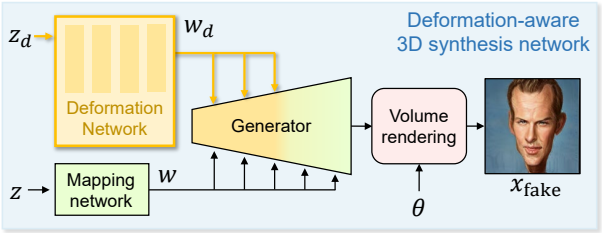

Figure 2: Network architecture of a deformation-aware 3D synthesis network. The network consists of a deformation network, a mapping network, a feature generator, and a volume rendering module. The network takes latent codes z_d and z, and a camera pose parameter θ as inputs, and synthesizes an image in a multi-view consistent way.

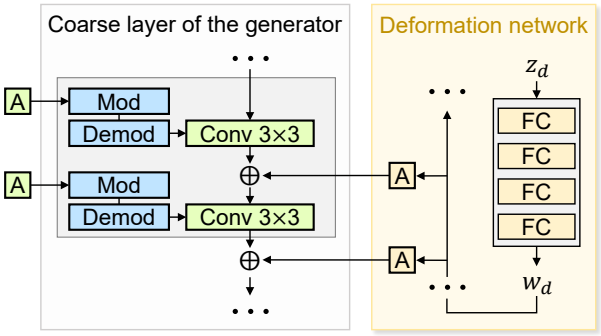

Figure 3: Network architectures of a generator and a deformation network. The generator network is based on the StyleGAN2 generator [Karras et al. 2020b]. FC: a fully-connected (FC) layer. A: an affine layer consisting of a single FC layer. Mod: a modulation layer. Demod: a demodulation layer.

with semantic features learned in 2D GANs for image editing. Unfortunately, parametric shape models are not applicable to drawings as the diverse 3D geometries in drawings often deviate from the representation space of existing parametric shape models. Another research direction is to reconstruct the 3D geometry of an input image based on the physics-based priors of light transport, where Unsup3d [Wu et al. 2020], GAN2Shape [Pan et al. 2021], and StyleGANRender [Zhang et al. 2021] show promising results. However, these methods often fail to handle drawings, because of their restrictive physics-based priors that assume the accurate decomposition of an image into illumination, appearance, and shape, which does not hold in drawings.

3 BACKGROUND ON EG3D

Before introducing our approach, we first provide a brief review of the network architecture of EG3D [Chan et al. 2022], a state-of-the-art 3D GAN network upon which our network is built. Specifically, it starts with a randomly sampled GAN latent code z, which turns into 2D convolutional features after passing through a StyleGAN-based feature generator [Karras et al. 2019]. Generated 2D feature maps are then rearranged into 3D orthogonal feature planes, from which any 3D point can be described with the projected features. Given the features, a multi-layer perceptron (MLP) decoder predicts the color and density of a 3D point, which are subsequently used for volume rendering, resulting in an image x_{fake} and a depth map d_{fake} at a camera pose θ. The feature generator and the MLP decoder are trained using a discriminator D, which is conditioned with the input camera pose θ to promote the generator to synthesize images that accurately reflect the camera pose. In contrast to being successful as a 3D GAN model for realistic portrait images, directly applying EG3D to drawings results in catastrophic failures as shown in Figure 7, due to the fundamental ambiguity in drawings.

4 DOMAIN ADAPTATION TO DRAWINGS

Built upon EG3D [Chan et al. 2022], Dr.3D is equipped with three remedies that mitigate the ill-posedness of photo-to-drawing 3D-aware domain adaptation: (1) a deformation-aware 3D synthesis network, (2) an alternating adaptation scheme for image synthesis and pose estimation, and (3) geometric priors for adaptation to drawing. In this section, we introduce each remedy in detail.

4.1 Deformation-aware 3D Synthesis

Drawings may have local shape variations that do not exist in photographs taken by cameras. To handle such domain gaps effectively, we introduce a deformation-aware 3D synthesis network G. Our network architecture builds on top of the EG3D network as shown in Figure 2. To model diverse shape deformations in drawings, our network uses an additional latent code z_d. The deformation code z_d turns into residual features via an MLP-based deformation network as shown in Figure 3, which are then added to early convolutional features in the StyleGAN feature generator. Note that modulating early layers in a StyleGAN generator is known to provide large-scale changes to synthesized images [Jang et al. 2021; Yang et al. 2022]. This simple feature-modulation strategy allows us to model diverse shape variations in drawings.

The role of the deformation-aware network is twofold. First, it introduces additional dimensions to the latent space so that local shape variations that may uniquely exist in target artistic drawing domains can be more effectively handled. Second, the residual features generated by the deformation-aware network help model the domain gap between the source and target domains more effectively. Specifically, as the mapping network is an MLP and the generator consists of spatially-invariant convolution operations, finetuning them cannot effectively model local deformations. To resolve this, our deformation-aware network estimates spatially-variant residual features to better handle local feature differences between the source and target domains. Moreover, the residual features help retain the original weights of the generator so that the knowledge about 3D structures learned in the original networks can be better preserved for more successful domain adaptation. Refer to the Supplemental Document for implementation details and additional analysis of the deformation network.

4.2 Alternating Adaptation of Pose Estimation and Image Synthesis

EG3D [Chan et al. 2022], which our method builds upon, requires known camera poses associated with input training images for its training. While camera poses for real portraits can be readily

estimated using an off-the-shelf pose estimation network [Deng et al. 2019; Feng et al. 2021], it is not trivial to obtain camera poses for portrait drawings. Previous pose estimation networks trained on real portraits fail on drawings due to the large domain gap, and there exist no datasets with ground-truth poses of drawings to train a pose estimation network. To tackle this problem, we may also adapt a pose-estimation network trained on portrait photos to drawings so that we can estimate the poses of drawings to train a synthesis network. However, adapting a pose-estimation network and a 3D synthesis network is a chicken-and-egg problem. Adapting a pose-estimation network requires training data with ground-truth pose labels, which can be obtained by an adapted 3D synthesis network, while adapting a 3D synthesis network requires an accurately adapted pose-estimation network.

To resolve this, we propose an alternating adaptation approach that alternatingly updates the 3D synthesis network G and pose-estimation network P (Fig. 4). Specifically, at each iteration of the alternating adaptation, we synthesize a pseudo ground-truth dataset using the current G, and update P using the synthesized dataset. Then, using the updated P, we estimate the poses of the real drawings in a training dataset and update G using the estimated poses. In this way, we can progressively adapt both P and G to a target drawing domain. However, at early iterations of the alternating adaptation, the poses of training images are not accurately estimated by P due to the large domain gap, which may eventually lead to the failure of adaptation. To overcome this, we introduce training losses with geometric priors, which will be described in Section 4.3, to guide the adaptation process. In the following, we describe each step of our alternating adaptation in more detail.

Adapting 3D Synthesis Network. Given an input drawing x_{real} as a training sample, we estimate its camera pose θ using a *fixed* pose estimation network. With the estimated pose θ, our 3D synthesis network G generates an image x_{fake} and its corresponding depth map d_{fake}. To adapt G, we employ the adversarial loss $\mathcal{L}_a$ of the original EG3D [Chan et al. 2022], which is based on a conditional discriminator D. Specifically, D takes either a synthetic or real image, x_{fake} or x_{real}, with its corresponding camera pose θ and evaluates how realistic it is. We update both G and D an adversarial-learning manner by back-propagating the loss. However, using the adversarial loss alone is not enough as there is no guarantee that the camera pose θ is accurate especially at early iterations of the alternating adaptation. Inaccurate pose estimation typically leads to learning flattened geometries for drawings as shown in Figure 7. To address this issue, we introduce an additional loss $\mathcal{L}_g$ based on geometric priors, described in Section 4.3. The 3D synthesis network G is then updated by minimizing a loss defined as:

$$\mathcal{L} = \mathcal{L}_a(x_{\text{fake}}, x_{\text{real}}, \theta) + \mathcal{L}_g(x_{\text{fake}}, d_{\text{fake}}, \theta). \tag{1}$$

Adapting Pose-estimation Network. We adapt the pose-estimation network P while fixing the 3D synthesis network G. To adapt P, we first generate a pseudo training dataset Ω that consists of multiple pairs of randomly sampled camera poses θ and their corresponding images x^{θ}_{fake}. We synthesize x^{θ}_{fake} as $x^{\theta}_{\text{fake}} = G(z, \theta)$ where z is a randomly sampled GAN latent code. On the pseudo dataset, we finetune our pose-estimation network P by minimizing the pose-estimation loss $\mathcal{L}_p$ defined as:

$$\mathcal{L}_p = \frac{1}{|\Omega|} \sum_{\{\theta, x^{\theta}_{\text{fake}}\} \in \Omega} \left\| \theta - P(x^{\theta}_{\text{fake}}) \right\|_2^2. \tag{2}$$

As our deformation-aware 3D synthesis network G continuously adapts to a drawing domain thanks to the adversarial and geometric-prior-based losses, our pose-estimation network P can coordinately adapt to a drawing domain through alternating adaptation.

4.3 Additional Losses with Geometric Priors

In order to guide the alternating adaptation process to a proper solution, the loss $\mathcal{L}_g$ is defined as a combination of three losses:

$$\mathcal{L}_g = \alpha \mathcal{L}_d + \beta \mathcal{L}_n + \gamma \mathcal{L}_p, \tag{3}$$

where α, β and γ are balancing weights. $\mathcal{L}_d$ is a depth similarity loss, $\mathcal{L}_n$ is a normal smoothness loss, and $\mathcal{L}_p$ is a pose loss defined in Equation (2). The pose loss $\mathcal{L}_p$ guides the 3D synthesis network to synthesize an image that matches the input camera pose θ. $\mathcal{L}_d$ and $\mathcal{L}_n$ correspond to geometric priors that guide G to synthesize a valid 3D geometry and an image correctly reflecting the input camera pose θ. In the following, we describe geometric priors $\mathcal{L}_d$ and $\mathcal{L}_n$ in detail, and discuss how the loss terms guide the alternating adaptation process to a proper solution.

Depth Similarity Loss. Even though portrait drawings have intrinsic geometric ambiguity, there are still similarities between drawings and real photographs because the category of subjects is still the same as human face. This incurs our first observation: the geometry of a subject depicted in a drawing is similar to the geometry in a photograph *at a high level*. We implement such prior by penalizing the *low-frequency* difference between the depth of a synthesized drawing d_{fake} and that of a synthesized photo $d_{\text{fake,photo}}$:

$$\mathcal{L}_d = \left\| k * d_{\text{fake}} - k * d_{\text{fake,photo}} \right\|_2^2, \tag{4}$$

where k is a 15×15-sized Gaussian low-pass filter of standard deviation 5. We use a synthesis network G_{photo} trained on real FFHQ photos [Karras et al. 2019] to generate its depth $d_{\text{fake,photo}} = G_{\text{photo}}(z, \theta)$. Note that latent code z and pose θ are the ones used for the drawing sample: $d_{\text{fake}} = G(z, \theta)$.

Normal Smoothness Loss. We further penalize abrupt changes of a synthesized geometry, which is implemented as a loss function:

$$\mathcal{L}_n = \| \nabla n_{\text{fake}} \|_2^2, \tag{5}$$

where ∇ is the spatial gradient operator and n_{fake} is a surface normal map computed from a synthesized depth map d_{fake}.

Effect on Alternating Adaption. The additional losses are crucial in guiding the alternating adaptation toward a proper solution. At early iterations of the alternating adaptation process, the 3D synthesis network G produces images that are close to real portrait images. As the pose estimation network P can accurately estimate the camera poses of such synthesized images at early iterations, the pose loss $\mathcal{L}_p$ can enforce G to produce images of the correct camera poses. On the other hand, the depth similarity loss $\mathcal{L}_d$ promotes G to synthesize 3D geometries that are close to their corresponding source-domain geometries. As the source-domain geometries

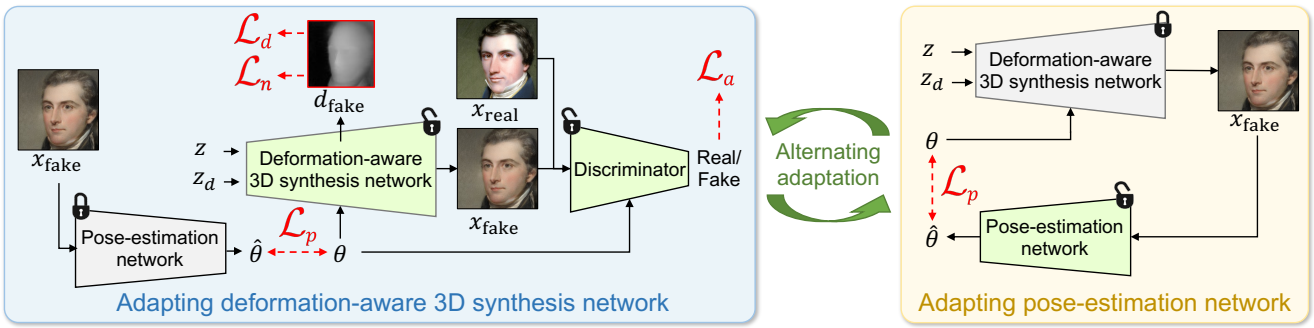

Figure 4: Alternating adaptation. Our approach alternatingly adapts the deformation-aware 3D synthesis network and the pose-estimation network. x_{real}: Portrait of Benjamin Moore McVickar, 1825 by Charles Cromwell Ingham, MetMuseum [Public Domain] via (https://bit.ly/3c9skiy).

Figure 5: 3D-aware drawing synthesis results of our 3D synthesis network adapted to different datasets by Dr.3D (from left to right: historical art, ukiyo-e, caricature, and anime).

have valid geometric structures and correctly reflect the camera poses, $\mathcal{L}_d$ guides G to synthesize valid unflattened geometries that correctly reflect the camera poses during the entire adaptation process. Finally, $\mathcal{L}_n$ helps avoid degenerate 3D structures with high-frequency artifacts. Thanks to the pose loss and geometric priors, the 3D synthesis network can be adequately adapted without drifting to an improper solution, which also helps the adaptation of the pose-estimation network.

4.4 Training Details

We pretrain the 3D synthesis network G and the pose-estimation network P on the real portrait images of the FFHQ dataset [Karras et al. 2019]. We apply horizontal flip for data augmentation. We use the Adam optimizer [Kingma and Ba 2014] with learning rates of 0.0001 and 0.00125 for optimizing P and G, respectively. The learning rate for the discriminator D is 0.00075. The 3D synthesis and pose-estimation networks G and P are alternatively trained within a mini-batch of 32 images. We freeze the first 10 layers of the discriminator D for stable domain adaptation [Mo et al. 2020]. We use the weights α, β and γ differently for target drawing domains as provided in the Supplemental Document.

5 ASSESSMENT

We conduct extensive validation of our method on four datasets of different drawing styles: historical art [Karras et al. 2020a], ukiyo-e [Pinkney 2020], anime [Anonymous et al. 2019], and caricature [Huo et al. 2018]. For the anime dataset, we crop and align face regions using an off-the-shelf face detection method [King 2009]. We apply Dr.3D to each dataset and obtain an adapted 3D GAN model separately. Figure 5 shows curated examples of 3D-aware drawing synthesis for the different drawing styles, demonstrating our 3D-aware synthesis capability for diverse drawing styles. Refer to the Supplemental Document for uncurated results.

5.1 Comparison

We compare Dr.3D to recent GAN-based 3D synthesis methods: StyleNeRF [Gu et al. 2022], π-GAN [Chan et al. 2021] and EG3D [Chan et al. 2022]. In the case of EG3D, we directly adapt EG3D from real photos to artistic drawings using camera poses estimated by an off-the-shelf pose-estimation network [Feng et al. 2021]. For the results of parametric fitting [Feng et al. 2021] and physics-based decomposition methods [Pan et al. 2021; Wu et al. 2020], refer to the Supplemental Document. Figure 6 shows a qualitative comparison between previous 3D GANs and ours. Naïve domain adaptation of

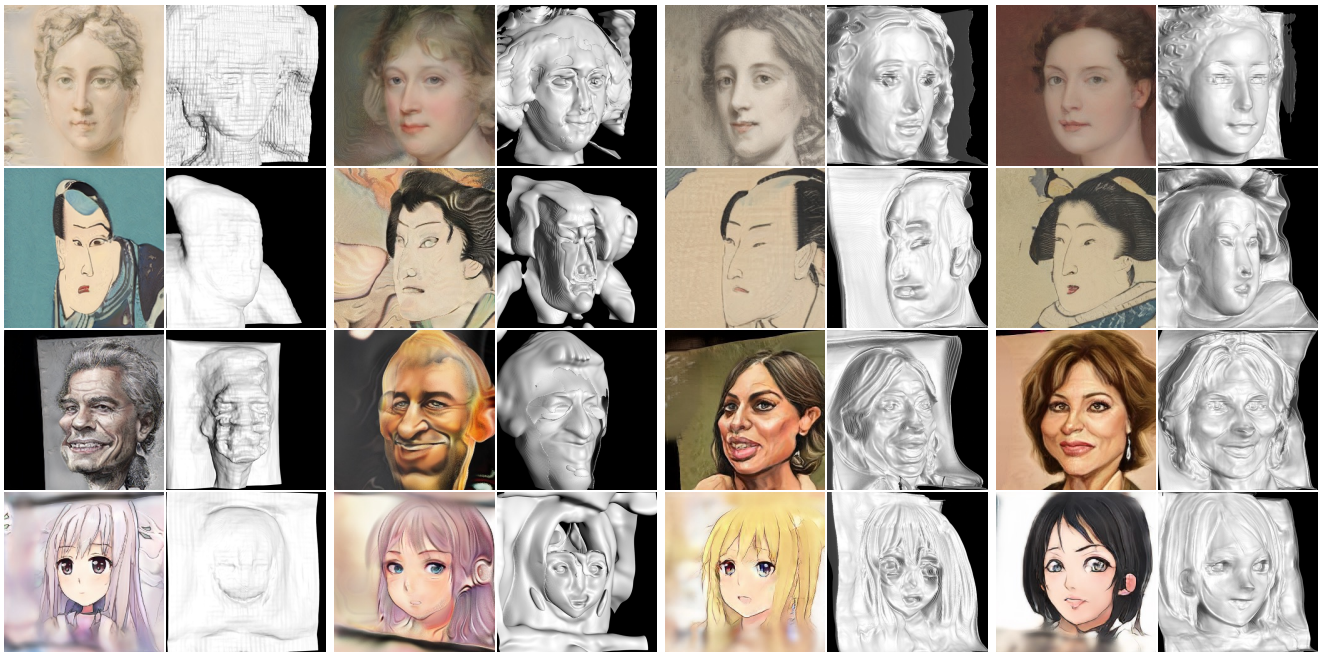

Figure 6: Qualitative comparison among StyleNeRF [Gu et al. 2022], π-GAN [Chan et al. 2021], EG3D [Chan et al. 2022] and ours. The contents of the images are different as they are generated by differently trained generator models. StyleNeRF, π-GAN and EG3D produce corrupted 3D geometries and unnatural-looking images especially for challenging styles such as ukiyo-e and anime, while our method produces more plausible shapes and images.

the previous 3D GANs fails to handle diverse drawing shapes and appearances, resulting in low-fidelity images and flattened geometries. Dr.3D reconstructs plausible shapes and images of drawing, outperforming the previous methods.

We further conduct quantitative analysis on the fidelity of synthesized images and shapes. The qualities of synthesized images are evaluated using FID [Heusel et al. 2017] and KID [Bińkowski et al. 2018]. We use 256×256-sized images for all the methods except π-GAN, for which we use 128×128-sized images due to its large memory requirement. Table 1 shows the evaluation results where Dr.3D achieves the best image-synthesis fidelity except for caricatures, thanks to our effective adaptation scheme.

Quantitative evaluation of synthesized shapes mandates the ground-truth shapes of drawings, which are challenging to obtain in most cases. For the historical-art dataset, as done in EG3D [Chan et al. 2022], we obtain the *pseudo* ground-truth shapes and poses of randomly generated drawings using a parametric fitting method [Feng et al. 2021]. We measure depth and pose error by calculating MSE between generated sets and pseudo ground-truth depths and poses. For the evaluation of caricatures, we utilize the 3DCaricShop dataset, which provides paired images and 3D shapes created by artists. We reconstruct caricature images using GAN-inversion and measure depth and pose error with ground-truth geometries.

Table 2 shows that Dr.3D generally performs better than the other methods in terms of shapes and poses. While StyleNeRF achieves better depth accuracy than ours for the historical-art dataset, it shows the worst pose accuracy. Also, while the table shows that

Table 1: Quantitative comparison on the image quality among π-GAN [Chan et al. 2021], StyleNeRF [Gu et al. 2022], EG3D [Chan et al. 2022] and ours.

		Hist. art	Ukiyo-e	Anime	Caricature
π-GAN	FID ↓	46.40	65.91	48.78	73.25
	KID $\times 10^3$ ↓	26.14	53.79	28.29	52.15
StyleNeRF	FID ↓	34.99	58.52	27.94	22.53
	KID $\times 10^3$ ↓	14.51	58.72	12.41	11.72
EG3D	FID ↓	26.95	40.16	20.75	**15.71**
	KID $\times 10^3$ ↓	9.295	32.94	8.699	**7.123**
Dr.3D (Ours)	FID ↓	**23.42**	**37.38**	**18.74**	19.69
	KID $\times 10^3$ ↓	**5.916**	**29.65**	**6.335**	9.180

EG3D achieves comparable results to ours, it tends to produce noisy and flattened shapes as shown in Figure 6.

5.2 Ablation Study

Dr.3D effectively deals with the intrinsic ambiguity of drawing images by means of (1) a deformation-aware 3D synthesis network, (2) alternating adaptation of pose estimation and image synthesis, and (3) geometric priors. We assess the impact of each component by starting with our baseline network, EG3D [Chan et al. 2022]. Figure 7 shows an ablation result. Using the original EG3D model on drawings results in a flattened shape (Figure 7(a)). For training the EG3D model, we used the camera pose estimated from an off-the-shelf pose-estimation network [Feng et al. 2021]. Our alternating

Table 2: Quantitative comparison on the shape and pose quality among π-GAN [Chan et al. 2021], StyleNeRF [Gu et al. 2022], EG3D [Chan et al. 2022] and Dr.3D.

	Hist. art		Caricature	
	Depth	Pose	Depth	Pose
π-GAN	0.305	0.072	0.151	0.077
StyleNeRF	**0.169**	0.333	0.688	0.326
EG3D	0.215	0.054	0.033	**0.047**
Dr.3D (Ours)	0.217	**0.030**	**0.020**	0.070

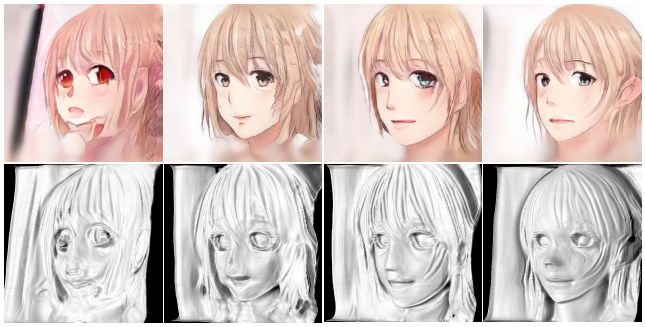

Figure 7: Ablation study. The baseline model (EG3D [Chan et al. 2022]) synthesizes a distorted image and a flattened geometry as shown in (a). While our alternating adaptation helps avoid flattened shapes as shown in (b), our deformation-aware 3D synthesis network, and geometric priors further improve the synthesis quality.

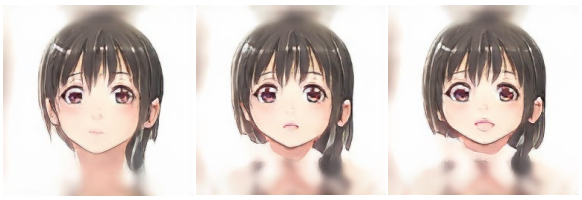

Figure 8: Image synthesis results from different deformation codes z_d. For all the results, the same latent code z is used.

adaptation of the pose-estimation network and the deformation-aware 3D synthesis network enables us to recover a better 3D geometry (Figure 7(b)). Adding our deformation-aware 3D synthesis network further improves the shape-reconstruction fidelity and the quality of synthesized images as it helps capture shape and style variations in drawings (Figure 7(c)). Our full method, Dr.3D, with the geometric priors, results in the best synthesis quality for both image and shape (Figure 7(d)).

As discussed in Section 4.1, drawings have a larger distribution of potentially-feasible 3D shapes than photos. Our deformation network helps model such a larger distribution of drawings by expanding the representation space with an additional latent code z_d, which leads to higher-quality adaptation results as shown in Figure 7(c). Figure 8 shows another example of the deformation network. In the figure, while the same latent code z synthesizes all the images, they exhibit different details due to different deformation codes z_d, proving the larger representation space expanded by the deformation network. More analysis on the deformation-aware network is provided in the Supplemental Document.

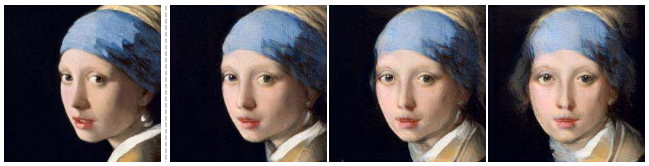

Figure 9: Novel view synthesis of a real-world drawing. Input: Girl with a Pearl Earring, 1665 by Johannes Vermeer, WikiArt [Public Domain] via (https://bit.ly/3PE66CT).

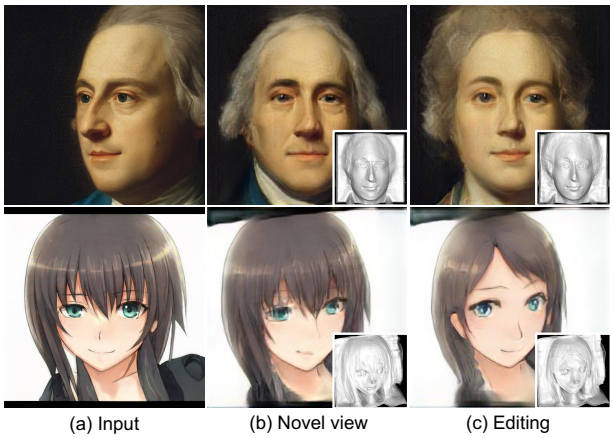

Figure 10: Semantic editing of input drawings. Top: male to female. Bottom: hairstyle change. Input on the top row: Gulian Verplanck, 1771 by John Singleton Copley, WikiArt [Public Domain] via (https://bit.ly/3PE6vVV).

5.3 3D-aware Semantic Editing of Drawing

Combining GAN inversion with Dr.3D enables multi-view consistent editing of real-world drawings such as novel-view synthesis and semantic editing. Figure 9 shows examples of novel-view synthesis of real-world drawings. In these examples, we estimate the camera poses of the input images using our domain-adapted pose-estimation network and invert the images to GAN latent codes using the pivotal tuning inversion method [Roich et al. 2021]. Then, we synthesize novel views of the input images by feeding their latent codes and new camera poses to the 3D synthesis network.

Dr.3D also enables multi-view consistent semantic editing on real-world drawings. Figure 10 shows examples of semantic editing. In these examples, we use editing vectors found by applying InterfaceGAN [Shen et al. 2020] using the original EG3D network trained on the FFHQ dataset.

6 CONCLUSION

This paper presented Dr.3D, a novel 3D GAN adaptation method from real portraits to artistic drawings. To handle the intrinsic geometric ambiguity of drawings, we proposed alternating adaptation of the pose estimation and image synthesis, a deformation-aware network, and geometric priors. We experimentally validated that our approach can successfully adapt 3D GANs to drawings for the first time. Dr.3D allows to edit an artistic drawing in consideration of its 3D geometric structure and semantics of the content.

Limitations. While Dr.3D can produce superior results to previous methods, it may still produce flattened geometries for some latent codes for challenging domains such as anime. Refer to the Supplemental Document for a failure example. Also, our method is limited in dealing with the background region in which 3D-consistent shared geometric features do not exist in training images. We note that this limitation also applies to existing 3D-GAN methods including EG3D [Chan et al. 2022]. One potential way to resolving this would be to divide the feature-generation procedure into two: one for the foreground and the other for the background [Gu et al. 2022]. Extending Dr.3D to diverse target domains including non-human faces would also be an interesting future direction.

ACKNOWLEDGMENTS

This research was supported by IITP grants funded by the Korea government (MSIT) (2021-0-02068, 2019-0-01906), an NRF grant funded by the the Korea government (MOE) (2022R1A6A1A03052954), and Pebblous.

REFERENCES

Anonymous, the Danbooru community, Gwern Branwen, and Aaron Gokaslan. 2019. Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2018. https://www.gwern.net/Danbooru2018 Accessed: DATE.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401* (2018).

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proc. of SIGGRAPH*. 187–194.

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. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proc. of IEEE/CVF CVPR*. 16123–16133.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. PI-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proc. of IEEE/CVF CVPR*. 5799–5809.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Proc. of IEEE/CVF CVPR Workshops*. 0–0.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. on Graphics (TOG)* 40, 4 (2021), 1–13.

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

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2022. StyleNeRF: A Style-based 3D Aware Generator for High-resolution Image Synthesis. In *Proc. of ICLR*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. of NeurIPS*.

Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2018. WebCaricature: a benchmark for caricature recognition. In *British Machine Vision Conference*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. of IEEE/CVF CVPR*. 1125–1134.

Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *ACM Trans. on Graphics (TOG)* 40, 4 (2021), 1–16.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training generative adversarial networks with limited data. In *Proc. of NeurIPS*. 12104–12114.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. of IEEE/CVF CVPR*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and improving the image quality of StyleGAN. In *Proc. of IEEE/CVF CVPR*. 8110–8119.

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. on Graphics (TOG)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. 2020. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *Proc. of IEEE/CVF CVPR*. 5871–5880.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. of ECCV*. Springer, 405–421.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2020. Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs. In *CVPRW*.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. HoloGAN: Unsupervised learning of 3D representations from natural images. In *Proc. of IEEE/CVF ICCV*. 7588–7597.

Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. of IEEE/CVF CVPR*. 11453–11464.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot image generation via cross-domain correspondence. In *Proc. of IEEE/CVF CVPR*. 10743–10752.

Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. 2021. Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. In *Proc. of ICLR*.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*. 296–301.

Justin NM Pinkney and Doron Adler. 2020. Resolution dependent GAN interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334* (2020).

Justin N. M. Pinkney. 2020. Aligned Ukiyo-e faces dataset. https://www.justinpinkney.com/ukiyoe-dataset.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744* (2021).

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. GRAF: Generative radiance fields for 3D-aware image synthesis. In *Proc. of NeurIPS*. 20154–20166.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of GANs for semantic face editing. In *Proc. of IEEE/CVF CVPR*. 9243–9252.

Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in GANs. In *Proc. of IEEE/CVF CVPR*. 1532–1540.

Attila Szabó, Givi Meishvili, and Paolo Favaro. 2019. Unsupervised generative 3D shape learning from natural images. *arXiv preprint arXiv:1910.00287* (2019).

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *Proc. of IEEE/CVF CVPR*. 6142–6151.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. 2020. Unsupervised Learning of Probably Symmetric Deformable 3D Objects From Images in the Wild. In *Proc. of IEEE/CVF CVPR*.

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for StyleGAN image generation. In *Proc. of IEEE/CVF CVPR*. 12863–12872.

Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. 2022. StyleAlign: Analysis and Applications of Aligned StyleGAN Models. In *Proc. of ICLR*.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7693–7702.

Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. 2021. Image {GAN}s meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering. In *International Conference on Learning Representations*.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-aware generator of GANs based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788* (2021).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of IEEE/CVF ICCV*. 2223–2232.