Cross-domain Adaptation for Few-shot 3D Shape Generation

Anonymous authors
Paper under double-blind review

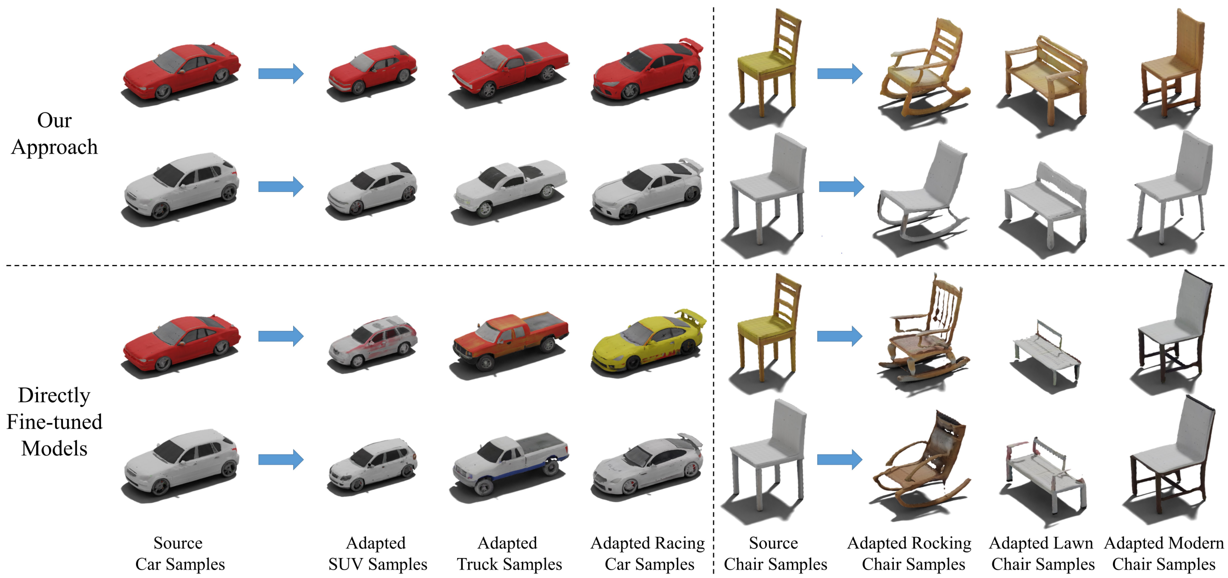

Figure 1: Given pre-trained 3D shape generative models, we propose to adapt them to target domains using a few target samples while preserving diverse geometry and texture information learned from source domains. Compared with directly fine-tuned models which tend to replicate the few-shot target samples instead of producing novel samples, our approach only needs the silhouettes of target samples as training data and achieves diverse generated shapes following target geometry distributions but different from target samples.

Abstract

Realistic and diverse 3D shape generation is helpful for a wide variety of applications such as virtual reality, gaming, and animation. Modern generative models learn from large-scale datasets and generate new samples following similar distributions. However, when training data is limited, deep neural generative networks overfit and tend to replicate training samples. Prior works focus on few-shot image generation to produce high-quality and diverse results using a few target images. Unfortunately, abundant 3D shape data is typically hard to obtain as well. In this work, we make the first attempt to realize few-shot 3D shape adaptation by adapting generative models pre-trained on large source domains to target domains. To relieve overfitting and keep considerable diversity, we propose to maintain the probability distributions of the pairwise relative distances between adapted samples at feature-level and shape-level during domain adaptation. Our approach only needs the silhouettes of few-shot target samples as training data to learn target geometry distributions and achieve generated shapes with diverse topology and textures. Moreover, we introduce several metrics to evaluate generation quality and diversity. The effectiveness of our approach is demonstrated qualitatively and quantitatively under a series of few-shot 3D shape adaptation setups.

1 Introduction

In recent years, 3D content has played significant roles in many applications, such as gaming, robotics, films, and animation. Currently, the most common method of creating 3D assets depends on manual efforts using specialized 3D modeling software like Blender and Maya, which is very time-consuming and cost-prohibitive to generate high-quality and diverse 3D shapes. As a result, the need for automatic 3D content generation becomes apparent.

During the past decade, image generation has been widely studied and achieved great success using generative models, including generative adversarial networks (GANs) (Goodfellow et al., 2014; Brock et al., 2019; Karras et al., 2019; 2020b; 2021), variational autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014; Vahdat & Kautz, 2020), autoregressive models (Van den Oord et al., 2016; Chen et al., 2018; Henighan et al., 2020), and diffusion models (Ho et al., 2020; Song & Ermon, 2020; Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021; Kingma et al., 2021). Compared with 2D images, 3D shapes are more complex and have different kinds of representations for geometry and textures. Inspired by the progress in 2D generative models, 3D generative models have become an active research area of computer vision and graphics and have achieved pleasing results in the generation of point clouds (Achlioptas et al., 2018; Yang et al., 2019; Zhou et al., 2021a), implicit fields (Chen & Zhang, 2019; Mescheder et al., 2019), textures (Pavllo et al., 2020; 2021; Richardson et al., 2023), and shapes (Gao et al., 2022; Liu et al., 2023). In addition, recent works based on neural volume rendering (Mildenhall et al., 2020) tackle 3D-aware novel view synthesis (Chan et al., 2021; 2022; Gu et al., 2022; Hao et al., 2021; Niemeyer & Geiger, 2021; Or-El et al., 2022; Schwarz et al., 2020; Xu et al., 2022; Zhou et al., 2021b; Schwarz et al., 2022).

Similar to 2D image generative models like GANs and diffusion models, modern 3D generative models require large-scale datasets to avoid overfitting and achieve diverse results. Unfortunately, it is not always possible to obtain abundant data under some circumstances. Few-shot generation aims to produce diverse and high-quality generated samples using limited data. Modern few-shot image generation approaches (Wang et al., 2018; Karras et al., 2020a; Mo et al., 2020; Wang et al., 2020; Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b; Zhu et al., 2022b;a; Zhao et al., 2023) adapt models pre-trained on large-scale source datasets to target domains using a few available training samples to relieve overfitting and produce adapted samples following target distributions. Nevertheless, few-shot 3D shape adaptation has yet to be studied, constrained by the complexity of 3D shape generation and the limited performance of early 3D shape generative models.

In this paper, we make the first attempt to study few-shot 3D shape adaptation pursuing high-quality and diverse generated shapes using limited data. We follow prior few-shot image generation approaches to adapt pre-trained source models to target domains using limited data. Since 3D shapes contain geometry and texture information, we need to clarify two questions: (i) what to learn from limited training data, and (ii) what to adapt from pre-trained source models to target domains. Naturally, we define two 3D shape domain adaptation setups: (i) geometry and texture adaptation (Setup A): the adapted models are trained to learn the geometry information of target data only and preserve the diversity of geometry and textures from source models, and (ii) geometry adaptation only (Setup B): the adapted models are trained to learn both the geometry and texture information of target data and preserve the diversity of geometry from source models only.

We design a few-shot 3D shape adaptation approach based on modern 3D shape GANs, which synthesize textured meshes with randomly sampled noises requiring 2D supervision only. Source models directly fine-tuned on limited target data cannot maintain generation diversity and produce results similar to training samples. As shown in Fig. 1, two different source samples become analogous after few-shot domain adaptation, losing diversity of geometry and textures. Therefore, we introduce a pairwise relative distances preservation approach to keep the probability distributions of geometry and texture pairwise similarities in generated shapes at both feature-level and shape-level during domain adaptation. In this way, the adapted models are guided to learn the common properties of limited training samples instead of replicating them. As a consequence, adapted models maintain similar generation diversity to source models and produce diverse results.

The main contributions of our work are concluded as follows:

- To our knowledge, we are the first to study few-shot 3D shape adaptation and achieve diverse generated shapes with arbitrary topology and textures.
- We propose a few-shot 3D shape adaptation approach to learn target geometry distributions using 2D silhouettes of extremely limited data (e.g., 10 shapes) while preserving diverse information of geometry

and textures learned from source domains. Our approach can also be adjusted to learn both target and geometry information using few-shot 2D RGB images as training data.

- We introduce several metrics to evaluate the quality and diversity of few-shot 3D shape generation and demonstrate the effectiveness of our approach qualitatively and quantitatively.

2 Related Work

3D Generative Models Early works (Wu et al., 2016; Smith & Meger, 2017; Lunz et al., 2020; Gadelha et al., 2017; Henzler et al., 2019) extend 2D image generators to 3D voxel grids directly but fail to produce compelling results with high resolution due to the large computational complexity of 3D convolution networks. Other works explore the generation of alternative 3D shape representations, such as point clouds (Achlioptas et al., 2018; Yang et al., 2019; Zhou et al., 2021a) and implicit fields (Chen & Zhang, 2019; Mescheder et al., 2019). Following works generate meshes with arbitrary topology using autoregressive models (Nash et al., 2020) and GANs (Luo et al., 2021). Meshdiffusion (Liu et al., 2023) first applies diffusion models to generate 3D shapes unconditionally. These works produce arbitrary topology only and need post-processing steps to achieve textured meshes that are compatible with modern graphics engines. DIBR (Chen & Zhang, 2019) and Textured3DGAN (Pavllo et al., 2020; 2021) synthesize textured 3D meshes based on input templated meshes, resulting in limited topology. GET3D (Gao et al., 2022) first proposes a 3D generative model to achieve arbitrary and diverse 3D geometry structures and textures using 2D images for supervision only. GET3DHuman (Xiong et al., 2023) depends on prior knowledge of human from external sources to synthesize textured 3D humans. DreamFusion (Poole et al., 2022) depends on a pre-trained 2D text-to-image diffusion model to perform text-to-3D synthesis. HoloDiffusion (Karnewar et al., 2023b) proposes a 3D-aware generative diffusion model to reconstruct 3D-consistent objects using 2D posed images. HoloFusion (Karnewar et al., 2023a) generates photo-realistic 3D radiance fields by combining HoloDiffusion with a 2D super-resolution network.

3D Shape Translation LOGAN (Yin et al., 2019) and UNIST (Chen et al., 2022) realize 3D shape translation based on VAEs trained on abundant data from two domains. Then translators are trained to transfer samples from one domain to the other based on the latent space provided by the VAEs. They tackle a different task from this work and aim to build a translation between two domains. Our approach aims to produce diverse results given few-shot data. Besides, LOGAN and UNIST are not qualified for few-shot data since both VAEs and translators need enough data to avoid overfitting.

Few-shot Generation Modern generative models need large amounts of data to achieve high-quality and diverse results. When training data is limited to a few samples, deep generative models tend to overfit and replicate them instead of generating novel results. Few-shot generation aims to solve the overfitting problem of generative models when training data is limited. Domain adaptation is a mainstream choice to realize few-shot generation. The key idea is to preserve the diverse information provided by source models while learning the common features of a few real target samples. In this way, a generative model for target domains is obtained to avoid overfitting or replicating training samples. The network structures of adapted models are consistent with source models in most cases.

Few-shot Image Generation Existing few-shot image generation methods aim to produce high-quality images with great diversity utilizing a few samples. Most modern approaches follow the TGAN (Wang et al., 2018) method to adapt generative models pre-trained on large source domains, including ImageNet (Deng et al., 2009), FFHQ (Karras et al., 2019), and LSUN (Yu et al., 2015) et al., to target domains with limited data. Following methods can be roughly divided into data augmentation approaches (Tran et al., 2021; Zhao et al., 2020a;b; Karras et al., 2020a), model regularization (Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b; Zhu et al., 2022b; Xiao et al., 2022), and trainable parameters fixing (Noguchi & Harada, 2019; Mo et al., 2020; Wang et al., 2020). CDC (Ojha et al., 2021) proposes a cross-domain consistency loss for generators and patch-level discrimination to build a correspondence between source and target domains. MaskDis (Zhu et al., 2022b) proposes to regularize the discriminator using masked features. DDPM-PA (Zhu et al., 2022a) first realizes few-shot image generation with diffusion models. Besides, other recent works have provided different research perspectives. RSSA (Xiao et al., 2022) proposes a relaxed spatial structural alignment method using compressed latent space derived from inverted GANs (Abdal et al., 2020). AdAM

(Zhao et al., 2022a) and RICK (Zhao et al., 2023) achieve improvement in the adaptation of unrelated source/target domains. Research including MTG (Zhu et al., 2021), OSCLIP (Kwon & Ye, 2022), GDA (Zhang et al., 2022b), and DIFA (Zhang et al., 2022a) et al. explore single-shot GAN adaptation with the guidance of pre-trained CLIP (Radford et al., 2021a) image encoders. This work first explores few-shot 3D shape adaptation and shares similar ideas of preserving diverse information provided by source models, achieving diverse textured 3D shapes using limited data. We design losses sharing similar formats with CDC but apply to both feature-level and shape-level information and make them adaptive to 3D shapes with a series of modifications.

3 Method

Given 3D generative models pre-trained on large source domains, our approach adapts them to target domains by learning the common geometry properties of limited training data while maintaining the generation diversity of geometry and textures. Directly fine-tuned models tend to replicate training samples instead of producing diverse results since the deep generative networks are vulnerable to overfitting, especially when training data is limited. To this end, we propose to keep the probability distributions of the pairwise relative distances between adapted samples similar to source samples.

We employ the 3D shape generative model GET3D (Gao et al., 2022) to illustrate the proposed approach. We first introduce GET3D briefly in Sec. 3.1. The silhouettes of target shapes are used as training data to learn target geometry distributions only under Setup A while RGB images are employed as training data to learn both geometry and texture distributions under Setup B. Adapted models are guided to realize geometry adaptation (Sec. 3.2) and texture adaptation (Sec. 3.3) using source models as reference for Setup A. Under Setup B, adapted models only preserve the diversity of geometry learned from source domains. As for textures, we guide adapted models to fit the distributions of training samples directly. Overall optimization targets under Setup A and B are provided in Sec. 3.4.

3.1 Preliminary: GET3D

GET3D is a 3D shape GAN trained on 2D images to generate 3D textured shapes. GET3D realizes arbitrary generation of topology and textures using the combination of geometry and texture generators. Both generators are composed of mapping networks M and synthesis networks S. We empirically fix the mapping networks M during domain adaptation in our approach. Ablations can be found in Appendix E. GET3D utilizes the differentiable surface representation DMTet (Shen et al., 2021) to describe geometry with signed distance fields (SDF) defined on deformation fields (Gao et al., 2020b;a). The texture generator uses mapped geometry and texture codes as inputs and generates texture fields for explicit meshes obtained by adopting DMTet for surface extraction. GET3D is trained with two 2D discriminators applied to RGB images and silhouettes, respectively.

GET3D is different from 3D-aware GANs and 3D diffusion models. Both 3D-aware GANs and GET3D need 2D images only as training data. 3D-aware GANs (Schwarz et al., 2020; Chan et al., 2021; 2022) generate novels views of 3D shapes but cannot extract 3D shapes directly. GET3D is the first randomly generative model trained on 2D images and synthesizing textured 3D shapes. Most 3D diffusion models (Liu et al., 2023; Nichol et al., 2022; Gupta et al., 2023; Chou et al., 2022; Shue et al., 2023) need 3D training data like meshes and point clouds since they need 3D ground truth to compute the reconstruction loss. Diffusion-based methods take up significantly larger computational costs, memory occupancy, and inference time. Our approach is implemented based on GET3D in this paper. However, it is not bound by certain network architectures of GET3D and can be applied to more powerful 3D shape GANs in the future to achieve higher-quality results. As an analogy, early few-shot image generation works are implemented with BigGAN (Brock et al., 2019), but they can be applied to StyleGANs (Karras et al., 2019; 2020b) as well.

3.2 Geometry Adaptation

We aim to guide adapted models to learn the common geometry properties of limited training samples while maintaining geometry diversity similar to source models. We propose to keep the probability distributions of

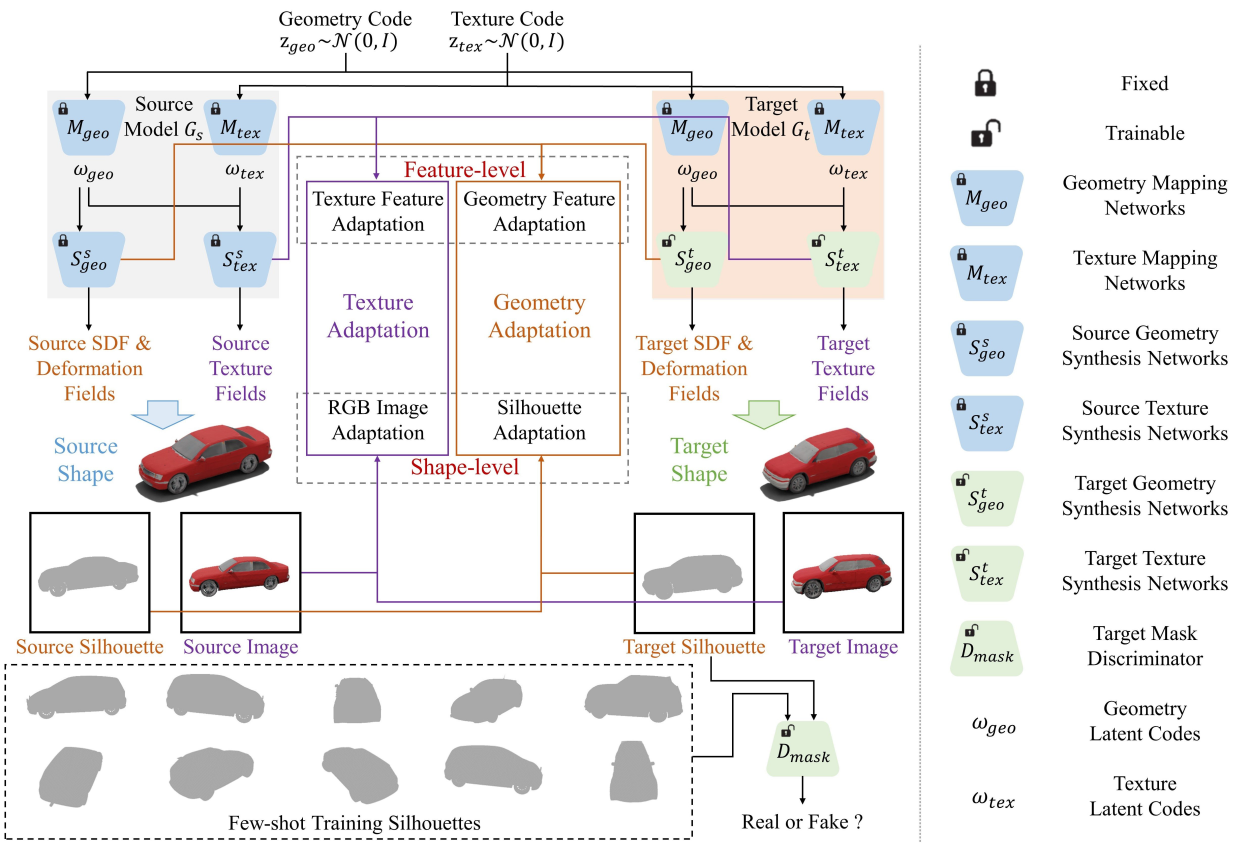

Figure 2: Overview of the proposed few-shot 3D shape generation approach using Cars → SUVs as an example: We maintain the distributions of pairwise relative distances between the geometry and textures of generated samples at feature-level and shape-level to keep diversity during domain adaptation. Only the silhouettes of few-shot target samples are needed as training data to learn target geometry distributions.

pairwise relative distances between the geometry structures of adapted samples at feature-level and shape-level. We first sample a batch of geometry codes $\{z^n_{geo}\}^N_0$ following the standard normal distribution $\mathcal{N}(0, I)$ and get mapped geometry latent codes $\{\omega^n_{geo}\}^N_0$ using fixed geometry mapping networks M_{geo}. The probability distributions for the i^{th} noise vector z^i_{geo} in the source and target geometry generators at feature-level can be expressed as follows:

$$p^{s,l}_{geo,i} = sfm(\{sim(S^{s,l}_{geo}(\omega^i_{geo}), S^{s,l}_{geo}(\omega^j_{geo}))\}_{\forall i \neq j}), \tag{1}$$

$$p^{t,l}_{geo,i} = sfm(\{sim(S^{t,l}_{geo}(\omega^i_{geo}), S^{t,l}_{geo}(\omega^j_{geo}))\}_{\forall i \neq j}), \tag{2}$$

where sfm and sim represent the softmax function and cosine similarity between activations at the l^{th} layer of the source and target geometry synthesis networks (S^s_{geo} and S^t_{geo}) which generate SDF and deformation fields. Then we guide target geometry synthesis networks to keep similar probability distributions to source models during domain adaptation with the feature-level geometry loss:

$$\mathcal{L}_{geo}(S^s_{geo}, S^t_{geo}) = \mathbb{E}_{z^i_{geo} \sim \mathcal{N}(0,I)} \sum_{l,i} D_{KL}(p^{t,l}_{geo,i} || p^{s,l}_{geo,i}), \tag{3}$$

where D_{KL} represents KL-divergence. Similarly, we use source and target silhouettes in place of the features in geometry synthesis networks to keep the pairwise relative distances of adapted samples at shape-level. For this purpose, we further sample a batch of texture codes $\{z^n_{tex}\}^N_0$ for shape generation. The probability distributions of shapes generated from the i^{th} noise vectors (z^i_{geo} and z^i_{tex}) by the source and target generators are given by:

$$p^s_{mask,i} = sfm(\Big\{sim(Mask(G_s(z^i_{geo}, z^i_{tex})), Mask(G_s(z^j_{geo}, z^j_{tex})))\Big\}_{\forall i \neq j}), \tag{4}$$

$$p^t_{mask,i} = sfm(\Big\{sim(Mask(G_t(z^i_{geo}, z^i_{tex})), Mask(G_t(z^j_{geo}, z^j_{tex})))\Big\}_{\forall i \neq j}), \tag{5}$$

where G_s and G_t are the source and target shape generators, $Mask$ represents the masks of 2D rendered shapes. We have the shape-level mask loss for geometry adaptation as follows:

$$\mathcal{L}_{mask}(G_s, G_t) = \mathbb{E}_{z^i_{geo}, z^i_{tex} \sim \mathcal{N}(0,I)} \sum_i D_{KL}(p^t_{mask,i} || p^s_{mask,i}). \tag{6}$$

3.3 Texture Adaptation

In addition, we also encourage adapted models to preserve the texture information learned from source domains and generate target shapes with diverse textures. We still apply the pairwise relative distances preservation approach to relieve overfitting and keep the generation diversity of textures. Since the generated textures for explicit meshes contain both geometry and texture information, we propose to use textures in regions shared by two generated shapes to compute the pairwise relative distances of textures while alleviating the influence of geometry. In the same way, we use the randomly sampled geometry codes $\left\{z^n_{geo}\right\}^N_0$ and texture codes $\{z^n_{tex}\}^N_0$ and get mapped latent codes $\left\{\omega^n_{geo}\right\}^N_0$ and $\{\omega^n_{tex}\}^N_0$ with fixed geometry and texture mapping networks M_{geo} and M_{tex}, respectively. The shared regions of two generated shapes produced by the source and adapted models are defined as the intersection of the masks of the 2D rendered shapes:

$$M^s_{i,j} = Mask(G_s(z^i_{geo}, z^i_{tex})) \wedge Mask(G_s(z^j_{geo}, z^j_{tex})) \ (i \neq j), \tag{7}$$

$$M^t_{i,j} = Mask(G_t(z^i_{geo}, z^i_{tex})) \wedge Mask(G_t(z^j_{geo}, z^j_{tex})) \ (i \neq j). \tag{8}$$

The probability distributions for the i^{th} noise vectors (z^i_{geo} and z^i_{tex}) in the source and target texture generators (S^s_{tex} and S^t_{tex}) at feature-level can be expressed as follows:

$$p^{s,m}_{tex,i} = sfm(\Big\{sim(S^{s,m}_{tex}(\omega^i_{geo}, \omega^i_{tex}) \otimes M^s_{i,j}, S^{s,m}_{tex}(\omega^j_{geo}, \omega^j_{tex}) \otimes M^s_{i,j})\Big\}_{\forall i \neq j}), \tag{9}$$

$$p^{t,m}_{tex,i} = sfm(\Big\{sim(S^{t,m}_{tex}(\omega^i_{geo}, \omega^i_{tex}) \otimes M^t_{i,j}, S^{t,m}_{tex}(\omega^j_{geo}, \omega^j_{tex}) \otimes M^t_{i,j})\Big\}_{\forall i \neq j}), \tag{10}$$

where $\otimes$ and sim represent the element-wise multiplication of tensors and cosine similarity between activations at the m^{th} layer of the source and target texture synthesis networks. For shape-level texture adaptation, we use 2D rendered shapes of RGB formats in place of the features in texture synthesis networks to compute the probability distributions:

$$p^s_{rgb,i} = sfm(\Big\{sim(RGB(G_s(z^i_{geo}, z^i_{tex})) \otimes M^s_{i,j}, RGB(G_s(z^j_{geo}, z^j_{tex})) \otimes M^s_{i,j})\Big\}_{\forall i \neq j}), \tag{11}$$

$$p^t_{rgb,i} = sfm(\Big\{sim(RGB(G_t(z^i_{geo}, z^i_{tex})) \otimes M^t_{i,j}, RGB(G_t(z^j_{geo}, z^j_{tex})) \otimes M^t_{i,j})\Big\}_{\forall i \neq j}), \tag{12}$$

where RGB represents the rendered RGB images of generated shapes. We have the feature-level texture loss and shape-level RGB loss for texture adaptation as follows:

$$\mathcal{L}_{tex}(S^s_{tex}, S^t_{tex}) = \mathbb{E}_{z^i_{geo}, z^i_{tex} \sim \mathcal{N}(0,I)} \sum_{m,i} D_{KL}(p^{t,m}_{tex,i} || p^{s,m}_{tex,i}), \tag{13}$$

$$\mathcal{L}_{rgb}(G_s, G_t) = \mathbb{E}_{z^i_{geo}, z^i_{tex} \sim \mathcal{N}(0,I)} \sum_i D_{KL}(p^t_{rgb,i} || p^s_{rgb,i}). \tag{14}$$

3.4 Overall Optimization Target

3.4.1 Setup A: Learning Geometry Only

Since adapted models are guided to learn the geometry information of training data, we only use the mask discriminator and apply the above-mentioned pairwise relative distances preservation methods to preserve

Figure 3: 10-shot generated shapes of our approach using ShapeNetCore Cars as the source domain.

diverse geometry and texture information learned from source domains. In this way, our approach only needs the silhouettes of few-shot target shapes as training data. The overall optimization target $\mathcal{L}$ of adapted models is defined as follows:

$$\mathcal{L} = \mathcal{L}(D_{mask}, G_t) + \mu\mathcal{L}_{reg} + \mu_1\mathcal{L}_{geo}(S^s_{geo}, S^t_{geo}) + \mu_2\mathcal{L}_{mask}(G_s, G_t) + \mu_3\mathcal{L}_{tex}(S^s_{tex}, S^t_{tex}) + \mu_4\mathcal{L}_{rgb}(G_s, G_t). \tag{15}$$

Here $\mathcal{L}(D_{mask}, G_t)$ and $\mathcal{L}_{reg}$ represent the adversarial objective of silhouettes and regularization term of generated SDFs used in GET3D. More details of these two losses are added in Appendix B. $\mu, \mu_1, \mu_2, \mu_3, \mu_4$ are hyperparameters set manually to control the regularization levels.

3.4.2 Setup B: Learning Geometry and Textures

The proposed adaptation approach under setup B has two differences compared with setup A. Firstly, the feature-level texture loss and shape-level RGB loss are no longer needed. Secondly, generators are guided by the RGB discriminator to learn target texture distributions. Therefore, we need RGB images of rendered real samples as inputs for the RGB discriminator. The overall optimization target of adapted models under setup B is defined as follows:

$$\mathcal{L} = \mathcal{L}(D_{mask}, G_t) + \mathcal{L}(D_{rgb}, G_t) + \mu\mathcal{L}_{reg} + \mu_1\mathcal{L}_{geo}(S^s_{geo}, S^t_{geo}) + \mu_2\mathcal{L}_{mask}(G_s, G_t). \tag{16}$$

Here $\mathcal{L}(D_{rgb}, G_t)$ represents the adversarial objective of rgb images used in GET3D. Details are added in Appendix B. μ, μ_1, μ_2 are hyperparameters set manually to control the regularization levels.

4 Experiments

We employ a series of few-shot 3D shape adaptation setups to demonstrate the effectiveness of our approach. We first show qualitative results in Sec. 4.1. Then we introduce several metrics to evaluate quality and diversity quantitatively in Sec. 4.2. We ablate our approach in Sec. 4.3 and provide additional experiments on larger domain gaps in 4.4.

Basic Setups The hyperparameter of SDF regularization μ is set as 0.01 for all experiments. We empirically find $\mu_1 = 2e+4, \mu_2 = 5e+3, \mu_3 = 5e+3, \mu_4 = 1e+4$ to work well for the employed adaptation setups. We conduct experiments with batch size 4 on a single NVIDIA A40 GPU. The learning rates of the generator and discriminator are set as 0.0005. Adapted models are trained for 40K-60K iterations. The resolution of 2D RGB images and silhouettes is 1024×1024. More details of implementation are added in Appendix G.

Datasets We use ShapeNetCore Cars, Chairs, and Tables Chang et al. (2015) as source datasets and sample several 10-shot shapes as target datasets, including (i) Trucks, (ii) Racing Cars, (iii) Sport Utility Vehicles (SUVs), (iv) Ambulances, (v) Police Cars corresponding to Cars,(vi) Rocking Chairs, (vii) Modern

Figure 4: 10-shot generated shapes of our approach using ShapeNetCore Chairs as the source domain.

Figure 5: Visualized samples comparison on 10-shot Cars → SUVs, Cars → Racing Cars, and Chairs → Rocking Chairs. The results of different approaches are synthesized with fixed noise inputs.

Chairs, (viii) Lawn Chairs corresponding to Chairs, and (ix) Round Tables, (x) School Tables corresponding to Tables. The few-shot 3D shapes are processed to RGB images (for Setup B only) and silhouettes using 24 evenly distributed camera poses as training data.

Figure 6: 10-shot generated shapes of our approach on Cars → Ambulances and Police Cars.

Datasets	Approach	CD (↓)	Intra-CD (↑)	Pairwise-CD (↑)	Intra-LPIPS (↑)	Pairwise-LPIPS (↑)
Cars → SUVs	DFTM	1.401	0.316 ± 0.002	0.513 ± 0.001	0.062 ± 0.001	0.063 ± 0.012
	FreezeT	1.553	0.240 ± 0.005	0.326 ± 0.002	0.055 ± 0.002	0.060 ± 0.014
	Ours	**1.323**	$\mathbf{0.511 \pm 0.006}$	$\mathbf{0.814 \pm 0.007}$	$\mathbf{0.109 \pm 0.026}$	$\mathbf{0.095 \pm 0.022}$
Cars → Trucks	DFTM	4.014	0.441 ± 0.003	0.689 ± 0.003	0.112 ± 0.002	0.119 ± 0.024
	FreezeT	4.175	0.412 ± 0.006	0.766 ± 0.002	0.120 ± 0.003	0.128 ± 0.027
	Ours	**3.940**	$\mathbf{1.061 \pm 0.014}$	$\mathbf{1.175 \pm 0.004}$	$\mathbf{0.145 \pm 0.022}$	$\mathbf{0.146 \pm 0.033}$
Chairs → Lawn Chairs	DFTM	40.559	4.001 ± 0.005	13.598 ± 0.013	0.165 ± 0.029	0.141 ± 0.047
	FreezeT	39.422	4.671 ± 0.022	19.269 ± 0.024	0.120 ± 0.032	0.165 ± 0.040
	Ours	**38.661**	$\mathbf{5.852 \pm 0.031}$	$\mathbf{22.989 \pm 0.022}$	$\mathbf{0.278 \pm 0.040}$	$\mathbf{0.166 \pm 0.054}$
Chairs → Rocking Chairs	DFTM	18.996	7.405 ± 0.022	15.312 ± 0.011	0.202 ± 0.039	0.203 ± 0.037
	FreezeT	18.503	5.541 ± 0.014	11.977 ± 0.009	0.203 ± 0.046	0.204 ± 0.036
	Ours	**17.598**	$\mathbf{8.773 \pm 0.029}$	$\mathbf{16.165 \pm 0.015}$	$\mathbf{0.289 \pm 0.062}$	$\mathbf{0.222 \pm 0.063}$

Table 1: Quantitative evaluation of our approach. We fix noise inputs for different methods to conduct fair comparison. CD scores are multiplied by 10^3. Our approach performs better on both generation quality and diversity.

Baselines Since few existing works explore few-shot 3D shape generation, we compare the proposed approach with directly fine-tuned models (DFTM) and fine-tuned models using fixed texture generators (FreezeT), including fixed texture mapping and texture synthesis networks.

4.1 Qualitative Evaluation

Setup A We visualize samples produced by our approach using source models pre-trained on ShapeNetCore Cars and Chairs in Fig. 3 and 4, respectively. Our approach only needs the silhouettes of few-shot training samples as target datasets to adapt source models to target domains while maintaining generation diversity of geometry and textures. In addition, we compare our approach with baselines using fixed noise inputs for intuitive comparison in Fig. 5. DFTM models replicate training samples and fail to keep generation diversity. FreezeT also fails to produce diverse textures since the mapped geometry codes influence the fixed texture synthesis networks. As a result, FreezeT models produce textured meshes similar to training samples under the guidance of RGB discriminators. Therefore, we further train FreezeT models without RGB discriminators or using source RGB discriminators. However, these approaches still fail to preserve the diverse geometry and texture information of source models and cannot produce reasonable shapes. We maintain the pairwise relative distances between generated shapes at feature-level and shape-level and achieve high-quality and diverse adapted samples. Supplemental visualized results are provided in Appendix I.

Setup B We employ RGB images of 10-shot Ambulances and Police Cars as training data for Setup B. As shown in Fig. 6, our approach produces diverse ambulances and police cars with diverse topology.

Quality Analysis The proposed approach realizes few-shot domain adaptation of pre-trained models. Our approach achieves generation quality similar to pre-trained GET3D models. We provide visualized samples of GET3D in Appendix B, in which incomplete textures of tires and failure of detailed structures can be found. As a result, our approach produces some samples with incomplete textures of tires and cannot

Datastes	Approach	FID (↓)	CD (↓)	Datastes	Approach	FID (↓)	CD (↓)
Cars → Ambulances	DFTM	101.583	6.896	Cars → Police Cars	DFTM	86.833	6.440
	Ours	**93.697**	**5.963**		Ours	**74.958**	**5.616**

Table 2: Quantitative evaluation of our approach on generation quality of geometry and textures.

Datastes	Approach	Intra-CD (↑)	Pairwise-CD (↑)	Inra-LPIPS (↑)	Pairwise-LPIPS (↑)
Cars → Ambulances	DFTM	0.300 ± 0.002	$\mathbf{1.027 \pm 0.007}$	0.079 ± 0.009	0.083 ± 0.017
	Ours	$\mathbf{0.558 \pm 0.004}$	0.638 ± 0.006	$\mathbf{0.093 \pm 0.018}$	$\mathbf{0.086 \pm 0.016}$
Cars → Police Cars	DFTM	0.426 ± 0.003	$\mathbf{0.926 \pm 0.008}$	0.109 ± 0.002	0.108 ± 0.017
	Ours	$\mathbf{0.902 \pm 0.005}$	0.902 ± 0.006	$\mathbf{0.115 \pm 0.009}$	$\mathbf{0.120 \pm 0.020}$

Table 3: Quantitative evaluation of our approach on generation diversity of geometry and textures.

synthesize some detailed structures similar to some training samples. Our approach can be combined with more powerful 3D shape GANs in the future to achieve better visual effects.

4.2 Quantitative Evaluation

Evaluation Metrics The generation quality of adapted models represents their capability to learn target geometry distributions. Chamfer distance (CD) (Chen et al., 2003) is employed to compute the distances of geometry distributions between 5000 adapted samples and target datasets containing relatively abundant target data to obtain reliable results. Besides, we design several metrics based on CD and LPIPS (Zhang et al., 2018) to evaluate the diversity of geometry and textures in adapted samples, which are computed in two ways: (i) pairwise-distances: we randomly generate 1000 shapes and compute the pairwise distances averaged over them, (ii) intra-distances (Ojha et al., 2021): we assign the generated shapes to one of the training samples with the lowest LPIPS distance and then compute the average pairwise distances within each cluster averaged over all the clusters. LPIPS results are averaged over 8 evenly distributed views of rendered shapes. Adapted models tending to replicate training samples may achieve fine pairwise-distances but only get intra-distances close to 0. However, adapted models with great generation diversity should achieve large values of both pairwise and intra-distances.

Setup A The quantitative results of our approach under Setup A are compared with baselines on several few-shot adaptation setups, as listed in Table 1. Our approach learns target geometry distributions better in terms of CD. Moreover, our approach also performs better on all the benchmarks of diversity, indicating its strong capability to produce diverse shapes with different geometry structures and textures.

Setup B We further add FID Heusel et al. (2017) to evaluate the generation quality under Setup B. To produce stable and reliable FID results, we use 73 ambulances samples and 133 police cars samples from ShapeNetCore Chang et al. (2015) as target datasets. FID results are averaged over 24 views of rendered shapes. The quantitative results are listed in Tables 2 and 3. Our approach achieves better results than DFTM models on the employed two setups. Compared with DFTM models, our approach also performs better in learning target geometry distributions in terms of CD. Besides, our approach achieves greater generation diversity in terms of Intra-CD and Intra-LPIPS. DFTM models get better results on Pairwise-CD and results close to our approach on Pairwise-LPIPS but get apparently worse results on intra-distances, indicating that they are overfitting to few-shot training samples and tend to replicate them instead of producing diverse results. We do not include FreezeT models for comparison under Setup B since the adapted models need to learn the texture information from limited training samples.

4.3 Ablation Analysis

We provide ablation analysis to show the roles played by each component of our approach. In Fig. 7, we show the qualitative ablation analysis using 10-shot Chairs → Rocking Chairs as an example. Our full approach adapts source samples to target domains while preserving diverse geometry and texture information. Adapted models only using GAN loss with mask discrimination fail to maintain geometry diversity or produce

Figure 7: Qualitative ablations of our approach using 10-shot Chairs → Rocking Chairs as an example. Results of different approaches are synthesized with fixed noise inputs for intuitive comparison.

Approach	CD (↓)	Intra-CD (↑)	Pairwise-CD (↑)	Intra-LPIPS (↑)	Pairwise-LPIPS (↑)
w/o Texture loss	18.178	8.054 ± 0.028	13.533 ± 0.010	0.221 ± 0.013	0.210 ± 0.045
w/o Geometry loss	18.409	7.551 ± 0.019	12.549 ± 0.009	0.271 ± 0.023	0.217 ± 0.057
w/o RGB loss	17.762	7.207 ± 0.018	13.124 ± 0.010	0.211 ± 0.006	0.213 ± 0.034
w/o Mask loss	18.275	6.878 ± 0.014	12.435 ± 0.008	0.248 ± 0.010	0.208 ± 0.010
Full Approach	**17.598**	$\mathbf{8.773 \pm 0.029}$	$\mathbf{16.165 \pm 0.015}$	$\mathbf{0.289 \pm 0.062}$	$\mathbf{0.222 \pm 0.063}$

Table 4: Quantitative ablations of the proposed approach using 10-shot Chairs → Rocking Chairs as an example.

high-quality shapes. Adding fixed source RGB discriminators results in texture degradation. Absence of the feature-level texture loss makes it harder for adapted models to preserve the texture information learned from source domains. Absence of shape-level RGB loss leads to repetitive textures and discontinuous shapes. As for the feature-level geometry and shape-level mask losses, their absence results in adapted samples sharing similar geometry structures and incomplete shapes. We also add ablations using geometry and mask losses, texture and RGB losses, feature-level losses, and shape-level losses, respectively. None of these approaches get compelling results. Incomplete geometry structures and low-quality textures can be found in their adapted samples. As shown in Table 4, the full approach achieves the best quantitative results on both generation quality and diversity. Without feature-level geometry loss or shape-level mask loss, adapted models perform worse on geometry diversity in terms of Intra-CD and Pairwise-CD. Similarly, adapted models perform worse on texture diversity in terms of Intra-LPIPS and Pairwise-LPIPS without feature-level texture loss or shape-level RGB loss. More detailed ablations for the number of training samples, number of views, fixed mapping networks, shared masks proposed in Sec. 3.4, and hyperparameters are added in Appendix E.

4.4 Larger Domain Gaps

We have conducted abundant experiments on related source/target domain adaptation like Cars → Trucks. In this section, we further add experiments on source/target domains with larger domain gaps. We employ

Figure 8: Visualized samples on 10-shot Tables → Modern Chairs and Lawn Chairs.

Adaptation	Intra-CD (↑)	Intra-LPIPS (↑)
Chairs → Modern Chairs (directly fine-tuned)	3.582 ± 0.004	0.149 ± 0.023
Chairs → Modern Chairs (ours)	$\mathbf{5.011 \pm 0.022}$	$\mathbf{0.254 \pm 0.045}$
Tables → Modern Chairs (ours)	4.735 ± 0.021	0.226 ± 0.030
Chairs → Lawn Chairs (directly fine-tuned)	4.001 ± 0.005	0.165 ± 0.029
Chairs → Lawn Chairs (ours)	$\mathbf{5.852 \pm 0.031}$	$\mathbf{0.278 \pm 0.040}$
Tables → Lawn Chairs (ours)	5.247 ± 0.018	0.242 ± 0.036

Table 5: Quantitative comparison between different adaptation setups. CD scores are multiplied by 10^3.

two adaptation setups: Tables → Modern Chairs and Lawn Chairs trained on 10-shot silhouettes. As shown in Fig. 8, our approach is qualified for domain gaps like Tables → Chairs. Our approach adapts source table samples to chairs and retains considerable diversity. In addition, we add quantitative results to evaluate the generation diversity under different adaptation setups and report results in Table 5. Our approach achieves relatively lower diversity when adapting Tables to Modern and Lawn Chairs compared with adapting from Chairs. Despite that, our approach still achieves apparently greater generation diversity than directly fine-tuned models even with larger domain gaps, showing its ability to maintain diversity in few-shot 3D shape generation. Additional visualized samples are added in Appendix I.

5 Conclusion and Limitations

This paper first explores few-shot 3D shape adaptation. We introduce a novel domain adaptation approach to produce 3D shapes with diverse topology and textures using limited 2D data. The relative distances between generated samples are maintained at both feature-level and shape-level. We only need the silhouettes of few-shot target samples as training data to learn target geometry distributions while keeping diversity. Our approach is implemented based on GET3D to demonstrate its effectiveness. However, it is not constrained by specific network architectures and can be combined with more powerful 3D shape generative models to produce higher-quality results in the future. Despite the compelling results of our approach, it still has some limitations. For example, it is mainly designed for relatively related source/target domains. Extending our approach to unrelated domain adaptation would be promising. Nevertheless, we believe this work takes a further step towards democratizing 3D content creation by transferring knowledge in available source models to fit target distributions using few-shot data.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8305, 2020.

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pp. 40–49. PMLR, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

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*, pp. 5799–5809, 2021.

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*, pp. 16123–16133, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. *Computer Graphics Forum*, 22(3):223–232, 2003.

Qimin Chen, Johannes Merz, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, and Hao Zhang. Unist: unpaired neural implicit shape translation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18614–18622, 2022.

Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, pp. 864–872. PMLR, 2018.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.

Gene Chou, Yuval Bahat, and Felix Heide. Diffusionsdf: Conditional generative modeling of signed distance functions. *arXiv preprint arXiv:2211.13757*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *International Conference on 3D Vision*, pp. 402–411. IEEE, 2017.

Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3d reconstruction. *Advances In Neural Information Processing Systems*, 33:9936–9947, 2020a.

Jun Gao, Zian Wang, Jinchen Xuan, and Sanja Fidler. Beyond fixed grid: Learning geometric image representation with a deformable grid. In *Proceedings of the European Conference on Computer Vision*, pp. 108–125. Springer, 2020b.

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. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022.

Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023.

Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14072–14082, 2021.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9993, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. *arXiv preprint arXiv:2308.14244*, 2023a.

Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18423–18433, 2023b.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020b.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863, 2021.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34:21696–21707, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *arXiv preprint arXiv:2203.09301*, 2022.

Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *Advances in Neural Information Processing Systems*, 2020.

Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *International Conference on Learning Representations*, 2023.

Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020.

Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16238–16248, 2021.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: A simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020.

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8280–8290, 2022.

Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International Conference on Machine Learning*, pp. 7220–7229. PMLR, 2020.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.

Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2750–2758, 2019.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2021.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13503–13513, 2022.

Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. *Advances in Neural Information Processing Systems*, 33:870–882, 2020.

Dario Pavllo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13879–13889, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021b.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286. PMLR, 2014.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 2022.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20875–20886, 2023.

Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*, pp. 87–96. PMLR, 2017.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 33:12438–12448, 2020.

Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021.

Turbosquid. https://www.turbosquid.com/. Accessed: 2022-05-19.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*, 29, 2016.

Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: Generating images from limited data. In *Proceedings of the European Conference on Computer Vision*, pp. 218–234, 2018.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9332–9341, 2020.

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11204–11213, 2022.

Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9287–9297, 2023.

Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18430–18439, 2022.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4541–4550, 2019.

Kangxue Yin, Zhiqin Chen, Hui Huang, Daniel Cohen-Or, and Hao Zhang. Logan: Unpaired shape transform in latent overcomplete space. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yabo Zhang, Mingshuai Yao, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Towards diverse and faithful one-shot adaption of generative adversarial networks. *Advances in Neural Information Processing Systems*, 2022a.

Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. Generalized one-shot domain adaption of generative adversarial networks. *Advances in Neural Information Processing Systems*, 2022b.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020a.

Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. *Advances in Neural Information Processing Systems*, 2022a.

Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9140–9150, 2022b.

Yunqing Zhao, Chao Du, Milad Abdollahzadeh, Tianyu Pang, Min Lin, Shuicheng YAN, and Ngai-Man Cheung. Exploring incompatible knowledge transfer in few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020b.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5826–5835, 2021a.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021b.

Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Few-shot image generation with diffusion models. *arXiv preprint arXiv:2211.03264*, 2022a.

Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Few-shot image generation via masked discrimination. *arXiv preprint arXiv:2210.15194*, 2022b.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations*, 2021.