DATID-3D: Diversity-Preserved Domain Adaptation Using Text-to-Image Diffusion for 3D Generative Model

Gwanghyun Kim[1] Se Young Chun[1,2,†]
[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{gwang.kim, sychun}@snu.ac.kr

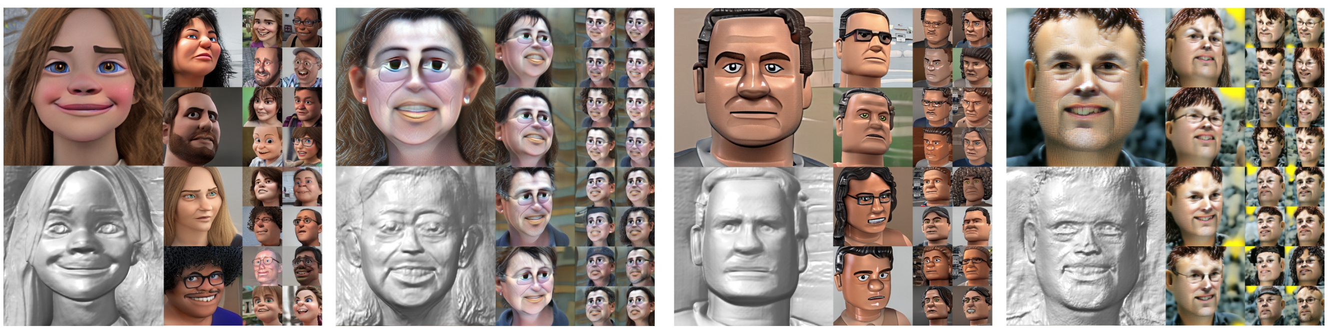

Figure 1. Our DATID-3D succeeded in domain adaptation of 3D-aware generative models without additional data for the target domain while preserving diversity that is inherent in the text prompt as well as enabling high-quality pose-controlled image synthesis with excellent text-image correspondence. However, StyleGAN-NADA*, a 3D extension of the state-of-the-art StyleGAN-NADA for 2D generative models [16], yielded alike images in style with poor text-image correspondence. See the supplementary videos at gwang-kim.github.io/datid_3d.

Abstract

Recent 3D generative models have achieved remarkable performance in synthesizing high resolution photorealistic images with view consistency and detailed 3D shapes, but training them for diverse domains is challenging since it requires massive training images and their camera distribution information. Text-guided domain adaptation methods have shown impressive performance on converting the 2D generative model on one domain into the models on other domains with different styles by leveraging the CLIP (Contrastive Language-Image Pre-training), rather than collecting massive datasets for those domains. However, one drawback of them is that the sample diversity in the original generative model is not well-preserved in the domain-adapted generative models due to the deterministic nature of the CLIP text encoder. Text-guided domain adaptation will be even more challenging for 3D generative models not only because of catastrophic diversity loss, but also because of inferior text-image correspondence and poor image quality. Here we propose DATID-3D, a domain adaptation method tailored for 3D generative models using text-to-image diffusion models that can synthesize diverse images per text prompt without collecting additional images and camera information for the target domain. Unlike 3D extensions of prior text-guided domain adaptation methods, our novel pipeline was able to fine-tune the state-of-the-art 3D generator of the source domain to synthesize high resolution, multi-view consistent images in text-guided targeted domains without additional data, outperforming the existing text-guided domain adaptation methods in diversity and text-image correspondence. Furthermore, we propose and demonstrate diverse 3D image manipulations such as one-shot instance-selected adaptation and single-view manipulated 3D reconstruction to fully enjoy diversity in text.

1. Introduction

Recently, 3D generative models [5, 6, 13, 18, 19, 22, 31, 40–42, 59, 60, 65, 69, 74, 75] have been developed to extend 2D generative models for multi-view consistent and explicitly pose-controlled image synthesis. Especially, some of them [5, 18, 74] combined 2D CNN generators like StyleGAN2 [28] with 3D inductive bias from the neural ren-

†Corresponding author.

dering [38], enabling efficient synthesis of high-resolution photorealistic images with remarkable view consistency and detailed 3D shapes. These 3D generative models can be trained with single-view images and then can sample infinite 3D images in real-time, while 3D scene representation as neural implicit fields using NeRF [38] and its variants [3, 4, 8, 10, 14, 17, 20, 32–34, 36, 45, 47, 50, 53, 54, 64, 66, 70–73] require multi-view images and training for each scene.

Training these state-of-the-art 3D generative models is challenging because it requires not only a large set of images but also the information on the camera pose distribution of those images. This requirement, unfortunately, has restricted these 3D models to the handful domains where camera parameters are annotated (ShapeNetCar [7, 61]) or off-the-shelf pose extractors are available (FFHQ [27], AFHQ [9, 26]). StyleNeRF [18] assumed the camera pose distribution as either Gaussian or uniform, but this assumption is valid only for a few pre-processed datasets. Transfer learning methods for 2D generative models [30, 39, 43, 44, 48, 55, 67, 68] with small dataset can widen the scope of 3D models potentially for multiple domains, but are also limited to a handful of domains with similar camera pose distribution as the source domain in practice.

Text-guided domain adaptation methods [1, 16] have been developed for 2D generative models as a promising approach to bypass the additional data curation issue for the target domain. Leveraging the CLIP (Contrastive Language-Image Pre-training) models [51] pre-trained on a large number of image-text pairs with non-adversarial fine-tuning strategies, these methods perform text-driven domain adaptation. However, one drawback of them is the catastrophic loss of diversity inherent in a text prompt due to the deterministic embedding of the CLIP text encoder so that the sample diversity of the source domain 2D generative model is not preserved in the target domain 2D generative models.

We confirmed this diversity loss with experiments. A text prompt "a photo of a 3D render of a face in Pixar style" should include lots of different characters' styles in Pixar films such as Toy Story, Incredible, etc. However, CLIP-guided adapted generator can only synthesize samples with alike styles as illustrated in Figure 1 (see StyleGAN-NADA*). Thus, we confirmed that naive extensions of these for 3D generative models show inferior text-image correspondence and poor quality of generated images in diversity. Optimizing with one text embedding yielded almost similar results even with different training seeds as shown in Figure 2(a). Paraphrasing the text for obtaining different CLIP embeddings was also trained, but it also did not yield that many different results as illustrated in Figure 2(b). Using different CLIP encoders for a single text as in Figure 2(c) did provide different samples, but it was not an option in general since only a few CLIP encoders have been released, and retraining them requires massive servers in practice.

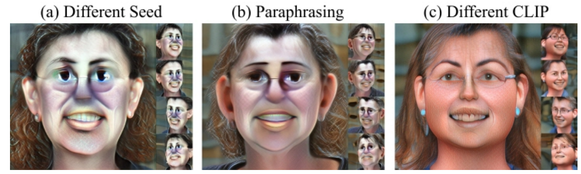

Figure 2. Existing text-guided domain adaptation [1, 16] did not preserve the diversity in the source domain for the target domain.

We propose a novel DATID-3D, a method of Domain Adaptation using Text-to-Image Diffusion tailored for 3D-aware Generative Models. Recent progress in text-to-image diffusion models enables to synthesize diverse high-quality images from one text prompt [52, 56, 58]. We first leverage them to convert the samples from the pre-trained 3D generator into diverse pose-aware target images. Then, the target images are rectified through our novel CLIP and pose reconstruction-based filtering process. Using these filtered target images, 3D domain adaptation is performed while preserving diversity in the text as well as multi-view consistency. We apply our novel pipeline to the EG3D [5], a state-of-the-art 3D generator, enabling the synthesis of high-resolution multi-view consistent images in text-guided target domains as illustrated in Figure 1, without collecting additional images with camera information for the target domains. Our results demonstrate superior quality, diversity, and high text-image correspondence in qualitative comparison, KID, and human evaluation compared to those of existing 2D text-guided domain adaptation methods for the 3D generative models. Furthermore, we propose one-shot instance-selected adaptation and single-view manipulated 3D reconstruction to fully enjoy diversity in the text by extending useful 2D applications of generative models.

2. Related Works

2.1. 3D generative models

Recent advances in 3D generative models have achieved multi-view consistent and explicitly pose-controlled image synthesis. Mesh-based [31, 65], voxel-based [13, 22, 40, 41, 69, 75], block-based and fully implicit representation-based [6, 40, 42, 59] 3D generative models have been proposed, but suffer from low image quality, view inconsistency, and inefficiency. Recently, efficient models [5, 18, 74] have been developed to combine the state-of-the art 2D CNN generator (*e.g.* StyleGAN2 [28]) with neural rendering [38]. Especially, EG3D utilizes tri-plane hybrid representation and poses conditioned dual discrimination to generate images with the state-of-the-art quality, view-consistency and 3D shapes in real-time. Such 3D generative models can be trained using single-view images and then can sample infinite 3D images in real-time whereas 3D scene representation as neural implicit fields using Neural Radiance Field (NeRF) [38] and its variants [3, 4, 8, 10, 14, 17, 20, 32–34, 36, 45, 47, 50, 53, 54, 64, 66, 70–73] requires multi-view images and training time for each scene.

Training recent 3D generative models is more difficult than training 2D generative models since it requires not only a large number of images but also the information on the camera parameter distribution of those images. To broadly leverage the state-of-the-art 3D generative models to cover wider domains, we propose a method of text-guided domain adaptation without additional images for the target domain and construct our novel pipeline so that the EG3D, a state-of-the-art 3D generator, can be fine-tuned to perform the synthesis of high-resolution multi-view consistent images in text-guided targeted domains.

2.2. Text-guided domain adaptation using CLIP

CLIP [51] is composed of the image encoder E_I^C and the text encoder E_T^C that translate their inputs into vectors in a shared multi-modal CLIP space. StyleGAN-NADA [16] fine-tunes a pre-trained StyleGAN2 G^θ [28] to shift the domain towards a target domain using a simple textual prompt guided by directional CLIP loss as follows:

$$\mathcal{L}^{\theta}_{\text{direction}}\left(\boldsymbol{x}^{\text{gen}}, y^{\text{tar}}; \boldsymbol{x}^{\text{src}}, y^{\text{src}}\right) := 1 - \frac{\langle \Delta I, \Delta T\rangle}{\|\Delta I\| \|\Delta T\|}, \quad (1)$$

where $\Delta I = E_I^C(\boldsymbol{x}^{\text{gen}}) - E_I^C(\boldsymbol{x}^{\text{src}}), \Delta T = E_T^C(y^{\text{tar}}) - E_T^C(y^{\text{src}})$. Here, the CLIP space direction between the source and target images ΔI and the direction between the source and target text descriptions ΔT are encouraged to align. HyperDomainNet [1] additionally proposes a domain-modulation technique to reduce the number of trainable parameters and the in-domain angle consistency loss to avoid mode collapse.

A critical drawback of these methods are that diversity inherent in a text prompt is catastrophically lost, resulting in alike samples to represent only one instance per text prompt due to the deterministic embedding of the CLIP encoder $E_T^C(\boldsymbol{x})$. Moreover, naive extensions of these methods to 3D models exhibit inferior text-image correspondence and poor image quality. Our proposed DATID-3D aims to achieve superior quality, diversity, and high text-image correspondence to existing 2D text-guided domain adaptation methods for 3D generative models qualitatively and quantitatively.

2.3. Text-guided diffusion models

Diffusion models have achieved great success in image generation [11, 23, 24, 62, 63]. Recently, these models have been extended to image-text multi-modal settings, showing promising results [2, 29, 52, 56, 58]. Especially, text-to-image diffusion models trained on billions of image-text pairs [52, 56, 58] enables to synthesize outstanding quality of diverse 2D images with one target text prompt through the stochastic generation process. Furthermore, recent progress [15, 57] enables the synthesis of text-guided novel scenes for a given subject using only a few images.

In the meanwhile, the text-guided diffusion models for 3D scenarios are underexplored. A concurrent work, DreamFusion [49], performs text-to-3D synthesis by optimizing a randomly-initialized NeRF guided by Imagen, a huge text-to-image diffusion model, that is not publicly available. Their results are impressive but tend to be blurry and lack fine details. Moreover, it requires a long time to generate one scene due to the an inherent drawback of NeRF-like methods. Also, if Stable diffusion [56], a publicly available lightweight text-to-image diffusion model, is used, DreamFusion generation, but often failed to reconstruct 3D scenes. In contrast, our proposed method enables the real-time synthesis of diverse high-resolution samples once trained, even with relatively small-scale, efficient diffusion models.

3. DATID-3D

We aim to transfer EG3D [5], a state-of-the-art 3D generator G_θ trained on a source domain, to a new target domain specified by a text prompt y while preserving multi-view consistency and diversity in text. We employ a pre-trained text-to-image diffusion model $\boldsymbol{\epsilon}_\phi$ as a source of supervision, but no additional image for the target domain is used. Firstly, we use our novel pipeline to construct a target dataset $\mathcal{D}(G_\theta, \boldsymbol{\epsilon}_\phi, y) = \{(\boldsymbol{c}_i, \boldsymbol{x}_i^{\text{src}}, \boldsymbol{x}_i^{\text{trg}})\}_{i=1}^N$ that consists of random latent vectors, camera parameters and corresponding target images in a text-guided domain as illustrated in Figure 3(a). Secondly, we refine the dataset to obtain $\mathcal{D}_f$ through our CLIP and pose reconstruction-based filtering process for improved image-text correspondence and pose-consistency, respectively, as shown in Figure 3(b). Lastly, with the rectified dataset, we fine-tune the generator G_θ with adversarial and density-regularization losses to preserve diversity and multi-view consistency as presented in Figure 3(c). In addition, we propose a one-shot instance-selected adaptation to let users fully enjoy diversity in the text as illustrated in Figure 4.

3.1. Text-guided target dataset generation

Here, we generate a source image $\boldsymbol{x}^{\text{src}} = G_\theta(\boldsymbol{z}, \boldsymbol{c})$ with random latent vector $\boldsymbol{z} \in \mathcal{Z}$ and camera parameter $\boldsymbol{c} \in \mathcal{C}$. Then, we manipulate the source image $\boldsymbol{x}^{\text{src}}$ to yield the target image $\boldsymbol{x}^{\text{trg}}$ guided by a text prompt $\boldsymbol{y}$ using the ideas in [37] using a text-to-image diffusion model $\boldsymbol{\epsilon}_\phi$, constructing a set of $(\boldsymbol{c}, \boldsymbol{x}^{\text{src}}, \boldsymbol{x}^{\text{trg}})$. Stable diffusion [56] is selected and employed in this work since a latent-based model is lightweight and publicly available while others [52, 58] are not. $\boldsymbol{x}^{\text{src}}$ is encoded to the latent representation $\boldsymbol{q}_0 = E^V(\boldsymbol{x}^{\text{src}})$ using VQGAN [12] encoder E^V. Then, the latent is perturbed through the stochastic forward DDPM (Denoising Diffusion Probabilistic Models) process [23] with noise schedule parameters $\{\bar{\alpha}_t\}_{t=1}^T$ until the return step $t_r < T_p$, where T is the number of total diffusion steps used in training and T_p is the pose-consistency step which is the last diffusion step

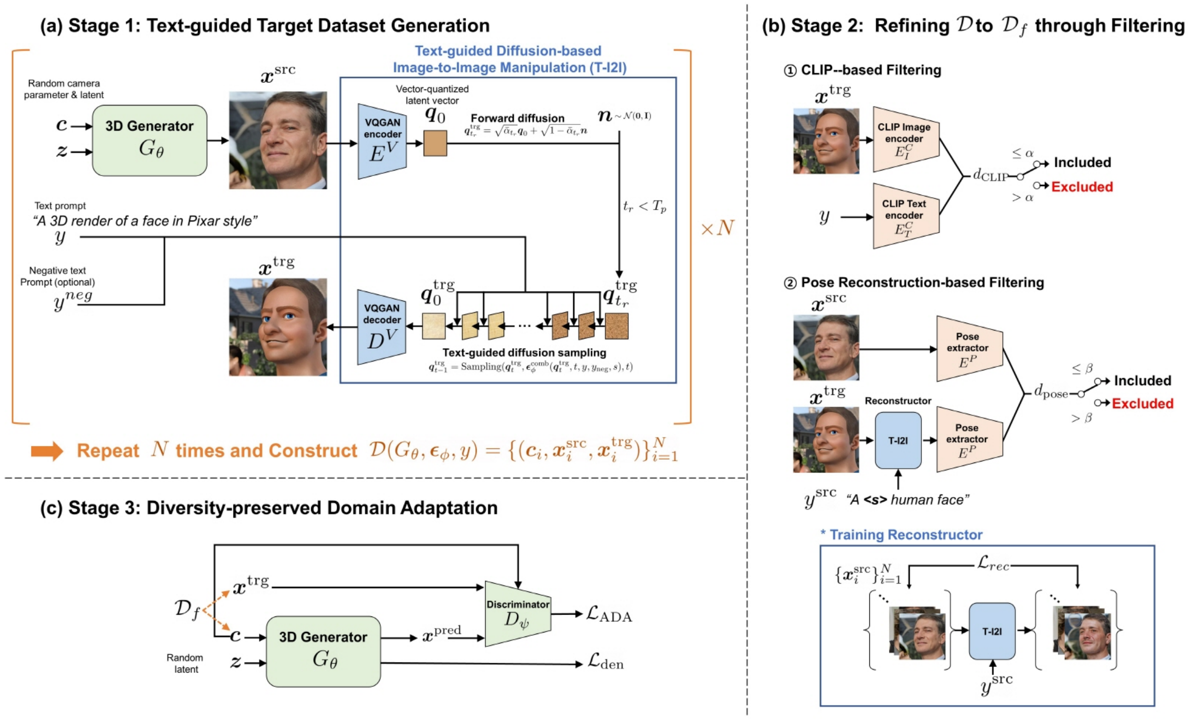

Figure 3. Overview of DATID-3D. We construct target dataset using the pre-trained text-to-image diffusion models, followed by refining the dataset through filtering process. Finally, we fine-tune our models using adversarial loss and density regularization.

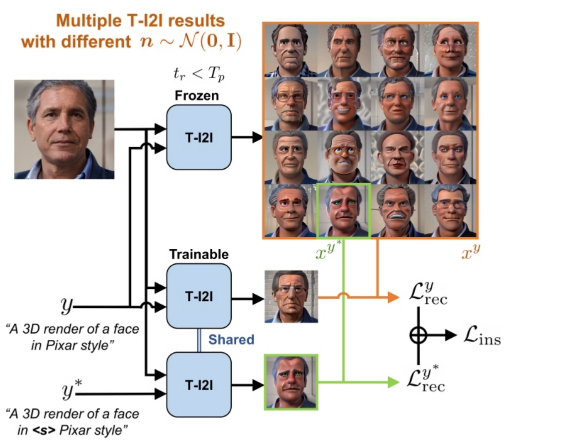

Figure 4. One-shot fine-tuning of text-to-image diffusion models for instance-selected domain adaptation. Resulting text-to-image diffusion models are applied to the Stage 1 in Figure 3(a).

when the pose information is preserved as follows:

$$\boldsymbol{q}_{t_r}^{\text{trg}} = \sqrt{\bar{\alpha}_{t_r}}\boldsymbol{q}_0 + \sqrt{1 - \bar{\alpha}_{t_r}}\boldsymbol{n}, \text{ where } \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (2)$$

We set $T = 1000$ following [11, 23, 62] and $T_p = 850$ based on the experimental results. Then, we generate the manipulated target latent $\boldsymbol{q}_0^{trg}$ from the perturbed latent $\boldsymbol{q}_{t_r}^{\text{trg}}$ through text-guided sampling process as follows:

$$\boldsymbol{q}_{t-1}^{\text{trg}} = \text{Sampling}(\boldsymbol{q}_t^{\text{trg}}, \boldsymbol{\epsilon}_\phi^{\text{comb}}(\boldsymbol{q}_t^{\text{trg}}, t, y, y_{\text{neg}}, s), t) \quad (3)$$

where s is a guidance scale parameter that controls the scale of gradients from a target prompt y and a negative prompt y_{neg} and $\boldsymbol{\epsilon}_\phi^{\text{comb}} = s\boldsymbol{\epsilon}_\phi(\boldsymbol{q}_t^{\text{trg}}, t, y) + (1 - s)\boldsymbol{\epsilon}_\phi(\boldsymbol{q}_t^{\text{trg}}, t, y_{\text{neg}})$. Any sampling method such as DDPM [23], DDIM [62], PLMS [35] can be used. We can specify y_{neg} to prevent the manipulated image from being contaminated by an unwanted concept or can just leave it as an unconditional text token. Then, we can obtain the target image $\boldsymbol{x}^{\text{trg}} = D^V(\boldsymbol{q}_0^{\text{trg}})$ using the VQGAN decoder that represents one of the diverse concepts inherent in the text prompt. By repeating this process N times, we can get a target dataset $\mathcal{D}(G_\theta, \boldsymbol{\epsilon}_\phi, y) = \{(\boldsymbol{c}_i, \boldsymbol{x}_i^{\text{src}}, \boldsymbol{x}_i^{\text{trg}})\}_{i=1}^N$.

3.2. CLIP and pose reconstruction-based filtering

We found that the raw target dataset $\mathcal{D}$ may sometimes include target images that may not correspond to the target text or that the camera pose is changed during the stochastic process. To resolve this issue, we propose CLIP-based and pose reconstruction-based filtering processes for enhanced image-text correspondence and pose consistency.

CLIP-based filtering. A CLIP distance score d_{CLIP} is the cosine distance in the CLIP space between a potential target image $\boldsymbol{x}^{\text{trg}}$ and a text prompt y and if

$$d_{\text{CLIP}}(\boldsymbol{x}^{\text{trg}}, y) = 1 - \frac{\langle E_I^C(\boldsymbol{x}^{\text{trg}}), E_T^C(y)\rangle}{\|E_I^C(\boldsymbol{x}^{\text{trg}})\|\|E_T^C(y)\|} > \alpha \quad (4)$$

where α is a chosen threshold, then $\boldsymbol{x}^{\text{trg}}$ is removed from $\mathcal{D}$.

Pose reconstruction-based filtering. A pose difference score d_{pose} is the l_2 distance between the poses extracted from $\boldsymbol{x}$ and $\boldsymbol{x}'$ using an off-the-shelf pose extractor E^{P}:

$$d_{\text{pose}}(\boldsymbol{x}, \boldsymbol{x}') = \|E^{\text{P}}(\boldsymbol{x}) - E^{\text{P}}(\boldsymbol{x}')\|_2^2. \tag{5}$$

To calculate pose difference, we leverage the universal Reconstructor that converts the target images from any shifted domain to the source domain (*e.g.*, human face) where the pose extractor is available. We fine-tuned the pre-trained text-to-image diffusion models to generate the source domain images $\{\boldsymbol{x}_i^{\text{src}*}\}_{i=1}^N$ using the following diffusion-based reconstruction loss $\mathcal{L}_{\text{rec}}^{\phi}$:

$$\mathbb{E}_{\boldsymbol{q}\in\{E^V(\boldsymbol{x}_i^{\text{src}})\}_{i=1}^N, \boldsymbol{\epsilon}\sim\mathcal{N}(0,1), t}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t, t, y^{\text{src}}))\|_2^2] \tag{6}$$

where y^{src} is a text prompt representing the source domain with a specifier word <s> (*e.g.*, "A photo of <s> face" in FFHQ [27]). The Reconstructor can be re-used for any target domain if the source domain is the same. Using the Reconstructor, we first convert the target image $\boldsymbol{x}^{\text{trg}}$ to a reconstructed image $\boldsymbol{x}^{\text{rec}}$. Then, if $d_{\text{pose}}(\boldsymbol{x}^{\text{rec}}, \boldsymbol{x}^{\text{src}}) > \beta$ (β is a threshold), then $\boldsymbol{x}^{\text{trg}}$ is excluded from $\mathcal{D}$ and another target image with same $(\boldsymbol{z}, \boldsymbol{c})$ is generated to supplement $\mathcal{D}$.

3.3. Diversity-preserved domain adaptation

With the filtered dataset $\mathcal{D}_f = \{(\boldsymbol{c}_i, \boldsymbol{x}_i^{\text{src}}, \boldsymbol{x}_i^{\text{trg}})\}_{i=1}^N$, we can perform either non-adversarial fine-tuning using CLIP-based loss like StyleGAN-NADA [51] and HyperDomainNet [1] or adversarial fine-tuning like StyleGAN-ADA [25]. As analyzed Section 4.5, we found that non-adversarial fine-tuning makes the generator lose diversity for the target text prompt and generates the samples representing one averaged concept among diverse concepts as well as showing sub-optimal quality because the cosine similarity loss can be saturated near the optimal point. In contrast, we found that adversarial fine-tuning preserves diverse concepts in the text. Our total loss to train pose-conditioned generator G_θ and discriminator D_ψ consists of ADA loss $\mathcal{L}_{\text{ADA}}$ and density regularization loss $\mathcal{L}_{\text{den}}$, which were used in EG3D [5].

ADA loss. ADA loss $\mathcal{L}_{\text{ADA}}$ [25] is an adversarial loss with adaptive dual augmentation and R1 regularization as follows:

$$\begin{aligned} \mathcal{L}_{\text{ADA}}^{\theta,\psi} &= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}, \boldsymbol{c}\sim\mathcal{C}}[f(D_\psi(A(G_\theta(\boldsymbol{z}, \boldsymbol{c})), \boldsymbol{c})] \\ &+ \mathbb{E}_{(\boldsymbol{c}, \boldsymbol{x}^{\text{trg}})\in\mathcal{D}_f}[f(-D_\psi(A(\boldsymbol{x}^{\text{trg}}), \boldsymbol{c}) + \lambda\|\nabla D_\psi(A(\boldsymbol{x}^{\text{trg}}), \boldsymbol{c})\|^2)] \end{aligned} \tag{7}$$

where $f(u) = -\log(1 + \exp(-u))$, A is a stochastic non-leaking augmentation operator with probability p. For more detailed information and algorithms of our pipelines, see the supplementary Section B.

Density regularization loss. Additionally, we use density regularization, which has been effective in reducing the occurrence of unwanted other shape distortions by promoting the smoothness of the density field [5]. We randomly select points v from the volume $\mathcal{V}$ for each rendered scene and also select additional perturbed points that have been slightly distorted by Gaussian noise δv. Then, we calculate the L1 loss between the predicted densities as follows:

$$\mathcal{L}_{\text{den}}^{\theta} = \mathbb{E}_{v\in\mathcal{V}}[\|\sigma_\theta(v) - \sigma_\theta(v + \delta v)\|]. \tag{8}$$

3.4. Instance-selected domain adaptation

Our method yields a domain-shifted 3D generator to synthesize samples to represent diverse concepts in the text. However, consider a scenario to pick up one of those diverse concepts and adapt the generator to this specific concept. Is it possible to manipulate our single 2D image guided by text and lift it to 3D with multiple versions?

To enable these applications and help users fully enjoy diversity in text, we propose a one-shot instance-selected adaptation. We first manipulate a source domain image in N_d multiple versions from a single text prompt as shown in Figure 4. Then, we choose one instance among those diverse instances and fine-tune our text-to-image diffusion models. However, unlike the prior work [57] to personalize the object and use it guided by the text, our goal is to inject a specific concept among many concepts, that are implicit in one text prompt, to the text-to-image diffusion model. Also, we fine-tune the text-to-image diffusion model only with a single image by limiting the diffusion time step from 0 to the pose-consistency step T_p using the following loss $\mathcal{L}_{\text{ins}}^{\phi}$:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,1), t\in[0,T_p]}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(E^V(\boldsymbol{x}^{y^*}), t, y^*))\|_2^2] \\ &+ \mathbb{E}_{\boldsymbol{z}\in\{E^V(\boldsymbol{x}_i^y)\}_{i=1}^{N_d}, \boldsymbol{\epsilon}\sim\mathcal{N}(0,1), t\in[0,T_p]}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\boldsymbol{z}_t, t, y))\|_2^2] \end{aligned} \tag{9}$$

where y and $\boldsymbol{x}^y$ are the target text prompt and manipulated image with the prompt, respectively. y^* and $\boldsymbol{x}^{y^*}$ are the target text prompt with a specifier word <s> and the selected instance image, respectively. The next step is just replacing the original text-to-diffusion model with our specified model in text-guided target generation stage in Figure 3(a).

Additionally, we can now perform single-view manipulated 3D reconstruction representing our chosen concept by combining the 3D GAN inversion method with instance-selected domain adaptation.

Table 1. Quantitative comparisons with the baselines in diversity, text-image correspondence and photo realism.

	Text-Corr.↑	Realism↑	Diversity↑
StyleGAN-NADA*	2.583	2.550	2.587
HyperDomainNet*	2.530	2.520	2.557
Ours	**3.573**	**3.437**	**3.347**

4. Experiments

For the experiments, we employ the publicly available lightweight Stable diffusion [56] as our pre-trained text-to-image diffusion model. We apply our novel pipeline to the

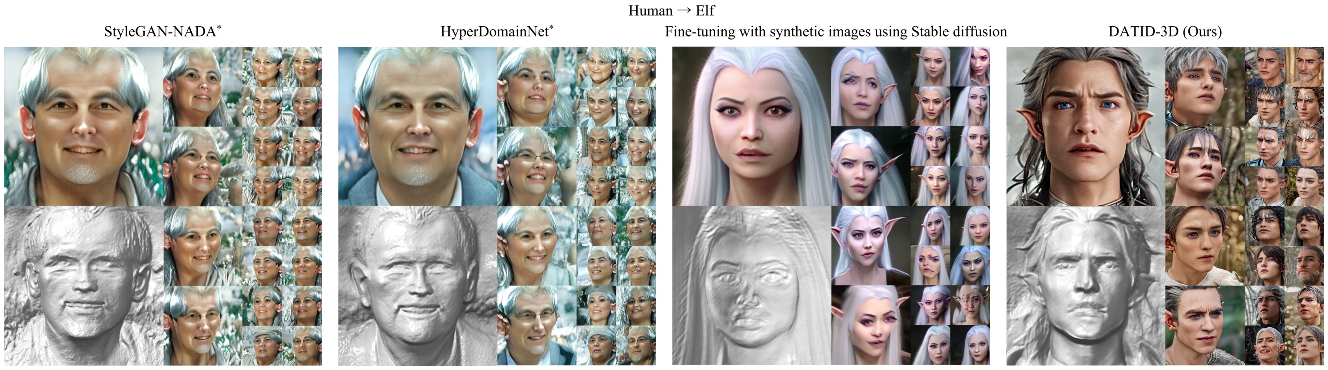

Figure 5. Qualitative comparison with existing text-guided domain adaptation methods. The star mark (*) refers to the 3D extension of each method that is developed for 2D generative models. Our DATID-3D yielded diverse samples while other baselines did not. Naively fine-tuning 3D generators with the synthetic images using T2I diffusion resulted in losing pose-controllability and 3D shapes. For more results, see the supplementary Figure S4.

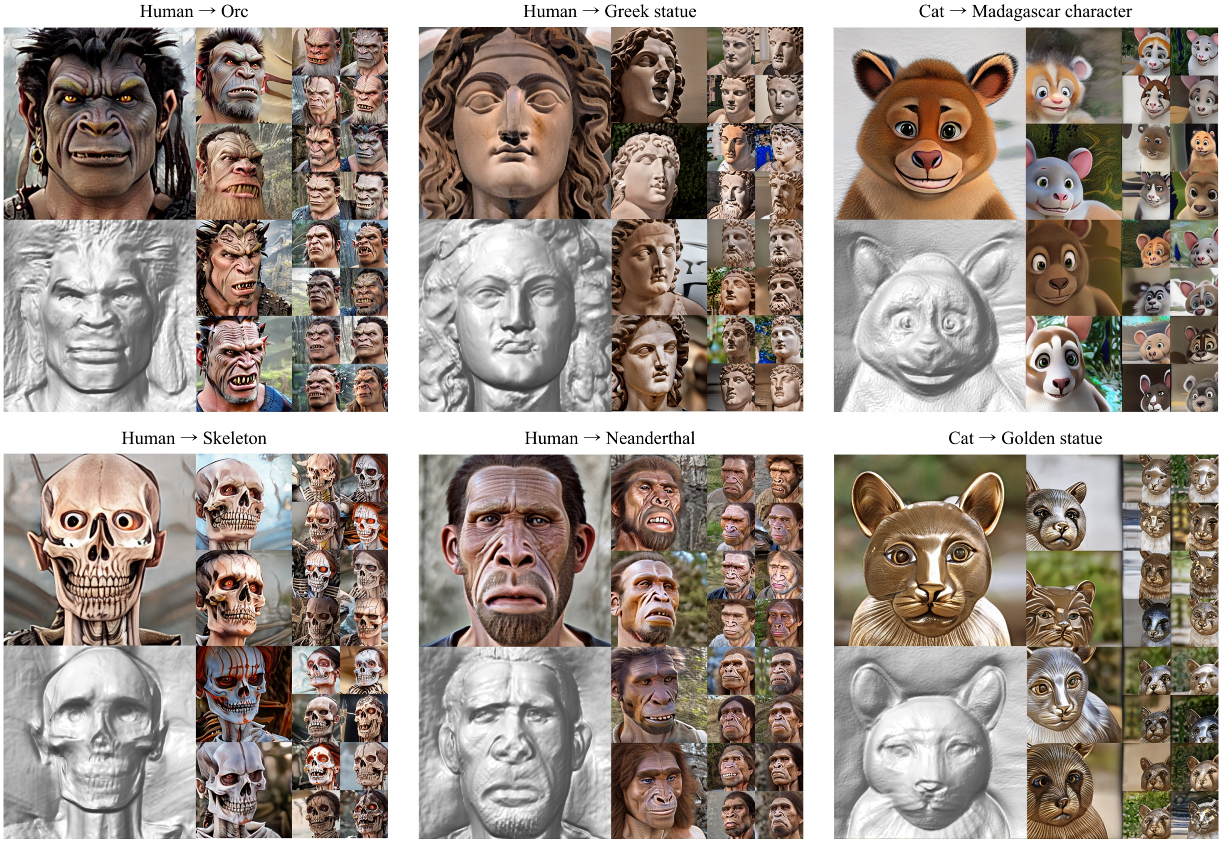

Figure 6. Wide range of out-of-domain text-guided adaption results. We fine-tuned EG3D [5], pre-trained on 512^2 images in FFHQ [27] and AFHQ Cats [9, 26], respectively, to generate diverse samples for a wide range of concepts. For more results, see the supplementary Figure S1, S2 and S3.

state-of-the-art 3D generators, EG3D [5], pre-trained on 512^2 images in FFHQ [27] and AFHQ Cats [9, 26], respectively. The pre-trained pose-extractor [21] was used. For fine-tuning the generator, 3,000 target images were used. We set $\alpha = 0.7$ and $\beta = 150$. For more detailed information about the setup of experiments, see the supplementary Section C and D.

4.1. Evaluation

Baselines. To the best of our knowledge, our method is the first method of text-guided domain adaptation tailored for 3D-aware generative models. Thus, we compare our method with CLIP-based method for 2D generative models, StyleGAN-NADA [51] and HyperDomainNet [1]. The star mark (*) refers to the extension of these method to 3D models. To achieve this, we just add random sampling of camera parameters, followed by the random latent sampling. In StyleGAN-NADA*, the directional CLIP loss is used to encourage the correspondence between the rendered 2D image with the text prompt. In addition to this, in-domain angle loss is used for HyperdomainNet*.

Qualitative results. As shown in Figures 1 and 5, the generator shifted by the baseline methods fail to generate high-quality samples, preserving diversity implicit in the text prompt. Even though in-domain angle loss in HyperDomainNet* [1] is proposed to improve sample diversity, it shows similar inferior results because the fundamental issue, the deterministic text embedding of the CLIP encoder, is not resolved. On the contrary, our DATID-3D enables the shifted generator to synthesize photorealistic and diverse images, leveraging text-to-image diffusion models and adversarial training. In addition, we present the results of naively fine-tuning the 3D generator with synthetic images generated from random noise using Stable diffusion [56]. However, this approach loses 3D shapes, depth, and pose-controllability, whereas our method, which uses a pose-aware target dataset, preserves 3D geometry effectively.

Quantitative results. We perform a user study to assess the perceptual quality of the produced samples. To quantify opinions, we requested users to rate the perceptual quality on a scale of 1 to 5, based on the following questions: (1) Do the generated samples accurately reflect the target text's semantics? (text-correspondence), (2) Are the samples realistic? (photorealism), (3) Are the samples diverse in the image group? (diversity). We use the EG3D pre-trained on 512^2 images in FFHQ [27] and choose four text prompts converting a human face to 'Pixar', 'Neanderthal', 'Elf' and 'Zombie' styles, respectively, for evaluation as these prompts are used in the previous work, StyleGAN-NADA [51]. As presented in Table 1, our results demonstrate the superior quality, high diversity, and high text-image correspondence of our proposed method as compared to the baselines. For more results and details on the comparison, see the supplementary Section A and D.

4.2. Results of 3D out-of-domain adaptation

We display a wide range of text-driven adaptation results through our methods in Figure 6, which are applied to the generators pre-trained on FFHQ [27] or AFHQ Cats [9, 26]. Our model enables the synthesis of high-resolution multi-view consistent images in various text-guided out-of-domains beyond the boundary of the trained domains, without additional images and camera information. For more results, see the supplementary Section A.

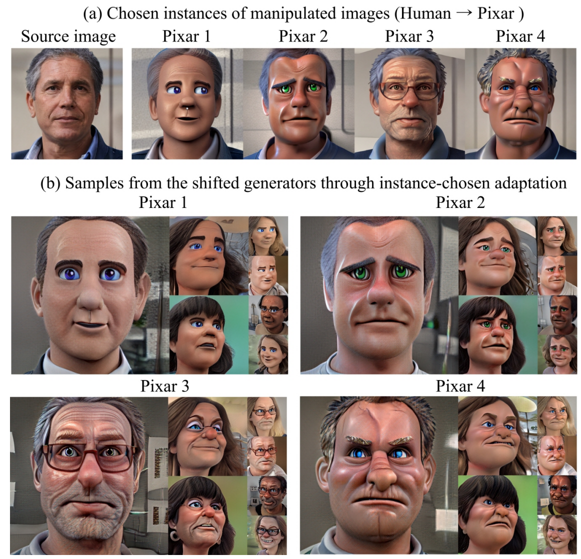

Figure 7. Results of instance-selected domain adaptation, selecting one Pixar sample to generate more diverse samples for it.

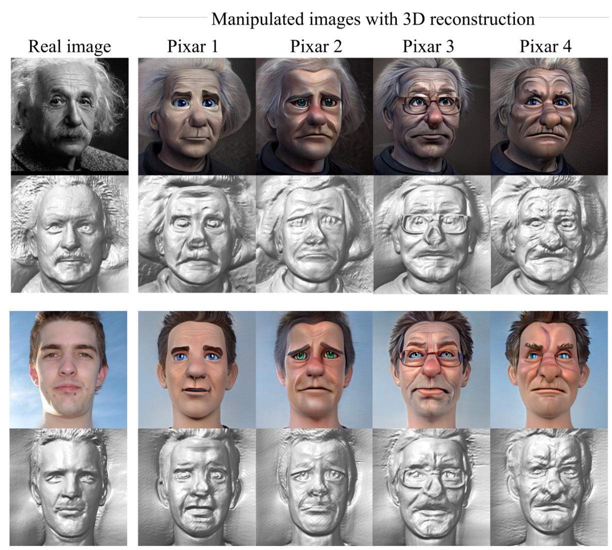

Figure 8. Results of single-view manipulated 3D reconstruction, generating diverse 3D images on other domains with view consistency for a given single real image.

4.3. Instance-selected domain adaptation

In Figure 7, we adapt our generator guided by text prompt, "a photo of 3D render of a face in Pixar style", in four different versions. Figure 7(a) presents the selected 4 cases. Figure 7(b) displays the images sampled from the generator adapted to each instance. We can further utilize these for single-view 3D manipulated reconstruction.

4.4. Single-view 3D manipulated reconstruction

As advancements of prior 2D text-guided image manipulation [16, 29, 46], our method enables (1) lifting the text-

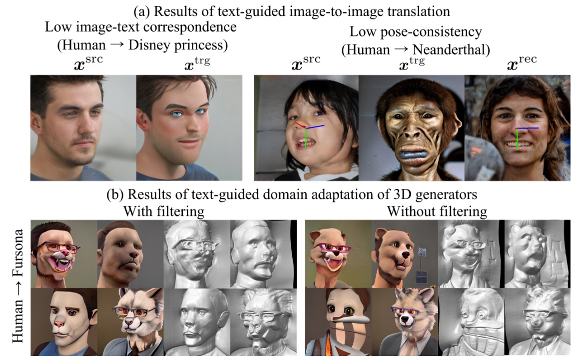

Figure 9. Discarded cases through our filtering process (a) and results of domain adaptation of 3D generative models with and without filtering (filtering rate $= 0.529$)

Figure 10. Diversity preservation can be ensured not by non-adversarial training loss, but by adversarial training loss.

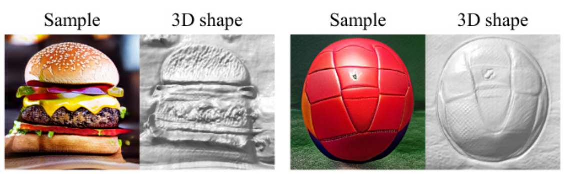

Figure 11. Rotation-invariant objects lose 3D shapes during fine-tuning due to the lack of information about directions.

guided manipulated images to 3D and (2) choosing one among diverse results from one text prompt. Figure 8 shows the results of single-view manipulated 3D reconstruction where instance-selected domain adaptation is combined with the 3D GAN inversion method [5].

4.5. Ablation studies

Effectiveness of filtering process. Frequent failure cases of image manipulation by text-to-image manipulation were observed as shown in Figure 9(a). Our filtering process improves perceptual quality and the quality of 3D shape, especially with filtering rate > 0.5 Fig. 9(b).

Adversarial vs non-adversarial fine-tuning. We compare the results of fine-tunings using the adversarial ADA loss with those using CLIP-based non-adversarial loss for the target images. For the non-adversarial loss, we employ the image directional CLIP loss that tries to align the direction between source and generated images with the direction between source and target images. As illustrated in Figure 10, we found that non-adversarial fine-tuning makes the generator lose diversity in the target text prompt and generates the samples representing one averaged concept among diverse concepts. Furthermore, it shows sub-optimal quality because the cosine similarity loss can be saturated near the optimal point. In contrast, we observed that adversarial fine-turning preserves diverse concepts in text with excellent quality. For more results of ablation studies, see the supplementary Section E.

5. Discussion and Conclusion

Limitation. We found that for the successful text-driven 3D domain adaptation, the important condition is that the target images generated in Stage 1 should preserve pose information. However, there are some unavoidable cases to meet this condition. One of the cases for pose information loss is that the target object is rotation-invariant or in 2D space. As shown in Figure 11, domain adaptations of 'Human face' → 'Cheeseburger' or 'Human face' → 'Soccer ball' failed because pose information is lost during the manipulation, reporting high pose-difference score.

Societal risks in our methods exist. We advise you to use our method carefully for proper purposes. Details on limitations and negative social impacts are given in the supplementary Section F.

Conclusion. We propose DATID-3D, a method of domain adaptation tailored for 3D generative models leveraging text-to-image diffusion models that can synthesize diverse images per text prompt. Our novel pipeline with the 3D generator has enabled excellent quality of multi-view consistent image synthesis in text-guided domains, preserving diversity in text and outperforming the baselines qualitatively and quantitatively. Our pipeline was able to be extended for one-shot instance-selected adaptation and single-view manipulated 3D reconstruction to meet user-intended constraints.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (NRF-2022R1A4A1030579, NRF-2022M3C1A309202211), Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B05035810) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

[1] Aibek Alanov, Vadim Titov, and Dmitry Vetrov. Hyperdomainnet: Universal domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2210.08884*, 2022. 2, 3, 5, 6, 7

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 3, 5, 6, 8

[6] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 1, 2

[7] 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. 2

[8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 6, 7

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 3, 4

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[13] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 1, 2

[14] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[16] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 1, 2, 3, 7

[17] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 2

[18] 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. 1, 2

[19] 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*, pages 14072–14082, 2021. 1

[20] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 2

[21] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. *arXiv preprint arXiv:2202.12555*, 2022. 6

[22] 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*, pages 9984–9993, 2019. 1, 2

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 3, 4

[24] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020. 3

[25] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 5

[26] 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. 2, 6, 7

[27] 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*, pages 4401–4410, 2019. 2, 5, 6, 7

[28] 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*, pages 8110–8119, 2020. 1, 2, 3

[29] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3, 7

[30] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 2

[31] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2020. 1, 2

[32] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[33] David B Lindell, Julien NP Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14556–14565, 2021. 2

[34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[35] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 4

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[37] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[38] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020. 2

[39] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 2

[40] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 1, 2

[41] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020. 1, 2

[42] 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*, pages 11453–11464, 2021. 1, 2

[43] 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*, pages 2750–2758, 2019. 2

[44] 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*, pages 10743–10752, 2021. 2

[45] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[46] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 7

[47] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2

[48] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 2

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

[50] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3, 5, 6, 7

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[53] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14161, 2021. 2

[54] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 2

[55] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943*, 2020. 2

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 5, 7

[57] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3, 5

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[59] 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. 1

[60] Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6258–6266, 2021. 1

[61] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[64] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[65] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 1, 2

[66] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 2

[67] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021.

[68] 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*, pages 9332–9341, 2020. 2

[69] 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. 1, 2

[70] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[71] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[72] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[73] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[74] 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*, 2021. 1, 2

[75] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018. 1, 2