StyleGAN-Fusion: Diffusion Guided Domain Adaptation of Image Generators

Kunpeng Song[1] Ligong Han[1] Bingchen Liu[2] Dimitris Metaxas[1] Ahmed Elgammal[1,3]
[1]Rutgers University [2]Bytedance Inc. [3]Playform AI

Abstract

Can a text-to-image diffusion model be used as a training objective for adapting a GAN generator to another domain? In this paper, we show that the classifier-free guidance can be leveraged as a critic and enable generators to distill knowledge from large-scale text-to-image diffusion models. Generators can be efficiently shifted into new domains indicated by text prompts without access to groundtruth samples from target domains. We demonstrate the effectiveness and controllability of our method through extensive experiments. Although not trained to minimize CLIP loss, our model achieves equally high CLIP scores and significantly lower FID than prior work on short prompts, and outperforms the baseline qualitatively and quantitatively on long and complicated prompts. To our best knowledge, the proposed method is the first attempt at incorporating large-scale pre-trained diffusion models and distillation sampling for text-driven image generator domain adaptation and gives a quality previously beyond possible. Moreover, we extend our work to 3D-aware style-based generators and DreamBooth guidance. For code and more visual samples, please visit our Project Webpage.

1. Introduction

Diffusion models have witnessed a remarkable rise in image-generation tasks with their ability to cover a wide range of visual semantics from different image domains [5, 11, 14, 16, 37, 45, 46, 48, 51, 58]. These models usually require lots of iterative generative steps, which is computationally demanding and makes it undesirable in many application scenarios. On the other hand, GANs [9, 10, 12, 13, 15, 21–23, 25] preserve their advantage over diffusion models in terms of generation speed and is more computational-friendly to train. Within a single image domain, GANs can have an accessible latent space with the expressive power to synthesize images with fine-grained variations. Leveraging a pre-trained text-to-image diffusion model, such as StableDiffusion [48], we propose a new training objective that can quickly shift a pre-trained GAN model into another image domain. We take advantage of StableDiffusion's prior knowledge learned from enormous text-and-image pairs and use such prior guiding the GAN model to shift its generation behavior. With the developed training objective, are able to shift the output of GAN to a totally different image domain without the need for any training images in that domain.

The availability of large-scale text-to-image models unleashes the potential of zero-shot domain shifting for GANs. Prior works, such as StyleGAN-NADA [8], take advantage of CLIP's [44] power to relate visual features to textual semantics. Via a quick training of minimizing CLIP's image-to-text similarity on a certain set of prompts, StyleGAN-NADA is able to generate cartoon avatars from a model trained only on realistic faces and oil paintings from a model initially trained only on photographs. However, these methods rely on CLIP, which has been shown to proceed with a misaligned text and image latent space [30]. CLIP loss is known hard to be minimized in previous work [33] as it tends to be trapped in a local minimum. This limits the effectiveness of the selected prompts to drive the GAN toward the desired image domain, leading to image artifacts and undermined decreased generation diversity.

In this work, we explore using diffusion to improve the performance of a text-driven image generator for domain adaptation. We leverage the power of pre-trained large-scale diffusion models and build on the recently proposed Score Distillation Sampling technique [42], where text-to-image diffusion acts as a frozen, efficient critic that predicts image-space edits. Our new domain adaptation method takes advantage of a pre-trained image diffusion model, providing well-aligned guidance directly from the image domain to help train the GAN model.

Intuitively, distillation of a generative model (eg. diffusion model) can provide more informative signals for a generator than a discriminative model (such as CLIP). In this paper, we investigate two techniques that combine diffusion with style-based generators to explore this idea further:

- We introduce the diffusion model score distillation sampling (SDS) into domain adaptation of style-based image generators and achieve better performance than prior art.
- To regularize the network and prevent model col-

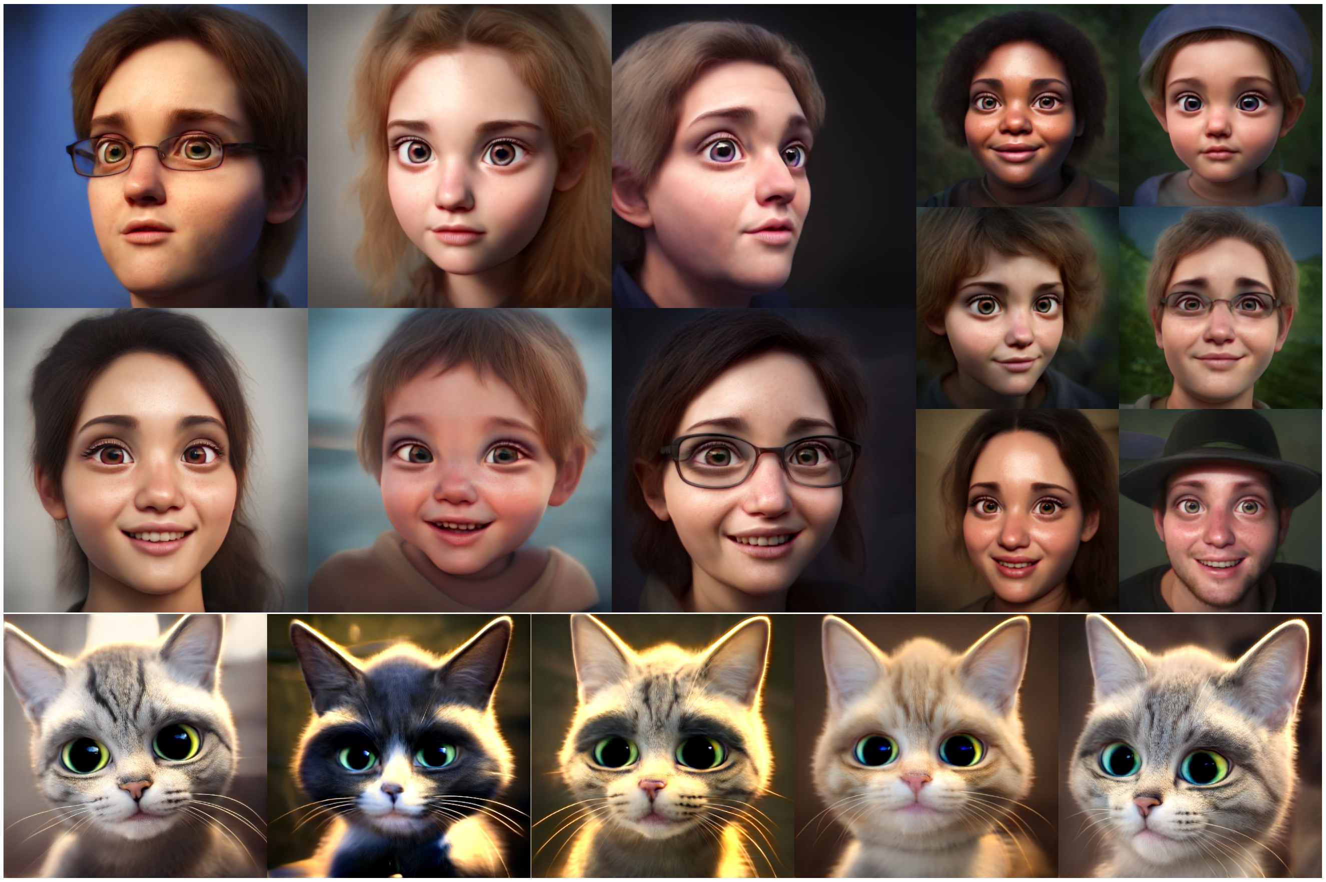


Figure 1. Example images after adapting generator to a domain specified by a text description. The first section is a photo from the FFHQ dataset to 3D stylized Anime, the second section is from cats to 3D rendering cats. Detailed text prompts can be founded in the appendix.

lapse, we propose a diffusion directional regularizer and adapt the reconstruction guidance to SDS. To solve blurry issues, we adapt the layer selection into our framework.

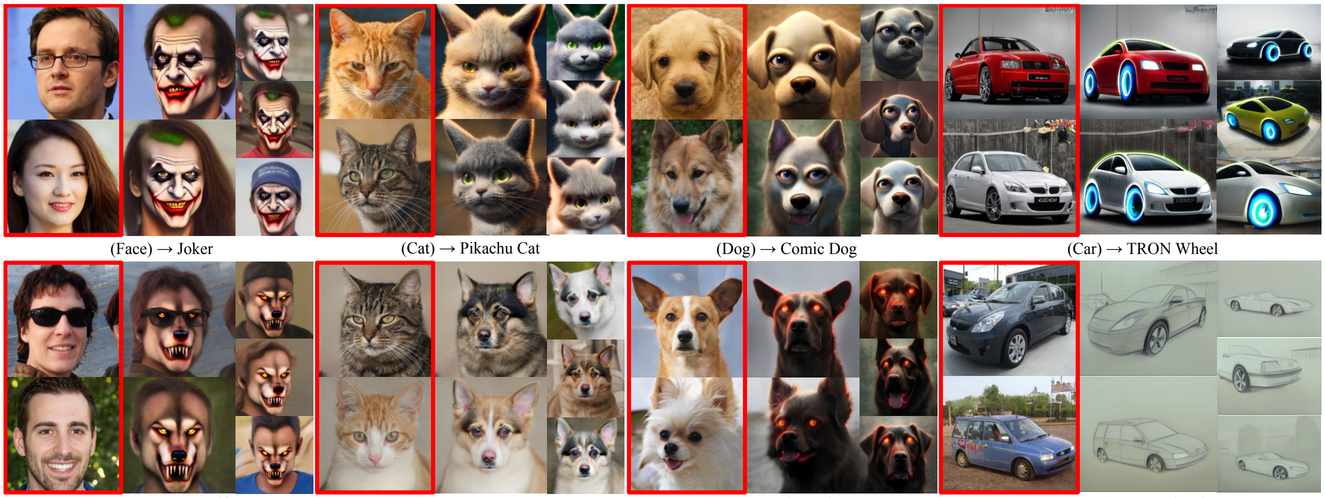
2. Related Work

Image generator domain adaptation. How can we get a generator without having access to enough real data? The goal of domain adaptation is to shift the data distribution of image generators to a desired new domain different from what it is trained on. Prior works branches into two directions: few-shot and text-guided zero-shot fine-tuning.

Few-shot models are trained with several hundred or fewer [32, 39] image samples. To better capture the target domain, some control channel statistics [38] or sampling process [55] in the latent space. Regularizer to prevent model collapse issue [29, 35, 41, 47, 54]. Some use auxiliary tasks [31, 57] to alleviate overfitting. Text-guided zero-shot fine-tuning uses only text as guidance. Prior works exploited the semantic power of large-scale CLIP models [44] to find editable latent space directions in a pre-trained StyleGAN2 [24]. For example, StyleCLIP [40] optimizes the latent code for the generator and minimizes the text-image similarity score from CLIP. StyleGAN-NADA [8] takes a step further by directly fine-tuning the generator using the CLIP text-image directional objective.

More recently, diffusion models show great potential in text-guided fine-tuning. similar to StyleCLIP [40], DiffusionCLIP [28] applies CLIP [44] objective to diffusion generators. Some fine-tune the text embedding [7] or the full diffusion model [50] on a few personalized images.

Text to image diffusion model. Diffusion models [17, 52] have achieved state-of-the-art image synthesis quality [37, 43, 51, 53], especially on large-scale text-to-image synthesis tasks. Introduced by [17], diffusion models use an iterative denoising process, which enables them to iteratively convert Gaussian noise into fine-grained images from a diverse and complicated image distribution. Latent diffusion models (LDMs) such as StableDiffusion [49] is a class of diffusion models that operates on a latent space of a pre-trained autoencoder. Instead of learning directly from the image space, learning from the latent space greatly reduces



(Face) → Joker

(Cat) → Pikachu Cat

(Dog) → Comic Dog

(Car) → TRON Wheel

(Face) → Werewolf

(Cat) → Dog

(Dog) → Epic Dark Dog

(Car) → Car Sketch

Figure 2. Generated images from experiments on FFHQ face, AFHQ-Cat, Car and Dog [2]. The text below each section is the driving prompt. Notice our model only takes in a target prompt and does not need the source prompt.

the data sample dimension. The latent space comes with well-compressed semantic features and visual patterns that are already learned by the autoencoder, thus saving the cost of the diffusion model to learn everything from scratch.

Score distillation sampling. Prior works use diffusion models as critics to optimize an image or a Differentiable Image Parameterization (DIP) [36] and bring it toward the distribution indicated by a text prompt. DreamFusion [42] is a recent work that proposed a Score Distillation Sampling (SDS) loss to utilize a pre-trained text-to-image diffusion model [51] to guide the training of NeRF [34]. Their proposed method can efficiently bypass the score-predicting module and approximate the gradient with the difference between the classifier-free guidance score and the ground-truth noise. DreamFusion [42] performs SDS on image pixel space. We adopt the same gradient trick but extend it to the latent space of StableDiffusion [49] and use it as guidance for StyleGAN2 [24] generator domain adaptation. We also include experiments with 3D-aware domain adaptation on EG3D [1] image generators.

3. Methods

3.1. Background

Latent diffusion model. We use the publicly available latent diffusion model(LDM) StableDiffusion [49] in this paper as our guidance model. A LDM encodes images $\mathbf{x}$ into latent space $\mathbf{z}$ with an encoder $\mathcal{E}$, $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$, and the denoising process is preformed in the latent space $\mathcal{Z}$. Briefly, a latent diffusion model ϵ_θ can be trained on a denoising objective of the following form:

$$\mathbb{E}_{\mathbf{z}_0,\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\epsilon_\theta(\mathbf{z}_t|\mathbf{c},t)-\boldsymbol{\epsilon}\|_2^2\right] \tag{1}$$

where $(\mathbf{x}, \mathbf{c})$ are data-conditioning pairs, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \text{Uniform}(1, T)$ and w_t is a weighting term.

Classifier-free guidance. In the denoising sampling process, *Classifier guidance* is an effective method to guide the synthesis better toward the desired direction, *e.g.* a class or a text prompt [4]. The method uses gradients from a pre-trained model $p(\mathbf{c}|\mathbf{z}_t)$ during sampling. *Classifier-free guidance* (CFG) [18] is an alternative technique that avoids this pre-trained classifier. During the training of the conditional diffusion model, randomly dropping the condition $\mathbf{c}$ lets the model learns to generate an image even without a condition. Therefore, a well-conditioned image can be generated by pushing the synthetic results under condition $\mathbf{c}$ further away from the unconditioned results during the diffusion process, where

$$\hat{\boldsymbol{\epsilon}}_{\theta,\mathbf{c}}(\mathbf{z}_t) = s \cdot \epsilon_\theta(\mathbf{z}_t|\mathbf{c},t) + (1-s)\cdot\epsilon_\theta(\mathbf{z}_t|\emptyset,t). \tag{2}$$

Here, $\epsilon_\theta(\mathbf{z}_t|\mathbf{c},t)$ and $\epsilon_\theta(\mathbf{z}_t|\emptyset,t)$ are conditional and unconditional $\boldsymbol{\epsilon}$-predictions. s is the guidance weight and increasing $s > 1$ strengthens the effect of guidance.

3.2. Model Structure and Diffusion Guidance Loss

An image $\mathbf{x}$ is generated with generator $\mathcal{G}$ from a style code $\mathbf{w} \sim P_\mathbf{w}$, where $P_\mathbf{w}$ is the pushforward measure from $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ to $\mathcal{W}$ through a mapping network g. The generated image $\mathbf{x} = \mathcal{G}(\mathbf{w})$ is then encoded into the latent space of the StableDiffusion model using its encoder $\mathcal{E}$, $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{c\times h\times w}$. Following the standard diffusion training schema, we sample a time-step $t \sim \text{Uniform}(1, T_x)$, with $0 < T_x < T$, and perform the forward process (namely, "q sample") to get a noisy latent: $\mathbf{z}_t = q(\mathbf{z}_0, t) := \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$.

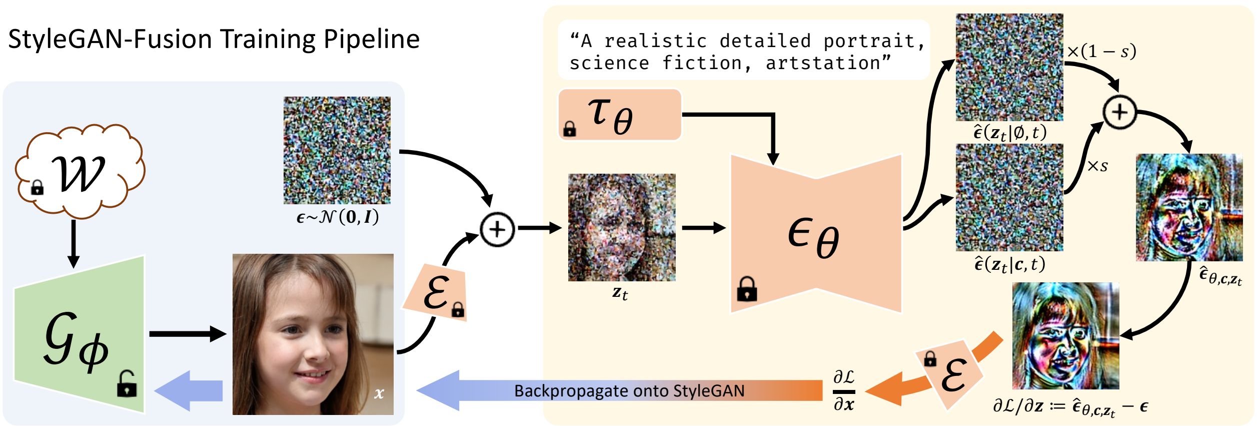

StyleGAN-Fusion Training Pipeline
$\mathcal{W}$
$\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$
$\mathcal{G}_\phi$
$\boldsymbol{x}$
$\mathcal{E}$
"A realistic detailed portrait, science fiction, artstation"
τ_θ
$\boldsymbol{z}_t$
ϵ_θ
$\hat{\boldsymbol{\epsilon}}(\boldsymbol{z}_t|\emptyset,t)$
$\times(1-s)$
$\times s$
$\hat{\boldsymbol{\epsilon}}(\boldsymbol{z}_t|\boldsymbol{c},t)$
$\hat{\boldsymbol{\epsilon}}_{\theta,c,z_t}$
$\partial\mathcal{L}/\partial\boldsymbol{z} := \hat{\boldsymbol{\epsilon}}_{\theta,c,z_t}-\boldsymbol{\epsilon}$
$\frac{\partial\mathcal{L}}{\partial\boldsymbol{x}}$
Backpropagate onto StyleGAN

Figure 3. Overview of our StyleGAN-Fusion framework. The style-based generator $\mathcal{G}_\phi$ receives the gradient $\frac{\partial\mathcal{L}}{\partial\mathbf{x}}$ backpropagated from $\frac{\partial\mathcal{L}}{\partial\mathbf{z}}$ through encoder $\mathcal{E}$. $\hat{\boldsymbol{\epsilon}}_{\theta,\mathbf{c},\mathbf{z}_t}$ is the classifier-free guidance score. All noises and noisy images are the decoded corresponding latents for visualization purposes.

Then the denoising process takes in $\mathbf{z}_t$ and returns the predicted noise $\hat{\boldsymbol{\epsilon}}_{\theta,\mathbf{c}}(\mathbf{z}_t)$ (classifier-free guidance score), conditioned on time step t and text prompt embedding y. Ideally, if $\mathbf{z}_t$ is faithfully rendered according to the given text condition, the diffusion model ϵ_θ should be able to correctly recover the true noise $\boldsymbol{\epsilon}$. We follow the gradient trick proposed by DreamFusion [42] to directly used the difference between the predicted and the ground-truth scores as gradient and backpropagated through $\mathcal{E}$ to the generator $\mathcal{G}$, $\nabla_\phi\mathcal{G}_\phi = \nabla_{\mathbf{z}}\mathcal{L}_{SDS}\frac{\partial\mathbf{z}}{\partial\phi}$, and

$$\nabla_{\mathbf{z}}\mathcal{L}_{SDS} := \mathbb{E}_{\mathbf{c},\boldsymbol{\epsilon},t}[w_t\left(\hat{\boldsymbol{\epsilon}}_{\theta,\mathbf{c}}(\mathbf{z}_t)-\boldsymbol{\epsilon}\right)]\,. \tag{3}$$

The generator parameters are updated accordingly.

3.3. Directional and Reconstruction Regularizer

The diffusion guidance loss provides the generator with an informative direction to evolve, improving the image fidelity at the cost of diversity. It does not encourage image diversities. In fact, we observe that after a sufficient amount of iterations, the generator will collapse to a fixed image pattern regardless of its input noise. We assume this happens because the loss is sufficiently minimized when the unconditional $\hat{\boldsymbol{\epsilon}}_\emptyset$ is equal to the conditional $\hat{\boldsymbol{\epsilon}}_{\mathbf{c}}$ for all $\mathbf{w}$. It is searching for an image $\mathbf{x}$ that makes diffusion ϵ_θ to predict the same noise for both unconditional and conditional inputs. This phenomenon is also known as differentiable image parameterization (or DIP). In such a case, the generator will lose image diversities, leading to a mode collapse.

To address this issue, we regularize the generator optimization process with an additional loss term, which we defined as a diffusion directional regularizer. Denote the original frozen generator as $\mathcal{G}_{frozen}$ and the current one $\mathcal{G}_{train}$, the classifier-free guidance scores are given by,

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_{train} &= \hat{\boldsymbol{\epsilon}}_{train,\emptyset} + s(\hat{\boldsymbol{\epsilon}}_{train,\mathbf{c}} - \hat{\boldsymbol{\epsilon}}_{train,\emptyset})\\ \hat{\boldsymbol{\epsilon}}_{frozen} &= \hat{\boldsymbol{\epsilon}}_{frozen,\emptyset} + s(\hat{\boldsymbol{\epsilon}}_{frozen,\mathbf{c}} - \hat{\boldsymbol{\epsilon}}_{frozen,\emptyset})\end{aligned} \tag{4}$$

The proposed directional regularizer is the cosine similarity between $\hat{\boldsymbol{\epsilon}}_{train}$ and $\hat{\boldsymbol{\epsilon}}_{frozen}$, maintaining a low directional difference. To efficiently implement it, we leverage the fact that a high dimensional Gaussian random variable lies on a sphere with high probability [20] and minimize their L_2 distance instead. To do so, we normalize each score tensor according to its expected radius $r = \sqrt{c\times h\times w}$, and add a regularization gradient term defined as,

$$\nabla_{\mathbf{z}}\mathcal{L}_{SDS}^{dir} := r\left(\frac{\hat{\boldsymbol{\epsilon}}_{train}}{||\hat{\boldsymbol{\epsilon}}_{train}||_2} - \frac{\hat{\boldsymbol{\epsilon}}_{frozen}}{||\hat{\boldsymbol{\epsilon}}_{frozen}||_2}\right). \tag{5}$$

We use $\mathcal{L}_{SDS}^{dir}$ as a constraint on the optimization of the generator. Note that during the fine-tuning, $\hat{\boldsymbol{\epsilon}}_{frozen}$ is a fixed starting point for a given style code. If the generator only gives a single image for all $\mathbf{w}$, the gradient from all fixed starting points will be different. Such regularizer encourages $\mathcal{G}_{train}(\mathbf{w})$ to maintain its initial optimization direction, adding a force in preventing model collapse. Experiments show the directional regularizer can efficiently prevent model collapse. It is a plug-in module that is compatible with other regularization methods.

Additionally, we extend the score distillation framework to *reconstruction guidance* [19] and introduce a reconstruction regularization. Intuitively, we want the current estimation of the clean latent image $\hat{\mathbf{z}}_0 = (\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}})/\sqrt{\bar{\alpha}_t}$, given by the Tweedie's formula [3], to be similar to the latent image $\mathbf{z}_0$ given by $\mathcal{G}_{frozen}$,

$$\nabla_{\mathbf{z}}\mathcal{L}_{SDS}^{rec} := r\left(\frac{\hat{\boldsymbol{\epsilon}}_{train}}{||\hat{\boldsymbol{\epsilon}}_{train}||_2} - \frac{\nabla_{\hat{\boldsymbol{\epsilon}}}\mathcal{L}_{rec}}{||\nabla_{\hat{\boldsymbol{\epsilon}}}\mathcal{L}_{rec}||_2}\right), \tag{6}$$

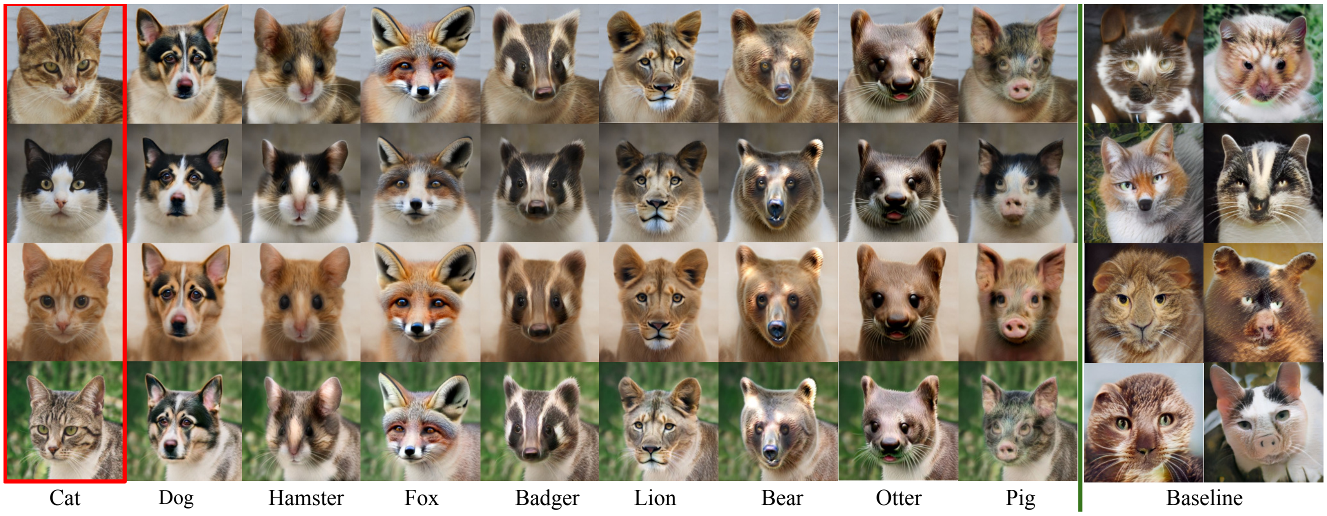


Figure 4. Uncrated samples from our method on Cat-to-8-animals (left). For each animal type, we show 1 uncurated sample from baseline (right). Notice the significant visual quality differences. Our method generated more visually realistic and natural results, including undistorted facial components, cleaner backgrounds, diverse poses, and higher pose faithfulness. We highlight all $\mathcal{G}_{frozen}$ in red box.

where $\nabla_{\hat{\epsilon}}\mathcal{L}_{rec}$ is the gradient of reconstruction loss $\mathcal{L}_{rec} = \|\hat{\mathbf{z}}_0 - \mathbf{z}_0\|_2^2$. We provide comparisons between the regularizers in the experiment section. The total loss is $\mathcal{L} = \mathcal{L}_{SDS} + \lambda_{dir}\mathcal{L}_{SDS}^{dir} + \lambda_{rec}\mathcal{L}_{SDS}^{rec}$, with λ's the weighting coefficients.

3.4. Timestep Range and Layer Selection

The range of denoising timestep (T_{SDS}) from which timestep $t \sim \text{Uniform}(T_{min}, T_{max})$ is sampled closely relates to the model behavior. Increasing T_{SDS} value leads to an amplified noise level in the resulting latent code following q sampling process. This, in turn, provides more scope for ϵ_θ to undertake modifications. The guidance signal from ϵ_θ is thus more related to high-level image structures. A smaller T_{SDS}, on the other hand, leaves less scope for ϵ_θ and is more related to local structures and details. T_{SDS} range configuration allows us to control the scale of changes (see Sec. 4.3).

Recall that style-based generators has a similar property: deeper layers control image composition and shallower layers the image details. Intuitively, if we optimize generator layers altogether, unsatisfied scenarios could occur where a high-level overall-structure guidance loss is used to update a shallow and detailed generator layer, resulting in blurry generated images. We use layer selection to overcome such issues. Inspired by StyleGAN-NADA [8], we perform N iterations of optimization on the $\mathcal{W}^+$ style code space based on the SDS objective and select k layers that correspond to the most significantly changed style codes. Ablation study (see Sec. 4.3) shows the quality boost of multiple k settings, especially in terms of reducing blurry vagueness.

4. Experiments

We begin by showing result images. In Fig. 1, the upper section contains generated face images in a 3D rendering style described by the text promt. we take a StyleGAN2 generator pre-trained on the photorealistic FFHQ dataset and fine-tune it using our method. The lower section contains generated cat images, fine-tuned from the AFHQ-Cat [2] checkpoint. Fig. 2 shows diverse results from werewolves, Joker, photorealistic or artistic rendering of cats/dogs and cars. Full text prompts in each experiment and additional image galleries are included in the appendix.

4.1. Qualitative evaluation

We compare our method and the baseline, StyleGAN-NADA, in multiple Cat-to-animals experiments. Fig. 4 shows 4 uncurated generated samples from our method for each animal type and 1 sample from the baseline. Our method is able to generate images with much higher visual quality and fidelity in these experiments. Baseline results are optimized to minimize CLIP metric in an adversarial manner [33] and have much lower visual quality. Our results look realistic and detailed whereas baseline results are flat with distorted details. We provide quantitative evaluations for these experiments in the following section.

Our method consistently outperforms the baseline. We conduct experiments on the FFHQ face model with a prompt that requires detailed artistic styles on rendering and lighting. Fig. 5 shows the results from the baseline and ours. Baseline images have distorted facial components like asymmetrical eyes, and a rough texture which directly conflicts with the prompt. Our results have much higher quality

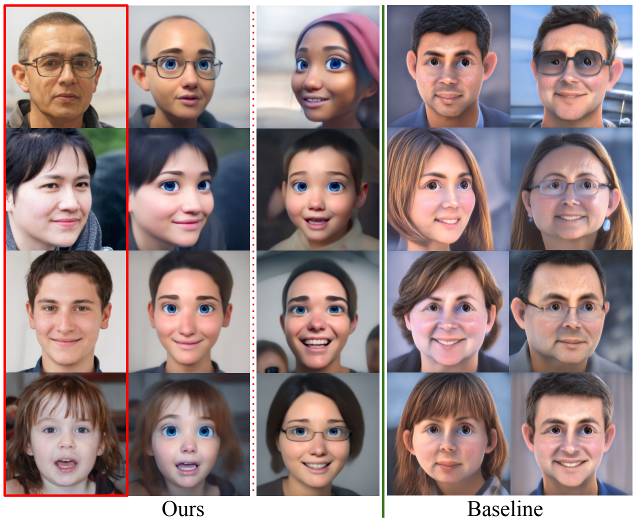


Figure 5. Compare our method with the baseline on ”FFHQ face”. Our method generates higher quality results than the baseline and better matches the prompt.

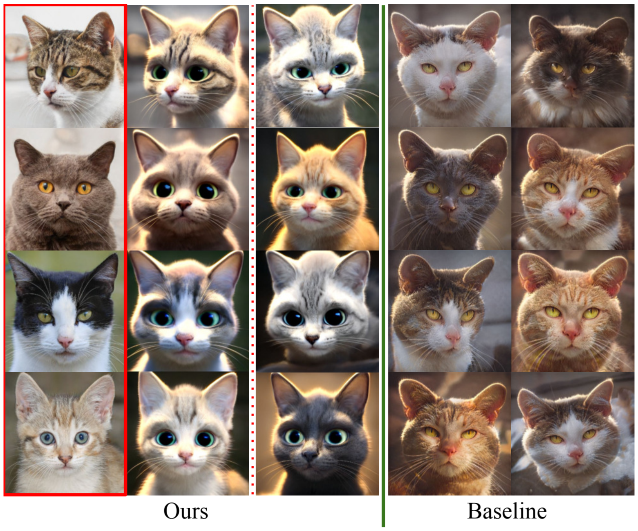


Figure 6. Compare our method with the baseline on ”AFHQ-Cat”. Our method generated significantly more realistic and natural results than the baseline.

and better match the prompt, especially in terms of natural and undistorted face layouts. Moreover, results from our model have more realistic 3D lighting and rendering which better match the text prompts.

Similar results are observed when adapting the StyleGAN2-Cat [24] model. Fig. 6 shows generated images from our method and the baseline. The baseline model does not properly follow the text description and fails in many aspects. The requested outcome was to have cute, circular, and large reflective eyes. However, the baseline images do not seem to exhibit these features. The rendering appears artificial and of low quality, lacking the appearance of a true 3D representation. Our model perfectly satisfies the prompt requests with cinematic-like lighting, smoother textures, a stronger 3D style, and high-quality details. We address our performance superiority from two perspectives:

- Prior work [33] has demonstrated that the CLIP loss is prone to be adversarially minimized, and be confined to a local minimum. Researchers also observed that optimization overcomes the CLIP loss by adding pixel-level perturbations to the image [8]. Our model, however, utilizes Stable Diffusion guidance and operates within the latent space. The generator and diffusion guidance are separated by an image Encoder, increasing the likelihood of semantically meaningful optimization rather than adversarially on a pixel level.

- Baseline model uses CLIP text encoders which return one single embedding vector for the entire text prompt. The vector space limits the capacity and forces information compression, reducing the embedding quality especially when the text prompts are long and detailed. Our method utilizes StableDiffusion as guidance which uses a sequence of text embeddings and cross-domain attention, improving the capability to capture multiple key constraints outlined in lengthy text prompts. We provide quantitative evaluations in the following section.

4.2. Quantitative Evaluation

This section quantitatively compares the baseline and our method. We conduct experiments adapting an AFHQ-Cat [2] generator to the domain of 8 other animals indicated by prompts. We manually extracted ground truth images from AFHQ-Wild subclass and calculate the FIDs in Tab. 1. We kindly remind these FIDs are achieved in a zero-shot manner, as both methods are trained solely on text without having an access to a single ground-truth image. Our method significantly outperforms baselines in FID score in all cat-to-8-animals experiments.

Such performance gains are persistent when evaluating the CLIP matching scores between images and the text prompt. Baseline encodes the entire text prompt into a fixed-length vector via CLIP text encoder, which inevitably compresses information. The issue becomes more noticeable and severe as the length of the text prompt increases. To provide quantitative evidence of the superiority of our approach, we evaluate how well our method and baseline capture each *keyword requirement* mentioned by the prompt. Specifically, we separate the prompt into keyword pieces and calculate the CLIP [44] image-text matching score between the generated image and keywords. We conduct experiments with our method and baseline in adapt-

Table 1. FID scores of Cat/Dog-to-Animals. Ground-truth images are extracted from the AFHQ dataset [2]. Our models consistently outperform the baseline in FIDs by a large margin.

	Cat		Dog	
	Ours	NADA	Ours	NADA
Dog/Cat	**150.76**	206.93	**124.72**	139.35
Fox	**51.51**	90.40	**61.10**	129.58
Lion	**30.34**	153.82	**52.52**	173.81
Tiger	**19.29**	115.46	**31.15**	223.33
Wolf	**45.33**	139.66	**71.29**	160.00

ing the StyleGAN2-Face and Cat generators, as shown in Fig. 5 and Fig. 6. The prompt requests a rendering style that contains diverse constraints including eyes, texture, and lighting. Both methods are trained for 2000 iterations after which we sample 2000 images for metric evaluation. Tab. 2 shows that the baseline has difficulty capturing all key constraints mentioned in the long text prompt. Our model, in contrast, better captures almost all keywords in CLIP scores. This fits our intuition as diffusion guidance uses cross-attention between images and each text token and has greater information capacity.

Table 2. CLIP between each keyword and generated images. Our model outperforms the baseline in almost all keywords.

	Face		Cat	
Prompt Keywords	Ours	NADA	Ours	NADA
3d cute face/cat	**0.303**	0.301	**0.315**	0.307
closeup cute and adorable	**0.247**	0.222	**0.256**	0.242
cute big circular reflective eyes	**0.276**	0.268	**0.276**	0.267
Pixar render	0.280	**0.282**	**0.251**	0.247
unreal engine	**0.273**	0.271	**0.264**	0.262
cinematic smooth	**0.227**	0.208	**0.221**	0.217
intricate detail	**0.213**	0.208	0.213	**0.218**
cinematic lighting	**0.231**	0.210	**0.230**	0.229

4.3. Timestep Range and CLIP-LPIPS Trade-off

Recall we sample a timestep t from range $T_{SDS} = (T_{min}, T_{max})$, in each iteration, based on which we apply the q sample process and update the generator. We experiment with the influence of its configuration on our model behavior. We observe that a large T_{SDS} enables more global structural modifications, while a small T_{SDS} only allows local detail modifications. Fig. 7 right side shows an example of adapting a cat generator to an otter generator. As T_{SDS} increases, the generated otter's ears become smaller and more realistic This property provides a smooth transition from the original domain to the target domain and adds more controllability to the optimization process.

For quantitative results, we calculate CLIP scores for image-text alignment and LPIPS [59] scores to address image diversity [2], as shown in Fig. 7 scatter plot. We fix $T_{min} = 0$ and conduct experiments with different T_{max} settings. Similar to previous work [26] [6], we notice a CLIP-LPIPS trade-off with T_{SDS}. A larger range enables structure changes and increases image fidelity to the target domain, resulting in high CLIP scores but lower LPIPS diversity. In contrast, a smaller range focuses on local changes and prefers faithfulness to the original domain, resulting in high LPIPS diversity but lower CLIP scores.

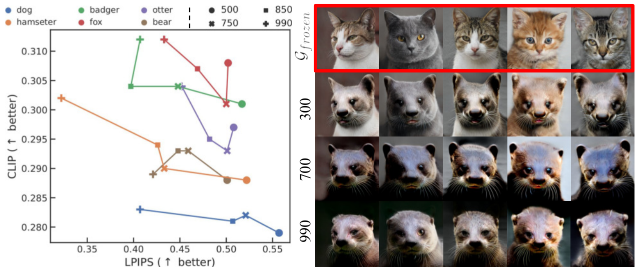


Figure 7. Denoising time step controls. The left figure shows the trade-off between fidelity and diversity. Right figure visually shows the effect. As T_{max} increases, generated images gradually show more otter features but lose distinctive fur and color patterns.

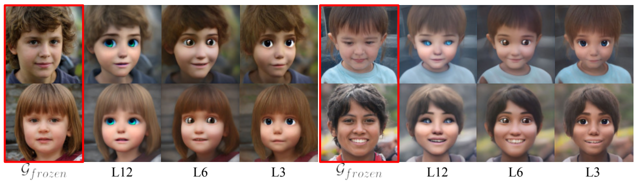


Figure 8. Ablation on layer selection. With fewer layers selected, vagueness disappears and image details become clear and sharp.

Table 3. Quantitative ablation on regularizer and layer selection. We calculate FIDs scores for face experiments.

FIDs	w/o Layer Selection		w/ Layer Selection 12 (L12)		L6	L3	L1
T	No Reg	Regularize	No Reg	Regularize	Regularize		
0.75	143.9836	**130.9132**	140.1852	**119.8765**	114.6877	105.0439	**102.7785**
0.80	135.5151	**120.8958**	126.5048	**104.189**	110.6044	97.9570	**93.2015**
0.85	119.1711	**107.4251**	108.306	**90.8212**	89.8302	81.322	**71.2215**

4.4. Ablation on Layer Selection and Regularizers

This section evaluates the effect of the layer Selection and our diffusion-guidance regularizers. For quantitative evidence, we show FIDs scores for face experiments in Tab. 3. Since we don't have ground-truth images, we utilize StableDiffusion img2img to approximate them with 3 strength values (T). The left part shows the effectiveness of our proposed regularization. Regularized models are consistently better in all experiments with or without layer selection. The right part of the table is a quantitative comparison of layer selection with different layer settings. From L12 to L1, we select fewer layers for each training

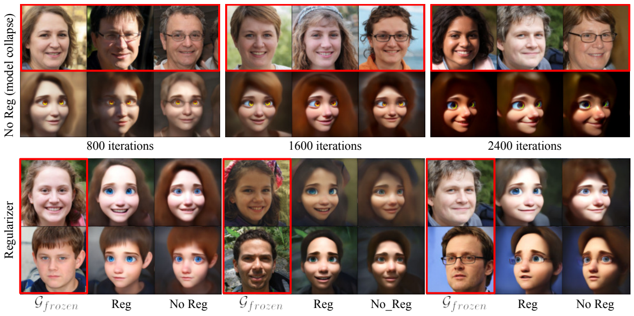


Figure 9. Regularization serves a key role. The regulated ones better preserve the details including facial expressions, hairs, and backgrounds, whereas the non-regulated approach ignores them.

iteration and FID performance improves Continuously.

Fig. 8 visually shows the effect of layer selection on our method. We select 12/6/3 layers to optimize for each iteration. Selecting fewer layers slows down the training speed and requires more iterations of training. We show results with the best visual quality for each layer configuration. By selecting fewer layers, the blurry and vague appearance of the image is mitigated, resulting in sharper and more distinct details, such as hair and eyes.

We visually show the importance of our regularizers for the model performance using the face experiment. The text prompt is the same as Fig. 1, requesting a rendering of anime faces with cute circular reflective eyes. Fig. 9 upper panel shows training snapshots of non-regularized baseline. Without regularization, the generator gradually collapses into a few very similar images and eventually completely ignores its input. It overfits on image features mentioned in the text prompt (large eyes, Pixar render, etc) but ignores all other aspects that are important for visual quality but are not mentioned in the prompt, such as background, gender, facial expression, etc. In another word, the generator is guided by L_{SDS} to focus on image features specified by the prompt while disregarding all other image features, leading to an overemphasis on the prompt-specified features. Given enough training iterations, the non-regulated approach degrades to very similar images and ignores its input $\mathbf{z}$, resulting in a model collapse. Fig. 10 further inspects the difference between regularizers.

4.5. Extension to 3D Generators and DreamBooth

We extend our method to 3D Geometry-aware generators from EG3D [1] on the face and cat models. During optimization, all parameters are frozen except the weights of Conv layers in the tri-plane generator. We add LPIPS [59] to the loss function to stabilize training. We show additionally extend to DreamBooth [50] guidance in Fig. 12 and the appendix. We tried public available DreamBooth checkpoints “Wa-vy” style [56] and “Woolitize” style [27]

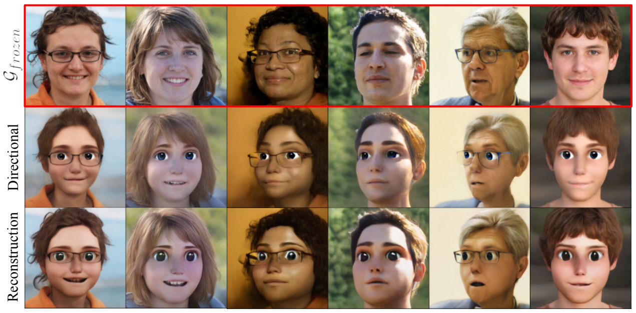


Figure 10. Compare directional and reconstruction regularizers. $\mathcal{L}_{SDS}^{rec}$ is a stronger constraint and keeps more details intact, whereas $\mathcal{L}_{SDS}^{dir}$ is mild and allows more modifications.

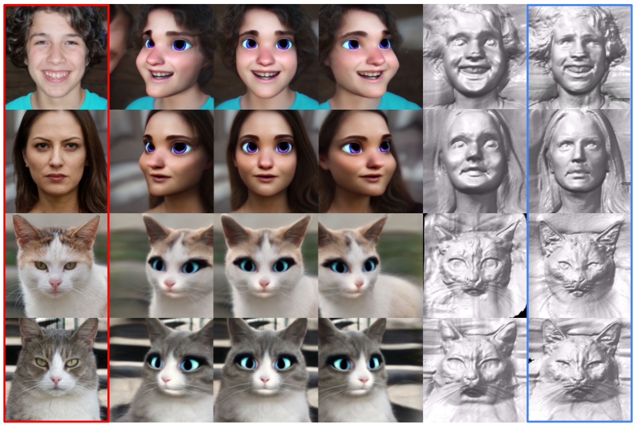


Figure 11. 3D domain adaptation on EG3D-Face and Cat [1] . Results from $\mathcal{G}_{train}$ are located in the middle; $\mathcal{G}_{frozen}$ on the left and its mesh on the right highlighted with blue box.

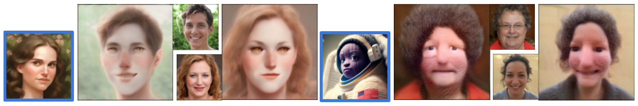


Figure 12. DreamBooth guided results. We show the original DreamBooth StableDiffusion samples in blue boxes.

5. Conclusion

We presented a novel domain adaptation method for image generators that uses StableDiffusion guidance and Score Distillation Sampling. Our method allows flexible control of the magnitude of modifications by selecting the value of T_{SDS}. With the introduced diffusion-guidance directional regularizer and layer selection techniques, our model is able to shift the generator to generate new images from a target domain indicated by the text prompt, with improved quality compared to existing methods. We also show that our method can be extended to 3D-aware style-based generators and used with DreamBooth models as guidance.

References

[1] 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. 3, 8

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. CoRR, abs/1912.01865, 2019. 3, 5, 6, 7

[3] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687, 2022. 4

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34, 2021. 3

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 1

[6] Adham Elarabawy, Harish Kamath, and Samuel Denton. Direct inversion: Optimization-free text-driven real image editing with diffusion models. arXiv preprint arXiv:2211.07825, 2022. 7

[7] 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. 2

[8] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG), 41(4):1–13, 2022. 1, 2, 5, 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, pages 2672–2680, 2014. 1

[10] Ligong Han, Ruijiang Gao, Mun Kim, Xin Tao, Bo Liu, and Dimitris Metaxas. Robust conditional gan from uncertainty-aware pairwise comparisons. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10909–10916, 2020. 1

[11] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. arXiv preprint arXiv:2303.11305, 2023. 1

[12] Ligong Han, Martin Renqiang Min, Anastasis Stathopoulos, Yu Tian, Ruijiang Gao, Asim Kadav, and Dimitris N Metaxas. Dual projection generative adversarial networks for conditional image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14438–14447, 2021. 1

[13] Ligong Han, Sri Harsha Musunuri, Martin Renqiang Min, Ruijiang Gao, Yu Tian, and Dimitris Metaxas. Ae-stylegan: Improved training of style-based auto-encoders. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3134–3143, 2022. 1

[14] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3615–3625, 2022. 1

[15] Ligong Han, Anastasis Stathopoulos, Tao Xue, and Dimitris Metaxas. Unbiased auxiliary classifier gans with mine. arXiv preprint arXiv:2006.07567, 2020. 1

[16] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. arXiv preprint arXiv:2306.05414, 2023. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020. 2

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. 3

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. arXiv preprint arXiv:2204.03458, 2022. 4

[20] Ferenc Huszár. Gaussian distributions are soap bubbles, Nov 2017. 4

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017. 1

[22] 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. 1

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019. 1

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of styleGAN. arXiv preprint arXiv:1912.04958, 2019. 2, 3, 6

[25] 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

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276, 2022. 7

[27] Ahsen Khaliq. Dreambooth woolitize. Hugging Face, 2022. https://huggingface.co/plasmo/woolitize-768sd1-5. 8

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In CVPR, pages 2426–2435, 2022. 2

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

[30] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. arXiv preprint arXiv:2203.02053, 2022. 1

[31] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations, 2020. 2

[32] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis, 2021. 2

[33] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. arXiv preprint arXiv:2112.01573, 2021. 1, 5, 6

[34] 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. Communications of the ACM, 65(1):99–106, 2021. 3

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

[36] Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. Distill, 2018. https://distill.pub/2018/differentiable-parameterizations. 3

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021. 1, 2

[38] 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

[39] 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

[40] 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. 2

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

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988, 2022. 1, 3, 4

[43] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10619–10629, 2022. 2

[44] 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. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 1, 2, 6

[45] 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. 1

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092, 2021. 1

[47] 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

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arXiv preprint arXiv:2112.10752, 2021. 1

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022. 2, 3

[50] 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. 2, 8

[51] 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. 1, 2, 3

[52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265. PMLR, 2015. 2

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations, 2021. 2

[54] 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. 2

[55] 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

[56] wavymulder. Dreambooth wa-vy style. Hugging Face, 2022. https://huggingface.co/wavymulder/wavyfusion. 8

[57] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. Advances in Neural Information Processing Systems, 34:9378–9390, 2021. 2

[58] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2022. 1

[59] 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, pages 586–595, 2018. 7, 8