LaDiffGAN: Training GANs with Diffusion Supervision in Latent Spaces

Xuhui Liu[1]*, Bohan Zeng[1]*, Sicheng Gao [1], Shanglin Li[1], Yutang Feng[1],
Hong Li[1], Boyu Liu[1], Jianzhuang Liu[2],Baochang Zhang[1,3,4†]
[1]Beihang University [2]Shenzhen Institute of Advanced Technology, Shenzhen, China
[3]Zhongguancun Laboratory, Beijing, China [4]Nanchang Institute of Technology, Nanchang, China

Abstract

Diffusion models have recently become increasingly popular in a number of computer vision tasks, but they fail to achieve satisfactory results for unsupervised image-to-image translation, since they require massive training data and rely heavily on extra guidance. In this scenario, GANs can alleviate these issues existing in diffusion models, albeit with suboptimal quality. In this paper, we leverage the advantages of both GANs and diffusion models by training GANs with diffusion supervision in latent spaces (LaDiffGAN) to solve the unsupervised image-to-image translation task. Firstly, to promote style transfer quality, we encode the data in specific latent spaces with styles of the target and source domains. Secondly, we introduce the diffusion process with different amounts of Gaussian noise to enhance the modeling capability of GANs on the complex data distribution. We accordingly design a latent diffusion GAN loss to align the latent features between generated and training images. Lastly, we introduce a heterogeneous conditional denoising loss that incorporates image-level supervision to further improve the quality of generated results. Our LaDiffGAN significantly alleviates the drawbacks associated with diffusion models, such as data leakage, high inference cost, and high dependence on large training data sets. Extensive experiments show that LaDiffGAN outperforms previous GAN models and delivers comparable or even better performance than diffusion models.

1. Introduction

Image-to-image translation, especially unsupervised with unpaired images, is one of the hottest research fields in machine learning and computer vision, and has a variety of far-reaching applications including photography, colorization, image inpainting, and style transfer [20, 29, 76]. As essentially a mapping function across two image domains, image-to-image translation is usually considered a generative task. Accordingly, many models have emerged obtaining impressive results, such as generative adversarial networks (GANs) [20], variational autoencoders (VAEs) [35] and flow models [34]. Nonetheless, due to the lack of paired samples coupled with semantic disparities across different domains, unsupervised image-to-image translation remains a challenging problem.

GANs employing a shared latent space and the cycle consistency assumption [76] have been the prevailing choice for unsupervised image-to-image translation in the past several years. However, GANs are notoriously susceptible to training instability and even mode collapse, and prone to generate unnatural details, which reveals the limited capability of GANs to capture complex data distributions. Recently, the great success of diffusion probabilistic models (DMs) [57] in image synthesis has attracted significant attention to explore their potential in image translation. Promptomania [1] attains impressive performance on image style transfer, yet it requires massive training data and relies heavily on the text-to-image model [8, 22]. CycleDiffusion [66] successfully employs DMs to solve the unpaired image translation without extra guidance (e.g., text description and semantic maps), but it shows unsatisfactory results when there are limited training data and a large gap across translation domains.

DMs' performances degrade with small training sets, and may regenerate training examples leading to copyright infringement issues. For instance, NovelAI [2] has been sued for copyright infringement due to the striking similarity between the results produced by Stable Diffusion [53] and the data in the training set. These drawbacks mainly attribute to the fact that the neural networks of DMs are directly fed with the ground truth, causing insufficient generalization with small training sets. Moreover, as verified by [9], such networks tend to memorize the information from the training set. As shown in Figs. 1 (d) and (e), the result of CycleDiffusion is similar to the sample from the train-

*These authors contributed equally.
†Corresponding Author: bczhang@buaa.edu.cn.

[1]https://promptomania.com
[2]https://novelai.net

Figure 1. (a) Input image. (b) Result generated by our LaDiffGAN. (c) Result generated by U-GAT-IT. (d) Result generated by CycleDiffusion. (e) A training image similar to (d). It is evident that our model generates images of higher-quality stylization than the state-of-the-art GAN model. Moreover, while CycleDiffusion can generate images that match the style of the target domain, as demonstrated in (d), quite similar images can be directly identified from the training dataset, as shown in (e).

ing set, and we observe that such privacy leakage is particularly evident in small training sets. In contrast, GANs only use the ground truth to supervise the generation process, thereby mitigating the problem of small data sets and reducing the likelihood of severe data leakage. Moreover, DMs faces the problem of long inference time in practice, which imposes a severe constraint in situations that require real-time processing.

This paper comprehensively considers **safety**, **convenience**, **quality**, and **inference speed** for unsupervised images-to-image translation. To address these critical concerns, we propose LaDiffGAN, a new framework that enables GANs to be optimized in latent spaces using the diffusion process, providing a deeper exploration of the potential of GANs in handling complex data distributions and generating better images. For **safety** and **inference speed**, we select GANs over DMs as the base model to mitigate the risk of data leakage and high inference cost. For **quality**, we encode the input and generated data into specific latent spaces with the styles of the target and source domains and employ the latent diffusion Markov chain to model richer distributions to facilitate the generators and discriminators to capture the essential representations of the target and source domains and be more resistant to overfitting. As such, LaDiffGAN does not rely on extra guidance nor massive training data to handle the unsupervised images-to-image translation problem, meeting the **convenience** requirement. We emphasize that LaDiffGAN is different from DiffusionGAN [65], which only injects instance noise in the image space during forward diffusion for stable training of GANs. LaDiffGAN supervises the styles of generated results in the latent spaces and leverages the pre-trained denoising networks conditioned on the downsampled generated images to further improve the quality of generation. It incorporates both image-level and latent-level supervision in a unified framework. We achieve that by developing two new loss functions, latent diffusion GAN (LDG) loss, and heterogeneous conditional denoising (HCD) loss. Overall, these designs enhance the optimization of the GAN model and promote the generation of better images with improved quality and fidelity, as shown in Fig. 1.

The main contributions of this paper are summarized as follows:

- We comprehensively analyze the pros and cons of GANs and diffusion models, and then present LaDiffGAN to explore the ability of GANs in handling unsupervised image-to-image translation by training GANs in the latent spaces using diffusion process.
- We design a latent diffusion GAN loss to optimize the latent representations of generated results, and a heterogeneous conditional denoising loss to introduce image-level supervision for high-quality generation.
- We conduct extensive experiments on various unsupervised image-to-image tasks. The proposed LaDiffGAN exhibits state-of-the-art qualitative and quantitative results compared with previous methods without the assistance of extra guidance such as text description and sketch.

2. Related Work

2.1. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [20] and their variants are trained this way: the generator of GANs tries to generate results that can fool the discriminator, while the discriminator is updated by distinguishing real and fake images. So far, GANs have been applied to a variety of generative tasks, such as DALL-E text-to-image generation [52], super-resolution (SRGAN [38], ESRGAN [64], SFTGAN [73], and GLEAN [10]), and style editing (styleGAN [29] and styleGAN2 [30]). To address some of the challenges associated with GAN training, several techniques have been introduced. For example, Diffusion GAN [65], DiffAug [75], and ADA [28] aim to improve the stability of GAN

training. Meanwhile, ProjectGAN [56] incorporates a projection step to supervise the generated results, projecting both generated and real samples into the feature space of EfficientNet [60]. BigGAN [7] realizes high-quality multi-resolution image synthesis with big models and large batch sizes. However, many experimental results show that it is difficult for GANs to capture complex data distributions [65].

2.2. Diffusion Probabilistic Models

Deep diffusion models (DMs) are first introduced by Sohl-Dickstein et al. [57] as a novel generative model that generates samples by gradually denoising images corrupted by Gaussian noise. Recent DMs advances have demonstrated their superior performance in image synthesis, including DDPM [25], DDIM [58], ADM [17], LSGM [62], LDM [53], and DiT [50]. DDGAN [68] can reduce the number of sampling steps by directly predicting the ground truth in each timestep. DMs also have achieved state-of-the-art performance in other synthesis tasks, such as text-to-image generation demonstrated in GLIDE [48], speech synthesis [36, 43], and super-resolution (SRDiff [40] and SR3 [55]). Moreover, DMs have been applied to text-to-3D synthesis in DreamFusion [51] and Magic3D [42], and other 3D object synthesis in RenderDiffusion [3], diffusion SDF [41], and 3D point cloud generation [46]. More applications include video synthesis [24, 26], semantic segmentation [6], text-to-motion generation [61] and object detection [13]. Besides, SinDDM [37] and SinDiffusion [63] generate diverse results by learning the internal patch distribution from a single image. Overall, DMs have shown promising results and have been widely adopted in various synthesis tasks.

2.3. Unsupervised Image-to-Image Translation

Unsupervised image-to-image translation involves generating images in domain A based on input images in domain B without paired training data.

Several approaches have been proposed for unsupervised image-to-image translation. CycleGAN [76], DiscoGAN [32], and Dual-GAN [71] incorporate a cycle-consistency loss to preserve key attributes of input images during style transfer. Inspired by cVAE-GAN [5] and cLR-GAN [18, 19], BicycleGAN [77] jointly adopts the latent code and bi-directional generation to achieve performance improvement. MUNIT [27] and DRIT [39] decompose the latent codes of images into a domain-invariant content space and a domain-specific style space to obtain diverse outputs. StarGAN [16] enhances CycleGAN by enabling translation across multiple domains simultaneously. CoupledGAN [45], UNIT [44], ComboGAN [4], and XGAN [54] encode images from different domains into a shared latent space. AGGAN [1] leverages an attention model to distinguish foreground from background to improve the quality of generated images, but it cannot aid object shape transfer. CartoonGAN [14] performs well in cartoon-style synthesis but is unsuitable for object shape modification in images. U-GAT-IT [31] employs attention modules for feature selection. TransGAGA [67] and TravelGAN [2] represent latent features through Cartesian product and preserving vector arithmetic. NiceGAN [12] proposes a training strategy in which the discriminator's encoder is reused for the generator and only trained when maximizing the adversarial loss. DRB-GAN [69] and [11, 70, 72] utilize one model for multiple style image synthesis. CUT [49] and DCLGAN [21] attempt to alleviate the cycle loss's restrictiveness in GAN models. However, compared with the improvements brought by the cycle loss, its shortcomings are tolerable.

Recent methods, Prompt-to-Prompt [22], Instruct-Pix2Pix [8], and [52], use text models for object editing. CycleDiffusion [66] explores the shared latent space of diffusion models and can achieve unsupervised image-to-image translation by adopting pre-trained ILVR [15], SDEdit [47], and EGSDE [74]. However, when two domains differ greatly, it is challenging for CycleDiffusion to achieve style transfer. SinDDM [37] can achieve style transfer faithful to the internal statistics of training data. Nevertheless, these diffusion models face some issues, such as slow inference, large amounts of training data, and disclosure of dataset information.

3. Method

In this section, we present LaDiffGAN in detail, an efficient GAN model that effectively addresses the critical challenges posed by diffusion models and significantly improves the fidelity of GAN's outputs for unsupervised image-to-image translation. LaDiffGAN achieves this by encoding images into latent spaces and introducing two new loss functions that iteratively supervise the training process in the diffusion manner.

3.1. Overview of LaDiffGAN

Given the source domain X_s and target domain X_d, the goal of LaDiffGAN is to map images from X_s to high-fidelity results with the style of X_d. The overall training framework of LaDiffGAN is shown in Fig. 2. Like most classic GAN models for the image-to-image translation problem, LaDiffGAN consists of two generators $\mathcal{G}_{s\to d}$ and $\mathcal{G}_{d\to s}$ and two discriminators $\mathcal{D}_s$ and $\mathcal{D}_d$, which provide image-level supervision. Inspired by LDM [53] and Diffusion-GAN [65], we introduce a pre-trained autoencoder $\mathcal{E}_d$ to map X_d and the generated result $X_{s\to d}$ into the target latent space, and another pre-trained autoencoder $\mathcal{E}_s$ to map X_s and the generated result $X_{d\to s}$ into the source latent space. A diffusion Markov chain is then employed to iteratively inject Gaussian noise into the latent features, which helps to enrich the distributions of the generators' outputs. Moreover,

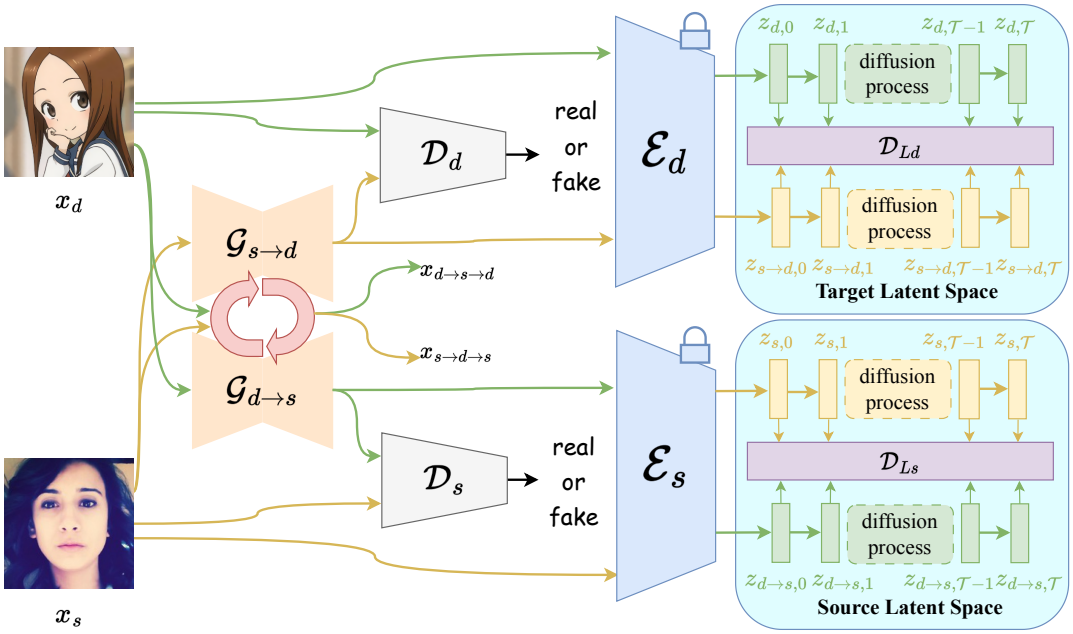

Figure 2. Overview of our LaDiffGAN. Initially, we obtain translated images using the generators $\mathcal{G}_{s\to d}$ and $\mathcal{G}_{d\to s}$ of a GAN model. We then encode these images into the latent spaces of the target and source domains to ensure the style consistency between the generated and real images. Finally, we employ the latent diffusion models to supervise the generated results in both the image and latent spaces, as depicted in Fig. 3. $\mathcal{E}_d$ and $\mathcal{E}_s$ are two autoencoders [53] pre-trained on the target and source data sets, respectively, and are fixed during LaDiffGAN training. $\mathcal{D}_d$ and $\mathcal{D}_s$ are two discriminators in the image space, while $\mathcal{D}_{Ld}$ and $\mathcal{D}_{Ls}$ are two discriminators in the latent spaces, respectively.

we incorporate two latent discriminators $\mathcal{D}_{Ls}$ and $\mathcal{D}_{Ld}$ to distinguish whether a latent feature is from a real or fake image. To obtain style-matching and high-quality results, we design two loss functions, latent diffusion GAN loss, and heterogeneous conditional denoising loss. These loss functions unify the image-level and latent-level supervision in one framework, providing further constraint for both the generators and the latent discriminators.

In essence, LaDiffGAN is a GAN model trained in the latent spaces using the diffusion process, and only $\mathcal{G}_{s\to d}$ is used to generate images with the style of X_d during inference. Accordingly, LaDiffGAN can achieve high-quality synthesis without extra guidance in real-time applications, and alleviate the risk of data leakage.

3.2. Latent Diffusion GAN Loss

Different from previous methods such as UNIT [44], ComboGAN [4] and XGAN [54] that encode images from different domains into a sharing latent space for generating the outputs conditioned on the latent code, LaDiffGAN encodes both generated results and training images into the target and source latent spaces. Moreover, the new latent diffusion GAN (LDG) loss is introduced to penalize the latent codes of generated images that are not well aligned with those of the training images, ensuring the style consistency between the generated and real images. Consequently, the latent spaces enable LaDiffGAN to capture the unique style characteristics of the target and source domains, facilitating the generation of high-quality results.

Specifically, given samples x_s and x_d from domains X_s and X_d, respectively, we first use the autoencoder $\mathcal{E}_d$ pre-trained on the target domain to map x_d and the generated result from x_s into the target latent space and extract their style features:

$$z_{s\to d},\ z_d = \mathcal{E}_d(\mathcal{G}_{s\to d}(x_s)),\ \mathcal{E}_d(x_d). \tag{1}$$

After that, we iteratively add Gaussian noise to $z_{s\to d}$ and z_d through the forward diffusion chain with an adaptive length $\mathcal{T}$. Formally, we define the distributions of the latent codes in timestep t by:

$$\begin{aligned} q(z_{s\to d,t}|z_{s\to d}) &= \mathcal{N}(z_{s\to d,t}|\sqrt{\gamma_t}z_{s\to d}, (1-\gamma_t)I), \\ q(z_{d,t}|z_d) &= \mathcal{N}(z_{d,t}|\sqrt{\gamma_t}z_d, (1-\gamma_t)I), \end{aligned} \tag{2}$$

where $\gamma_t \in (0,1)$ are the variances of the Gaussian noise in $\mathcal{T}$ iterations.

Then, we utilize the latent discriminator $\mathcal{D}_{Ld}$ to distinguish the fake and real latent codes at each timestep t.

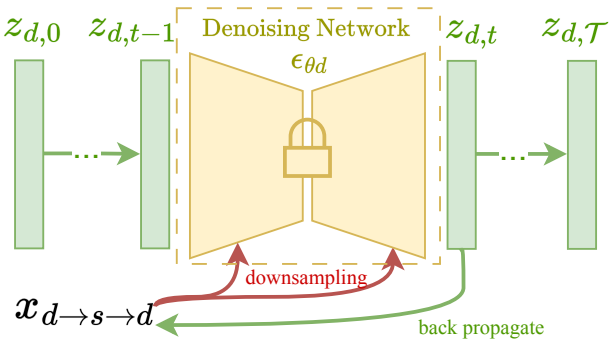

Figure 3. HCD loss computation. The denoising network $\epsilon_{\theta d}$ is pre-trained on the target data set and fixed during LaDiffGAN training.

Hence, the objective of the LDG loss $L_{LDG}^{s\to d}$ is for the generator $\mathcal{G}_{s\to d}$ to fool $\mathcal{D}_{Ld}$, so that the style of the generated results well aligns with the target style, which is defined below to optimize $\mathcal{D}_{Ld}$ and $\mathcal{G}_{s\to d}$:

$$L_{LDG}^{s\to d} = \mathbb{E}_{z_{d,t}\sim q(z_{d,t}|z_d)}[(\mathcal{D}_{Ld}(z_{d,t},t))^2] + \mathbb{E}_{z_{s\to d,t}\sim q(z_{s\to d,t}|z_{s\to d})}[(1-\mathcal{D}_{Ld}(z_{s\to d,t},t))^2]. \tag{3}$$

Supervising the latent codes on the target domain with different Gaussian noise levels enables the generator $\mathcal{G}_{s\to d}$ to effectively produce better images with the desired style, while making the latent discriminator $\mathcal{D}_{Ld}$ more robust to overfitting. Furthermore, the maximum timestep $\mathcal{T}$ is adaptively adjusted in every iteration during training. It is noteworthy that, since z_t tends to approach pure noise as t increases, we restrict $\mathcal{T}$ to $\{0,1,...,T/4\}$ to avoid undesirable interference with the training of the latent discriminator $\mathcal{D}_{Ld}$, where T is the step length of the diffusion chain in the pre-training of the denoising network $\epsilon_{\theta d}$ (see Section 3.3) and is set to 100 as DDIM [58]. $\mathcal{T}$ is adjusted as follows:

$$\mathcal{T} = \frac{T\cdot i}{4K},\ i\in\{0,1,...,K\}, \tag{4}$$

where K denotes the number of maximum training iterations. Following [65], we also sample t from $\{0,1,...,\mathcal{T}\}$ uniformly for the computation of $L_{LDG}^{s\to d}$.

In the above description of the LDG loss, we only define half of it $L_{LDG}^{s\to d}$. Another half $L_{LDG}^{d\to s}$ can be obtained similarly and is omitted here. Finally, $L_{LDG} = L_{LDG}^{s\to d} + L_{LDG}^{d\to s}$.

3.3. Heterogeneous Conditional Denoising Loss

In addition to the LDG loss which gives latent-level supervision, we accomplish image-level supervision by designing the heterogeneous conditional denoising (HCD) loss to further improve the quality of the generated results.

Conditional Latent Denoising Models. Before the training of LaDiffGAN, we follow LDM [53] to train a conditional latent denoising module with 100 DDIM denoising timesteps on the target latent space. Let $\epsilon_{\theta d}$ be the denoising network, which is trained by predicting the noise ϵ from $z_{d,t}$ conditioned on the downsampled images $x_d\downarrow$. The pre-training objective is defined as:

$$L_{LDM}^{s\to d} = \mathbb{E}_{z_{d,t},x_d,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_{\theta d}(z_{d,t},x_d\downarrow,t)\|_2^2], \tag{5}$$

where $\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, and $\downarrow$ represents the downsampling operation. Similarly, another denoising network $\epsilon_{\theta s}$ is pre-trained with x_s, whose objective $L_{LDM}^{d\to s}$ is defined similar to Eq. 5

HCD Loss. With the pre-trained conditional latent diffusion models, we design the HCD loss to construct a relation between the latent and image spaces. Unlike the commonly used cycle loss on image translation tasks, which directly matches the target image x_d with the image $x_{d\to s\to d}$ generated by the target-to-source and source-to-target mappings in the image space, the HCD loss leverages the pre-trained diffusion model $\epsilon_{\theta d}$ to match the generated image space with the latent space of the target domain, as shown in Fig. 3. The generation of $x_{d\to s\to d}$ is expressed as:

$$x_{d\to s\to d} = \mathcal{G}_{s\to d}(\mathcal{G}_{d\to s}(x_d)), \tag{6}$$

and half of the HCD loss is defined as:

$$L_{HCD}^{s\to d} = \mathbb{E}_{z_{d,t},\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_{\theta d}(z_{d,t},x_{d\to s\to d}\downarrow,t)\|_2^2]. \tag{7}$$

$L_{HCD}^{s\to d}$ uses $x_{d\to s\to d}$ as the image-level condition in the latent denoising network $\epsilon_{\theta d}$. Since $\epsilon_{\theta d}$ is fixed during LaDiffGAN training, minimizing $L_{HCD}^{s\to d}$ will update the generators such that their generated images $x_{d\to s\to d}$ can help the generation of the target style.

Similarly, we can obtain the other half of the HCD loss $L_{HCD}^{d\to s}$. Finally, $L_{HCD} = L_{HCD}^{s\to d} + L_{HCD}^{d\to s}$.

Note that compared with the cycle loss that imposes strict image alignment between the target/source and generated images, our HCD loss allows for more variations in the generated images due to the diffusion process.

3.4. Other Loss Functions

Following [31], we also introduce the commonly used loss functions for unsupervised image-to-image translation, including adversarial loss L_{GAN}, cycle loss L_{cycle}, identity loss $L_{identity}$ and CAM loss L_{CAM}. On the whole, with the pre-trained autoencoders and denoising networks, we train LaDiffGAN to optimize the total objective:

$$\begin{aligned} L_{total} =& \alpha_1 L_{LDG} + \alpha_2 L_{HCD} + L_{other}, \\ L_{other} =& \lambda_1 L_{GAN} + \lambda_2 L_{cycle} + \\ & \lambda_3 L_{identity} + \lambda_4 L_{CAM}, \end{aligned} \tag{8}$$

Table 1. Kernel Inception Distances obtained by difference image translation models. Lower is better.

Methods	selfie2anime	horse2zebra	cat2dog	vangogh2photo
DRIT	4.38 ± 0.40	3.35 ± 0.74	2.69 ± 0.87	6.47 ± 0.89
MUNIT	3.71 ± 0.39	2.68 ± 0.63	2.70 ± 0.47	5.54 ± 0.82
UNIT	5.14 ± 0.41	5.09 ± 0.92	2.65 ± 0.48	4.05 ± 0.43
U-GAT-IT	2.62 ± 0.30	2.77 ± 0.70	0.85 ± 0.28	2.33 ± 0.44
DCLGAN	12.7 ± 0.55	2.71 ± 0.43	1.27 ± 0.29	3.11 ± 0.46
LaDiffGAN (ours)	**1.97 ± 0.23**	**2.61 ± 0.73**	**0.75 ± 0.26**	**2.16 ± 0.44**
Methods	anime2selfie	zebra2horse	dog2cat	photo2vangogh
DRIT	4.51 ± 0.48	3.40 ± 0.37	3.58 ± 0.47	6.75 ± 0.80
MUNIT	3.49 ± 0.37	4.78 ± 0.65	3.46 ± 0.34	23.5 ± 0.85
UNIT	4.04 ± 0.36	4.30 ± 0.56	2.05 ± 0.30	3.78 ± 0.51
U-GAT-IT	1.57 ± 0.26	2.70 ± 0.50	0.82 ± 0.27	2.71 ± 0.60
DCLGAN	7.99 ± 0.62	5.74 ± 0.63	1.06 ± 0.23	18.8 ± 1.16
LaDiffGAN (ours)	**1.52 ± 0.31**	**2.68 ± 0.49**	**0.76 ± 0.26**	**2.70 ± 0.61**

where $\alpha_1 = 0.2$, $\alpha_2 = 0.01$, $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 10$, and $\lambda_4 = 1000$ are set empirically. More details of these loss functions are provided in the supplementary materials.

4. Experiments

In addition to the extensive experiments described in this section, we also provide more results in the supplementary materials.

4.1. Implementation Details

Datasets. We evaluate the performance of LaDiffGAN on four datasets: selfie2anime, horse2zebra, cat2dog, and vangogh2photo. (1) The selfie2anime dataset is presented in U-GAT-IT [31], both the selfie and anime styles of which have 3400 and 100 images for training and testing, respectively. (2) The horse2zebra and vangogh2photo datasets are built in CycleGAN [76], with the amounts of the classes for training being: 1,067 (horse), 1,334 (zebra), 6,287 (photo), and 400 (vangogh), and for testing: 120 (horse), 140 (zebra), 751 (photo), and 400 (vangogh). Note that the training data and the testing data of vangogh2photo are the same. (3) The cat2dog dataset is used in DRIT [39], with the amounts of the classes for training: 871 (cat) and 1,364 (dog), and for testing: 100 (cat) and 100 (dog). All images are resized to 256 × 256 for training.

Training Details. We adopt 4 A100-SXM4-40GB for training. All of the models are trained by Adam [33] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. To accelerate the training, we adopt the pre-trained U-GAT-IT as the based model, which was trained for 400,000 iterations, and is further trained for 100,000 iterations in LaDiffGAN.

Evaluation Metric. We adopt the Kernel Inception Distance (KID) as the quantitative metric to evaluate the image quality, which computes the squared Maximum Mean Discrepancy between the features of real and fake images, where the features are extracted by the Inception model [59]. Unlike Frechet Inception Distance (FID) [23], KID is an unbiased metric, making it more reliable than FID, especially with limited testing images. The lower KID indicates that the generated images look more similar to the real images [1].

4.2. Comparison with State-of-the-Art Methods

Methods. We compare our LaDiffGAN with state-of-the-art (SOTA) GAN-based methods including DRIT [39], MUNIT [27], UNIT [44], U-GAT-IT [31], and DCLGAN [21], and one diffusion-based method CycleDiffusion (LDM+DDIM) [66]. We use their official code to train these models on each dataset for 500,000 iterations.

Quantitative Comparison. The comparison on unsupervised image-to-image translation is shown in Table 1. LaDiffGAN achieves the best performance on all four benchmark datasets, each of which contains two tasks, such as selfie2anime and anime2selfie. Especially, LaDiffGAN obtains pronounced improvements on the selfie2anime task. Although the latest approach DCLGAN shows impressive performance on the cat2dog and horse2zebra datasets, it fails to obtain satisfactory results on both selfie2anime and vangogh2photo. In contrast, LaDiffGAN succeeds in generating best results regardless of the amounts of changes in both shape and texture between different domains.

Qualitative Comparison. A qualitative comparison is given in Fig. 4, demonstrating that our LaDiffGAN generates better images with high visual quality and consistent styles with the target domains, particularly when tackling the challenging selfie2anime task. Although U-GAT-IT pro-

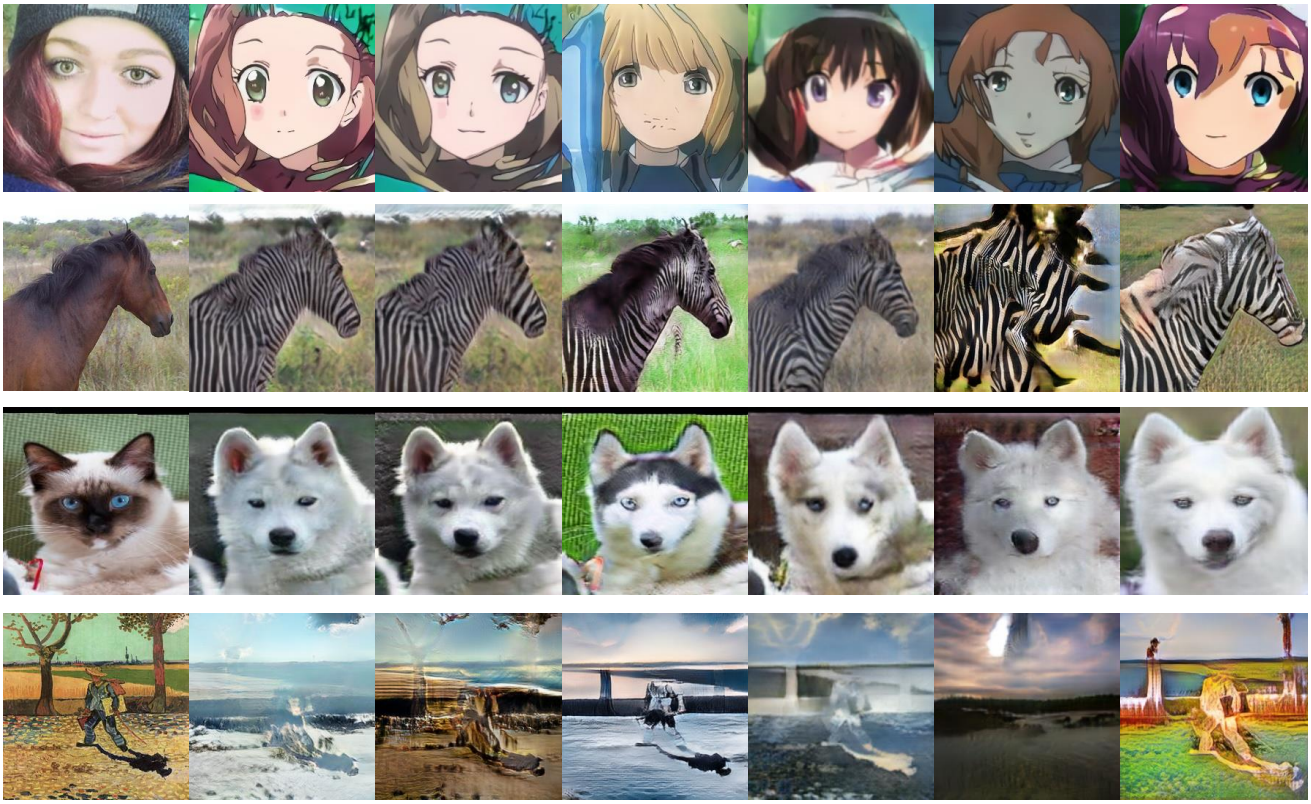

Figure 4. Qualitative comparison with state-of-the-art GAN-based models.

Table 2. Ablation study on the selfie2anime dataset.

Methods	selfie2anime	anime2selfie
Baseline	2.62 ± 0.30	1.57 ± 0.26
Baseline + LDG loss	2.12 ± 0.26	1.58 ± 0.31
Baseline + HCD loss	2.44 ± 0.24	1.53 ± 0.32
LaDiffGAN	**1.97 ± 0.23**	**1.52 ± 0.31**

Table 3. Quantitative comparison with CycleDiffusion and DiffusionGAN on selfie2anime and anime2selfie.

Methods	selfie2anime	anime2selfie
CycleDiffusion	4.11 ± 0.59	8.94 ± 0.67
U-GAT-IT+DiffGAN	2.61 ± 0.28	1.75 ± 0.32
LaDiffGAN (ours)	**1.97 ± 0.23**	**1.52 ± 0.31**

duces comparable quantitative results with LaDiffGAN, it suffers from distortions (e.g., the eyes in the first row of Fig. 4 (c)) or undesired styles (e.g., the fourth row of Fig. 4 (c)).

4.3. Ablation Study

We conduct a comprehensive ablation study to verify the effectiveness of the proposed LDG loss and HCD loss. Specifically, we utilize the selfie2anime dataset for evaluation, which exhibits significant stylistic differences across the domains. We use U-GAT-IT [31] as the baseline, and construct three comparison models: (1) the baseline trained with the LDG loss, (2) the baseline trained with the HCD loss, (3) and the full model of LaDiffGAN with both the losses. The results are shown in Table 2. We can see that the two losses both have significant effects. Moreover, Fig. 5 intuitively shows how the two losses work on the selfie2anime task. Concretely, compared with the baseline U-GAT-IT, the generated result of the baseline trained with the LDG loss attains more high-quality style, while that of the baseline trained with the HCD loss exhibits better visual quality with fine textures. Meanwhile, our full model LaDiffGAN produces the best overall result and outperforms other models.

4.4. Further Analysis

We further analyze the limitations of diffusion models for unsupervised image synthesis when the amount of training data is limited. To this end, we conduct a comparative analysis between the performances of LaDiffGAN and CycleDiffusion (LDM+DDIM) [66]. CycleDiffusion is designed to perform multiple tasks such as text-to-image gen-

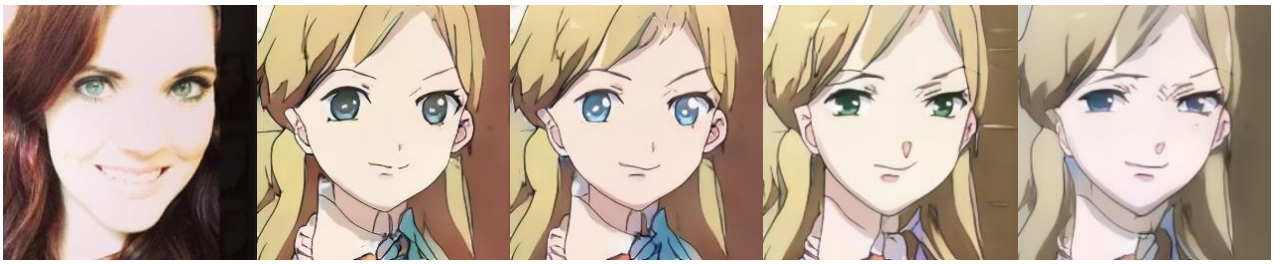

Figure 5. Visual results of the ablation study.

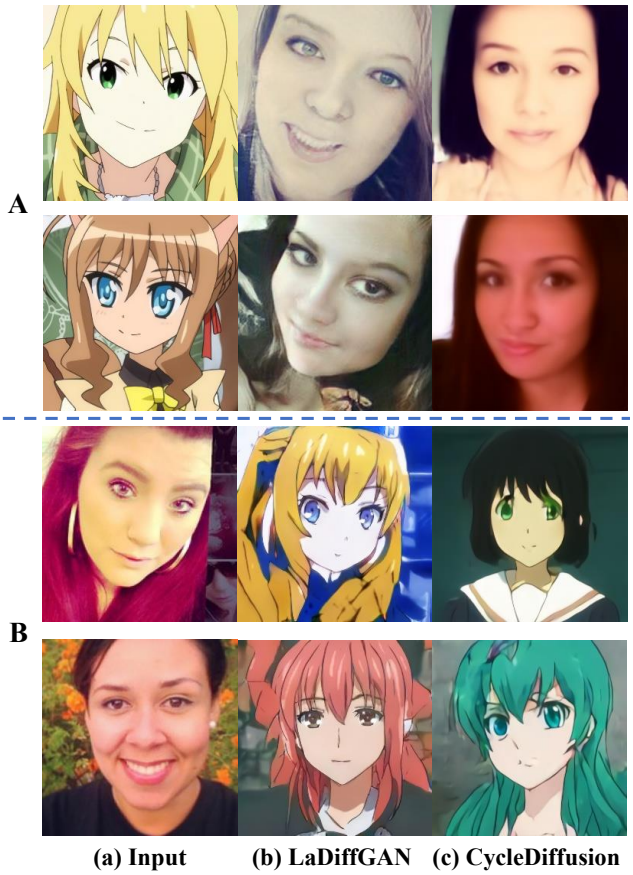

Figure 6. Row A: anime2selfie; Row B: selfie2anime.

eration and wild-animal-to-dog translation. We train it on the selfie2anime dataset. The quantitative results are given in Table 3. Evidently, CycleDiffusion performs poorly on the image translation task with a large domain gap. We also provide the qualitative comparison in Fig. 6, it is obvious that LaDiffGAN exhibits much more consistent results with the input than CycleDiffusion.

Moreover, we compare our LaDiffGAN with U-GAT-IT trained by Diffusion-GAN [65] on the selfie2anime dataset. As shown in Table 3, LaDiffGAN delivers better performance on both the selfie2anime and anime2selfie tasks, demonstrating its superior ability to Diffusion-GAN. Further qualitative comparisons are provided in the supplementary materials.

4.5. Inference Time Comparison

We verify the efficiency of LaDiffGAN by comparing its inference time with CycleDiffusion. Taking frames-per-second (FPS) as the evaluation metric, we randomly select 100 samples for testing. As a result, the FPS of LaDiffGAN is 35.54, much faster than that of CycleDiffusion (0.21), demonstrating the great efficiency of LaDiffGAN. Note that since LaDiffGAN only uses the generator $\mathcal{G}_{s\to d}$ for inference, its time cost is the same as U-GAT-IT.

5. Conclusion

Despite the great success of diffusion models in image synthesis, GANs still matter for unsupervised image-to-image translation. In this paper, we propose LaDiffGAN which trains GANs with diffusion supervision in latent spaces. We employ two autoencoders to map images into two specific latent spaces with styles of the target and source domains, and then introduce a latent diffusion GAN loss to align the latent codes between generated and training images. Moreover, we design a heterogeneous conditional denoising loss that incorporates image-level supervision to further enhance the quality of generated results. Extensive experiments illustrate that our LaDiffGAN achieves state-of-the-art performance.

Acknowledgements. The work was supported by the National Key Research and Development Program of China (2023YFC3300029), Zhejiang Provincial Natural Science Foundation of China(LD24F020007), Beijing Natural Science Foundation (L223024), National Natural Science Foundation of China (62076016), "One Thousand Plan" projects in Jiangxi Province (Jxsg2023102268), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Z231100005923035).

References

[1] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, 2018. 3, 6

[2] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *CVPR*, 2019. 3

[3] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. *arXiv:2211.09869*, 2022. 3

[4] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *CVPRW*, 2018. 3, 4

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *ICCV*, 2017. 3

[6] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2021. 3

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 3

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv:2211.09800*, 2022. 1, 3

[9] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv:2301.13188*, 2023. 1

[10] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[11] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *CVPR*, 2017. 3

[12] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *CVPR*, 2020. 3

[13] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv:2211.09788*, 2022. 3

[14] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *CVPR*, 2018. 3

[15] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 3

[16] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[18] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2016. 3

[19] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2016. 3

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 2020. 1, 2

[21] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *CVPR*, 2021. 3, 6

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022. 1, 3

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022. 3

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 3

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3, 6

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 2

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[31] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2019. 3, 5, 6, 7

[32] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 3

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[34] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 1

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[36] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv:2009.09761*, 2020. 3

[37] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model. *arXiv:2211.16582*, 2022. 3

[38] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken,

Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[39] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3, 6

[40] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022. 3

[41] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. *arXiv:2212.03293*, 2022. 3

[42] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv:2211.10440*, 2022. 3

[43] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Diffusion acoustic model for singing voice synthesis. *arXiv:2105.02446*, 2021. 3

[44] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 3, 4, 6

[45] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NeurIPS*, 2016. 3

[46] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 3

[47] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 3

[48] 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:2112.10741*, 2021. 3

[49] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 3

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv:2212.09748*, 2022. 3

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

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2, 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 4, 5

[54] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*. 2020. 3, 4

[55] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 3

[56] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *NeurIPS*, 2021. 3

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 3

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 3, 5

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[60] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 3

[61] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv:2209.14916*, 2022. 3

[62] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *NeurIPS*, 2021. 3

[63] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Sindiffusion: Learning a diffusion model from a single natural image. *arXiv:2211.12445*, 2022. 3

[64] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 2

[65] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv:2206.02262*, 2022. 2, 3, 5, 8

[66] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv:2210.05559*, 2022. 1, 3, 6, 7

[67] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *CVPR*, 2019. 3

[68] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *ICLR*, 2022. 3

[69] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *CVPR*, 2021. 3

[70] Zheng Xu, Michael Wilber, Chen Fang, Aaron Hertzmann, and Hailin Jin. Learning from multi-domain artistic images for arbitrary style transfer. *arXiv:1805.09987*, 2018. 3

[71] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 3

[72] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. In *ECCVW*, 2018. 3

[73] Yutian Zhang, Xiaohua Li, and Jiliu Zhou. Sftgan: a generative adversarial network for pan-sharpening equipped with spatial feature transform layers. *JARS*, 2019. 2

[74] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. In *NeurIPS*, 2022. 3

[75] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 2

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 3, 6

[77] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 3