Latent-CycleGAN for Domain Shift [Group 7]

Heyuan Chi
4742393
Scientific Computing

Yutong Zeng
4732327
Data and Computer Science

Abstract

Generative Adversarial Networks (GANs) have shown remarkable success in unpaired image-to-image translation tasks, exemplified by the CycleGAN framework. However, GAN-based methods often encounter issues such as mode collapse and insufficient preservation of fine-grained details. On the other hand, diffusion models exhibit robust structural fidelity but are limited by high computational overhead and slower inference speeds. To address these challenges, we propose Latent-CycleGAN, a novel architecture that integrates a pre-trained Stable Diffusion model into the CycleGAN framework. By leveraging LoRA-based fine-tuning and introducing additional latent-space discriminators alongside cycle-consistency constraints, our approach effectively combines the strengths of GAN- and diffusion-based methodologies. Experimental results demonstrate that Latent-CycleGAN produces higher-fidelity translations with enhanced detail retention, improved style consistency, and greater training stability compared to conventional CycleGAN. This work provides a promising new direction for unpaired image translation and broader applications in unsupervised domain adaptation.

All source code and additional information are available: https://github.com/HeyuanChi/latent-cycle-gan

1. Introduction

Image-to-image (I2I) conversion is the task of transforming an image from one domain to another while preserving key content. A famous example is image style transfer, where the goal is to change the style of an image without altering its underlying content structure. Early work by Gatys et al. [3] demonstrated that deep convolutional features could be disentangled into separate representations for content and style, making it possible to generate images that combine the content of one source with the style of another. This finding established the groundwork for style transfer and, more broadly, for domain shift problems in computer vision, which are often formulated as special cases of I2I transformation.

In recent years, many new methods employing deep generative models have been investigated for I2I translation, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models. Among them, GAN-based approaches are notably successful, thanks to adversarial objectives that guide the generator to produce outputs resembling real images in the target domain. Furthermore, recent developments in diffusion models show promise for stable training and producing diverse outputs, though with higher computational overhead. These advancements have opened up new possibilities, such as zero-shot editing—using powerful pretrained models to perform style transformations without additional fine-tuning.

However, the main problem of diffusion models is that it is hard to ensure the consistence of between target and source image. In details, the structure of object in image could be lost and the edge of object might be distorted. In this case, the process could be considered as create a new image rather than domain shift.

In this paper, we propose a novel image-to-image translation framework, Latent-CycleGAN, which enhances the classic CycleGAN by integrating a pre-trained Stable Diffusion model and employing LoRA-based fine-tuning. Compared with using diffusion model to generate target image directly, our method introduces additional latent space discriminators and cycle-consistency constraints alongside traditional image-space adversarial training, thereby preserving structural fidelity and ensuring style consistency. Extensive experiments on datasets such as vangogh2photo, horse2zebra, and cat2dog demonstrate that our approach significantly improves visual quality and detail retention. This work offers a new perspective for unsupervised domain adaptation and style transfer, setting the stage for further advancements in the field.

2. Related Works

2.1. GAN-based Image Translation

Traditional I2I methods generally rely on paired data, which could be expensive or impractical to collect. CycleGAN [23] addressed this by introducing a cycle consistency loss, ensuring that an image translated from domain A to B and then back to A remains unchanged. Subsequent dual-generator architectures, such as DiscoGAN [11] and DualGAN [22], adopted similar cycle constraints. UNIT [15] further proposed a shared latent space for unsupervised translation. Although these GAN-based methods are effective, they often encounter issues like mode collapse and may not preserve fine-grained details when the domains differ significantly in semantics or style.

Still, GAN-based models are widely used and have shown impressive results in tasks such as photo coloring, image-to-image conversion, and style transformation. For instance, pix2pix [10] used paired training images, while CycleGAN [23] succeeded even without paired data by leveraging a cyclic loss. Enhanced techniques—like StarGAN [2] for multi-domain translation, BicycleGAN [25] for multi-modal outputs, and frameworks like MUNIT [9] or DRIT [14] for decomposing content and style—have made GAN-based methods increasingly flexible and controllable.

2.2. Diffusion-Based Approaches

Denoising Diffusion Probabilistic Models (DDPMs) [6] have emerged as a strong alternative to GANs, capable of generating high-fidelity images by reversing a progressive noising process. Improvements such as DDIM [20] and classifier-free guidance [7] further enhance generation speed and quality. More recently, latent diffusion models, as exemplified by Stable Diffusion [17], reduce computational cost by operating in a compressed latent space. These models can be conditioned on text or images to produce new outputs that preserve source content while adopting the desired style. Compared with GANs, diffusion methods tend to be more stable during training and yield diverse outputs, but often require longer sampling times.

2.3. Combining Diffusion and GANs

Recent efforts have explored hybrid approaches that integrate diffusion models with GANs to exploit the strengths of both. Methods like VQ-Diffusion [4] and DiffusionGAN [21] incorporate diffusion priors to regularize GAN outputs, improving training stability and structural consistency. Along this line, our work inserts a latent diffusion module into the CycleGAN framework, using a pretrained Stable Diffusion UNet as a noise predictor in the latent space. This additional constraint enforces structural integrity beyond pixel-level losses.

2.4. Low-Rank Adaptation

Large models such as Stable Diffusion contain billions of parameters, making full fine-tuning infeasible in many applications. Low-Rank Adaptation (LoRA) [8] addresses this issue by inserting low-rank matrices into the attention layers of pretrained networks, enabling task-specific adaptation with minimal computational overhead. In our framework, we adopt LoRA to fine-tune the diffusion UNet modules, retaining the powerful generative prior of the original model while achieving effective, domain-specific customization.

3. Method

In this section, we present our method in detail. Building upon the CycleGAN framework [24], we incorporate two pretrained Stable Diffusion models [18] (fine-tuned by LoRA [8]), and propose novel latent-space constraints to improve the translation quality.

3.1. CycleGAN (baseline)

CycleGAN [24] aims to learn a mapping between two image domains, denoted as A and B. It does not require to use paired training data. As shown in the middle of figure 1, the framework includes two generators and two discriminators:

- G_A: a generator translating images from domain A to domain B.
- G_B: a generator translating images from domain B to domain A.
- D_A: a discriminator distinguishing real B-domain images from translated (fake) images produced by G_A.
- D_B: a discriminator distinguishing real A-domain images from translated (fake) images produced by G_B.

3.1.1 Generator Losses

1. Image-Space Adversarial Loss

   CycleGAN uses adversarial objectives in the image space. For generator G_A, which maps $x \in A$ to $\tilde{x} = G_A(x) \in B$, the corresponding discriminator D_A tries to classify real images $y \sim B$ as real and $\tilde{x}$ as fake. In a least-squares GAN (LSGAN) formulation [16]:

   $$\mathcal{L}_{\text{GAN}}(G_A, D_A) = \mathbb{E}_{x\sim A}\big[(D_A(G_A(x)) - 1)^2\big].$$

   Similarly for G_B and D_B:

   $$\mathcal{L}_{\text{GAN}}(G_B, D_B) = \mathbb{E}_{y\sim B}\big[(D_B(G_B(y)) - 1)^2\big].$$

2. Cycle-Consistency Loss

   To ensure learned mappings are reversible, CycleGAN enforces

   $$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_A, G_B) = \mathbb{E}_{x\sim A}\big[\|G_B(G_A(x)) - x\|_1\big] \\ +\mathbb{E}_{y\sim B}\big[\|G_A(G_B(y)) - y\|_1\big].\end{aligned}$$

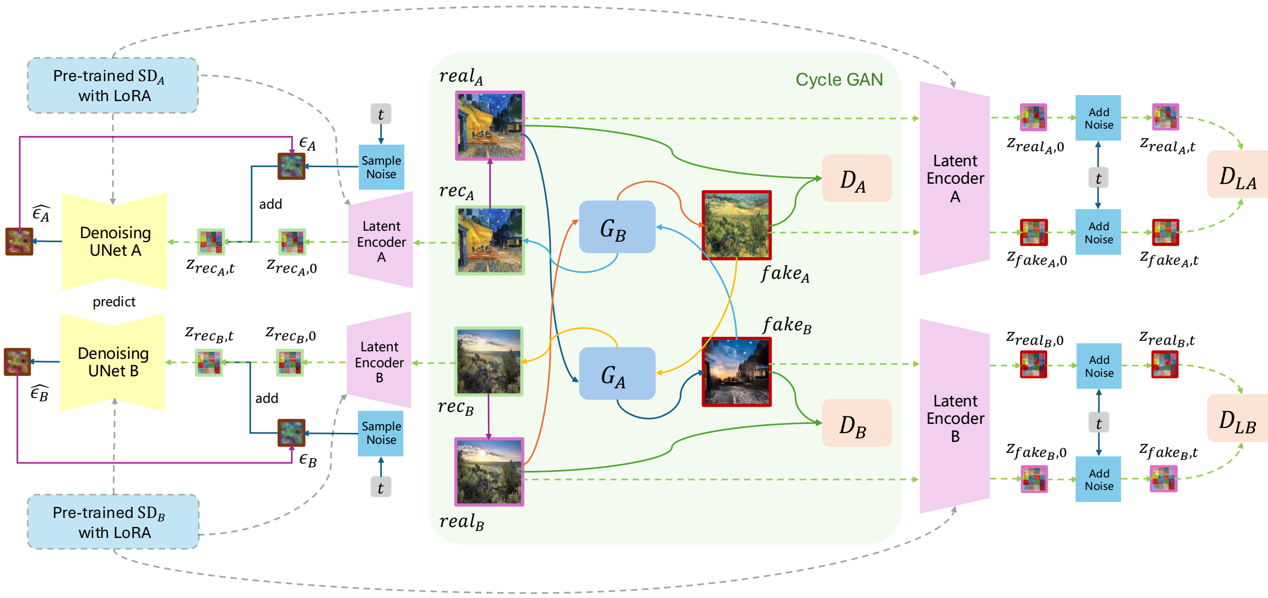


Figure 1: **Method Overview.** We start with the standard CycleGAN setup: two generators (G_A, G_B) translating between domains A and B, and two discriminators (D_A, D_B) for real-vs-fake image supervision. We then integrate a pretrained Stable Diffusion (SD) model [18] (optionally fine-tuned via LoRA [8]) and introduce two new latent-space discriminators (D_{LA}, D_{LB}). Each image is encoded by the SD VAE into latent space, partially noised at a small diffusion timestep, and fed to the latent discriminators. This combined image- and latent-level training enforces cycle consistency in both spaces, yielding more faithful and style-consistent image translations.

This penalizes large deviations between the reconstructed image and the original.

3. Identity Loss

An optional identity loss [24] can be used to encourage generators to preserve color or other low-level features:

$$\mathcal{L}_{\text{id}}(G_A, G_B) = \mathbb{E}_{y\sim B}\big[\|G_A(y) - y\|_1\big] + \mathbb{E}_{x\sim A}\big[\|G_B(x) - x\|_1\big].$$

3.1.2 Discriminator Losses

Each discriminator is trained with a real vs. fake objective. For D_A,

$$\mathcal{L}_{D_A} = \tfrac{1}{2}\Big[\mathbb{E}_{y\sim B}\big[(D_A(y)-1)^2\big]+\mathbb{E}_{x\sim A}\big[(D_A(G_A(x)))^2\big]\Big],$$

and similarly for D_B,

$$\mathcal{L}_{D_B} = \tfrac{1}{2}\Big[\mathbb{E}_{x\sim A}\big[(D_B(x)-1)^2\big]+\mathbb{E}_{y\sim B}\big[(D_B(G_B(y)))^2\big]\Big].$$

3.2. Stable Diffusion and LoRA

In our work, we incorporate Stable Diffusion (SD) [18] as a powerful pretrained model for guidance in the latent space. Unlike conventional image-to-image translation methods, SD encodes images into a latent representation and applies a denoising UNet to generate high-quality samples.

The core of Stable Diffusion models consists of:

- A Variational Autoencoder (VAE): encodes an input image into a latent code and decodes a latent code back into the image space.

- A Denoising UNet: trained via a diffusion process to remove noise from latent representations at various timesteps.

When an image x is provided, the VAE encoder produces a latent code $z = \text{VAE.encode}(x)$. During sampling or inference, the UNet iteratively denoises a noisy latent until it converges on a clean sample, which the VAE decoder then maps back to image space.

To adapt the pretrained SD model to specific styles or domains, we can apply LoRA [8], a parameter-efficient fine-tuning technique. LoRA injects low-rank adaptations into the weight matrices of the UNet, reducing the number of additional trainable parameters while preserving most of the original SD capabilities. This fine-tuning step allows SD to capture domain-specific cues without overfitting or forgetting the general knowledge from the original training.

3.3. Latent-CycleGAN: Our Proposed Method

We now introduce **Latent-CycleGAN** as shown in figure 1, which augments the standard CycleGAN training with latent-space discriminators and latent cycle-consistency constraints. This leverages Stable Diffusion (optionally fine-tuned with LoRA) to better guide the translation process at the latent level.

In addition to the image-space discriminators D_A and D_B, we introduce new latent discriminators:

- D_{LA}: operates in the SD latent space for domain-B images, evaluating real vs. fake latents.
- D_{LB}: operates in the SD latent space for domain-A images, evaluating real vs. fake latents.

These discriminators aim to enforce that the translated samples (once encoded by the SD VAE) remain realistic in the diffusion latent space.

3.3.1 Latent Generator Losses

1. Latent Adversarial Loss

   For $x \in A$, let $\tilde{y} = G_A(x)$. We encode both real y (from B) and fake $\tilde{y}$ with the SD VAE, obtaining $z_{\text{real}} = \text{VAE.encode}(y)$ and $z_{\text{fake}} = \text{VAE.encode}(\tilde{y})$. To add mild stochasticity, we choose a small diffusion timestep t and inject noise according to the forward diffusion equation:

   $$z_{\text{real}}^{(t)} = \sqrt{\alpha_t}\, z_{\text{real}} + \sqrt{1-\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

   and analogously for $z_{\text{fake}}^{(t)}$. The latent discriminator D_{LA} is trained with:

   $$\mathcal{L}_{\text{GAN-latent}}(G_A, D_{LA}) = \mathbb{E}\big[(D_{LA}(z_{\text{fake}}^{(t)}) - 1)^2\big].$$

   Similarly for G_B and D_{LB}.

2. Latent Cycle-Consistency Loss

   To maintain consistency through the two-stage translation, we apply the SD forward diffusion to the reconstructed images and then measure alignment in noise space. For example, let $\hat{x}_A = G_B(G_A(x))$. The VAE encodes $\hat{x}_A$ into $z_{\hat{x}_A}$, which is then diffused to timestep t. The diffusion UNet predicts the noise ϵ_{pred}, which we compare to the actual noise ϵ added:

   $$\mathcal{L}_{\text{cyc-lat-A}} = \mathbb{E}\big[\|\hat{\epsilon}(z_{\hat{x}_A}) - \epsilon_A\|_2^2\big],$$

   with an analogous term for $\hat{x}_B$,

   $$\mathcal{L}_{\text{cyc-lat-B}} = \mathbb{E}\big[\|\hat{\epsilon}(z_{\hat{y}_B}) - \epsilon_B\|_2^2\big].$$

3.3.2 Latent Discriminator Losses

The discriminators D_{LA} and D_{LB} are optimized similarly to standard LSGAN:

$$\mathcal{L}_{D_{LA}} = \tfrac{1}{2}\Big[\mathbb{E}\big((D_{LA}(z_{\text{real}_A}^{(t)}) - 1)^2\big) + \mathbb{E}\big((D_{LA}(z_{\text{fake}_A}^{(t)}))^2\big)\Big],$$

$$\mathcal{L}_{D_{LB}} = \tfrac{1}{2}\Big[\mathbb{E}\big((D_{LB}(z_{\text{real}_B}^{(t)}) - 1)^2\big) + \mathbb{E}\big((D_{LB}(z_{\text{fake}_B}^{(t)}))^2\big)\Big].$$

3.3.3 Full Objective

Finally, we combine the image-space and latent-space losses with weighting factors. Denote λ_{GAN}, λ_{cyc}, λ_{id}, α_{GAN}, α_{cyc}. The total generator objective is:

$$\begin{aligned}\mathcal{L}_G = \; & \lambda_{\text{GAN}}\Big(\mathcal{L}_{\text{GAN}}(G_A, D_A) + \mathcal{L}_{\text{GAN}}(G_B, D_B)\Big) \\ & + \lambda_{\text{cyc}}\, \mathcal{L}_{\text{cyc}}(G_A, G_B) \; + \; \lambda_{\text{id}}\, \mathcal{L}_{\text{id}}(G_A, G_B) \\ & + \alpha_{\text{GAN}}\Big(\mathcal{L}_{\text{GAN-lat}}(G_A, D_{LA}) + \mathcal{L}_{\text{GAN-lat}}(G_B, D_{LB})\Big) \\ & + \alpha_{\text{cyc}}\, \Big(\mathcal{L}_{\text{cyc-lat-A}}(G_A, G_B) + \mathcal{L}_{\text{cyc-lat-B}}(G_B, G_A)\Big).\end{aligned}$$

Each discriminator has a matching real-vs-fake loss:

$$\mathcal{L}_D = \{\mathcal{L}_{D_A}, \mathcal{L}_{D_B}, \mathcal{L}_{D_{LA}}, \mathcal{L}_{D_{LB}}\}.$$

3.3.4 Implementation Details

We implement our approach in PyTorch, building upon the original CycleGAN codebase [24]. The generators (G_A, G_B) and image-space discriminators (D_A, D_B) use ResNet or U-Net backbones. The latent discriminators (D_{LA}, D_{LB}) are adapted from PatchGAN [10] to operate on four-channel latent inputs from Stable Diffusion's VAE. The SD model can remain frozen or be partially updated using LoRA [8] to capture domain-specific styles. We randomly sample a diffusion timestep t from a small range (early in the diffusion schedule) so that some semantic information remains while still challenging the latent discriminators with mild noise.

4. Experiments and Results

In this section, we present our experiments and evaluations. We begin by introducing the datasets and metrics, followed by a description of training details. We then detail the LoRA fine-tuning approach and results, continue with our main Latent-CycleGAN framework along with different hyperparameter configurations, and conclude with additional experiments, a full set of quantitative metrics, and qualitative comparisons.

4.1. Datasets

We primarily evaluate the proposed Latent-CycleGAN on the three dataset: vangogh2photo, horse2zebra and cat2dog. The vangogh2photo and horse2zebra datasets are built in CycleGAN [24]. The amounts of the classes for training are 6,287 (photo), and 400 (vangogh) in vangogh2photo; and 1,067 (horse), 1,334 (zebra) in horse2zebra. And for testing are : 751 (photo), and 400 (vangogh) in vangogh2photo; 120 (horse), 140 (zebra) in horse2zebra. (The training data and the testing data of vangogh2photo are the same.) The cat2dog dataset is used in DRIT [14]. The amounts for training are 871 (cat) and 1,364 (dog). And for testing are 100 (cat) and 100 (dog). All images are resized to 256 × 256 for training.

4.2. Evaluation Metrics

To evaluate the performance of our method, we employ Kernel Inception Distance (KID) as the quantitative measure. KID introduced by Bińkowski et al.[1], is an unbiased alternative to the widely used FID [5]. FID often produces biased estimates, especially with smaller datasets, since it relies on Gaussian assumptions and covariance matrix inversion. In contrast, KID adopts a polynomial-kernel MMD, making it distribution-free, more robust in low-data scenarios, and numerically simpler. Empirical studies show that KID correlates with FID on large samples but outperforms it with limited data[13][19]. Thus, KID is recommended when evaluating generative models with small datasets.

4.3. Training Details

All experiments were conducted on a server equipped with One NVIDIA A100-SXM4-80GB GPU and 120 GB of system RAM (AutoDL Cloud Server). We implemented our models in PyTorch using Python 3.8. The Stable Diffusion VAE and UNet components were loaded from CompVis Stable Diffusion v1.5[18] and remained partially frozen unless otherwise stated.

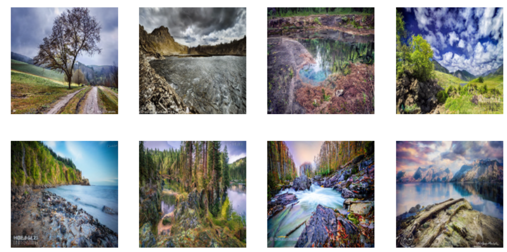
We used the Adam optimizer [12] ($\beta_1 = 0.5$, $\beta_2 = 0.999$) for all networks and applied standard data augmentation (random cropping, flipping) for both the A and B domains. Batch size was set to 16 with an initial learning rate of 1e-4 (decayed after some epochs to zero).

4.4. LoRA Fine-Tuning

To explore the benefit of stylistic adaptation within the diffusion model, we incorporated LoRA adapters [8] into the Stable Diffusion UNet. Specifically, we inserted low-rank adapters ($r = 8$) into each cross-attention layer to capture domain-specific style cues for images. We fine-tuned LoRA parameters on each dataset for 5000 epochs, with a reduced learning rate of 1e-5, while freezing most of the pretrained diffusion weights.

Figure 2 highlights representative outputs on vangogh2photo datasets, obtained by simply sampling from the LoRA-adapted diffusion model. These LoRA-enhanced diffusion modules are then integrated into our final pipeline to serve as the latent encoder/decoder backbone.

(a) Generated Van Gogh's paints by LoRA-SD.



(b) Generated real photots by LoRA-SD.

Figure 2: The generated figures on two domains by LoRA-SD models which are trained on corresponding datasets.

By fine-tuning the original SD models through the LoRA method, we can obtain models with stronger generalization ability in the corresponding field. In this way, we can use them to guide CycleGAN later.

4.5. Latent-CycleGAN

First, we trained the models on vangogh2photo dataset. We tested different hyperparameter configurations for our Latent-CycleGAN framework. Then we did additional experiments on other datasets.

4.5.1 Experiments on vangogh2photo

Recall that we introduced two additional weighting factors: α_{GAN} for latent-space adversarial loss and α_{cyc} for latent cycle-consistency. We examined three settings:

1. **Set A:** $\alpha_{\text{GAN}} = 1$, $\alpha_{\text{cyc}} = 0.1$, $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{id}} = 10$.

2. **Set B:** $\alpha_{\text{GAN}} = 1$, $\alpha_{\text{cyc}} = 0.1$, $\lambda_{\text{GAN}} = 0$, $\lambda_{\text{cyc}} = 0$, $\lambda_{\text{id}} = 0$.

3. **Set C:** $\alpha_{\text{GAN}} = 0.1$, $\alpha_{\text{cyc}} = 0.01$, $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{id}} = 10$.

For the first two sets, we keep the original image-space adversarial weight $\lambda_{\mathrm{GAN}} = 1$, cycle-consistency weight $\lambda_{\mathrm{cyc}} = 10$, and identity weight $\lambda_{\mathrm{id}} = 10$. We also match the architecture of the image discriminators and generators to the standard CycleGAN [24].

4.5.2 Qualitative Observations

We first trained the base CycleGAN model for 200 epochs (about 100K iterations). And then we trained the Latent-CycleGAN model with different parameter sets for 50 epochs (about 25K iterations) based on the original CycleGAN states.

Figure 3 (a) and (b) illustrate examples of photo to Van Gogh and Van Gogh to photo translations, respectively, under each parameter set. We compare our Latent-CycleGAN results against a baseline model that applies vanilla CycleGAN directly on images (with no latent-space adversarial or cycle terms).

From the results of translating real photos to Van Gogh style, sets A and C show only marginal improvements over the baseline. Among these, set C exhibits more accurate colors, making it slightly better than the baseline. When translating Van Gogh paintings to real photos, both sets A and C outperform the baseline, with set C once again producing more realistic color rendering. In contrast, set B, which completely removes the original loss, performs quite poorly. It suggests that the latent loss can provide assisted guidance, but it cannot fully replace the original loss.

4.6. Additional Experiments

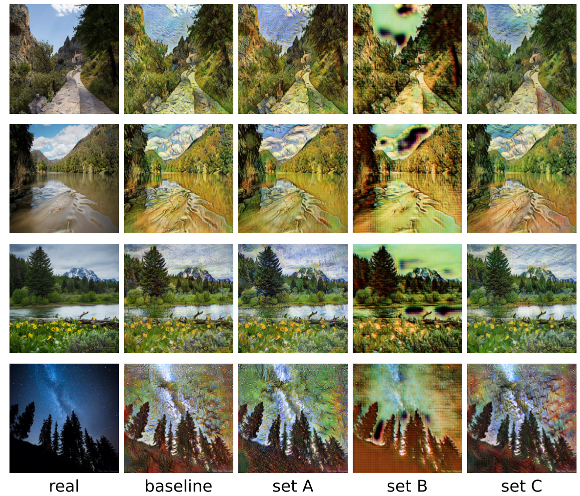
To further validate the robustness of our approach, we also conducted experiments on other datasets (horse2zebra and cat2dog). All hyperparameters follow the same setting as in the set C. Figure 4 illustrates some translated examples.

4.7. Quantitative Comparison

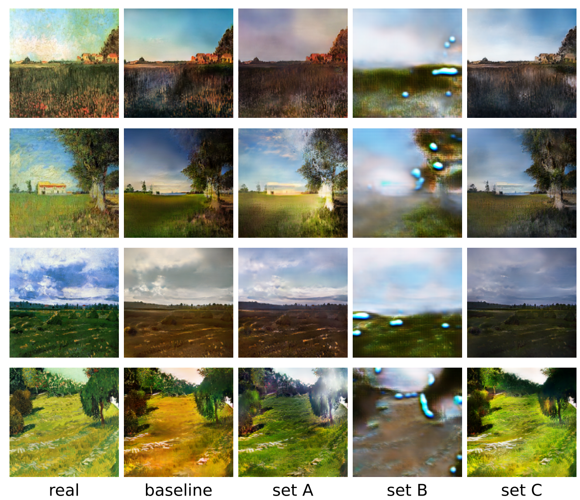
We computed the aforementioned quantitative metric KID on each dataset and compared our Latent-CycleGAN against the standard CycleGAN models (baseline).

Table 1 presents the comprehensive metrics on different datasets. Our models achieved a lower KID except on each dataset for vangogh2photo. From these experiments, we observe that incorporating LoRA fine-tuning and latent-space adversarial/cycle-consistency consistently improves both perceptual quality and style accuracy across diverse domains.

Based on these results, we conclude that LoRA-based fine-tuning and latent-space adversarial training together enhance both the fidelity and the style-consistency of unpaired image translation across multiple datasets. This highlights the general applicability of our proposed framework.



(a) Example of translation results from real photos to Van Gogh style.



(b) Example of translation results from Van Gogh's Paintings to real style.

Figure 3: Visualization results of image translation based on the Latent-CycleGAN framework.

Table 1: Comparison of Kernel Inception Distances on different datasets. Lower is better.

Dataset	Baseline	Latent-CycleGAN(Ours)
vangogh2photo	**0.028 ± 0.004**	0.031 ± 0.004
photo2vangogh	0.029 ± 0.005	**0.027 ± 0.005**
horse2zebra	0.090 ± 0.006	**0.076 ± 0.005**
zebra2horse	0.129 ± 0.004	**0.125 ± 0.004**
cat2dog	0.226 ± 0.005	**0.166 ± 0.005**
dog2cat	0.161 ± 0.003	**0.133 ± 0.003**

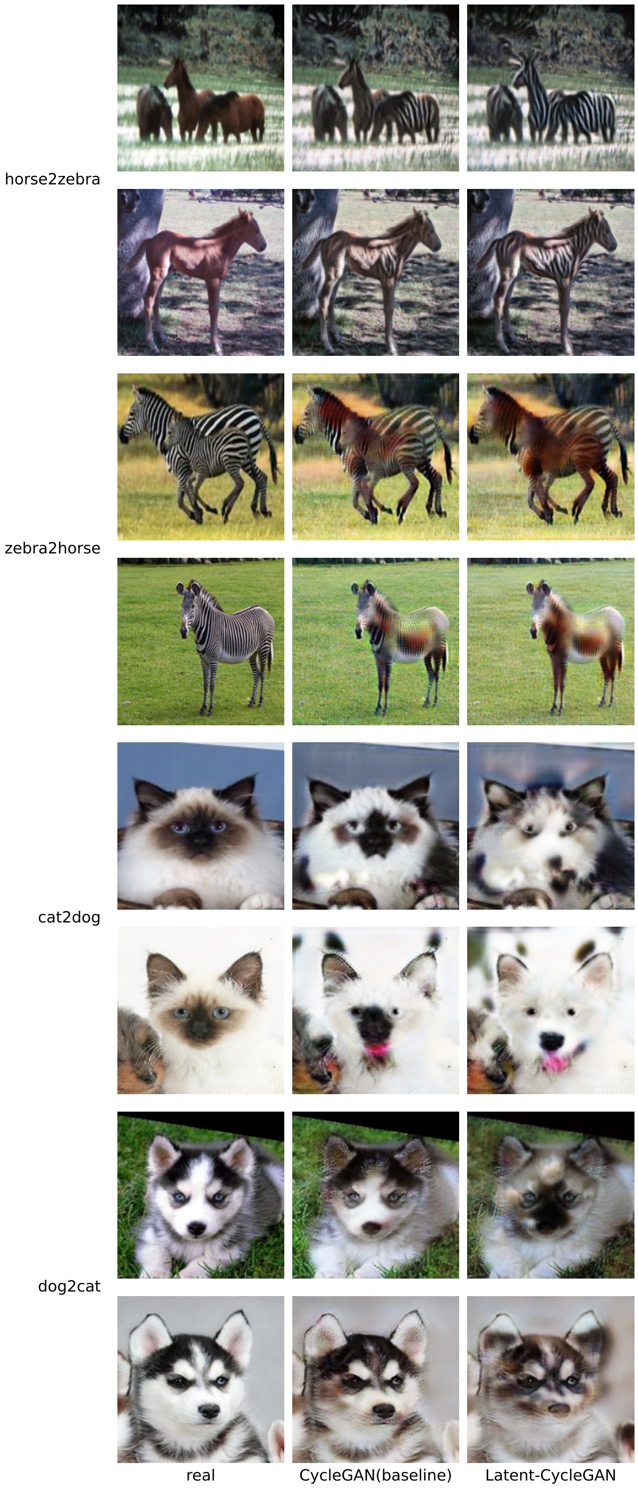


Figure 4: Visualization results of image translation on datasets horse2zebra and cat2dog based on the Latent-CycleGAN framework.

5. Conclusion

In this work, we propose **Latent-CycleGAN**, a framework that combines the classic CycleGAN with a pretrained Stable Diffusion model (optionally fine-tuned via LoRA) for unpaired image-to-image translation. Unlike the conventional CycleGAN, which relies solely on adversarial training in the image space, our method adds extra discriminators and cycle-consistency constraints in the latent space of Stable Diffusion, thereby preserving structural fidelity and enhancing visual quality while maintaining flexibility in style transformation.

Key findings are as follows:

1. LoRA fine-tuning helps the pretrained diffusion model adapt better to specific domain styles, providing more stable and accurate guidance for image translation tasks.

2. Latent-space adversarial loss and latent cycle-consistency loss, coupled with the image-space losses, consistently improve both perceptual quality and style accuracy across multiple datasets. Although the KID metric may not always change dramatically, the overall results are more balanced in terms of detail, color, and structure.

3. For domains with significant cross-domain variations, incorporating constraints in the latent space effectively mitigates mode collapse or loss of detail, preserving essential semantic and structural information.

In summary, our study demonstrates that integrating large-scale diffusion models into CycleGAN's latent-space learning can yield richer style transfer, more diverse outputs, and better image quality in unpaired domain mapping tasks. Going forward, we plan to explore the following directions:

1. Multimodal and multi-style integration: Incorporating text prompts or other auxiliary inputs to enable more controllable outputs.

2. Higher resolution and large-scale datasets: Investigating the potential for ultra-high-resolution generation and more diverse large datasets.

3. Training efficiency and adaptability: Further refining the synergy between LoRA-based parameter-efficient fine-tuning and latent-space constraints, aiming to reduce training costs and improve generality.

By preserving the ease of training and unpaired learning advantages of CycleGAN while leveraging the powerful latent-space modeling capabilities of Stable Diffusion, Latent-CycleGAN can offer a new perspective and stronger benchmarks for future research in unsupervised image translation and domain adaptation.

References

[1] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[2] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 2

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[4] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 2

[5] 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 *Advances in Neural Information Processing Systems (NIPS)*, pages 6626–6637, 2017. 5

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhuo Li, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. arXiv preprint arXiv:2106.09685. 2, 3, 4, 5

[9] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 2, 4

[11] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. Pmlr, 2017. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in GANs. In *International Conference on Machine Learning (ICML)*, pages 3581–3590, 2019. 5

[14] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision (ECCV)*, pages 36–52, 2018. 2, 5

[15] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–708, 2017. 2

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 2

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

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

[19] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5228–5237, 2018. 5

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

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

[22] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2, 3, 4, 5, 6

[25] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476, 2017. 2