Comparative Analysis of Image Style Transfer Using Diffusion and GAN Architectures

Chengke Zou, Yankai Mao, Zhongyuan Cao

1 Introduction

Generative AI is undeniably one of the most important and fastest-moving fields in computer science. Within this rapidly evolving landscape, neural style transfer represents a particularly well-defined and instructive problem. It combines the creativity of visual synthesis with the rigor of algorithmic design, offering a contained yet meaningful challenge for empirical exploration.

This project focuses on neural style transfer as a means to examine and compare key generative architectures. Specifically, it investigates two representative paradigms that have shaped the field in recent years: a Generative Adversarial Network (GAN) model and a modern Diffusion model. Both will be implemented and trained on the ArtBench dataset, which contains a diverse set of high-quality oil paintings. The resulting models will be evaluated through established quantitative metrics to assess their relative performance and to highlight their respective advantages and limitations in practical style transfer tasks.

2 Related Work

Image style transfer has long been a central topic in example-guided artistic image generation. Early approaches relied on handcrafted low-level features to match local patches between content and style images. More recently, pre-trained deep convolutional neural networks have been employed to capture feature distributions, enabling a more effective representation of complex style patterns.

2.1 Style Transfer with Diffusion model

Diffusion models are inspired by non-equilibrium thermodynamics (7). They define a Markov chain of diffusion steps that slowly add random noise to the data, and then learn to reverse the diffusion process to construct desired data samples from the noise. Building upon this principle, Denoising Diffusion Probabilistic Models (DDPM) (10) provided a tractable and scalable framework for training deep generative models using a simple Gaussian noise schedule and a reweighted variational objective, significantly improving image fidelity over prior likelihood-based methods.

Subsequent developments refined both training and sampling efficiency, such as Improved DDPM and Score-based Generative Models (SGMs). These advances enabled controllable and high-quality image generation, leading to large-scale diffusion systems such as GLIDE, Imagen, and Stable Diffusion, which further incorporate text conditioning via CLIP or transformer-based encoders.

For visual creativity, UnCLIP inverted the CLIP representation to guide diffusion sampling, allowing semantic control over generation consistent with textual and visual embeddings.

More recently, Inversion-Based Style Transfer with Diffusion Models (11) formalized the idea of reconstructing and editing specific diffusion trajectories to transfer styles between arbitrary images. By inverting the diffusion process to the latent noise space and re-synthesizing with altered conditions, these methods achieve fine-grained and faithful style transfer while preserving content structure.

2.2 Style Transfer with GANs

Research on Image-to-Image (I2I) translation, a task strongly shaped by Generative Adversarial Networks (GANs), started with supervised approaches. Pix2Pix (2) established an early baseline for paired translation, where each input image has a corresponding target.

The challenge of handling unpaired data, which is much more common in real-world settings, was later addressed by CycleGAN (3). Its introduction of the cycle-consistency loss made it possible to train models without aligned image pairs, as demonstrated in tasks such as photo-to-painting translation.

Follow-up research sought to improve both scalability and controllability. StarGAN v2 (4) proposed a multi-domain framework with an instance-level style encoder, enabling diverse image synthesis from reference examples.

More recently, studies have moved toward text-guided control, made possible by large-scale vision-language models. StyleGAN-NADA (5), for instance, showed that a pre-trained StyleGAN can be adapted to new domains (e.g., "a photo $\rightarrow$ a sketch") using only text prompts from CLIP, removing the need for explicit style references.

3 Method and Algorithm

3.1 CycleGAN

In first method, We adopt the classic GAN-based framework, CycleGAN(3), to address the style transfer problem between real photos and artworks. Unlike traditional style transfer methods requiring paired datasets, CycleGAN performs unpaired image-to-image translation by jointly learning two mappings $G : \mathcal{X} \rightarrow \mathcal{Y}$ and $F : \mathcal{Y} \rightarrow \mathcal{X}$, along with discriminators $D_\mathcal{X}$ and $D_\mathcal{Y}$ that enforce realism in each domain. The adversarial objective encourages generated samples to be indistinguishable from real ones:

$$\mathcal{L}_{GAN}(G, D_\mathcal{Y}) = \mathbb{E}_{y \sim p_\mathcal{Y}(y)}[\log D_\mathcal{Y}(y)] + \mathbb{E}_{x \sim p_\mathcal{X}(x)}[\log(1 - D_\mathcal{Y}(G(x)))].$$

To preserve content, a cycle-consistency loss is introduced:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_x[\|F(G(x)) - x\|_1] + \mathbb{E}_y[\|G(F(y)) - y\|_1].$$

The full objective is $\mathcal{L} = \mathcal{L}_{GAN}(G, D_\mathcal{Y}) + \mathcal{L}_{GAN}(F, D_\mathcal{X}) + \lambda \, \mathcal{L}_{cyc}(G, F)$, where λ balances realism and reconstruction.

In our project, we plan to use ArtBench(1) as the painting dataset (domain $\mathcal{Y}$) and ImageNet as the real-world photo dataset (domain $\mathcal{X}$) to train and evaluate the model.

3.2 Inversion-Based Style Transfer with Diffusion Models

The diffusion model we selected here is the Inversion-based Style Transfer model (11), which is built upon Stable Diffusion models (12). InST learns a textual embedding directly from a single reference image to generate new artistic images in the same style without fine-tuning the diffusion model. Uses the CLIP text encoder to represent a new, learnable concept token. The embedding for this concept, denoted $\hat{v}$ is optimized so that the generated image matches the target artistic image.

$$\hat{v} = \arg\min_v \mathbb{E}_{x,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(z_t, t, v(y))\|_2^2 \right]$$

To improve efficiency and generalization, the model introduces multi-layer cross attention: Image embeddings from the CLIP encoder are projected through attention layers to extract key image information. The cross-attention mechanism updates query, key, and value at each layer as:

$$Q_i = W_Q^{(i)} \cdot Q_{i-1}, \quad K_i = W_K^{(i)} \cdot \tau_\theta(y), \quad V_i = W_V^{(i)} \cdot \tau_\theta(y)$$
$$v_{i+1} = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i$$

Finally, a stochastic inversion step introduces controlled noise to preserve content and enable stylistic diversity, computed as:

$$\hat{\epsilon}_t = (z_{t-1} - \mu_T(z_t, t))\sigma_t$$

3.3 Evaluation

We evaluate results using two complementary metrics. The Structural Similarity Index (SSIM)(8) measures how well the generated image preserves the content and structure of the original, while a CLIP-based similarity score(9) assesses semantic and stylistic alignment between the output and target style description. Higher SSIM and CLIP similarity indicate better content retention and style consistency.

References

[1] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. *The ArtBench Dataset: Benchmarking Generative Models with Artworks*. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[3] 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.

[4] Yunjey Choi, Youngik Uh, Jaejun Yoo, and Jung-Woo Ha. *StarGAN v2: Diverse Image Synthesis for Multiple Domains*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8187–8197, 2020.

[5] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Zada, and Daniel Cohen-Or. *StyleGAN-NADA: CLIP-Guided Domain Adaptation of Pretrained StyleGANs*. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. *ImageNet: A Large-Scale Hierarchical Image Database*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[7] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. *Deep Unsupervised Learning Using Nonequilibrium Thermodynamics*. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[8] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. *ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 862–871, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. *Learning Transferable Visual Models from Natural Language Supervision*. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. *Inversion-Based Style Transfer With Diffusion Models*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] 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)*, 2022.