A Comprehensive Survey of Style Transfer: Techniques, Models, and Applications

Weiqi Wang[1,2,3], Weiting Wang[1,2,3,*], Ying Xu[1,2,3,*], Feilong Bao[1,2,3(✉)], Zhecong Xing[3], Zhiguo Zhang[3], Yuan Zhang[3]

[1] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot 010000, China
[2] Inner Mongolia Key Laboratory of Mongolian Information Processing Technology College of Computer Science, Inner Mongolia University, Hohhot 010000, China
[3] College of Computer Science, lnner Mongolia University, Hohhot 010000, China
csfeilong@imu.edu.cn

Abstract. Style transfer learning has garnered significant attention in the field of computer vision in recent years, with anime style transfer being particularly notable for its entertaining nature and widespread application. This fascinating feature has been integrated into various short video platforms, mini-programs, and photography applications. According to our survey, nearly one in four individuals has used applications based on anime style transfer models, with most users providing positive feedback, citing its novelty, fun, and high playability. This paper provides a comprehensive summary of the technological advancements in style transfer, focusing on four mainstream methods: Convolutional Neural Networks (CNN), Variational Autoencoders (VAE), Vision Transformers (ViT), and Generative Adversarial Networks (GAN). We detail the implementation of specific models for each method and systematically compare the performance of several representative models. Finally, we include links to the open-source code of these models to facilitate further research and application.

Keywords: Style Transfer, Convolutional Neural Networks (CNN), Variational Autoencoders (VAE), Vision Transformer (ViT), Generative Adversarial Networks (GAN), Image Processing.

1 Introduction

With the development of deep learning technology and the advent of the big data era, style transfer technology has made significant progress in the fields of computer vision and image processing. Style transfer is an image processing technique [1, 2] aimed at applying the artistic style of one image to another while maintaining the content consistency of the target image. This field has evolved through multiple stages, from early basic methods to current deep learning approaches, achieving substantial advancements.

The earliest style transfer techniques can be traced back to 1999 when Efros and Leung [3] proposed a texture synthesis method based on Markov Random Fields. This

*These autors contributed equally to this work.

method achieved texture synthesis by matching and stitching local image patches. It was capable of generating high-quality textures and was suitable for simple texture images. However, this approach was not suitable for complex images and styles, and the generation process was relatively slow.

As the information age progressed, visual information became increasingly abundant, and images grew more complex. Consequently, these early methods were gradually replaced by more advanced technologies. In 2001, Hertzmann et al. [4] proposed a filtering-based method for style transfer, which achieved style transfer through image filtering and matching. This method had the advantage of being applicable to a variety of style transfer tasks, but it heavily relied on input images and had limited effectiveness in handling complex styles.

In the following decade, style transfer technology did not attract widespread attention. It wasn't until 2013, when Kingma and Welling [5] introduced the Variational Autoencoder (VAE) method, and 2015, when Gatys et al. [6] proposed a neural network-based approach, that style transfer regained significant interest. These advancements brought renewed focus to the field, highlighting the potential of deep learning techniques in achieving more sophisticated and effective style transfer results.

Since 2017, style transfer technology has experienced a surge in development, leading to the emergence of several models based on Generative Adversarial Networks (GANs), such as CycleGAN [7] and DualGAN [8]. Simultaneously, van den Oord et al. optimized the VAE-based [9] method and proposed the VQ-VAE model [10]. These advancements have garnered significant attention and research interest from scholars and experts, further driving the evolution and application of style transfer techniques.

In recent years, Transformer-based models [11] have also begun to be applied in the field of style transfer, such as Vision Transformer (ViT) [12] and DALL-E [13]. Within the realm of Generative Adversarial Networks (GAN) [14,15], several noteworthy models have emerged, including UGATIT [16], DualStyleGAN [17], and Vtoonify [18]. These models have demonstrated outstanding performance in image style transfer, particularly in anime style transfer.

Anime style transfer [19,20,21] is a specialized application of style transfer, aiming to transform images or videos into those with a specific anime style. The implementation of anime style transfer typically relies on deep learning techniques such as Convolutional Neural Networks (CNN) [22,23] and Generative Adversarial Networks (GAN) [24,25,26]. By training models to learn and extract the characteristics of anime styles, these features can then be applied to target images to achieve the desired style transfer. The subsequent comparison of model performances will focus on this particular application.

In this paper, we introduce the convolutional neural network (CNN)-based, variational auto-encoder (VAE)-based, Vision Transformer (ViT)-based, and generative adversarial network (GAN)-based approaches for style migration in Sections 2, 3, 4, and 5, respectively. Each section lists representative models of the above methods and describes their structure and advantages and disadvantages. In Section 6, the performance of these representative models is compared and links to the open source code of the different models are provided. Section 7 summarizes the full text.

2 Methods Based on Convolutional Neural Networks (CNN)

With the development of deep learning, Convolutional Neural Networks (CNN) [27] have demonstrated powerful capabilities in image processing tasks. CNNs extract image features through multi-layer convolution operations, enabling style transfer while preserving the content of the image. The core of style transfer technology lies in effectively separating and recombining the content and style features of images.

2.1 Working Principle

The system uses neural representations to separate and recombine the content and style of arbitrary images, providing a neural algorithm for creating artistic images. Convolutional Neural Networks (CNNs) are composed of multiple layers of small computing units that process visual information in a hierarchical, feedforward manner. Each layer's units can be viewed as a set of image filters used to extract specific features from the input image. The output of each layer is a so-called feature map, which represents different filtered versions of the input image.

When a CNN is trained for object recognition, it gradually reveals the object's information through the processing layers, allowing the higher-level feature maps to focus more on the actual content of the image rather than on detailed pixel values. To obtain the style representation of an input image, a feature space is utilized, which was initially designed to capture texture information. This feature space is based on the correlations between the filter responses in each layer of the network. By including feature correlations from multiple layers, a multi-scale representation of the input image is obtained, capturing its texture information rather than its global arrangement.

Advantages of This Method:

- High-quality image generation: CNN-based style transfer methods can generate high-quality images with rich details.
- Strong feature extraction capability: CNNs can extract complex features from images, making the style transfer effect more natural.

Disadvantages of This Method:

- High computational complexity: The multi-layer structure of CNNs results in a large computational load and long training times.
- Dependence on training data: A large amount of high-quality training data is required to ensure the model's effectiveness.

2.2 Representative Model - Neural Style Transfer

Neural Style Transfer leverages Convolutional Neural Networks (CNNs) to achieve high-quality image style transfer. The core idea of this method is to separately extract the features of the content image and the style image, and then apply the style features to the content image through an optimization process. The model structure is shown in **Fig. 1**.

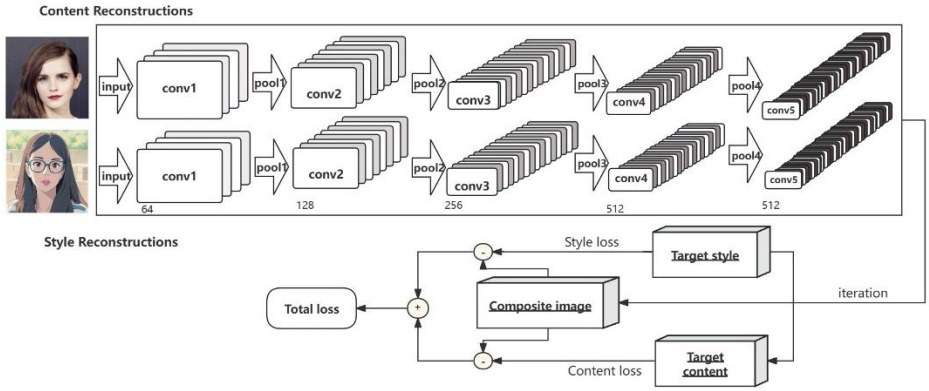

Fig. 1. Neural Style Transfer Model Structure

Content Representation: Using a specific layer of the pre-trained VGG network [28,29], the content features of the input image are extracted. These feature maps capture the high-level content information of the image at the deeper layers of the network. Style Representation: The style features of the style image are extracted using multiple layers of the VGG network, and the Gram matrices [30,31]of these feature maps are computed. These Gram matrices represent the correlations between feature maps at different layers, capturing the texture information of the style image. Image Synthesis: An initial white noise image is optimized to match the content features of the content image and the style features of the style image simultaneously. Specifically, a loss function is defined that includes both content loss and style loss components. The optimization process aims to minimize this loss function. The form of the loss function is as follows:

$$L_{total}(p, a, x) = \alpha L_{content}(p, x) + \beta L_{style}(a, x) \tag{1}$$

Where $L_{content}$ is the content loss, measuring the difference in content features between the generated image and the content image; L_{style} is the style loss, measuring the difference in style features between the generated image and the style image; α and β are weighting factors used to balance the importance of content and style.

3 Methods Based on Variational Autoencoders (VAE)

Variational Autoencoder (VAE) is a deep generative model that utilizes variational Bayesian inference to achieve efficient approximate inference and learning. By encoding input data into a continuous latent space and generating data samples from this latent space, VAEs have successfully demonstrated superior performance in various image generation tasks. VAEs are particularly well-suited for style transfer tasks, as their encoder-decoder structure can effectively extract and reconstruct the content and style features of images.

3.1 Working Principle

VAE consists of an encoder and a decoder. The encoder maps the input image to a distribution of latent variables, while the decoder generates images from these latent variables. Specifically, given a dataset $X = \{x^{(i)}\}_{i=1}^{N}$, the data is assumed to be generated by an unobserved continuous random variable z through the following process: first, z is sampled from a prior distribution $p_\theta(z)$; then, data x is generated from the conditional distribution $p_\theta(x|z)$.

Since the posterior distribution $p_\theta(z|x)$ is typically intractable, VAE introduces an approximate inference model $q_\phi(z|x)$ and uses variational inference for approximation. By optimizing the variational lower bound $L(\theta,\phi;x)$, the model parameters can be efficiently learned.

Advantages of This Method:

- Efficient inference and learning: VAE utilizes variational inference, achieving efficient approximate posterior inference and parameter learning.
- Powerful generative capability: VAE can generate high-quality images, making it suitable for various image generation and style transfer tasks.

Disadvantages of This Method:

- Reconstruction quality limitation: The reconstruction quality of VAE may be limited by its assumptions about the latent variable distribution, especially with high-dimensional data.
- High computational complexity: The training process involves a significant amount of parameter optimization and computation, particularly when handling large-scale datasets.

3.2 Representative Model - VQ-VAE

VQ-VAE (Vector Quantized Variational Autoencoder) is an improved version of the Variational Autoencoder that enhances the diversity and quality of generated images by introducing a vector quantization mechanism. The core idea of VQ-VAE is to use a discrete latent space, thereby avoiding the continuous assumption of latent variables in VAE, which improves the stability and diversity of the generated images. The model structure is shown in **Fig. 2**.

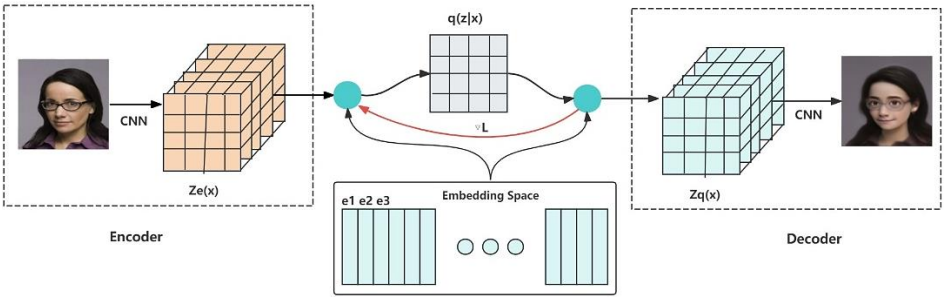

Fig. 2. VQ-VAE Model Structure

Encoding Process: The input image is encoded by the encoder (typically a Convolutional Neural Network) into a set of continuous latent representations. These latent representations are then quantized to the nearest vectorized code, generating discrete latent variables z_q. Decoding Process: The discrete latent variables are decoded by the decoder (typically a Convolutional Neural Network) to generate the image. Optimization Process: The model parameters are optimized by defining a loss function that includes reconstruction loss and vector quantization loss. The reconstruction loss measures the difference between the generated image and the original image, while the vector quantization loss ensures consistency between the continuous latent representations output by the encoder and the discrete vectorized codes. The form of the loss function is as follows:

$$L_{total} = L_{recon} + \beta L_{vq} \tag{2}$$

where L_{recon} is the reconstruction loss, L_{vq} is the vector quantization loss, and β is a weighting factor that balances the importance of the two loss components.

4 Methods Based on Vision Transformer (ViT)

Vision Transformer (ViT) is a pure Transformer model applied to computer vision tasks. Unlike traditional Convolutional Neural Networks (CNNs), ViT directly processes sequences of image patches and uses self-attention mechanisms to achieve image classification and style transfer. The design of ViT is inspired by the successful application of Transformers in the field of Natural Language Processing (NLP).

4.1 Working Principle

The ViT model divides the input image into fixed-size patches, then flattens and embeds these patches into a linear vector space. These embedded patch sequences are fed into a standard Transformer encoder for feature extraction and processing. Unlike CNNs, ViT does not rely on local receptive fields and convolution operations but captures contextual information of the image globally through the self-attention mechanism.

Advantages of This Method:

- Global Context Capture: The self-attention mechanism can capture contextual information of the image globally, enhancing the model's understanding of both image details and overall structure.
- High Scalability: ViT can handle large-scale datasets and performs exceptionally well after large-scale pre-training.
- Strong Generalizability: ViT can be directly applied to various vision tasks, such as image classification, image generation, and style transfer.

Disadvantages of This Method:

- High Data Requirements: Due to the lack of prior knowledge from convolution operations, ViT does not perform as well as CNNs on small-scale datasets and requires large-scale data for pre-training.

- High Computational Resource Requirements: The self-attention mechanism in Transformers has high computational complexity, demanding significant hardware resources.

4.2 Representative Model - Vision Transformer (ViT)

ViT was proposed by the Google Brain team. Its core idea is to transform the image processing problem into a sequence modeling problem. By processing sequences of image patches through a Transformer encoder, it achieves tasks such as image classification and style transfer. The model structure is shown in **Fig. 3**.

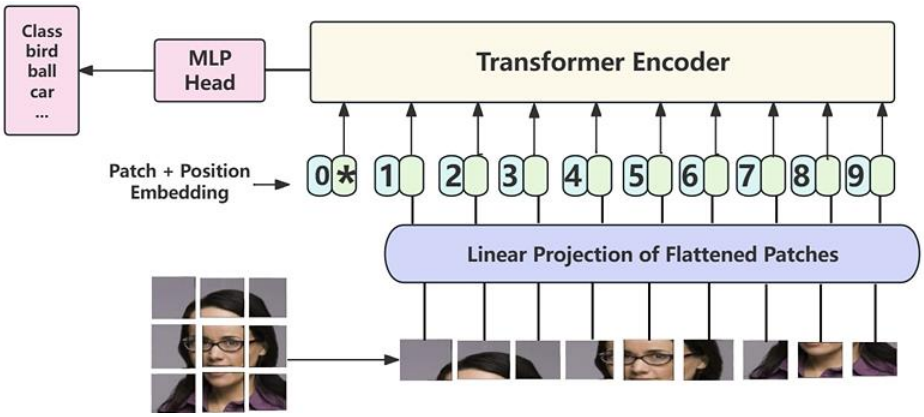

Fig. 3. Vision Transformer Model Structure

Image Patch Division: The input image is divided into fixed-size patches (e.g., 16x16 pixels). Patch Embedding and Position Encoding: Each image patch is flattened and embedded into a linear vector space. Position encodings are added to retain the positional information of each patch. Transformer Encoder: The sequence of embedded image patches is fed into a standard Transformer encoder, which performs feature extraction through multi-head self-attention mechanisms and feedforward neural networks. Classification Head: A learnable classification token is added to the output of the Transformer encoder, which is used for image classification or other vision tasks. The form of the loss function is as follows:

$$L_{total} = L_{cls} + \beta L_{reg} \tag{3}$$

Where L_{cls} is the classification loss, which measures the model's performance on the classification task; L_{re} is the regularization loss, which prevents the model from overfitting; and β is the weighting factor used to balance the importance of the classification loss and the regularization loss.

5 Methods Based on Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GAN) are a type of deep learning model that achieve high-quality image generation through adversarial training between two neural networks: a generator and a discriminator. GANs perform exceptionally well in style transfer tasks, capable of generating realistic style-transferred images.

5.1 Working Principle

GAN consists of a generator and a discriminator. The generator is responsible for creating realistic images, while the discriminator's task is to distinguish between real and fake images. Through the adversarial training process, the generator aims to produce images that are realistic enough to deceive the discriminator, while the discriminator continuously improves its ability to correctly identify real and generated images.

Advantages of This Method:

- High-quality image generation: GANs can generate high-quality, detail-rich images, making them suitable for various image generation and style transfer tasks.
- High flexibility: GANs can achieve multiple image generation and transformation tasks through different generator and discriminator structures.

Disadvantages of This Method:

- Training instability: The adversarial training process of GANs can lead to issues such as mode collapse and training instability.
- High computational resource requirements: The training process of GANs requires substantial computational resources and time, especially when handling high-resolution images.

5.2 Representative Model - CycleGAN

CycleGAN is a GAN model used for unsupervised image-to-image translation, achieving image style transfer through cycle consistency loss. CycleGAN does not require paired training data and can learn mappings from one domain to another without matched image pairs. The model structure is shown in **Fig. 4**.

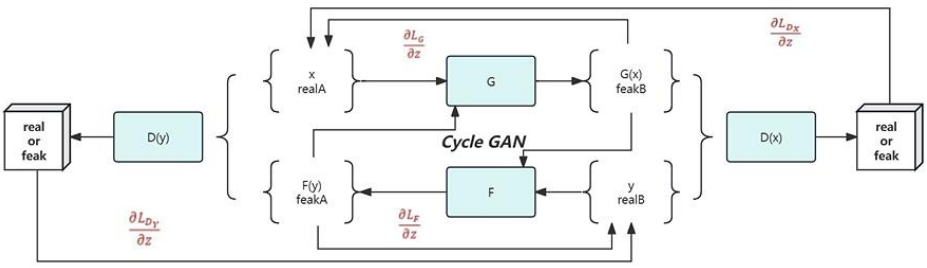

Fig. 4. CycleGAN Model Structure

Generators and Discriminators: Generator G: $X \rightarrow Y$ Maps the input image from domain X to the target domain Y. The discriminator DY attempts to distinguish between images generated by G, i.e., $G(x)$, and real images in the target domain y. Generator F: $Y \rightarrow X$ Maps images from the target domain Y back to the original domain X. The discriminator DX attempts to distinguish between images generated by F, i.e., $F(y)$, and real images in the original domain x. Adversarial Loss: The adversarial loss for generator G and discriminator DY is defined as:

$$\mathcal{L}_{GAN}(G, DY, X, Y) = \mathbb{E}_{y \sim P_{data}(y)}[log\, DY(\mathcal{Y})] + \mathbb{E}_{\mathcal{X} \sim P_{data}(\mathcal{X})}\left[log\left(1 - DY\big(G(\mathcal{X})\big)\right)\right] \quad (4)$$

Similarly, the adversarial loss for generator F and discriminator DX is defined as:

$$\mathcal{L}_{GAN}(F, DY, Y, X) = \mathbb{E}_{\mathcal{X} \sim P_{data}(\mathcal{X})}[log\, DX(x)] + \mathbb{E}_{y \sim P_{data}(\mathcal{Y})}\left[log\left(1 - DX\big(F(\mathcal{Y})\big)\right)\right] \quad (5)$$

Cycle Consistency Loss: The cycle consistency loss ensures that an image, after being translated by two generators, can be reconstructed back to the original image:

$$\mathcal{L}_{cyc}(D, F) = \mathbb{E}_{\mathcal{X} \sim p_{data}(\mathcal{X})}\left[\left\|F\big(G(x)\big) - x\right\|_1\right] + \mathbb{E}_{y \sim p_{data}(\mathcal{Y})}\left[\left\|G\big(F(\mathcal{Y})\big) - \mathcal{Y}\right\|_1\right] \quad (6)$$

This loss consists of two parts:

$\mathbb{E}_{\mathcal{X} \sim p_{data}(\mathcal{X})}\left[\left\|F\big(G(x)\big) - x\right\|_1\right]$: Ensures that an image x from domain X, when mapped to domain Y by generator G and then back to domain X by generator F, remains close to the original image x.

$\mathbb{E}_{y \sim p_{data}(\mathcal{Y})}\left[\left\|G\big(F(\mathcal{Y})\big) - \mathcal{Y}\right\|_1\right]$: Ensures that an image y from domain Y, when mapped to domain X by generator F and then back to domain Y by generator G, remains close to the original image $\mathcal{Y}$.

Full Loss Function: The full loss function for CycleGAN combines the adversarial loss and the cycle consistency loss. The complete loss function is given by:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \quad (7)$$

Where λ is a weighting factor that balances the importance of the adversarial loss and the cycle consistency loss.

5.3 Representative Model - DualstyleGAN

DualStyleGAN is a Generative Adversarial Network (GAN) model designed for exemplar-driven high-resolution portrait style transfer. It introduces dual style paths (intrinsic style path and extrinsic style path) to achieve flexible control over content and style. The model structure is shown in **Fig. 5**.

Dual Style Paths: Intrinsic Style Path: Responsible for controlling the style of the original domain, including facial structure and detailed features. Extrinsic Style Path: Responsible for controlling the style of the target domain, adjusting color and complex structural styles through a hierarchical structure.

Style Transfer Process: Intrinsic Style Representation: The intrinsic style features are extracted using a pre-trained StyleGAN model, retaining the structural information

of the original image. Extrinsic Style Representation: The extrinsic style path is introduced, using additional convolutional layers and residual blocks to adjust the color and structure of the target domain.

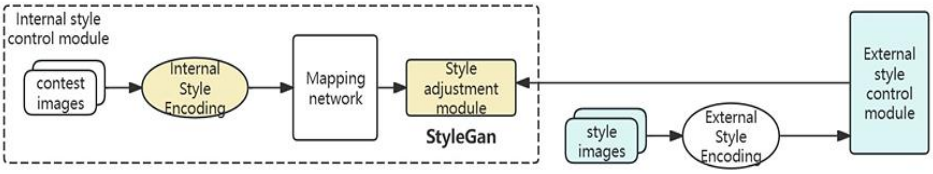

Fig. 5. DualstyleGAN Model Structure

Adversarial Training: The generator and discriminator undergo adversarial training, where the generator attempts to create realistic images to deceive the discriminator, while the discriminator continually improves its ability to distinguish real images from generated ones. Cycle consistency loss and style loss are used to ensure consistency between the content and style of the generated images.

Progressive Fine-Tuning: During the training process, the task difficulty is gradually increased to smoothly transition the generated space to the target domain. Progressive fine-tuning ensures that the model maintains high quality and diversity in the generated images.

$$L_{total} = L_{GAN} + \lambda_{cyc} L_{cyc} + \lambda_{style} L_{style} \quad (8)$$

Where L_{GAN} is the adversarial loss, measuring the similarity between the generated images and the real images; L_{cyc} is the cycle consistency loss, ensuring that the generated images can reconstruct the original images after passing through the dual style paths; L_{style} is the style loss, ensuring consistency in the style of the generated images; λ_{cyc} and λstyle are weighting factors used to balance the importance of each loss term.

6 Comparison of Mainstream Model Performance

In this section, we provide a detailed comparison of the performance of mainstream models based on different methods in the task of anime style transfer. Specifically, we examined methods based on Convolutional Neural Networks (CNN), Variational Autoencoders (VAE), Vision Transformers (ViT), and Generative Adversarial Networks (GAN). These models were trained and tested on the same anime dataset.

As shown in **Fig. 6**, we present the test results of different models trained on the same anime dataset. These results illustrate the specific performance and generated effects of each method in handling the anime style transfer task.

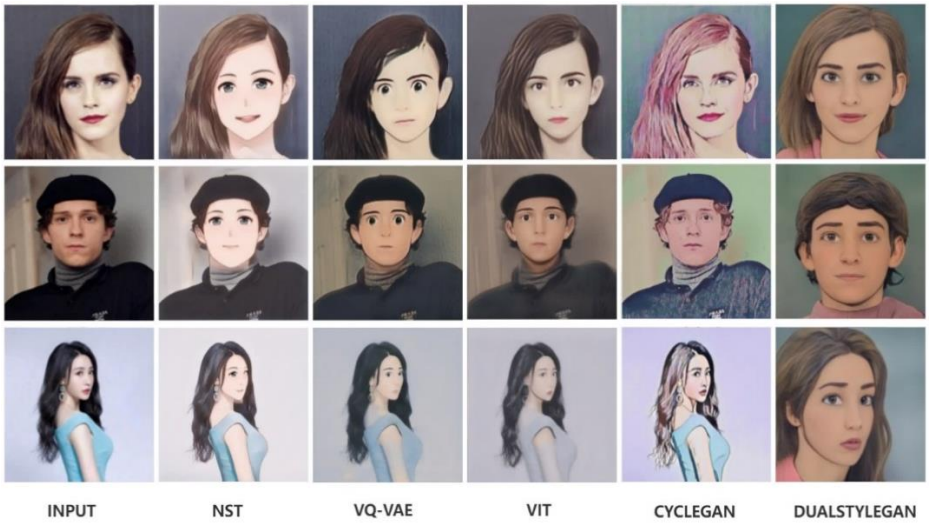

Fig. 6. Model Performance Comparison

To quantitatively evaluate the performance of these models, we calculated the Frechet Inception Distance (FID) score for each model. The FID score measures the similarity between the generated images and real images, with lower scores indicating higher similarity. The experimental results are summarized in **Table 1**, which lists the FID scores for the five models.

Table 1. Comparison of FID Scores for Relevant Methods

Method	Model	FID		
		Hayao Miyazaki	Pixar	American Comic
CNN	Neural Style Transfer	146.58	143.13	142.86
VAE	VQ-VAE	136.12	137.25	134.98
ViT	Vision Transformer	**107.11**	112.19	109.66
GAN	CycleGAN	124.03	119.34	127.53
	DulestyleGAN	110.25	**104.97**	**106.56**

Additionally, we have summarized the fundamental mechanisms of each mainstream style transfer model and provided relevant access links to facilitate further research and application. This information is listed in **Table 2**.

Table 2. Mechanisms and Access Links of Mainstream Models

Method	Model	Mechanism and Access Links
CNN	Neural Style Transfer [6]	Extract high-level image features, Gram matrices, VGG network Access Links: https://github.com/jcjohnson/neural-style
VAE	VQ-VAE [10]	Vector quantization representation, Encoder, Decoder Access Links: https://github.com/SingleZombie/DL-Demos/tree/master/dldemos/VQVAE
ViT	Vision Transformer [12]	Divided into fixed-size patches, Retain spatial information, Capture global context Access Links: https://github.com/google-research/vision_transformer
GAN	CycleGAN [7]	No paired training data required, Adversarial generator and discriminator, Cycle consistency Access Links: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix
	DulestyleGAN [8]	Intrinsic and extrinsic styles, High resolution, Exemplar-based Access Links: https://github.com/ williamyang1991/DualStyleGAN

7 Conclusion

Style transfer technology is garnering increasing attention due to the unpredictability of its generated images. Different parameter settings and input images produce varying effects, keeping researchers and users engaged with a sense of anticipation. In our study, we found that for the specific task of anime style transfer, methods based on Generative Adversarial Networks (GANs) and Vision Transformers (ViTs) demonstrated outstanding performance. Both techniques excelled in terms of FID scores and visual quality, positioning them at the forefront.

In this paper, we provided a detailed overview of four major style transfer techniques, including methods based on Convolutional Neural Networks (CNNs), Variational Autoencoders (VAEs), Vision Transformers (ViTs), and Generative Adversarial Networks (GANs). We highlighted classic models for each method and conducted experimental comparisons to analyze their performance in anime style transfer tasks.

We conducted a comprehensive summary and provided links to the open-source codes for all discussed models to facilitate further research. Future research can explore the following aspects:

- Model Architecture: Investigate more efficient and optimized network structures to enhance the effectiveness and speed of style transfer.
- Training Duration: Optimize the training process to reduce the time required while ensuring the quality of generated images.
- Data Volume: Study the performance of models under varying amounts of training data to determine the optimal data scale.

We hope this paper provides a clear direction for subsequent research in style transfer technology, fostering its development and application.

Acknowledgments. This research was supported in part by the National Natural Science Foundation project (No.62066033), Inner Mongolia Natural Science Foundation Outstanding Youth Fund project (No.2022JQ05), Inner Mongolia Autonomous Region Science and Technology Plan project (No.2021GG0158), Hohhot City University-Institute Collaborative Innovation project, and Inner Mongolia University Young Scientific and Technological Talent Cultivation project (No.21221505)

References

1. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzingand improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
2. Li, C., Wand, M.: Combining markov random fields and convolutional neural net works for image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2479–2486 (2016)
3. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1033–1038. IEEE (1999)
4. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 557–570 (2023)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
7. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
8. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. pp. 2849–2857 (2017)
9. Chen, J., Liu, G., Chen, X.: Animegan: A novel lightweight gan for photo animation. In: Artificial Intelligence Algorithms and Applications: 11th International Symposium, ISICA 2019, Guangzhou, China, November 16–17, 2019, Revised Selected Papers 11. pp. 242–256. Springer (2020)
10. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems 30 (2017)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

13. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
16. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
17. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: Exemplar-based high resolution portrait style transfer. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 7693–7702 (2022)
18. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Vtoonify: Controllable high-resolution portrait video style transfer. ACM Transactions on Graphics (TOG) 41(6), 1–15 (2022)
19. Gao, X., Zhang, Y., Tian, Y.: Learning to incorporate texture saliency adaptive attention to image cartoonization. In: ICML. vol. 2, p. 6 (2022)
20. Men, Y., Yao, Y., Cui, M., Lian, Z., Xie, X.: Dct-net: domain-calibrated translation for portrait stylization. ACM Transactions on Graphics (TOG) 41(4), 1–9 (2022)
21. Song, G., Luo, L., Liu, J., Ma, W.C., Lai, C., Zheng, C., Cham, T.J.: Agilegan: stylizing portraits by inversion-consistent transfer learning. ACM Transactions on Graphics (TOG) 40(4), 1–13 (2021)
22. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
23. Selim, A., Elgharib, M., Doyle, L.: Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (ToG) 35(4), 1–18 (2016)
24. Jang, W., Ju, G., Jung, Y., Yang, J., Tong, X., Lee, S.: Stylecarigan: caricature generation via stylegan feature map modulation. ACM Transactions on Graphics (TOG) 40(4), 1–16 (2021)
25. Huang, J., Liao, J., Kwong, S.: Unsupervised image-to-image translation via pretrained stylegan2 network. IEEE Transactions on Multimedia 24, 1435–1448 (2021)
26. Li, B., Zhu, Y., Wang, Y., Lin, C.W., Ghanem, B., Shen, L.: Anigan: Style guided generative adversarial networks for unsupervised anime face generation. IEEE Transactions on Multimedia 24, 4077–4091 (2021)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
29. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196(2018)
30. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. Advances in neural information processing systems 28 (2015)
31. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)