Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31076

Research and Analysis of Image Style Transfer

Jingyang Li

School of Engineering, Rensselaer Polytechnic Institute, Troy, USA
Lij53@rpi.edu

Abstract. Image style transfer is an emerging image processing technique that generates an image while preserving the structural content of the original image and incorporating the visual style of another image. This technology is widely applied in digital art, photography, animation, and visual media. Traditional methods of style transfer rely on manual modeling based on texture synthesis and physical rendering. This paper provides a systematic and comprehensive overview of the development of image style transfer technology, thoroughly reviewing the technical evolution from early optimization-based methods to modern deep learning frameworks. It begins by introducing traditional approaches based on texture synthesis and manual rendering models; then focuses on convolutional neural network (CNN)-based methods, represented by Gatys et al.'s neural style transfer approach; followed by an in-depth analysis of generative adversarial network (GAN)-based style transfer models, including representative works such as Pix2Pix, CycleGAN, NICE-GAN, and StyleGAN. These methods have demonstrated excellent performance in paired, unpaired, and controllable style transfer tasks. Additionally, the paper discusses the latest research advancements in diffusion models, particularly high-fidelity controllable image generation methods like latent space diffusion models and Stable Diffusion. Furthermore, it systematically summarizes commonly used network architectures, loss functions, evaluation metrics, and benchmark datasets in style transfer. Finally, it analyzes the main challenges and future research directions of current technologies from perspectives such as content consistency, semantic alignment, computational efficiency, and objective evaluation. This paper can provide researchers and practitioners in the field of image style transfer with structured references.

Keywords: Style transfer, Generative Adversarial Network, Convolutional Neural Network.

1. Introduction

Image style transfer is a technique that aims to generate an image that preserves the content of a source while adopting the style of another. This technique is widely used among multiple areas such as photography, digital art, animation. Due to the emergence of deep learning, style transfer has become an active research topic in computer vision, which bridges the area of arts and computer science.

Early style-transfer algorithms relied on texture synthesis and optimization-based approaches. The main theory of these methods is that all image styles could be described by multiple statistic

DOI: 10.54254/2755-2721/2026.TJ31076

models of local traits. As an example, the texture of an image of pistachio which has an open on it, can be statistically described as a high probability of two curves with certain curvature intercept. This kind of style transfer was mostly done by manually creating models. Such method requires analysis and creating a model for each type of style. It also needs a program for each single style or scene. Thus, this method is inflexible and computational intensive when training. The situation changes when a breakthrough was achieved by Gatys et al., who demonstrated that convolutional neural networks can separate and recombine content and style representations [1].

Convolutional Neural Network (CNN) methods represent a groundbreaking advancement in computer vision. This method utilizes CNNs' layered architecture to extract layered visual features from images, ranging from low-level textures and colors in surficial layers to high-level shapes and semantics in deeper layers. This hierarchical representation allows CNNs to capture both global structure and refined visual referring that are essential for tasks such as classification, segmentation, and stylization. In the context of image style transfer, CNNs provide a natural framework for disentangling content and style features.

Rombach, Robin, et al.'s work is a milestone work in the field of image generation, and its proposed Latent Diffusion Model (LDM) is the core foundation of the Stable Diffusion Model [2]. The traditional diffusion model directly performs the processes of "adding noise" and "removing noise" in the image pixel space, which requires a huge amount of computation. The core innovation of LDM lies in its ability to transfer this process from a high-dimensional pixel space to a low dimensional latent space

Nevertheless, several challenges remain, including maintaining content consistency, achieving semantic alignment between domains, improving model efficiency, and establishing objective evaluation metrics. In addition, rapid developments in the area continuously reshape the requirements, which indicates that no single fixed approach can fully address all these challenges.

However, there is a shortage of an updated review that systematically covers GAN, Transformer, and diffusion-based developments, along with evaluation and application perspectives. Thus, in order to provide a thorough view of current situation of style translation, this paper will focus on several recent developments in style transfer. Methods will be mainly focused on GANs, representative models and loss functions will be summarized, and evaluation metrics and datasets will be discussed.

2. Early methods

Before the wide use of deep learning, style transfer is mainly done by manually designed traits models and traditional graphic processing techniques. These methods can be mainly classified into two types: rendering methods based on physical models and transfer based on textural synthesis.

The rendering method mainly adopts image filtering and manual calculation. By simulating the physical process of art creation, such as brush strokes or color diffusion, the stylization can be realized. This method is done by designing specific image processing filters to abstract the images, and then generates cartoon, sketches, or other artistic effects. However, such method is usually complicated in computation, and hard to adapt to different requirements of varied artistic styles. The quality is rough and the robustness is poor.

The texture synthesis methods are realized by analyzing local texture features of the style images and apply them to the output images. A representative method is Image Quilting algorithm. This method divides images into small pieces, and then realizes style transfer by matching and combining these pieces. Although it has lots of improvements compared to the rendering method, it's still

DOI: 10.54254/2755-2721/2026.TJ31076

limited in style representation, difficult to capture and transfer complex art styles' advanced traits, and obtains discontinuity and artifacts when processing complex structures.

Yann LeCun et al.'s work is a truly milestone work in the field of deep learning and computer vision [3]. It systematically elaborates on gradient based learning methods and provides the first comprehensive introduction to the application of Convolutional Neural Networks (CNN) in document recognition tasks.

3. Convolutional Neural Network

Gatys et al. brought a revolutionary breakthrough in image style transfer and completely reconstructed the technical path of style transfer [1]. It's the first time that the layered feature extraction capability of CNN is utilized to decouple the structural information of content images from the artistic features such as texture and color of style images.

Generally, the implementation method is as follows.

First, choose a pre-trained model, such as VGG-19, to extract features from the image. Then define a loss function, which consists of two parts: content loss and style loss.

Content loss refers to retaining the structural information of the original content by calculating the difference between the output image and the content image on the CNN deep feature map. Style loss introduces the Gram matrix to quantify style features, ensuring that the output image matches the artistic characteristics such as strokes and color diffusion of the style image. Finally, it performs joint optimization, accomplished through a series of steps as follows:

Initialize the synthesized image. Usually, it is initialized with a white noise image. Using a copy of the content image is also viable. It converges faster by weighting the combination of content loss and style loss to generate a new image that preserves both content and style. Then input the content image style image and the current synthesized image separately into the pre trained VGG network. Calculate content loss and style loss at the specified layer, and weighted sum to obtain the total loss, then continue with backpropagation. The key point of this step is that only the gradients of the pixels in the synthesized image are calculated here. The weights of the VGG network are frozen throughout the entire process and do not participate in updates. Next is updating the image. Use gradient descent algorithms (such as L-BFGS or Adam) to update the pixel values of the synthesized image to reduce the total loss. Repeat the above steps until the loss converges or reaches the predetermined number of iterations.

Compared to traditional methods such as texture synthesis and stroke simulation, the Neural Style Transfer (NST) algorithm proposed by the Gatys team in this paper can handle images of any style and has a more stable output. Nevertheless, this method has obvious performance bottlenecks. Due to iterative optimization, its efficiency is low, and it takes minutes to generate a single image. This has given rise to subsequent rapid style transfer method, and Li et al. provided a very comprehensive and in-depth review in the field of convolutional neural networks [4].

Xia Zhao et al. systematically reviewed the development of CNN in computer vision [5]. It can help readers establish a clear understanding of CNN components and functions, and also provide readers with a comprehensive understanding of CNN's applications, current challenges, and future research directions in core fields such as image classification, object detection, semantic segmentation, super-resolution reconstruction, and video prediction.

DOI: 10.54254/2755-2721/2026.TJ31076

4. Generative Adversarial Networks

In recent years, technologies such as Generative Adversarial Networks (GANs) and Diffusion Models have been introduced to further enhance the quality and flexibility of style transfer. The emergence of Generative Adversarial Networks (GANs) provides a revolutionary approach: by training a feedforward generative network to directly learn the mapping from content space to stylization space, real-time style transfer can be achieved.

GAN introduces a new mode of rival between generators and discriminators. The generator acts like a learner who learns to draw, with the goal of creating stylized works that is as indistinguishable from reality as possible. The discriminator, on the other hand, is an experienced art appraiser whose task is to examine a work and determine whether it is truly a good work or just a poor work by a novice imitator. The entire training process is a constantly evolving game between the generator and discriminator. During the process of discrimination, the discriminator accumulates experience and becomes more powerful. In order to pass the evaluation of increasingly powerful discriminators, generators have to improve their skills, learn more sophisticated style strokes, color matching, and overall composition, rather than simply imitating textures. In the end, the ideal state is that the generator's artwork is so advanced that the discriminator cannot pick out any flaws, and at this point researchers have an excellent style transfer model. The key advantage of this adversarial training is that the generator learns not how to minimize a rigid mathematical formula, but how to capture the overall, indescribable "feeling" and "charm" in the target style domain, thereby generating visually more coordinated and realistic works.

Since the emergence of GAN technology, many GAN models have emerged. There are three typical GAN models: Pix2Pix, CycleGAN, and StyleGAN.

4.1. Pix2Pix

Before Pix2Pix, many image generation tasks (such as generating photos from sketches) were difficult because one input could correspond to countless "correct" outputs. Traditional pixel by pixel loss functions (such as L2 loss) can cause the model to learn the "average answer", resulting in blurry generated results. Thus, Phillip Isola et al. proposed a universal image processing framework, commonly known as Pix2Pix. It utilizes Conditional Generative Adversarial Networks (cGANs) as its core, providing a unified solution for various image to image conversion tasks [6]. It no longer asks the model to 'guess' what color each pixel should be, but instead assigns it a more advanced task:' Generate an image that looks sufficiently 'real' in the target domain '. It achieves this goal through an architecture called Conditional Generative Adversarial Network.

Pix2Pix uses U-Net as its generator. It can be understood as a very sophisticated "encoder decoder" system. The function of the encoder is like the "understanding" stage of a painter. It gradually compresses the input image through multiple layers of convolution, extracting advanced and abstract features. However, in this process, many low-level and accurate contour and position information are lost. And the decoder part is like the painter's "drawing" stage. It gradually samples and restores a complete image based on the advanced features understood by the encoder.

Pix2Pix's discriminator uses PatchGAN. Traditional discriminators give the entire image a total score, which may result in them only focusing on the macroscopic consistency of the image and ignoring the roughness of local details.

The PatchGAN discriminator used by Pix2Pix is like an appraiser holding a magnifying glass. It doesn't just look at the entire painting, but divides the image into many small image blocks, and

DOI: 10.54254/2755-2721/2026.TJ31076

finally combines the results of all the small blocks to give the final judgment. This method forces the generator to strive for excellence in every local area, generating high-quality textures and details.

4.2. CycleGAN

Before CycleGAN, image translation models like Pix2Pix were very powerful, but they had a fatal limitation: they required paired training data. This means that for each input image, there must be an output image that corresponds exactly to the content and serves as the 'standard answer'. In reality, such data is extremely difficult to obtain.

Jun Yan Zhu, Taesung Park solved this problem by proposing a new model [7]. Its innovation lies in achieving image translation without paired data. It only requires two independent, unrelated sets of images. For example, set A consists of a set of photos of horses. Set B consists of a bunch of photos of zebras. The goal of the model is to learn the mapping relationship between these two sets, so that it can convert any photo of a horse into a zebra, or any photo of a zebra into a horse.

The circular consistency loss is CycleGAN's most ingenious move, which solves the fundamental problem of "how to maintain content consistency without supervision". Its idea is very intuitive and powerful: translation should be reversible. It imposes constraints through a 'loop': first, forward loop consistency. A photo of a horse X, after being converted into a zebra GX by generator G, is then handed over to generator F for conversion back to horse F (GX). The final image obtained should be very similar to the original photo of the horse X. Then there is backward loop consistency: a photo Y of a zebra is converted into a horse FY by generator F, and then passed back to generator G for conversion back to zebra G (FY). The final image should be very similar to the original zebra photo Y. The constraint of this' circular reconstruction 'forces the generator to never lose or distort the core content information of the original image (such as pose, background, object shape) when changing the style (horse to zebra)

Chen, Runfa, et al. proposes a new framework called NICE-GAN, which uses a reusing discriminator as the encoder and proposes an effective decoupling training strategy to improve the unsupervised image conversion effect while reducing model complexity [8]. This approach has positive implications for future research.

4.3. StyleGAN

Karras, Tero, Samuli Laine, and Timo Aila.'s work is an important milestone in the field of image generation, and its proposed StyleGAN architecture greatly promotes the development of high-resolution and high-quality image generation [9]. It can be said that the revolution of StyleGAN is almost entirely reflected in the design of its generator, while the discriminator is relatively traditional, but there are also some corresponding adjustments.

The generator of StyleGAN includes a mapping network consisting of multiple fully connected layers. This network converts the input traditional noise vector (usually denoted as z) into an intermediate latent vector w. The advantage of doing so is that the intermediate latent space W does not have to follow the distribution of the input noise z (such as the standard normal distribution), thus learning a more decoupled and linear representation. This means that in the W space, different dimensions are more likely to correspond to independent visual features in the image.

In order to generate more natural images (such as skin texture, hair details), StyleGAN adds a pixel by pixel random noise after each convolution layer and before AdaIN. This noise is single channel and will be applied to all feature maps through a learnable scaling factor. This provides the generator with the ability to introduce subtle random variations.

DOI: 10.54254/2755-2721/2026.TJ31076

Unlike traditional GANs that use noise vectors as the initial input for the generator, StyleGAN's generator takes a learnable constant matrix as the initial input. This helps reduce feature entanglement and allows the network to focus more on controlling image generation through style vectors.

In recent years, new architectures such as Transformer have also been introduced to the field of style transfer. For instance, Gihyun Kwon proposes a Laplacian Pyramid Network, which achieves higher-quality stylization while maintaining high speed through two stages—"drafting" and "revision." [10].

5. Conclusion

This paper systematically elaborates on the developmental trajectory of image style transfer technology, comprehensively reviewing the evolution from early traditional methods based on texture synthesis and physical rendering to modern generative models centered on deep learning. Traditional approaches relied on manual feature design and physical modeling, resulting in high computational costs and limited flexibility. The neural style transfer method achieved the effective separation of content and style for the first time by leveraging hierarchical features of convolutional neural networks, marking the formal entry of style transfer into the deep learning era. Subsequently, generative adversarial networks significantly advanced the development of style transfer, enabling paired, unpaired, and controllable style transformations. Representative models such as Pix2Pix, CycleGAN, NICE-GAN, and StyleGAN have achieved remarkable breakthroughs in perceptual realism, diversity, structural preservation, and computational efficiency. In recent years, diffusion-based generative methods—particularly latent space diffusion models and Stable Diffusion—have further pushed the technical boundaries in terms of image quality and controllable generation. Meanwhile, this paper systematically summarizes commonly used network architectures, loss functions, evaluation metrics, and dataset frameworks in style transfer. Although significant progress has been made in current image style transfer technology, challenges persist in maintaining content consistency under complex stylistic conditions, achieving precise cross-semantic alignment, improving model computational efficiency, and unifying objective evaluation standards. Future research is expected to continue advancing in directions such as efficient network architecture design, semantically controllable generation, multimodal guidance, and standardized evaluation systems. By organizing existing research, this paper aims to provide clear technical references for researchers and practitioners in related fields while promoting the development of image style transfer toward higher quality, stronger controllability, and broader applications.

References

[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv: 1508.06576.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[3] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[4] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.

[5] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4), 99.

DOI: 10.54254/2755-2721/2026.TJ31076

[6] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
[7] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).
[8] Chen, R., Huang, W., Huang, B., Sun, F., & Fang, B. (2020). Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8168-8177).
[9] Karras, T., Laine, S., & Aila, T. (2019). 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).
[10] Lin, T., Ma, Z., Li, F., He, D., Li, X., Ding, E., ... & Gao, X. (2021). Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5141-5150).