A Review of Deep Learning-Based Image Style Transfer Methods: From Gatys to Diffusion Models

Liu Yuyan
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, Sichuan, China
2690383374@qq.com

ABSTRACT

This paper systematically reviews how the image style conversion technology based on deep learning has developed step by step, from the classic Gatys method to the current diffusion models. We propose a classification framework, which includes four technical types: iterative optimization method, feedforward network method, video timing method and diffusion model method. In this framework, we compare the representative methods, and reveal the fundamental differences and performance trade-offs between different technology types from the perspectives of content expression, style modeling, semantic controllability and computational efficiency. We also pointed out several key challenges, including the incomplete separation of style and content, the limited ability to deal with complex style conversion, the difficulty in balancing the consistency of video timing and computational efficiency, the insufficient controllability and efficiency of diffusion model, and the lack of standardized evaluation system. We also look forward to the future direction, such as enhancing interpretability, expanding to 3D scenes, realizing lightweight deployment and building ethical framework. This review hopes to provide a clear technical roadmap and method reference for researchers.

Keywords: Image Style Transfer; Deep Learning; Convolutional Neural Network; Video Style Transfer; Diffusion Model; Evaluation System.

1. INTRODUCTION

Image style transfer is to transfer the style of a reference picture to another content picture, while maintaining the original appearance of the content picture. This is an important research direction in the cross field of computer vision and computer graphics, and it has been widely used in short video special effects, digital art creation, virtual reality, and digital protection of cultural heritage. Since Gatys et al.' s pioneering work in 2015 [1], when they used convolutional neural networks to extract content and style features, this field has developed very rapidly, and many methods have emerged: slow methods based on iterative optimization, real-time methods based on feedforward networks, methods aimed at video and ensuring time coherence, and new methods based on diffusion models. However, most of the existing reviews [2-4] only pay attention to a specific technical stage, or only a single type of method, lacking a complete development thread from Gatys to diffusion model, and they also fail to deeply compare the fundamental differences and performance tradeoffs between different methods. Moreover, the evaluation system of style transformation is still very fragmented, different studies use many different subjective and objective standards, and there is no unified comparison framework. In order to solve these problems, this paper systematically reviews the technical progress of image style conversion from the classic Gatys method to the Difficulty Models, and puts forward an image style conversion method which includes four technical paradigms (iterative optimization, feed forward networks, video temporal methods, And difference models), and compares and analyzes the representative methods from the perspectives of content representation, style modeling, semantic controllability and computational efficiency. We also briefly introduced the evaluation system of style conversion. The rest of the paper is arranged as follows: the second part talks about the evolution and representative methods of the four technological paradigms; The third part discusses the evaluation system of style transformation and gives a cross-paradigm comparison. The fourth part summarizes the main challenges and points out the future research direction; The fifth part summarizes the whole paper.

2. TECHNICAL EVOLUTION OF IMAGE STYLE TRANSFER METHODS

This part will follow four technical routes to elaborate on those representative methods. These four routes are: iterative optimization method, feedforward network method, video timing method, and diffusion model method.

2.1 Iterative Optimization-Based Neural Style Transfer

Gatys et al. [1]' s pioneering work, using a pre-trained VGG network to do style conversion for the first time. They define the content loss as the mean square error between the generated picture and the content picture on the high-level feature map, while the style loss is calculated by Gram matrix to capture the correlation between style features. By constantly optimizing a random noise picture and minimizing the total loss, we can generate high-quality and artistic pictures. This method can get high-quality results with flexible style definition, but it takes hundreds of iterations to generate a picture, and the calculation cost is very high, so it can't be applied in real time.

Later improvements focused on speed, control and diversity. Li et al. [5] introduced Markov random fields (MRF) to replace the global Gram matrix, which improved the fidelity of local texture. Luan et al. [6] added semantic segmentation mask to realize the separate control of foreground and background styles. Gatys et al. [7] further allow users to guide stylized areas through brushes. These methods laid the foundation for later research, but their iterative characteristics made real-time interaction difficult. The main limitations include: (1) Gram matrix is essentially a second-order statistic, which is difficult to capture non-local texture structure and often leads to the artifacts of repeated patterns; (2) Iterative optimization needs hundreds of back propagation, and each picture takes several minutes; (3) Each generation is independent, and the calculation cannot be reused. It is these efficiency bottlenecks that directly urge researchers to turn to real-time methods based on feedforward generation networks.

2.2 Real-Time Style Transfer with Feedforward Generation Networks

In order to solve the problem that iterative optimization is too slow, Johnson et al. [8] proposed perceptual loss, and also trained a forward generation network to do real-time conversion of a single style. This network takes a content picture as input, and can directly output a stylized picture, and the reasoning speed is as fast as milliseconds. Ulyanov et al. [9] introduced texture network, which made the model smaller. Dumoulin et al. [10] proposed Conditional Instance Normalization (CIN), which allows a network to support multiple styles by switching normalization parameters. Chen et al. [11] put forward StyleBank, which expressed styles as learnable basic vectors, and further improved the efficiency of multi-style conversion.

The most significant breakthrough comes from adaptive case normalization (AdaIN) proposed by Huang and Belongie [12]. The core idea of AdaIN is to adjust the feature statistics (such as mean and variance) of content pictures to the same as those of style pictures. The formula is written as follows: Adain (x, y) = σ (y) (xμ (x))/σ (x)+μ (y). This method does not need to be trained once for every style; A model can realize the conversion of any style in real time. Li et al. [13] proposed Whitening and Coloring Transform (WCT), which directly matched the second-order statistics in the feature space, and further improved the quality of arbitrary style transformation. Park and Lee [14] proposed SANet with a style attention mechanism for finer local style control. Moreover, Zhu et al. [15] proposed CycleGAN with cycle-consistency loss to achieve unpaired image-to-image style transfer, especially suitable for scenarios where paired data are difficult to obtain.

These methods are fast in inference, but have limitations: the training process is complex and requires large amounts of paired data; the quality of arbitrary style transfer is still slightly inferior to iterative optimization methods; for extreme styles or high-resolution images, artifacts may sometimes appear.

2.3 Temporally Consistent Style Transfer for Videos

Extending image style transfer to videos leads to flickering when frames are processed independently. To solve this problem, Ruder et al. [16] proposed using optical flow constraints to maintain temporal consistency, warping the stylized result of the previous frame to the current frame as a temporal regularization term. Heusel et al. [17] proposed an end-to-end framework that jointly optimizes style loss and temporal loss. Chen et al. [18] further proposed a real-time video style transfer method that greatly improves speed via optical flow-guided feature propagation. Deng et al. [19] proposed a multi-channel correlation method to further enhance temporal stability of video style transfer.

These works enable stylized videos with smooth temporal transitions while preserving artistic effects. Video style transfer expands application scenarios, but the main challenge lies in balancing temporal consistency and computational efficiency: high-precision optical flow is computationally expensive, while lightweight methods tend to lose detail; cumulative errors in long video processing remain an open problem.

2.4 Large-Scale Pre-Trained Style Transfer with Diffusion Models

In recent years, diffusion models have demonstrated powerful capabilities in image generation and have been quickly applied to style transfer. Rombach et al. [20] proposed latent diffusion models (Stable Diffusion) that perform diffusion and denoising in a latent space, significantly reducing computational cost. In style transfer, two main strategies are adopted: (1) fine-tuning pre-trained models (e.g., LoRA) on a small set of style images to make the model generate a specific style; (2) guided sampling (e.g., StyleDiffusion) that injects style features during the denoising process, enabling arbitrary style transfer without training. Kwon and Ye proposed a style diffusion model based on disentangled representations. Zhang et al. proposed an inversion-based diffusion style transfer method, further improving the precision of style control. Meanwhile, the introduction of the multimodal model CLIP has given rise to new text-guided style transfer paradigms (e.g., CLIPstyler, StyleCLIP), allowing users to control style attributes through natural language descriptions.

Diffusion models excel in creativity and semantic controllability, but have notable drawbacks: slow inference (seconds), difficulty in meeting real-time interaction requirements; the precision of style control still needs improvement; large model parameter count and high deployment cost.

3. STYLE TRANSFER EVALUATION SYSTEM AND CROSS-PARADIGM COMPARISON

3.1 Evaluation Metrics and Methods

Evaluation methods for style transfer fall into two categories: subjective assessment and objective metrics. Subjective assessment collects human preferences on generated images through user studies, commonly using mean opinion scores (MOS), but it is time-consuming and difficult to reproduce.

Objective metrics include: Content preservation metrics: Pixel-level metrics such as PSNR and SSIM, as well as the perceptual-level LPIPS. LPIPS was proposed by Zhang et al., using a pre-trained AlexNet to extract multi-layer feature maps and computing weighted distances; its correlation with human judgment is more than 30% higher than that of PSNR.

Style consistency metrics: Traditional methods are based on Gram matrices; recent studies use the ViT image encoder of CLIP to extract features and compute cosine similarity (CLIP score), capturing high-level semantic style.

Aesthetic quality metrics: These include FID [17] and NIMA. NIMA predicts an aesthetic score distribution (1-10) using MobileNet. Furthermore, Wright and Ommer proposed ArtFID, which combines FID with content-aware metrics, providing a more comprehensive quantitative evaluation tool for style transfer.

Moreover, Ioannou and Maddock noted in their survey that current style transfer evaluation “lacks a unified evaluation procedure.” The meta-evaluation study by Pauli et al. further revealed biases in existing metrics for content preservation assessment. The OmniStyle framework proposed by Wang et al. constructs a quality assessment mechanism from three dimensions-content preservation, style consistency, and aesthetic appeal-offering a reference for systematic evaluation. Domestic surveys by Zhang et al. [2], Liu et al. [3], and Jing et al. [4] have also summarized evaluation methods for style transfer from different perspectives.

3.2 Cross-Paradigm Comparative Analysis

A comprehensive comparison of the four major paradigms is conducted from the perspectives of content representation, style modeling, semantic controllability, computational efficiency, and temporal consistency. The results are shown in Table 1.

Table 1. Comprehensive comparison of the four technical paradigms.

Dimension	Iterative Optimization	Feedforward Networks	Video Temporal	Diffusion Models
Content representation	VGG high-level features	Encoder features	Features + optical flow	Latent codes
Style modeling	Gram matrix	Instance normalization	Per-frame + temporal	Fine-tuning / guidance
Semantic controllability	Weak	Medium	Weak	Strong (text-guided)
Inference speed	Slow (minutes)	Fast (milliseconds)	Relatively slow (seconds)	Slow (seconds)
Style flexibility	Arbitrary	Arbitrary	Arbitrary	Arbitrary
Temporal consistency	N/A	N/A	Strong	Weak (needs extra handling)

Iterative optimization methods yield the highest quality but are the slowest; feedforward networks achieve the best balance between speed and quality, becoming the mainstream for real-time applications; video methods extend the temporal dimension but face efficiency challenges; diffusion models lead in creativity and semantic controllability but have high computational cost.

4. KEY CHALLENGES AND FUTURE DIRECTIONS

4.1 Key Challenges

Despite the progress made by existing research, the following common challenges remain:

Incomplete disentanglement of style and content: When style textures are highly correlated with content semantics, the stylization process tends to destroy the main structure of the image [1,12]. No existing method can perfectly transfer arbitrary style while preserving semantic structure.

Limited ability to transfer complex styles: Current methods perform poorly when transferring complex styles that require global composition understanding (e.g., Cubism, Surrealism), and struggle to capture the artist's overall creative logic [13,14].

Balancing video temporal consistency and computational efficiency: Real-time stylization of long videos remains challenging. The accuracy and speed of optical flow estimation are difficult to achieve simultaneously, and the problem of cumulative temporal errors has not been fundamentally solved [16,18].

Controllability and efficiency of diffusion models: The controllability, computational efficiency, and interpretability of diffusion models need improvement. Text-guided style transfer suffers from insufficient semantic alignment accuracy and slow inference speed [20].

Lack of standardized evaluation system: Different studies use different metrics, and there is no unified benchmark dataset or comparison protocol. LPIPS, CLIP score, user studies, etc., each have their own emphases, and comprehensive evaluation methods are still immature.

4.2 Future Directions

According to its development law and the latest trend, we put forward the following future research directions:

Enhance interpretability: explore an easy-to-understand model that can separate style from content, and find out how to express style inside the neural network. We can use causal representation to learn and oppose decoupling methods.

Extend to 3D and dynamic scenes: Extend the style transfer from 2D pictures to 3D models, point clouds, NeRFs and moving scenes. 3D style migration needs to keep the consistency of geometry and the invariance of viewing angle.

Lightweight model and hardware deployment: develop lightweight models for mobile phones and edge devices, and combine knowledge distillation, network pruning and quantification to realize real-time style migration.

Ethics and copyright framework: With the wide application of AIGC, the copyright ownership and ethical issues in style transfer have become more and more prominent. We urgently need to establish traceable and certifiable technology identification mechanism and laws and regulations.

5. CONCLUSION

This paper systematically reviews the development of image style transfer technology from the classic Gatys method [1] to the diffusion model [20], and establishes a classification framework including four paradigms: iterative optimization, feedforward network, video timing method and diffusion model. This is the main theoretical contribution of this paper. On this basis, we compare these four paradigms from the perspectives of content representation, style modeling, semantic controllability, computational efficiency and time sequence consistency (see Table 1), and summarize the advantages and limitations of each method. The paper also summarizes the evaluation system of style transfer and points out the shortcomings of the existing evaluation index [17].

Research limitations: Due to the limitation of time and space, this paper does not discuss the mathematical principle behind each evaluation index (such as LPIPS, CLIP score, FID) and the difference of their applicability in different scenarios. In addition, fine-tuning techniques for diffusion models (e.g., LoRA, ControlNet) were only briefly introduced without comparative analysis. Furthermore, the cited Chinese surveys [2,3] mainly cover literature up to 2022; the most recent domestic progress from 2023 to 2025 needs to be supplemented in future work.

Recommendations for researchers: Based on the comparative analysis presented here, the following recommendations are made: if real-time interactive experience is desired, feedforward network methods (e.g., AdaIN [12]) are the best choice; if the highest generation quality is desired regardless of time cost, iterative optimization methods [1] remain irreplaceable; if unpaired cross-domain transfer (e.g., photo to oil painting) is needed, CycleGAN [15] performs excellently; if creative style synthesis and text-guided control are desired, diffusion models [20] offer the greatest potential, but their computational overhead must be considered. Future research should focus on overcoming the theoretical problem of style and content separation, designing a lighter diffusion model, and unifying the evaluation criteria.

This review aims to provide clear technical development roadmap and method reference for researchers in the field of image style transfer.

REFERENCES

[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). A neural algorithm of artistic style. Journal of Vision, *16*(12), 326-326. https://doi.org/10.1167/16.12.326

[2] Zhang, W., Wang, L., & Liu, Q. S. (2020). A review of image style transfer based on deep learning. Chinese Journal of Computers, *43*(1), 1-26.

[3] Liu, H. F., Han, L., & Wang, B. (2022). A survey of image style transfer algorithms. Journal of Image and Graphics, *27*(6), 1701-1722.

[4] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. IEEE Transactions on Visualization and Computer Graphics, *26*(11), 3365-3385. https://doi.org/10.1109/TVCG.2019.2921336

[5] Li, C., & Wand, M. (2016). Combining Markov random fields and convolutional neural networks for image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2479-2486). IEEE.

[6] Luan, F., Paris, S., Shechtman, E., & Bako, K. (2017). Deep photo style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4990-4998). IEEE.

[7] Gatys, L. A., Ecker, A. S., Bethge, M., Hertzmann, A., & Shechtman, E. (2017). Controlling perceptual factors in neural style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3985-3993). IEEE.

[8] Johnson, J., Alahi, A., & Li, F.-F. (2016). Perceptual losses for real-time style transfer and super-resolution. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision-ECCV 2016 (Vol. 9906, pp. 694-711). Springer. https://doi.org/10.1007/978-3-319-46475-6_43

[9] Ulyanov, D., Vedaldi, A., & Lempitsky, V. S. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint, arXiv:1607.08022. https://arxiv.org/abs/1607.08022

[10] Dumoulin, V., Shlens, J., & Kudlur, M. (2017). A learned representation for artistic style. In International Conference on Learning Representations (ICLR).

[11] Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2017). StyleBank: An explicit representation for neural image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2770-2779). IEEE.

[12] Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 1510-1519). Venice, Italy. https://doi.org/10.1109/ICCV.2017.167

[13] Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017). Universal style transfer via feature transforms. In Advances in Neural Information Processing Systems (Vol. 30, pp. 386-396). Neural Information Processing Systems Foundation.

[14] Park, D. Y., & Lee, K. H. (2019). Arbitrary style transfer with style-attentional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5880-5888). IEEE.

[15] 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 (ICCV) (Vol. 2017-October, pp. 2223-2232). IEEE. https://doi.org/10.1109/ICCV.2017.244

[16] Ruder, M., Dosovitskiy, A., & Brox, T. (2016). Artistic style transfer for videos. In German Conference on Pattern Recognition (GCPR) (pp. 26-36). Springer.

[17] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6626-6637). Neural Information Processing Systems Foundation.
[18] Chen, D., Liao, J., Yuan, L., Yu, N., & Hua, G. (2017). Real-time neural style transfer for videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7831-7839). IEEE.
[19] Deng, Y., Tang, F., Dong, W., Huang, H., Ma, C., & Xu, C. (2021). Arbitrary video style transfer via multi-channel correlation. arXiv preprint, arXiv:2109.05886. https://arxiv.org/abs/2109.05886
[20] 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 (CVPR) (pp. 10684-10695). IEEE.