Dean&Francis

ISSN 2959-6157

The Methods and Mechanisms of Preserving Content and Structure in Image Style Transfer

Baosen Hou

Software Engineering, Dalian Jiaotong University, Dalian, 116045, China
3325521957@qq.com

Abstract:

As artificial intelligence (AI) technology continues to advance, image style transfer has gradually moved from research to practical applications, becoming widely integrated into fields such as art creation, advertising design, and virtual reality. However, existing research often struggles to maintain the fidelity of the original image's content and structure while pursuing stylistic effects. This contradiction has become a key bottleneck restricting the further development of image style transfer technology. With advances in image processing, a key challenge is preserving the original image's structure and content while achieving a stylized effect. This paper reviews and analyzes the design principles and underlying mechanisms of various AI models, including Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Transformer architectures, in addressing the challenge of content preservation. Through the integration of existing literature, the results indicate that there is currently no universal solution, and trade-offs and choices need to be made for specific scenarios. Based on the limitations of existing methods, this paper explores potential future advancements in the technology, aiming to offer valuable insights for further development.

Keywords: Image Style Transfer, Content Preservation, Structure Fidelity, CNN, Generative Models

1. Introduction

Image style transfer technology has shown broad application potential in fields such as artistic creation, film and television production, and visual design. However, while many existing methods generate images with distinct styles, they often suffer from problems such as blurred object edges and distorted details, that is, in the pursuit of "style intensity", "content fidelity" is sacrificed. This phenomenon reflects the obvious shortcomings of the technology in terms of content and structure preservation. Since the image style transfer technique was first introduced by Gatys et al., various optimization and acceleration

techniques have been proposed, such as the feedforward network proposed by Johnson et al. [1,2]. Nonetheless, these methods often involve a trade-off between style flexibility and content fidelity. Therefore, this paper investigates the issue of "content and structure preservation," and analyzes the differences in the mechanisms of content encoding and style fusion among different models. By exploring the technical details such as loss function design and feature space operations, this paper aims to reveal the underlying reasons for the retention or loss of content information during the transfer process. In addition, a comparative analysis of the performance and limitations of the three main methods, namely Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Transformer models, in terms of content preservation is provided, followed by a discussion on the current research bottlenecks and potential future directions.

2. Theoretical Foundations and Key Technologies of Image Style Transfer

2.1 Content Feature Extraction

By leveraging Deep Convolutional Neural Networks, a differentiable, end-to-end framework is established for the decoupled representation of image content. The VGG series networks, especially VGG-19, have become the de facto standard backbone for early style transfer research due to their clear structure and distinct feature hierarchy [1,3]. Notably, the encoding of image content by CNNs exhibits a significant hierarchical division of labor. Low-level convolutional layers, like conv1_1 and conv2_1, primarily activate local geometric primitives, including low-order visual cues such as edge orientation, corner responses, and color contrast. As the depth of the network increases, higher convolutional layers like conv4_2 and conv5_2 gradually integrate this local information to form globally discriminative semantic representations, such as object parts, categories, and scene layouts. This hierarchical feature extraction is analogous to human visual processing mechanisms and can maintain the semantic consistency of generated images in style transfer. For example, when using high-level features such as conv4_2 as supervisory signals for content loss, the main structure of the image can be effectively preserved, even when there are significant style differences. This approach relies on the semantic features learned by the network during pretraining tasks such as ImageNet classification, but these features may not be fully applicable to style transfer across different styles. Existing studies reveal that fine-tuning feature extractors or adopting self-supervised pre-training methods such as DINO and MoCo v3 can greatly improve content fidelity and detail consistency, suggesting that a "universal feature extractor" is not always optimal.

2.2 Style Algorithm Design

In image transfer, style does not refer to specific objects or narrative content, but rather acts as a visual grammar system, manifested in non-local statistical patterns like the periodicity of textures, the arrangement of brushstrokes, and the contrast of colors. Gatys et al. proposed quantifying style features via Gram matrices, using covariance between feature channels to approximate textures and brush strokes, achieving the disentanglement of style and content [1]. However, Gram matrices rely on global pooling, which easily leads to the loss of local structural information and a lack of spatial hierarchy in style expression. Besides, Gram matrices are biased toward low-level features, which may suppress high-level semantic style information, causing image structure distortion or semantic artifacts. Thus, style modeling has progressively shifted from static statistical methods to dynamic generative modeling. On the one hand, multi-layer weighted fused Gram matrices alleviate scale imbalance, while Markov Random Field methods explicitly model the spatial adjacency of textures, ensuring spatial consistency of the style. From the perspective of frequency domain analysis, the magnitude spectrum governs the texture, while the phase spectrum determines the structure, thereby effectively separating style from content. On the other hand, with the introduction of GANs, style discriminators are employed to learn style distributions, making style expression more diverse and flexible [4,5]. It is evident that style representation has evolved from fixed statistical features to generative models that can be dynamically adjusted and controlled in the latent space.

2.3 Transfer Effect Evaluation

The assessment of generated image quality has long struggled with the challenge of the so-called "metric-perception gap." Traditional metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are computationally efficient but were originally developed for image compression and denoising, rendering them inadequate for properly reflecting the influence of style transfer on visual perception [6]. For example, an oil painting-style Mona Lisa may have significant pixel-level differences from the original image (resulting in a low PSNR) but exhibits greater artistic value due to preserved structure and brush stroke rhythm. This indicates that the goal of style transfer is not only pixel reproduction but

ISSN 2959-6157

also the reconstruction of visual semantics; thus, evaluation criteria need to shift from mere "similarity" to "the degree of integration between content and style." In recent years, evaluation methods have progressed from relying on single numerical metrics to adopting multi-dimensional, collaborative verification approaches. As a deep perceptual metric, LPIPS (Learned Perceptual Image Patch Similarity) can simulate the non-linear response of the human visual system and exhibits high consistency with subjective evaluations such as Mean Opinion Score (MOS), making it the leading quantitative evaluation method at present [7]. However, image similarity alone is insufficient to fully measure style transfer quality, especially in practical applications. As a result, task-specific evaluation methods have emerged, such as testing the identity preservation of stylized faces or assessing semantic consistency in scene classification, directly linking the quality of generated images to their performance in specific applications. Furthermore, despite high costs, user research methods like A/B testing, Likert scale scoring, and eye-tracking remain crucial in high-impact research, as they more intuitively reflect the impact of style transfer on user perception.

3. Style Transfer Methods and Mechanisms under Content Fidelity

3.1 Convolutional Neural Network-based Approaches

CNNs-based methods are widely employed in image style transfer due to their ability to produce stable and controllable outputs. In these methods, hyperparameters in the loss function, such as alpha and beta, control the balance between content and style, ensuring that the generated image retains the semantic information of the content while applying the style transfer. This approach, which depends on fixed network structures for style transfer, guarantees relatively consistent results in practice, making it particularly suitable for applications that require high stability.

Specifically, feedforward networks, using pre-trained models, quickly map input images to the target style. The method proposed by Johnson et al. is based on this concept, where an optimized loss function allows the trained network to directly generate images in the target style [2]. The advantage of this approach lies in its efficiency and consistency, especially when dealing with a single style, as the network's output is predictable and stable. However, the core mechanism of this method also reveals a limitation: each network can only transfer a specific style. If different styles are required, a new network model must be trained for each additional style. Thus, in scenarios that demand multi-style adaptation, CNN-based methods lack flexibility. Nevertheless, this approach remains a key technology in image style transfer due to its efficient computation and stable transfer performance. The core advantage of its mechanism lies in the precise control of the loss function's weights, enabling a balance between content and style expression, ensuring both style consistency and content fidelity.

3.2 Generative Adversarial Network-based Approaches

GANs-based methods exhibit unique advantages in style transfer. Their core mechanism drives the generator to continuously adjust its output through an adversarial game between the generator and discriminator, thereby achieving more realistic stylistic effects. Specifically, the generator's task is to generate images, while the discriminator judges whether these images conform to the target style. As training progresses, the generator gradually optimizes its generation process, learning to produce stylized images while maintaining an adversarial relationship with the discriminator.

However, the mechanism of GANs introduces potential issues with content preservation. Since the discriminator solely assesses the realism of the image's style without enforcing content fidelity, the generator may excessively alter the image content during training in an attempt to enhance the discriminator's feedback. For example, the generator may alter the basic shape or details of the image to enhance the style effect, resulting in a mismatch between the content and the original image. Thus, the success of GAN methods often depends on the balance between the discriminator and generator. If the generator focuses too much on style and neglects the structure of the content, it may result in generated images that retain the style but lose the semantic information and detail consistency of the original image [4]. This mechanism allows GANs to achieve very strong stylistic effects in style transfer but may also cause content distortion. The generator's adjustment process prioritizes style matching over content preservation, granting it high flexibility in style expression but relatively loose control over content preservation [8].

3.3 Transformer-based Approaches

Transformers have shown distinct advantages in image style transfer, particularly in capturing global features and modeling long-range dependencies [9]. Through the self-attention mechanism, Transformers can effectively model the relationships between distant pixels in an image, enabling more efficient global coordination and main-

taining overall consistency. For instance, when handling complex scenes with both foreground and background, Transformers can preserve style consistency, enhancing the overall harmony and aesthetic appeal of the image.

Compared with traditional CNNs, Transformers can establish stronger global correlations when processing images, giving them distinct advantages in style transfer. However, the computational complexity of the self-attention mechanism increases exponentially with image resolution, leading to a significant surge in computational and memory requirements when processing high-resolution images, especially with limited hardware resources, which may degrade performance. Thus, various optimization methods have been proposed in recent years. Sparse attention mechanisms, such as Linformer, significantly reduce computational complexity by limiting the number of elements involved in calculations; Linformer optimizes resource consumption by reducing the computation of global self-attention matrices via low-rank approximation. Additionally, variants such as Vision Transformer adopt block processing and more efficient module designs to improve processing speed [9,10]. With these optimization technologies, Transformers can better adapt to style transfer tasks for high-resolution images.

4. Existing Limitations and Future Development Directions

4.1 Model Performance Constraints

The performance of current image style transfer methods is still constrained, mainly due to the inadequacy of evaluation standards and the inconsistency in generation quality. In particular, most existing evaluation methods focus on structural similarity and style consistency but overlook dimensions such as image details and content consistency, resulting in a lack of comprehensiveness and objectivity in the evaluation system [7,8]. The reliance on singular evaluation standards in style transfer technology limits both the comparability and advancement of models in real-world applications. Future research should focus on developing a multi-dimensional, objective standard that aligns with human aesthetics to more effectively assess style transfer quality. Besides, current methods still face challenges in generating high-quality images. Although methods such as GANs can produce high-quality stylized images, issues such as content distortion, excessive style fusion, or inconsistency persist in complex scenarios. These problems may be inconspicuous in static images but become particularly prominent when processing high-resolution images with intricate details.

4.2 Technical Adaptability Constraints

Although style transfer technology has achieved good results on static images, its application in cross-domain tasks, especially in video processing, still faces significant challenges. The “dynamic transfer” problem in video style transfer is mainly manifested as temporal instability of images, such as flickering and jitter [11]. Existing methods perform style conversion independently on each frame, failing to effectively handle temporal consistency and leading to visual incoordination. To address this, future research should focus on temporal modeling, ensuring temporal coherence during style transfer through technologies such as optical flow and temporal encoders, resolving stability issues in video style transfer. Additionally, the practical application of the technology faces challenges in computational resources and efficiency. While the academic community pursues extreme performance, the industrial sector often imposes stricter requirements on model efficiency and cost. Particularly on mobile devices, due to computational and memory constraints, existing high-performance style transfer models cannot meet real-time requirements [12].

4.3 Optimized Research Paths

Future research should focus on addressing key bottlenecks in style transfer. Optimizing the backbone network structure is crucial for improving performance. Although existing CNN and Transformer architectures have made progress, their high computational complexity prevents them from meeting the demands of high-resolution images and real-time applications. As an emerging generative model, diffusion models may bring breakthroughs in style-content disentanglement due to their advantages in denoising and fine-grained generation. Thus, future research can explore the integration of diffusion models with existing methods to enhance the effectiveness and efficiency of style transfer. Meanwhile, the “dynamic transfer” challenge in video style transfer remains unsolved, with traditional methods frequently causing flickering and jitter, which disrupt visual coherence. To this end, future research should combine optical flow estimation and temporal encoders with style transfer to ensure style consistency across frames. Additionally, handling style changes in long time series while maintaining consistency is another important direction. Currently, the lack of a unified evaluation standard hinders comparisons between different methods. In the future, a comprehensive evaluation benchmark covering image quality, style consistency, and temporal stability should be established to promote technological progress and application implementation [8].

ISSN 2959-6157

5. Conclusion

This paper analyzes the methods and mechanisms for content and structure preservation in image style transfer. And the results demonstrate that the three mainstream technical paths, CNN, GAN, and Transformer, each have unique advantages and limitations in addressing content preservation challenges, and there is no universal solution. Future technological advancements may depend on the fusion and complementarity of models, such as leveraging the global attention mechanism of Transformers to guide the generation process of GANs, striking a balance between style creativity and content controllability. Despite persistent challenges like an incomplete evaluation system, inadequate temporal consistency, and high computational demands, advancing the understanding of current mechanisms and fostering cross-model collaboration are anticipated to lead to transfer technology that preserves content integrity while enhancing stylistic richness, ultimately expanding its applicability across diverse fields.

References

[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

[2] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Springer, Cham.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[4] 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).

[5] Liu, M., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In Advances in neural information processing systems (pp. 700-708).

[6] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[7] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[8] Li, Y., Wang, N., & Liu, J. (2020). A comprehensive evaluation of image style transfer methods. IEEE Access, 8, 19220-19235.

[9] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. Advances in neural information processing systems, 31.

[10] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI technical report, 1(1).

[11] Ruder, M., Dosovitskiy, A., & Brox, T. (2016). Artistic style transfer for videos. In European conference on computer vision (pp. 689-705). Springer, Cham.

[12] Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.