Journal of Imaging Science and Technology® 69(3): 1–8, 2025.
© Society for Imaging Science and Technology 2025

Integrating Wavelet Transforms into Image Reconstruction Networks for Effective Style Transfer

Yunfei Chu†, Xin-Yu Xiao†, Longchen Han, Yaoshun Yue, and Maohai Lin
Key Laboratory of Paper Science and Technology of Ministry of Education, Faculty of Light Industry, Qilu University of Technology, Shandong Academy of Sciences, Jinan 250353, China
E-mail: mhlin@qlu.edu.cn

Abstract. *Image style transfer, which involves remapping the content of a specified image with a style image, represents a current research focus in the field of artificial intelligence and computer vision. The proliferation of image datasets and the development of various deep learning models have led to the introduction of numerous models and algorithms for image style transfer. Despite the notable successes of deep learning based style transfer in many areas, it faces significant challenges, notably high computational costs and limited generalization capabilities. In this paper, we present a simple yet effective method to address these challenges. The essence of our approach lies in the integration of wavelet transforms into whitening and coloring processes within an image reconstruction network (WTN). The WTN directly aligns the feature covariance of the content image with that of the style image. We demonstrate the effectiveness of our algorithm through examples, generating high-quality stylized images, and conduct comparisons with several recent methods.*
Keywords: *style transfer, wavelet transfer network, whitening and coloring transforms*
© 2025 Society for Imaging Science and Technology.
[DOI: 10.2352/J.ImagingSci.Technol.2025.69.3.000000]

1. INTRODUCTION

Image style transfer has emerged as a pivotal area of inquiry within the domain of computer vision, captivating researchers and artists alike with its potential to generate visually compelling and artistically enriched images. This innovative technique artfully melds the intrinsic content from one image with the stylistic attributes of another, effectively transplanting elements such as texture and color schemes to forge captivating composite creations [1, 2]. As illustrated in Figure 1, this process involves the complex fusion of the content features from image A with the distinct stylistic elements from the lower-left corner of images B, C, D, E, and F, ultimately generating unique transformed images that correspond to the styles of B, C, D, E, and F. This not only preserves the original content's integrity but also imbues it with a new aesthetic essence, showcasing a remarkable blend of creativity and technology.

The complexity and diversity of images pose significant challenges in achieving optimal results in image style transfer. Consequently, many scholars [3, 4] have strived to expand and refine the theoretical foundations of image style transfer by introducing new algorithms and models derived from mathematics, physics, and computer science to enhance its effectiveness. With the rapid advancement of deep learning algorithms, particularly the emergence of convolutional neural networks (CNNs), the field of style transfer has experienced significant breakthroughs and progress.

Despite the swift advancement of style transfer algorithms based on CNNs, current methods often involve trade-offs among generalization, quality, and efficiency. Optimization-based approaches can handle various styles and yield visually pleasing outcomes but entail high computational costs. Conversely, feedforward methods are more efficient but are constrained to a predetermined number of styles or may compromise visual quality. As of now, achieving universal style transfer remains a formidable challenge. Developing neural networks capable of simultaneously achieving generalization, quality, and efficiency poses significant challenges. The primary challenge lies in accurately and effectively applying extracted style features (feature correlations) to render content images in a style-agnostic manner.

To address this issue, we propose a method capable of achieving versatile style transfer. The essence of this approach lies in integrating wavelet transforms into whitening and coloring processes within the image reconstruction network. Wavelet Transfer Network (WTN) aligns the covariance of content image features directly with that of style image features. It substitutes wavelet pooling and unpooling for the operations in the VGG encoder and decoder. Figure 2 illustrates the comprehensive framework of WTN.

Our motivation stems from the principle that a network's learned function should possess its inverse operation to enable precise signal recovery, thereby achieving authentic stylization. Once training is complete, the encoder and decoder remain fixed. Leveraging the advantageous property of wavelets, which minimizes information loss, WTN can fully reconstruct signals without requiring any post-processing steps. This learning-free approach fundamentally differs from the existing methods that necessitate predefined learning of feedforward network styles and fine-tuning of new styles.

The primary contributions of this study are as follows:

†The authors contributed equally to this work.

Received June 22, 2024; accepted for publication Oct. 16, 2024; published online xx xx, xxxx. Associate Editor: Min Xu.
1062-3701/2025/69(3)/000000/8/$25.00

Figure 1. Images combining the content of a photograph with the style of several well-known artworks. (A: *The original photograph* by Binyamin Mellish. B: *Woman III* by Roy Lichtenstein, 1982. C: *Landscape at L'Estaque* by Georges Braque, 1906. D: *Improvisation No. 30 (Cannons)* by Vasily Kandinsky, 1913. E: *The Artist Looks at Nature* by Charles Sheeler, 1943. F: *Starry Night and the Astronauts* by Alma Thomas, 1972.)

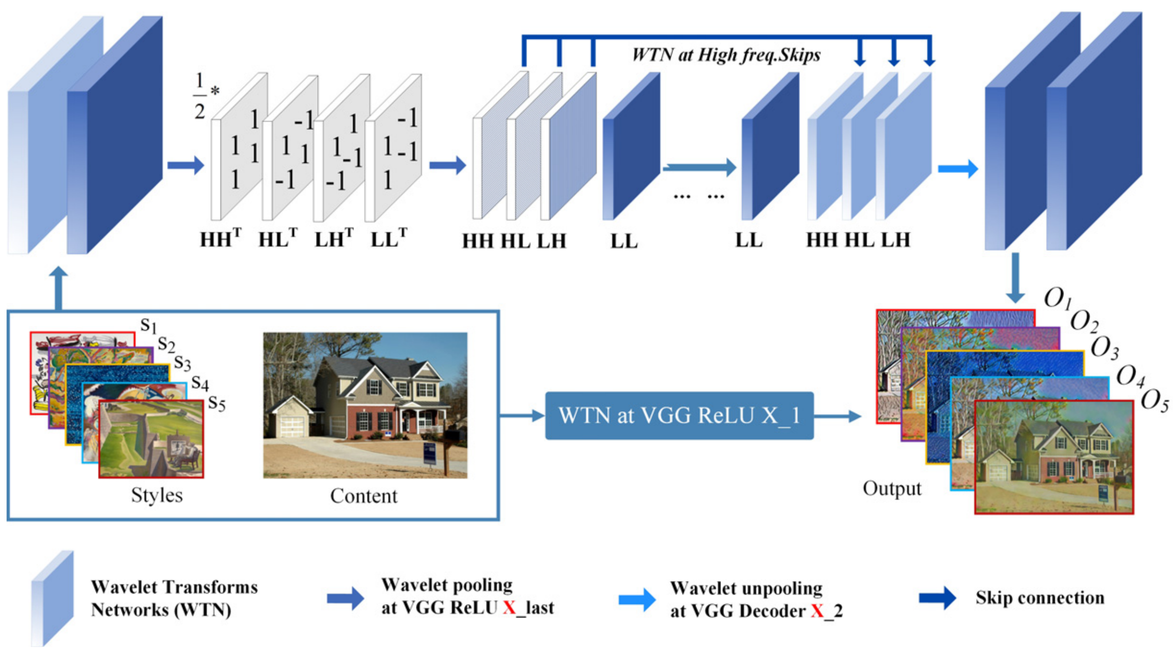

Figure 2. The overall framework of the Wavelet Transfer Network (WTN).

(a) We propose the Wavelet Transfer Network (WTN), an end-to-end photorealistic style transfer model. WTN removes the original style through whitening and introduces a new style through coloring.

(b) We integrate feature transformation with a pre-trained general encoder-decoder network, enabling the style transfer process to be implemented through straightforward feed-forward operations.

(c) We demonstrate the effectiveness of our method in universal style transfer, yielding high-quality visual outcomes. Furthermore, we showcase its application in universal texture synthesis.

2. RELATED WORK

Currently, image style transfer methods are widely applied both locally and globally. These innovative methods are broadly classified into two main categories: traditional image style transfer and neural network-based image style transfer [5], as detailed in Table I [2]. Traditional approaches, predominantly example-based, utilize the image analogy method to form a correlation between a pair of images. This correlation is then leveraged to artistically stylize additional images. However, a notable limitation of these traditional methods is their dependency on paired images portraying identical types of scenes. This requirement often renders

Table I. Summary of image style transfer methods.

Category		Method or Type	Representative	Applicable scenarios
Traditional image style migration		Brushstrokes render ideas	Refs. [6, 7]	Artistic Creation
		Image analogy ideas	Refs. [8, 9]	Image processing
		Filtering processing ideas	Refs. [10]	Real-time Image Processing
		Texture synthesis ideas	Refs. [11, 12]	Image Texture Generation
Neural network-based image style migration	Based on image iteration	Gram matrix	Refs. [1, 13]	Artistic Creation
		Maximum mean variance	Refs. [14, 15]	Image Processing
		Markov random field	Refs. [15, 16]	Image Synthesis
		Deep image analog	Refs. [17, 18]	
		Relaxation optimal transmission	Refs. [19]	
	Based on model iteration	Monostyle	Refs. [20]	Real-time Image Processing
		Multi-style	Refs. [21]	Industrial Applications
		Arbitrary style	Refs. [22, 23]	

them less effective for arbitrary style transfers where such specific scene congruence is lacking, thereby limiting their versatility in applications demanding a broader stylistic application.

Gatys et al. [1] proposed an algorithm based on the correlations between deep features, implemented within an iterative optimization framework, achieving arbitrary stylization. Following this, scholars have developed various methods to address different aspects such as speed, quality, user control, diversity, semantic understanding, and photo-realism [24]. These methods are simple to implement and can produce near real-time results, which is beneficial for applications requiring the rapid processing of large volumes of images [25].

Classical methods mainly match the colors and tones of images but are limited in scope. Research scientists have proposed methods such as Deep Photo Style Transfer (DPST) and a variant of photorealistic style transfer, WCT (PhotoWCT), to improve style transfer effects [24]. However, few of these methods demand significant computational resources and may result in blurred final outputs [26]. In contrast, our proposed method preserves the fine structure of images in an end-to-end manner with minimal spatial distortion, thus eliminating the need for additional post-processing steps.

3. WAVELET TRANSFER NETWORK

3.1 *Reconstruction Decoder*

We developed a self-encoder network tailored for general image reconstruction tasks. For this purpose, the VGG-19 model was selected to serve as the encoder; this component was kept static while a corresponding decoder network was trained specifically to invert the VGG features back to their original image formats, as depicted in Fig. 1(a). The architecture of the decoder mirrors that of VGG-19 up to the Relu_X_1 layer, incorporating layers of nearest-neighbor upsampling to effectively expand the feature maps. In an effort to thoroughly assess the utility of features extracted at various depths, we extracted feature maps from five distinct layers within the VGG-19 architecture, specifically at Relu_X_1 layers (where X = 1, 2, 3, 4, 5), and trained individual decoders for each layer [27]. To achieve high-fidelity reconstruction of the input images, we employed both pixel reconstruction loss and feature loss in our training process [25].

$$L = \|I_o - I_i\|_2^2 + \lambda \|\Phi(I_o) - \Phi(I_i)\|_2^2. \qquad (1)$$

In the context of this study, I_i and I_o represent the input image and reconstructed output, respectively, while Φ denotes the VGG encoder extracting Relu_X_1 features. Additionally, λ serves as the weight balancing the two losses. Upon the successful conclusion of the training phase, the decoder is firmly established in a static configuration—this means that no further fine-tuning is undertaken [28]. It is subsequently employed as a robust feature inverter, dedicated to reversing the encoded features back to their original form with precision and reliability [29].

Building upon this foundational architecture, we incorporate the Whitening and Coloring Transform (WCT) process, as detailed in Figure 3. This method employs a layered approach using the VGG network, where each level applies WCT to blend extracted content features with style features iteratively extracted from various layers. This sequential process aligns the feature covariance of the content with those of the style image at each corresponding level, followed by progressive reconstruction using dedicated decoders for each layer. This ensures a nuanced style application across multiple scales, maintaining the content's structural integrity while effectively infusing the style attributes.

3.2 *Whitening and Coloring Transforms*

Given a pair of content images I_c and style images I_s, we initially extract their vectorized VGG feature maps

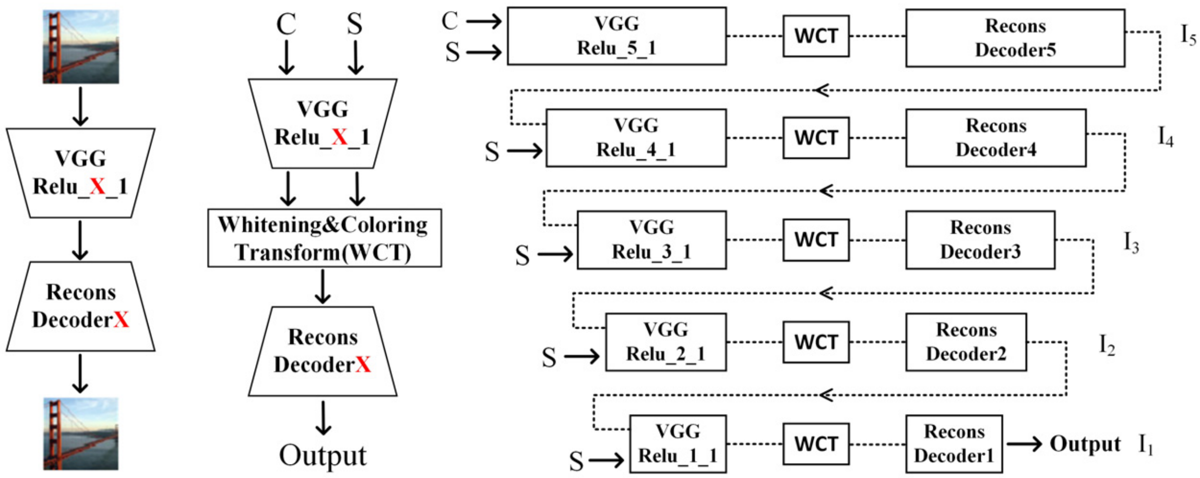

Figure 3. The workflow of WCT.

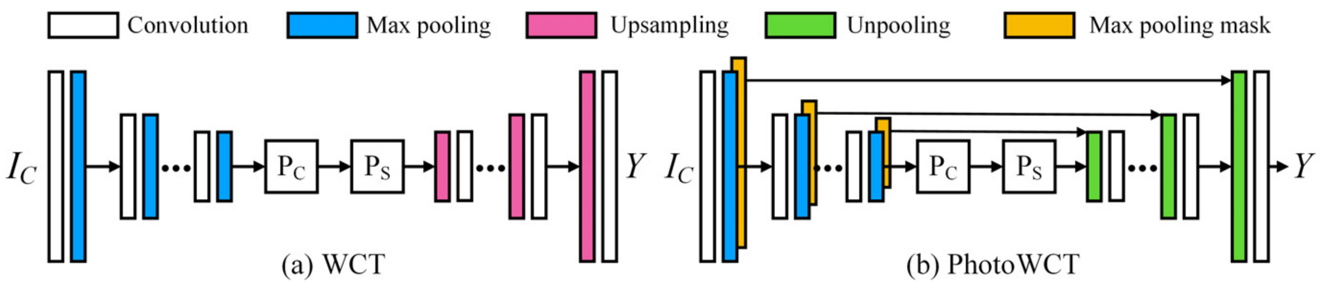

Figure 4. The differences between WCT and PhotoWCT.

$f_c \in \mathbb{R}^{C\times H_c\times W_c}$ and $\boldsymbol{f}_s \in \mathbb{R}^{C\times H_s\times W_s}$. H_c, W_c (as well as H_s, W_s) represent the height and width of the content (style) features, while C denotes the number of channels. If f_c is directly fed into the decoder, it will reconstruct the original image I_c. Subsequently, we propose the use of whitening and coloring transformations to adjust f_c to match the statistics of $\boldsymbol{f}_s$. The goal of WCT is to directly transform f_c to match the covariance matrix of $\boldsymbol{f}_s$. This process involves two steps: whitening and coloring transformations [30].

As depicted in Figure 4, the workflows of WCT and PhotoWCT differ significantly. WCT employs a direct approach, simply upsampling whitened and colored content features to match the style dimensions (panel a). In contrast, PhotoWCT (panel b) incorporates additional steps such as unpooling and the use of max pooling masks, designed to preserve more structural details and enhance photorealism during the style transfer process. These enhancements in PhotoWCT facilitate a more refined transformation, ensuring finer control over spatial details and help address the common loss of detail seen in traditional WCT applications, ultimately yielding more photorealistic outputs.

3.2.1 *Whitening Transformation*

Before whitening, we first center f_c by subtracting its mean vector $\mathbf{m}_c$ and linearly transform f_c to obtain $\widehat{f_c}$ as shown in Eq. (2), ensuring that the feature maps are uncorrelated ($\hat{f}_c\hat{f}_c^T = I$).

$$\hat{f}_c = \boldsymbol{E}_c D_c^{-\frac{1}{2}} E_c^T f_c. \qquad (2)$$

In this equation, $\boldsymbol{D}_c$ represents a diagonal matrix containing the eigenvalues of the covariance matrix $\hat{f}_c\hat{f}_c^T$, where $f_c f_c$ belongs to $\mathbb{R}^{C\times C}$, and $\boldsymbol{E}_c$ denotes the corresponding orthogonal matrix of eigenvectors. It satisfies the condition $f_c f_c^T = \boldsymbol{E}_c \boldsymbol{D}_c \boldsymbol{E}_c^T$.

3.2.2 *Color Transformation*

We begin by centering $\boldsymbol{f}_s$ through subtraction of its mean vector $\mathbf{m}_s$, followed by color transformation, which is essentially the inverse operation of whitening, linearly transforming $\hat{f}_c$ as shown in Eq. (3), yielding $\hat{f}_{cs}$ with the desired correlation between feature maps

$$\hat{f}_{cs} = E_s D_s^{\frac{1}{2}} E_s^T \hat{f}_c, \quad (\hat{f}_{cs}\hat{f}_{cs}^T = \boldsymbol{f}_s \boldsymbol{f}_s^T). \qquad (3)$$

$\boldsymbol{D}_s$ is a diagonal matrix containing eigenvalues of the covariance matrix $\boldsymbol{f}_s\boldsymbol{f}_s^T$ and $\boldsymbol{E}_s$ is the orthogonal matrix of corresponding eigenvectors. Finally, we recenter f_{cs} by adding the mean vector $\mathbf{m}_s$ of the style, i.e., $\hat{f}_{cs} = \hat{f}_{cs} + \mathbf{m}_s$.

After WCT, we can mix $\hat{f}_{cs}$ with the content feature f_c as shown in Eq. (4), then feed it into the decoder for the user to

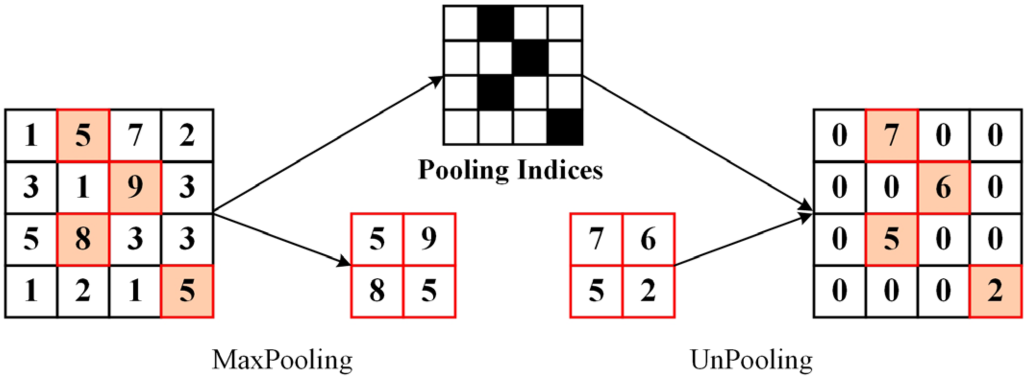

Figure 5. The process of pooling and unpooling.

control the intensity of stylization effect:

$$\hat{f}_{cs} = \alpha \hat{f}_{cs} + (1 - \alpha) f_c. \tag{4}$$

α serves as the style weight for the user to control the transfer effect.

3.3 *Wavelet Corrective Transfer*

3.3.1 *Haar Wavelet Pooling and Unpooling*

We introduce Haar wavelets, referred to as the main components of pooling and unpooling, to elucidate the primary constituents of our model. Haar wavelet pooling consists of four kernels: $\{LL^\top, LH^\top, HL^\top, HH^\top\}$, with low-pass ($L$) and high-pass ($H$) filters defined as

$$L^\top = \frac{1}{\sqrt{2}}(1 \quad 1), \quad H^\top = \frac{1}{\sqrt{2}}(-1 \quad 1). \tag{5}$$

As a result, unlike conventional pooling operations, Haar wavelet pooling outputs four channels. In this study, the low-pass filter captures smooth surfaces and textures, while the high-pass filter extracts information about vertical, horizontal, and diagonal edge styles. For simplicity, we denote the output of each kernel as LL, LH, HL, and HH, respectively.

As illustrated in Figure 6, this method combines wavelet-based pooling/unpooling with WTN to achieve sophisticated style transfer. Wavelet pooling decomposes images into different frequency components, enabling detailed manipulation at various scales. Multiple modules adjust the content features to match the style's covariance properties, while skip connections preserve high-frequency details, ensuring that the final output retains both the style's aesthetics and the content's structural integrity. This approach enhances photorealistic style transfer with a focus on detail retention.

A significant advantage of our wavelet pooling technique is its ability to precisely reconstruct the original signal through a process called wavelet unpooling [31]. By reversing the pooling operation, wavelet unpooling meticulously restores the signal to its initial form, utilizing element-wise transposed convolution followed by a summation of results to achieve complete signal recovery. (For an in-depth explanation, please refer to supplementary material.) This distinctive capability enables our model to stylize images while preserving their intrinsic details and substantially minimizing information loss and noise amplification. In stark contrast, traditional max pooling methods do not possess a precise inverse function, leading to a scenario where encoder-decoder networks, such as those employed in WCT and PhotoWCT, are unable to achieve full signal restoration [32]. This limitation underscores the superior functionality of our approach in maintaining the integrity and quality of the original imagery during the style transfer process.

It should be highlighted that while Haar wavelet pooling and unpooling is highly effective, it is not the only technique capable of flawlessly reconstructing the original signal. Fourier Transforms [33], for instance, also allow for perfect reconstruction of the original data. However, while Fourier Transforms analyze the signal in its entirety, Haar wavelets partition the original signal into channels that capture different constituent parts. This selective partitioning enables Haar wavelets to achieve superior stylization effects, allowing for more precise manipulation and analysis of specific signal components. Consequently, Haar wavelets are preferred in applications where such detailed control and stylization are critical.

4. EXPERIMENTAL RESULTS

4.1 *Decoder Training*

For the multi-level stylization method, we trained five reconstruction decoders corresponding to the Relu_X_1 ($X = 1, 2, \ldots, 5$) layers of VGG-19. These decoders were trained on the Microsoft COCO dataset, with the weight balancing the two losses in Eq. (1) set to 1.

4.2 *Style Transfer*

To substantiate the efficacy of our proposed algorithm, we provide a detailed comparative analysis against existing methodologies, as illustrated in Table I. Additionally, we showcase the stylization results achieved by our algorithm in Fig. 6. To ensure a fair comparison, we meticulously adjusted the style weights of competing methods to optimize

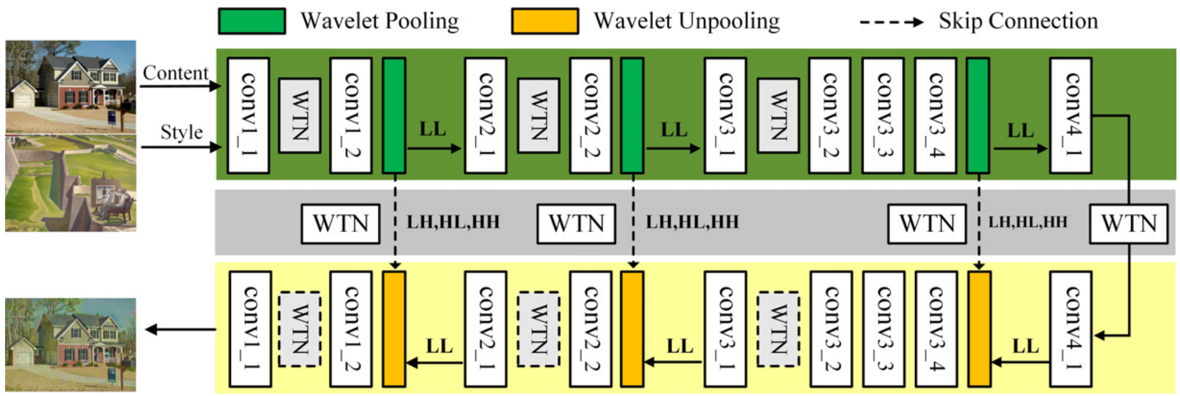

Figure 6. The details of pooling and unpooling in WTN network.

Table II. Differences between our approach and other methods.

	TNet	Gatys et al.	WCT	PhotoWCT	**Ours**
Arbitrary	✓	✓	×	✓	✓
Efficient	✓	✓	✓	×	✓
Learning-free	×	×	×	✓	✓

Table III. Quantitative comparisons between different stylization methods.

	WCT	PhotoWCT	TNet	Gatys et al.	Ours
log(Ls)	8.1	8.7	5.2	9.2	**7.1**
Preference/%	17.2	26.3	9.6	13.4	**29.9**
Time/sec	2.6	0.39	0.09	1.22	**0.93**

their stylization effects. The optimization-based method referenced in Refs. [10, 16] is adept at handling a wide array of arbitrary styles, yet it occasionally grapples with issues related to unexpected local minima. Conversely, while the technique mentioned in Ref. [14] markedly enhances stylization speed, it unfortunately compromises both quality and versatility. This often results in the generation of repetitive and predictable patterns that detract from the richness and depth of the image content.

Table II provides a comparative analysis of our method against TNet, Gatys et al., WCT, and PhotoWCT based on three criteria: arbitrary style transfer, efficiency, and being learning-free. Our method, along with TNet, Gatys et al., and PhotoWCT, supports arbitrary style transfer, offering flexibility for diverse styles, while WCT is more limited in this regard. In terms of efficiency, all methods, including ours, perform well, enabling fast processing. Additionally, our approach, like WCT and PhotoWCT, is learning-free, requiring no additional training post-deployment, unlike TNet and Gatys et al., which demand further fine-tuning. This comparison highlights our method's balance of flexibility, speed, and ease of use.

Our work closely aligns with recent approaches [3, 6, 17] in terms of generalization but offers more appealing stylization results. In Ref. [9], content features are replaced with style features based on patch similarity, limiting its ability to retain content, while failing to adequately reflect style when transmitting only low-level information. Similarly, in Ref. [15], content features are adjusted to match the mean and variance of style features, which proves ineffective in capturing high-level representations of style. Even when trained on a set of styles, it fails to generalize well
Q.2 to unseen styles.

Results shown in Table III demonstrate the ineffectiveness of the method in Ref. [16] in capturing and synthesizing significant style patterns, especially for complex styles with rich local structures and non-smooth regions. In contrast, Figure 5 vividly displays the superior stylization results achieved by our method. Remarkably, without the necessity of learning any specific style, our approach skillfully captures and replicates visually significant patterns found in style images.

Furthermore, our method excels in ensuring that key components within content images are not only preserved but are also beautifully stylized, enhancing the overall visual impact. This is a notable improvement over other techniques, which tend to merely overlay patterns onto the smoother areas of the image, often overlooking more textured or detailed regions. This nuanced approach to stylization underscores the advanced capabilities of our method, setting it apart in terms of both effectiveness and aesthetic fidelity.

In addition to qualitative assessment, we quantitatively evaluated the differences between different methods by calculating the covariance matrix differences (Ls) on all five VGG feature layers, including stylization results and given style images. We randomly selected 10 content images and 40 style images, calculated the average differences for all styles, and present the results. The quantitative results indicate that our stylization results have lower Ls, suggesting closer proximity to style statistical data. Figure 7 shows the actual effect of WTN after 200, 300, 400, and 500 training epochs. The figure illustrates the progressive improvement in style transfer quality as training advances, highlighting how increased training rounds result in more refined and coherent stylized outputs. Q.1

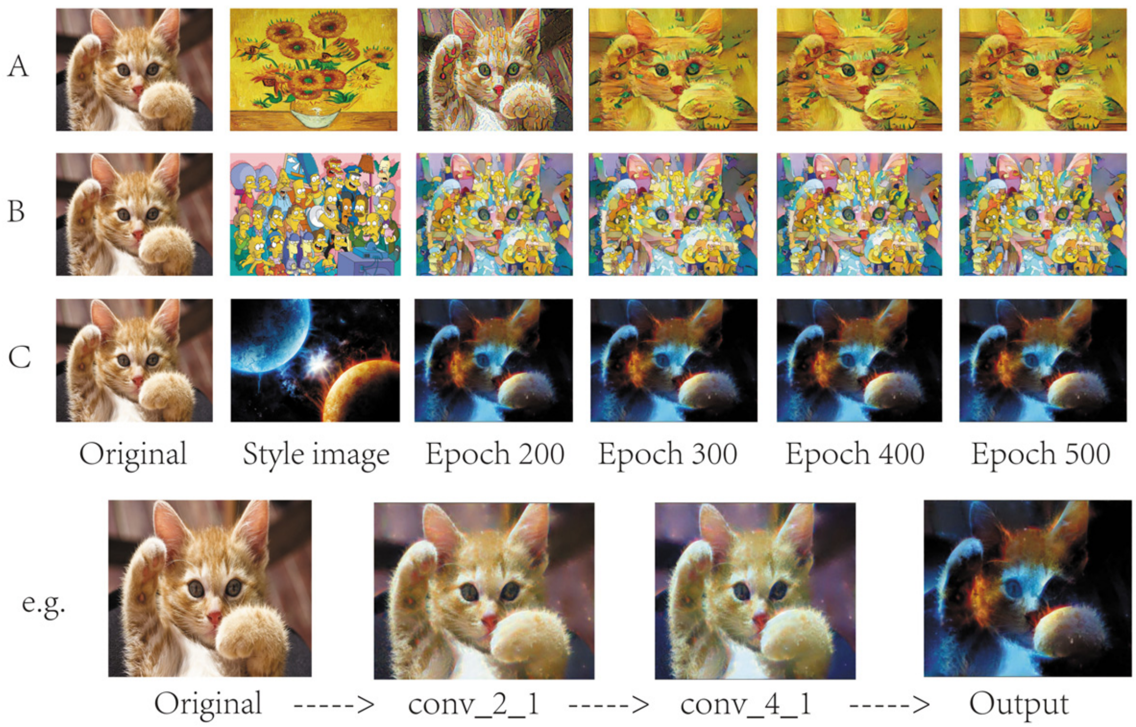

Figure 7. Actual effect of WTN (Epoch for training 200, 300, 400, 500 rounds).

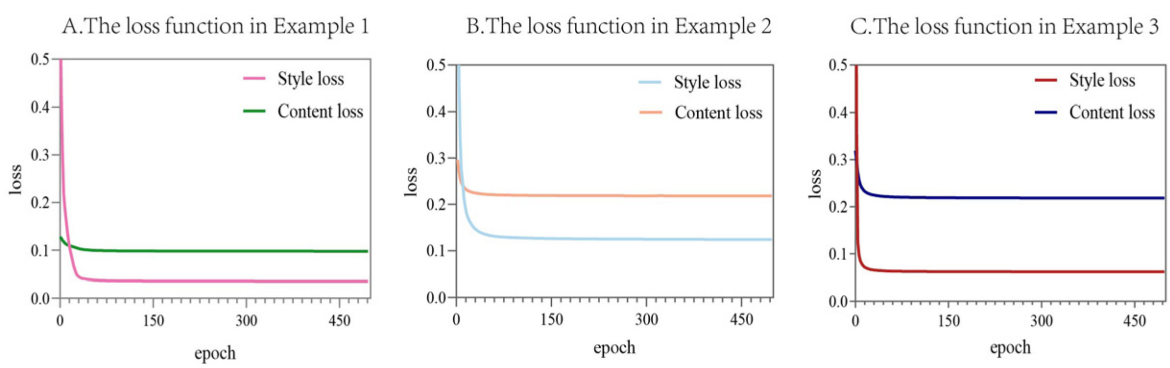

Figure 8. The loss function of Examples used in Fig. 5.

To further assess the effectiveness of our method, we conducted a user study to evaluate the subjective preferences of ours shown in Fig. 5. We used 5 content images and 30 style images, generating 150 results for each content/style pair for each method. We randomly selected 3 style images for each subject to evaluate. The stylized images were displayed side by side on a webpage in random order. Each subject was asked to select their favorite result for each style. The study indicates that our method received more votes for better stylization results. Exploring evaluation metrics based on human visual perception for general image synthesis problems may be an intriguing direction.

In our comparative analysis of efficiency, we meticulously evaluated our method against others in the field. Gatys et al.'s approach [1] is notably slower, primarily because it relies on iterative optimization loops that typically require at least 500 iterations to achieve satisfactory results. Conversely, methods [34] and [24] demonstrate higher efficiency as they operate based on a single pass through a pre-trained network. Method [2], while also leveraging a single forward pass, tends to be relatively slow; this is attributed to the extensive feature swapping operations that must be conducted across thousands of image patches.

Our method maintains a commendable level of efficiency, though it is marginally slower compared to alternatives [5, 26, 28, 30, 31]. This slight delay is largely due to the feature value decomposition step integral to the WCT. Importantly the computational load of this particular step does not scale with the size of the image. Instead, it is contingent solely upon the number of filters–or the dimensions of these filters–highlighting a significant advantage in terms of scalability and practical applicability in diverse contexts where image size can vary substantially.

5. CONCLUSION

In this study, we introduced a sophisticated universal style transfer algorithm designed to obviate the need for individual style learning. Our approach centers around the deployment of an autoencoder specifically trained for image reconstruction. This strategic training enabled us to meticulously unfold the image generation process. Within this framework, we incorporated whitening and coloring transformations during the forward pass, effectively aligning the statistical distribution and correlation of intermediate features between the content and style images.

Moreover, we developed a comprehensive multi-level stylization pipeline that systematically integrated style information at various levels, thereby significantly enhancing the final visual outcomes. Additionally, this innovative approach is not only limited to style transfer but is also highly effective for texture synthesis applications.

Empirical evaluations of our algorithm reveal its exceptional ability to generalize across a diverse range of arbitrary styles, distinctly outperforming existing state-of-the-art techniques. These results underscore the robustness and versatility of our method, setting a new benchmark in the field of style transfer and texture synthesis.

ACKNOWLEDGMENT

This study did not receive any funding support and the authors have no conflicts of interest. The code and images utilized in this paper are available upon request from the corresponding author.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style". Preprint, arXiv:1508.06576 (2015).
[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2016), pp. 2414–2423.
[3] M. Berning, K. M. Boergens, and M. Helmstaedter, "SegEM: efficient image analysis for high-resolution connectomics," Neuron **87**, 1193–1206 (2015).
[4] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," PLoS Comput. Biol. **10**, e1003963 (2014).
[5] Q. Cai, M. Ma, C. Wang, and H. Li, "Image neural style transfer: a review*," Comput. Electr. Eng. **108**, 108723 (2023).
[6] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell. **40**, 834–848 (2018).
[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: a deep convolutional activation feature for generic visual recognition," *Int'l. Conf. on Machine Learning* (PMLR, Bejing, 2014), pp. 647–655.
[8] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," *Proc. Seventh IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 1999).
[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, E. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," *ACM Conf. on Multimedia (MM)* (ACM, New York, NY, 2014), pp. 675–678.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition". Preprint, arXiv:arXiv:1409.1556 (2014).
[11] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," *Proc. 28th Annual Conf. on Computer Graphics and Interactive Techniques* (ACM, New York, NY, 2001), pp. 327–340.
[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).
[13] S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain IT cortical representation," PLoS Comput. Biol. **10**, e1003915 (2014).
[14] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 640–651 (2017).
[15] J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," Int. J. Comput. Vis. **40** (2000).
[16] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *28th Conf. on Neural Information Processing Systems* (NIPS, Montreal, 2014).
[17] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," *Proc. 27th Annual Conf. on Computer Graphics and Interactive Techniques* (ACM, New York, NY, 2000), pp. 479–488.
[18] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound-constrained optimization," ACM Trans. Math. Softw. **23**, 550–560 (1997).
[19] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2015).
[20] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *IEEE Int'l. Conf. on Computer Vision* (IEEE, Piscataway, NJ, 2015).
[21] U. Güçlü and M. A. J. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," J. Neurosci. **35**, 10005–10014 (2015).
[22] J. B. Tenenbaum and W.T. Freeman, "Separating style and content with bilinear models," Neural Comput. **12**, 1247–1283 (2000).
[23] Q. Wang, S. Li, Z. Wang, X. Zhang, and G. Feng, "Multi-source style transfer via style disentanglement network," IEEE Trans. Multimedia **26**, 1373–1383 (2024).
[24] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2017).
[25] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," *14th European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2016).
[26] Y. Li, M. Y. Liu, X. Li, M. H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," *15th European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2018).
[27] H. Lee and H. C. Choi, "Artifact-free image style transfer by using feature map clamping," IEEE Access **8**, 89489–89496 (2020).
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2016).
[29] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," *Computer Vision – ECCV 2014* (Springer, Cham, 2014).
[30] T. Y. Chiu, "Understanding generalized whitening and coloring transform for universal style transfer," *IEEE/CVF Int'l. Conf. on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2019).
[31] S. Mallat, "Understanding deep convolutional networks," Phil. Trans. R. Soc. A-Math. Phys. Eng. Sci. **374** (2016).
[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *13th European Conf. on Computer Vision (ECCV)* (Springer, Cham, 2014).
[33] Z. Jin, X. Shen, B. Li, and X. Xue, "Style spectroscope: improve interpretability and controllability through Fourier analysis," Mach. Learn. **113**, 3485–3503 (2024).
[34] H. Lee, S. Seo, S. Ryoo, and K. Yoon, "Directional texture transfer," *Proc. 8th Int'l. Symp. on Non-Photorealistic Animation and Rendering* (ACM, New York, NY, 2010), pp. 43–48.