MILESTONE FOR NEURAL STYLE TRANSFER

A PREPRINT

Hehao Qin
College of Urban and Environmental Science
Peking University
2100013272@stu.pku.edu.cn

Bokai Huang
Yuanpei College
Peking University
2100017828@stu.pku.edu.cn

Chaoran Liu
School of Earth and Space Science
Peking University
2300012622@stu.pku.edu.cn

January 13, 2025

ABSTRACT

This paper explores and compares different approaches to Neural Style Transfer (NST), focusing on traditional Gatys-style NST, Laplacian-steered NST, and the CycleGAN model. Gatys-style NST leverages a pre-trained VGG19 model to extract hierarchical features and achieve texture and style synthesis, while Laplacian-steered NST incorporates a Laplacian pyramid to better preserve semantic content and reduce artifacts. CycleGAN, a GAN-based method, is employed for unpaired image-to-image translation tasks. Experimental results demonstrate the effectiveness of each approach, highlighting the strengths of LapStyle in preserving semantic structures and mitigating artifacts, as well as the challenges of CycleGAN in object recognition and domain-specific style transfer. These findings provide valuable insights into the limitations and improvements needed for neural style transfer models.

Keywords Neural Style Transfer · Constrain Optimization · Convolution Neural Network

1 Introduction

Rendering the semantic content of an image in different styles has long been a challenging yet rewarding task. Prior to the advent of neural style transfer (NST), most style transfer methods relied on non-parametric approaches for texture synthesis while preserving semantic information. For instance, early works utilized correspondence maps to guide texture synthesis [Efros and Freeman, 2023] or focused on transferring high-frequency details while retaining coarse structures [Ashikhmin, 2003]. However, these methods were limited to low-level feature representations and often required handcrafted designs, restricting their application to controlled subsets of images, such as faces or text characters [Tenenbaum and Freeman, 2000].

The emergence of deep convolutional neural networks (CNNs) revolutionized computer vision by enabling models to separate content and style in images more effectively [Krizhevsky et al., 2012]. Gatys et al. [Gatys et al., 2016] proposed a seminal neural style transfer method that leveraged the hierarchical feature extraction capabilities of a pre-trained VGG network. Lower layers captured textures and patterns, while higher layers encoded semantic content. This method formulated style transfer as an optimization problem, where a generated image was iteratively refined to balance the preservation of content and the application of style.

Subsequent extensions to this method further improved performance. Gatys et al. [Gatys et al., 2017] introduced guided Gram loss for spatial consistency and techniques to preserve color distribution in stylized images. These advancements expanded the applicability of NST but still faced challenges in maintaining local structural details and smooth transitions.

Inspired by Poisson Image Editing [Gangnet et al., 2003], Laplacian-steered NST was introduced by Li et al. [Li et al., 2017] to address these limitations. This method incorporated a Laplacian pyramid to achieve multi-scale processing, allowing finer control over stylization while preserving local structures. By progressively applying style transfer across different spatial resolutions, Laplacian-steered NST ensured both global coherence and detailed accuracy in the stylized output.

In this paper, we investigate and compare the standard NST method and its Laplacian-steered extension to evaluate their respective strengths and limitations. We employ the pre-trained VGG19 network as the backbone and use a curated dataset of diverse content and style images for a comprehensive evaluation. This study provides insights into the capabilities and challenges of both methods, paving the way for future improvements in neural style transfer.

2 Model Structure

2.1 Neural Style Transfer (NST)

Gatys-style NST is based on the VGG19 network pretrained on ImageNet, known for its ability to extract hierarchical image features [Simonyan, 2014]. VGG19 consists of 19 layers, including 16 convolutional layers and 3 fully connected layers. Lower layers capture local textures and patterns, while higher layers encode semantic content.

NST employs three losses: content loss to preserve the semantic structure of the content image, style loss to capture stylistic patterns of the style image, and total variation (TV) loss to ensure smoothness in the output image.

2.1.1 Loss Functions

Content Loss: Measures differences in feature maps between the content and generated images:

$$\mathcal{L}_{\text{content}} = \sum \text{Mean}\left((F_c - F_t)^2\right).$$

Style Loss: Compares Gram matrices of style and generated images across layers:

$$\mathcal{L}_{\text{style}} = \sum_l w_l \cdot \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l\right)^2.$$

TV Loss: Encourages spatial smoothness:

$$\mathcal{L}_{\text{TV}} = \sum_{i,j} \left(|I_{i+1,j} - I_{i,j}| + |I_{i,j+1} - I_{i,j}|\right).$$

Total Loss: Combines all components:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{content}} + \beta \cdot \mathcal{L}_{\text{style}} + \gamma \cdot \mathcal{L}_{\text{TV}}.$$

2.2 Laplacian Neural Style Transfer (LapStyle)

LapStyle uses a Laplacian pyramid for multi-scale decomposition, allowing finer control over stylization. Content and style images are processed at multiple resolutions, with outputs progressively merged to ensure global coherence and local detail preservation.

2.2.1 Laplacian Style Model

The model integrates content and style features at each pyramid level using convolutional layers with instance normalization and ReLU activations. The stylized output is blended with the original content image:

$$\text{Output} = w_c \cdot \text{Content Features} + w_s \cdot \text{Style Output}.$$

Histogram matching is applied to ensure color consistency with the style image.

2.2.2 Optimization Process

The model iteratively refines the stylized output at each pyramid level. The final image is reconstructed by combining all levels, ensuring seamless integration of content and style. Multi-resolution processing helps retain semantic accuracy and avoid artifacts.

2.3 CycleGAN for Image-to-Image Translation

CycleGAN is a GAN-based model designed for unpaired image-to-image translation [Zhu et al., 2017]. It uses two generators (G and F) for bidirectional domain mapping and two discriminators (D_A and D_B) to distinguish real from generated images. Cycle-consistency loss ensures reversibility:

$$F(G(x)) \approx x, \quad G(F(y)) \approx y.$$

CycleGAN excels at tasks like style transfer and seasonal translation but relies heavily on dataset quality and accurate object recognition. In our implementation, a pretrained CycleGAN model translates images from the 'summer' domain to 'winter' and transforms horses into zebras.

3 Model Output

Figure 1: Strange Color Blocks in $output1\_5$

Figure 2: Neural Style Transfer Results: 3 content images, 6 style images, and their combinations.

The input content and style images and output blended images are aligned here. Our results show that NST successfully transforms three pictures with different content into our assigned style. The output images contain the texture and color details from the style input, but keep the semantics of the original content. But we also notice that some part in the output images are a little wierd compared to the original images such as part a and part b in $output1\_5$:

We blame the appearance of abnormally bright regions that do not exist in the original image is often to the over-activation of certain features in the hidden layers of the model. This is typically due to the mismatch between the content and style features or the biases in the Gram matrix computation. Additionally, optimization issues, such as high learning rates, and insufficient regularization may lead to pixel intensity imbalances, resulting in these artifacts.

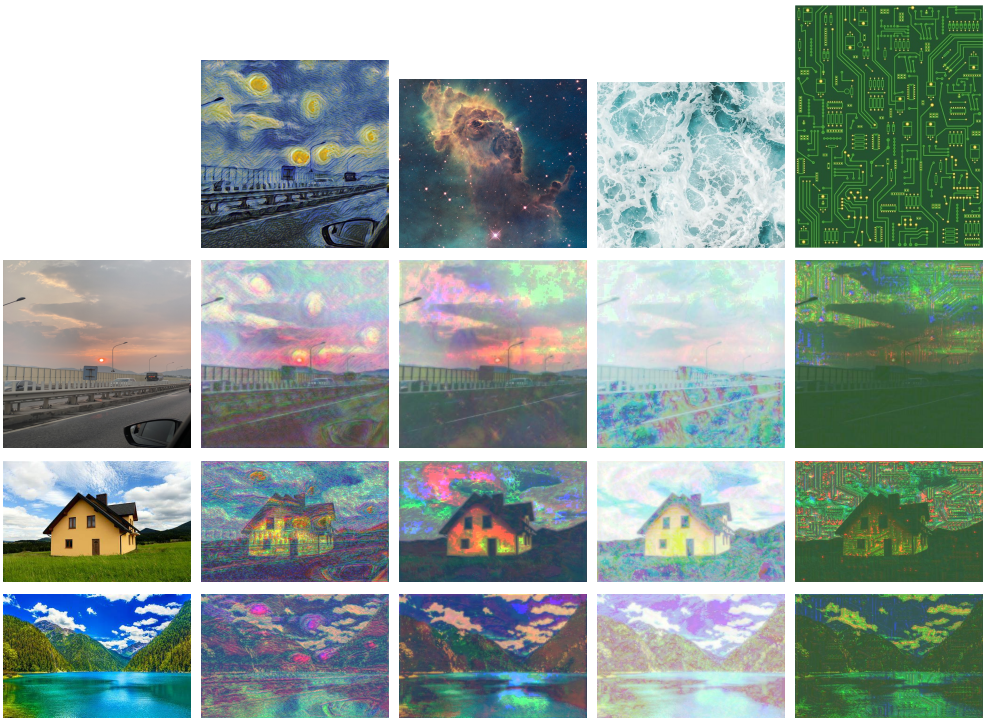


Figure 3: Laplacian Neural Style Transfer Results: 3 content images, 4 style images, and their combinations

To address this issue, we suggest adjusting the loss function weights, particularly reducing the style loss weight or increasing the regularization weight (e.g., Total Variation Loss), which can balance the optimization process. This can also be mitigated by selecting more suitable VGG layers for content and style representation, preprocessing the style image to reduce extreme features, and using more stable optimization methods with lower learning rates. Post-processing techniques, such as histogram matching or smoothing, can further refine the output and mitigate abnormal bright regions.

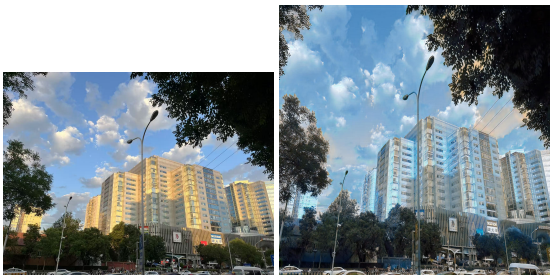
The output of LapStyle Transfer demonstrates excellent performance in preserving the semantic content of the input image. None of the original semantics are lost in the generated images, and the model effectively retains semantic details across all levels of representation. The objects in the generated images, such as mountains and buildings, exhibit minimal distortion in their contours. Notably, compared to traditional Neural Style Transfer (NST), LapStyle avoids introducing abnormally bright color artifacts that were not present in the original image. This is mainly owing to the employ of multi-scale processing, progressive optimization, stronger regularization, and smoother integration of style features, which in combination ensure a more balanced and natural output.

For a full comparison, we also list the results of CycleGAN model. We ran CycleGAN to test two style transfer tasks: horse2zebra and summer2winter. CycleGAN's style transfer involves first identifying object A, which essentially performs a form of semantic segmentation. The identified object A is then subjected to style transfer, while the regions outside of the object remain unchanged. However, we observed several limitations with this approach.

First, the accuracy of object identification depends on the model's feature extraction and segmentation precision, as well as the quality of the training dataset. For example, in our case, example (a) successfully identified the main body of the mule and applied the style transfer appropriately. However, in example (b), the model incorrectly identified non-object A regions as part of the object and altered their style. In example (d), the model failed to recognize the cartoon object, suggesting that the training dataset likely did not include cartoon-like images. These shortcomings could be addressed by improving the quality of the training dataset and fine-tuning the model.

(a) (a)

(b) (b)



(c) (c)

(d) (d)

Figure 4: CycleGAN results: a) a mule has been successfully transferred into zebra; b) a horse has been transferred into zebra, but the human's leg is also identified as a part of horse; c) an image taken in summer has been transferred into winter; d) Sunny Starscout form My Little Pony, but the model fails to recognize it as a horse

4 Discussion

Our experiments reveal significant differences between the traditional Gatys-style NST, Laplacian-steered NST, and CycleGAN methods in terms of semantic preservation, artifact reduction, and generalization. Gatys-style NST, though effective in capturing style features, often introduces artifacts like abnormal color blocks due to over-activation of features in certain layers. In contrast, Laplacian-steered NST mitigates these issues by employing multi-scale processing and Laplacian constraints, which ensure smoother and more balanced outputs. It excels at preserving semantic content while maintaining high stylistic fidelity.

CycleGAN, while offering flexibility for unpaired image-to-image translation, relies heavily on the accuracy of object recognition and segmentation. This dependency leads to limitations when the training dataset lacks diversity or when non-object regions are misclassified. For example, the model failed to recognize cartoon objects, suggesting a domain mismatch in the training data. Moreover, misclassification of non-target regions, as seen in some cases, highlights the need for improved feature extraction and segmentation mechanisms.

In summary, while NST and LapStyle provide a significant advancement in semantic preservation and artifact reduction, CycleGAN showcases the potential of domain translation but requires improvements in handling diverse and complex datasets. Future research should focus on combining the strengths of these approaches, incorporating robust training datasets, and exploring adaptive loss functions to address their respective limitations.

Acknowledgment

Code and data have been attached to this file.

Contribution

Bokai Huang contributed to the implementation and analysis of the NST section. Chaoran Liu focused on the LapStyle section. Hehao Qin handled the CycleGAN implementation and took responsibility for drafting the manuscript. All authors collaboratively worked on the formatting and revision of the final paper.

References

Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 571–576. 2023.

N Ashikhmin. Fast texture transfer. *IEEE computer Graphics and Applications*, 23(4):38–43, 2003.

Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3985–3993, 2017.

Michel Gangnet, Andrew Blake, et al. Poisson image editing. In *Acm Siggraph*, pages 313–318, 2003.

Shaohua Li, Xinxing Xu, Liqiang Nie, and Tat-Seng Chua. Laplacian-steered neural style transfer. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1716–1724, 2017.

Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.