Layer-Wise Weighting for Granular Control in Neural Artistic Stylization

Chetan Tyagi
ctyagi@ualberta.ca

Linh Le
lvle@ualberta.ca

February 16, 2026

Abstract

Neural Style Transfer (NST) synthesizes images by recombining structural content with artistic style using deep convolutional neural networks. In standard formulations, style loss is aggregated uniformly across multiple convolutional layers, limiting explicit control over spatial-scale contributions. We propose a Layer-Wise Style Weighting (LWSW) framework that introduces independent coefficients for each selected VGG-19 feature layer. Because shallow layers encode fine textures while deeper layers capture higher-level abstractions, redistributing their contributions enables controllable transitions between texture-dominant and structure-dominant stylization. We evaluate perceptual and structural impacts of layer-wise weighting and extend the framework to temporally consistent video stylization.

Introduction

Neural Style Transfer (NST) [8] demonstrated that deep convolutional networks implicitly separate content and style representations. Content is preserved through high-level feature activations, while style is encoded through second-order feature correlations captured by Gram matrices.

In conventional NST, style loss is computed across multiple convolutional layers and combined using a single global coefficient. Although visually compelling, this formulation assumes equal stylistic contribution across layers. However, convolutional neural networks are inherently hierarchical [22], with early layers capturing low-level textures and deeper layers encoding larger-scale abstractions.

Because receptive field sizes increase with depth, uniform aggregation of style losses restricts fine-grained control over spatial-scale stylization. We introduce a layer-wise weighting mechanism to explicitly exploit this hierarchy and enable interpretable multi-scale stylization control.

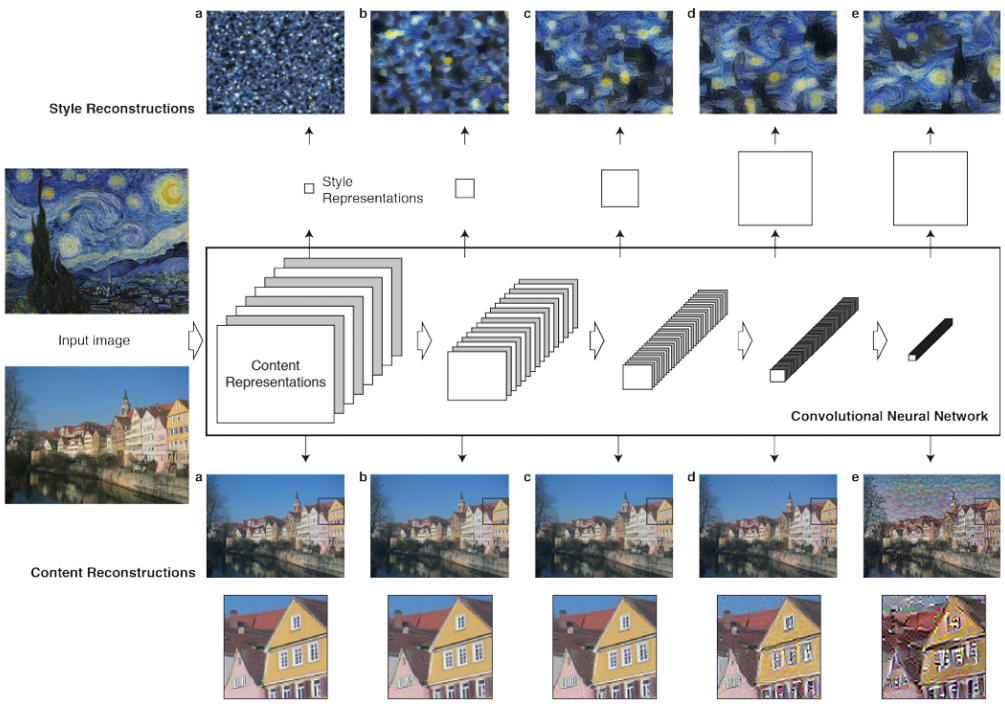

Figure 1: Hierarchical stylization behavior across VGG-19 layers [8]. Shallow layers capture fine textures; deeper layers capture structural abstraction.

Brief Summary of Existing Work

Classical Texture Modeling

Before deep learning, artistic stylization was formulated as texture synthesis. Multi-scale statistical models [10, 11] characterized texture via spatial frequency decompositions. Patch-based approaches [17, 15] transferred local structure under similarity constraints, with extensions improving efficiency and directionality [16, 18]. A comprehensive taxonomy is provided in [14]. Bilinear [12] and manifold-based models [13] attempted to separate style and content representations.

Hierarchical Deep Representations

Large-scale image classification advances [1, 22, 23] demonstrated that deep CNNs learn hierarchical features. Empirical and neuroscientific evidence suggests correspondence between CNN depth and representational abstraction [4, 5, 3, 7]. Reconstruction studies [9] show that intermediate features preserve multi-scale spatial information.

Neural Style Transfer Formulation

NST [8] formulates stylization as feature-space optimization:

$$\mathcal{L}(x) = \mathcal{L}_{\text{content}} + \mathcal{L}_{\text{style}}.$$

Content loss:

$$\mathcal{L}_{\text{content}} = \|\phi_{l_c}(x) - \phi_{l_c}(c)\|_2^2.$$

Style loss:

$$\mathcal{L}_{\text{style}} = \sum_{l \in \mathcal{S}} \|G_l(x) - G_l(s)\|_F^2, \quad G_l(x) = \phi_l(x)\phi_l(x)^\top.$$

Research Gap

Standard NST employs:

$$\mathcal{L}_{\text{style}} = \alpha \sum_{l \in \mathcal{S}} E_l.$$

This scalar aggregation neglects scale-specific control. We propose instead:

$$\mathcal{L}_{\text{style}} = \sum_{l \in \mathcal{S}} \alpha_l E_l,$$

enabling structured manipulation of stylistic granularity across convolutional depths.

Methodology

We build upon optimization-based NST using VGG-19 features. The content loss is defined as:

$$L_{\text{content}} = \frac{1}{2}\|F_l(x) - F_l(c)\|^2.$$

Our layer-wise weighting modifies style loss:

$$L_{\text{style}} = \sum_{l \in S} \alpha_l E_l.$$

Three stylization regimes are defined:

- **Fine-Grain:** Emphasis on shallow layers.
- **Balanced:** Uniform weighting.
- **Abstract:** Emphasis on deeper layers.

We evaluate perceptual similarity using LPIPS and SSIM. For video stylization, we introduce temporal consistency:

$$L_{\text{temporal}} = \|x_t - \text{warp}(x_{t-1})\|^2.$$

Objectives

Our primary objective is to introduce explicit multi-scale control into neural style transfer by replacing uniform style aggregation with layer-wise weighting. We aim to demonstrate that redistributing stylistic emphasis across convolutional layers leads to predictable and interpretable changes in stylization behavior.

A secondary objective is to evaluate whether this additional control improves perceptual quality and stability in both static image and video stylization scenarios.

Implementation Plan

We will first reproduce the baseline optimization-based neural style transfer implementation to establish a reference model. After verifying correctness and ensuring faithful reproduction of standard results, we will integrate the proposed layer-wise weighting modification into the style loss formulation.

We will experiment with multiple weight configurations and analyze outputs across defined stylization profiles. Both qualitative inspection and quantitative perceptual metrics will be used to compare results and evaluate differences in structural abstraction and texture fidelity.

For the video extension, we will incorporate temporal consistency loss and evaluate the impact of different layer-weight configurations on flicker reduction and motion coherence.

Timeline for CMPUT 414

- Week 1–2: Review literature and reproduce baseline neural style transfer implementation.
- Week 3: Implement layer-wise weighting mechanism and define stylization profiles.
- Week 4: Conduct controlled experiments and analyze perceptual differences.
- Week 5: Extend framework to video stylization and evaluate temporal stability.
- Week 6: Finalize report and prepare presentation and demo.

Conclusion

This project introduces Layer-Wise Style Weighting as a structured extension to neural artistic stylization. By leveraging the hierarchical nature of convolutional neural networks, we provide explicit control over stylistic granularity across spatial scales. The proposed framework enhances interpretability, improves artistic flexibility, and establishes a foundation for future exploration in controllable and temporally stable neural style transfer systems.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. NeurIPS, 2012.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. CVPR, 2014.

[3] U. Güçlü and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience, 2015.

[4] D. L. K. Yamins et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS, 2014.

[5] C. F. Cadieu et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol, 2014.

[6] M. Kümmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. ICLR Workshop, 2015.

[7] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol, 2014.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. A Neural Algorithm of Artistic Style. arXiv:1508.06576, 2015.

[9] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. arXiv:1412.0035, 2014.

[10] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. SIGGRAPH, 1995.

[11] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. IJCV, 2000.

[12] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. Neural Computation, 2000.

[13] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. CVPR, 2004.

[14] J. E. Kyprianidis et al. A taxonomy of artistic stylization techniques for images and video. IEEE TVCG, 2013.

[15] A. Hertzmann et al. Image analogies. SIGGRAPH, 2001.

[16] M. Ashikhmin. Fast texture transfer. IEEE CGA, 2003.

[17] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. SIGGRAPH, 2001.

[18] H. Lee et al. Directional texture transfer. NPAR, 2010.

[19] X. Xie, F. Tian, and H. S. Seah. Feature guided texture synthesis for artistic style transfer. DIMEA, 2007.

[20] S. Karayev et al. Recognizing image style. arXiv:1311.3715, 2013.

[21] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. JOSA A, 1985.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[23] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2014.

[24] Y. Jia et al. Caffe: Convolutional architecture for fast feature embedding. ACM Multimedia, 2014.