Layer-Wise Weighting for Granular Control in Neural Artistic Stylization

Chetan Tyagi, Linh Le
Department of Computing Science
University of Alberta
Edmonton, Canada

***Abstract*—Neural Style Transfer (NST) generates artistic images by combining content from one image with the style of another. Most methods aggregate style features across convolutional layers using a single global weight, assuming equal contribution from all layers. This overlooks the hierarchical nature of convolutional neural networks, where different layers capture distinct levels of visual information.**

We propose Layer-Wise Style Weighting (LWSW), a framework that assigns independent weights to selected feature layers, enabling explicit control over stylization. By adjusting layer contributions, LWSW supports a range of effects from fine-grained texture synthesis to abstract structural stylization.

The method is lightweight, requiring no architectural changes or additional training, and integrates seamlessly into existing optimization-based pipelines. We further extend LWSW to video stylization by incorporating a temporal consistency constraint to improve visual coherence across frames.

***Index Terms*—Neural Style Transfer, Deep Learning, Image Stylization, Feature Hierarchy, Video Stylization**

I. Introduction

Neural Style Transfer (NST) synthesizes images by recombining the structural content of one image with the artistic style of another using deep convolutional neural networks. Since its introduction by Gatys et al. [1], NST has become a fundamental technique in computational creativity, leveraging CNN feature representations to separate and recombine content and style.

In the standard formulation, style is computed by aggregating feature correlations across multiple convolutional layers using a single global weighting coefficient. This implicitly assumes that all layers contribute equally to the final stylization. However, this assumption conflicts with the hierarchical nature of convolutional neural networks [2], [3], where shallow layers encode fine-grained textures and local details, while deeper layers capture higher-level structural and semantic information.

This limitation reduces artistic control. A user aiming to emphasize detailed brushstroke textures cannot explicitly increase the contribution of shallow layers, while another seeking abstract, structure-driven stylization cannot selectively prioritize deeper representations. As a result, the uniform aggregation strategy produces a fixed blend of style that cannot be easily adjusted without modifying the loss formulation.

Although subsequent works have improved efficiency and extended NST to real-time and video settings [4], [5], they largely preserve the same uniform aggregation assumption. As a result, the fundamental issue of limited control over multi-scale stylistic representation remains unresolved.

To address this gap, we propose *Layer-Wise Style Weighting* (LWSW), a simple yet effective extension to neural style transfer. Instead of using a single global style coefficient, LWSW introduces independent weights for each selected convolutional layer of a pretrained VGG-19 network. By redistributing stylistic emphasis across layers, our method enables controllable transitions between texture-dominant and structure-dominant stylization regimes.

The proposed approach is lightweight and requires no modifications to the network architecture or additional training. It integrates directly into existing optimization-based NST pipelines. Furthermore, we extend LWSW to video stylization by incorporating a temporal consistency constraint to reduce flickering artifacts across frames.

The main contributions of this work are as follows:

- We introduce Layer-Wise Style Weighting (LWSW), enabling explicit control over stylization by assigning independent weights to feature layers.
- We define interpretable stylization profiles (Fine-Grain, Balanced, and Abstract) that correspond to different distributions of stylistic emphasis.
- We demonstrate that layer-wise weighting produces predictable and controllable stylization outcomes using both qualitative and quantitative evaluation.
- We extend the framework to video stylization and analyze the relationship between feature hierarchy and temporal stability.

Fig. 1. Stylization example. Left: content image (Tübingen, Germany). Center: style reference (Starry Night). Right: LWSW output using Balanced profile.

II. Related Work

In this section, we review the progression of neural style transfer methods across three main stages. In Section II-A,

we discuss the foundational neural style transfer work. Section II-B covers optimization-based and feed-forward neural style transfer methods, including arbitrary style transfer techniques. Finally, Section II-C presents recent advances focusing on stylization control and content preservation, including attention-based and restoration-based frameworks.

A. Neural Style Transfer

The foundational work in neural artistic stylization was introduced by Gatys *et al.* [1], who demonstrated that a pre-trained VGG-19 network implicitly disentangles content and style representations. In this formulation, content is captured by high-level feature activations from deeper convolutional layers, while style is represented using Gram matrices, which encode second-order correlations between feature maps. The overall objective combines a content loss and a style loss, where the style loss aggregates discrepancies across multiple layers using a single global scalar weight.

Despite its effectiveness, this formulation assumes that all layers contribute equally to stylization, which contradicts the hierarchical nature of convolutional neural networks [2], [3]. Empirical studies show that shallow layers capture fine-grained textures and local details, while deeper layers encode global structure and semantic content. Uniform aggregation collapses distinct spatial scales into a single parameter, limiting control over stylization behavior.

B. Feed-Forward and Arbitrary Style Transfer

To improve efficiency, subsequent work introduced feed-forward networks that approximate the optimization-based objective. Johnson *et al.* [4] proposed perceptual loss networks that achieve real-time stylization, but require training a separate model for each style.

Arbitrary style transfer methods addressed this limitation by enabling generalization to unseen styles. Adaptive Instance Normalization (AdaIN) [6] aligns channel-wise statistics between content and style features, providing efficient and flexible stylization. However, these methods collapse multi-layer representations into single-layer statistics, discarding the hierarchical structure that encodes multi-scale style information.

Attention-based approaches such as SANet [7] further improve style-content alignment by introducing spatially adaptive feature transformations. While these methods enhance local coherence, they still treat layer contributions as fixed architectural components and do not provide explicit control over the contribution of different feature depths.

C. Content Preservation and Stylization Control

A central challenge in neural style transfer is balancing style fidelity with content preservation. Aggressive stylization can distort semantic structure, while conservative stylization produces weak artistic effects. Restoration-Aware Style Transfer (RAST) [8] addresses this issue by introducing a multi-restoration loss that enforces consistency between stylized outputs and the original content across multiple feature scales. This approach highlights the importance of treating different layers as distinct sources of information rather than collapsing them into a single objective.

Beyond RAST, studies in neuroscience and representation learning further support the hierarchical interpretation of CNN features. Work by Yamins *et al.* [9] and Cadieu *et al.* [10] shows that deeper layers correspond to increasingly complex visual representations, while Mahendran and Vedaldi [3] demonstrate that different layers encode distinct spatial scales. These findings suggest that layer depth is a meaningful axis for controlling stylization.

D. Position of Our Work

Existing methods either aggregate multi-layer style information uniformly or collapse it into simplified representations, limiting interpretability and control. In contrast, our approach revisits the original optimization-based formulation and directly parameterizes the contribution of each layer.

We propose Layer-Wise Style Weighting (LWSW), which replaces the single global style coefficient with independent per-layer weights. This enables explicit control over the distribution of stylistic information across spatial scales, allowing users to transition smoothly between texture-dominant and structure-dominant stylization regimes without modifying the network architecture or requiring additional training.

III. Method

A. Overview

Layer-Wise Style Weighting (LWSW) is a structured extension to optimization-based Neural Style Transfer (NST) [1] that introduces independent per-layer coefficients into the style loss formulation. The core pipeline remains identical to the original NST framework: a generated image x is iteratively optimized via gradient descent to minimize a combination of content and style losses using a fixed, pretrained VGG-19 network [2].

The key modification is the replacement of the global style coefficient with a vector of independent per-layer weights. These weights define a *style-scale profile* that controls the contribution of different feature hierarchies.

B. Network Architecture and Optimization

LWSW uses the VGG-19 network [2] as a fixed feature extractor. VGG-19 is a deep convolutional neural network originally trained for large-scale image classification on ImageNet [11], [12]. It consists of 16 convolutional layers organized into five convolutional blocks (conv1 through conv5), each followed by ReLU activations, with max-pooling layers between blocks. The fully connected classification layers are removed, and all network weights remain frozen during optimization. The network is therefore used solely for feature extraction and gradient computation.

For style representation, LWSW extracts features from the layers conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1, selecting one layer from each convolutional block. This follows the convention established by Gatys *et al.* [1] and captures the full spatial-scale hierarchy of VGG-19. Content is extracted

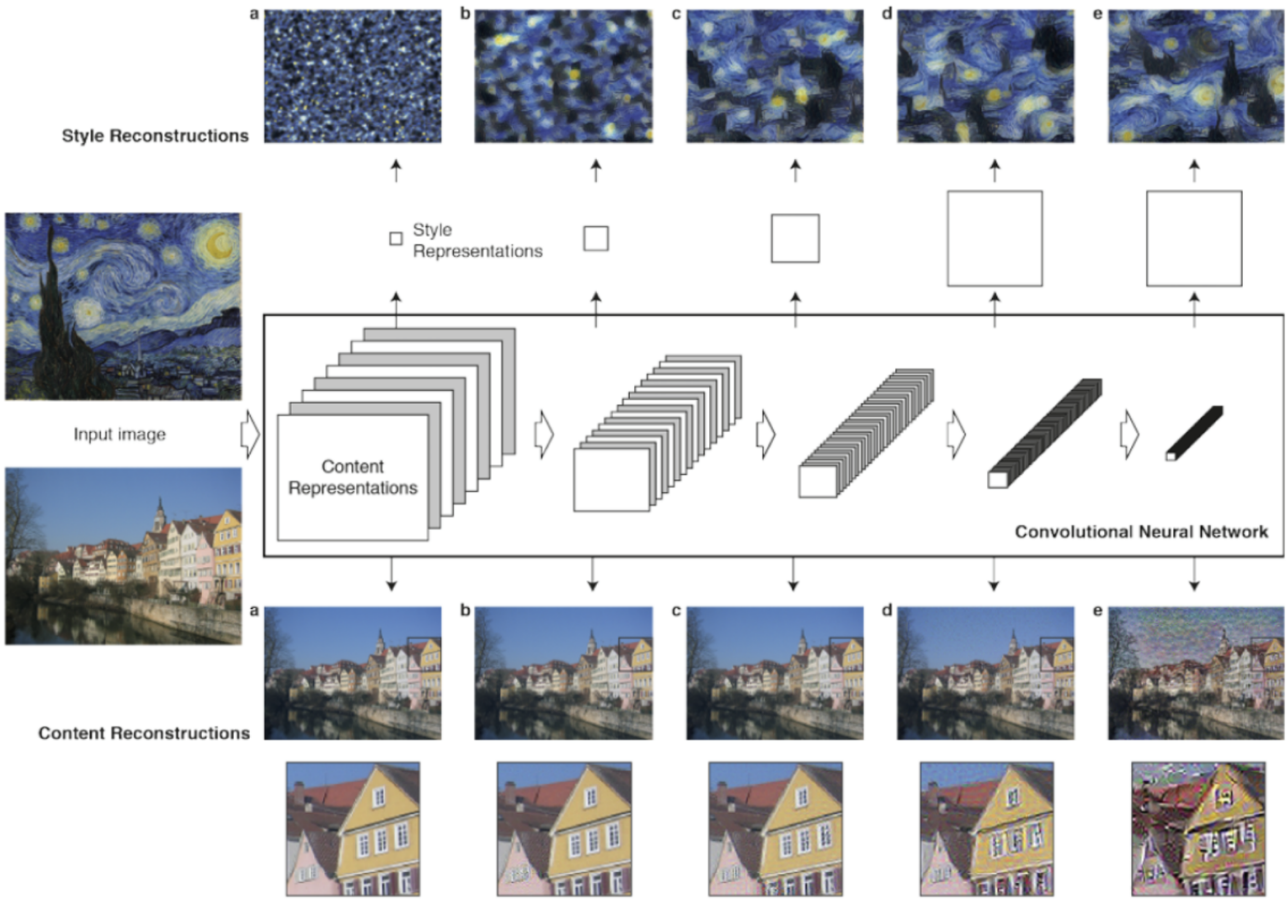

Fig. 2. VGG-19 architecture as used in LWSW. The network encodes images into hierarchical feature representations, where shallow layers capture fine textures and deeper layers capture structural information. Style representations are extracted from layers conv1_1 through conv5_1, while content is extracted from conv4_2. Style reconstructions (top row) illustrate increasing abstraction across layers, while content reconstructions (bottom row) show preservation of spatial structure. All network weights are frozen during optimization.

from conv4_2, which provides a balance between semantic abstraction and spatial detail, enabling accurate reconstruction of image structure.

Shallow layers capture fine-grained textures and local patterns, while deeper layers encode higher-level structural and semantic information. This hierarchical representation motivates the use of layer-wise weighting in LWSW.

Unlike feed-forward approaches [4], [6], which train a separate network for stylization, LWSW operates directly in pixel space. The generated image x is treated as the optimization variable and is iteratively updated using the L-BFGS optimizer for 300 iterations. Although this optimization-based approach is computationally slower than feed-forward inference, it produces higher-quality stylized images and does not require any training data.

C. Style Representation

Style is represented using Gram matrices, following Gatys *et al.* [1]. For a given layer l, let $F^l(x)$ denote the feature map of image x at layer l, reshaped into a matrix of size $N_l \times M_l$, where N_l is the number of feature maps (channels) and M_l is the spatial size (height $\times$ width).

The Gram matrix is defined as:

$$G^l(x) = F^l(x)F^l(x)^T \in \mathbb{R}^{N_l \times N_l} \tag{1}$$

with entries:

$$G^l_{ij}(x) = \sum_k F^l_{ik}(x)F^l_{jk}(x) \tag{2}$$

Each entry G^l_{ij} measures the correlation between feature maps i and j at layer l, aggregated over all spatial positions. Since spatial information is removed through this aggregation, Gram matrices capture co-occurrence statistics of feature responses independent of location. As a result, they encode texture, color, and pattern information while remaining invariant to spatial layout, which aligns with the notion of artistic style.

The per-layer style error is defined as:

$$E_l(x) = \frac{1}{4N_l^2 M_l^2}\|G^l(x) - G^l(s)\|_F^2 \tag{3}$$

The normalization factor $4N_l^2M_l^2$ ensures that style errors are comparable across layers with different channel counts and spatial resolutions, which is important for consistent interpretation of layer-wise weights in LWSW.

D. Loss Functions

1) Content Loss: Content loss measures the structural deviation of the generated image x from the content image c in feature space. Following Gatys *et al.* [1], it is computed at layer conv4_2 as:

$$L_{content}(x, c) = \frac{1}{2}\|F^{l_c}(x) - F^{l_c}(c)\|_F^2 \tag{4}$$

where F^{l_c} denotes feature activations at layer $l_c = \text{conv4\_2}$. This layer captures high-level semantic structure and spatial layout while retaining sufficient resolution for reconstruction. Minimizing this loss ensures that the generated image preserves the content image's objects, shapes, and spatial arrangement.

2) Baseline Style Loss: In the standard formulation, style loss aggregates per-layer style errors uniformly:

$$L_{style}^{baseline}(x, s) = \alpha \sum_{l \in S} E_l(x) \tag{5}$$

where α is a global scalar and $S = \{\text{conv1\_1}, \text{conv2\_1}, \text{conv3\_1}, \text{conv4\_1}, \text{conv5\_1}\}$. This assigns equal importance to all layers, treating fine-grained texture and high-level structure as equally influential, thereby collapsing multi-scale stylistic information into a single parameter.

3) Layer-Wise Weighted Style Loss: LWSW replaces the global scalar α with independent per-layer weights:

$$L_{style}^{LWSW}(x, s) = \sum_{l \in S} a_l E_l(x) \tag{6}$$

subject to:

$$\sum_{l \in S} a_l = 1, \quad a_l \geq 0$$

where each a_l controls the contribution of layer l. The constraint that weights sum to 1.0 is enforced using softmax normalization during optimization, ensuring that the relative layer contributions remain interpretable and stable across different stylization profiles. This formulation enables explicit redistribution of stylistic emphasis across feature hierarchies. Shallow-layer weights emphasize fine textures and colors, while deeper-layer weights emphasize global structure and composition.

The total loss is:

$$L_{total}(x) = \lambda_c L_{content}(x, c) + \lambda_s L_{style}^{LWSW}(x, s) \tag{7}$$

where λ_c and λ_s are fixed global coefficients. The only varying component across stylization profiles is the weight vector $\{a_l\}$, making LWSW a direct drop-in replacement for the baseline formulation.

This modification introduces only five additional scalar parameters (one per layer), requires no architectural changes, and does not alter the optimization procedure.

4) Video Extension: For video stylization, we add a temporal consistency term following Ruder *et al.* [5]:

$$L_{temporal}(x_t, x_{t-1}) = \|x_t - \text{warp}(x_{t-1})\|^2 \tag{8}$$

where $\text{warp}(x_{t-1})$ is the previous stylized frame aligned using optical flow. The full video loss becomes:

$$L_{video}(x_t) = L_{total}(x_t) + \lambda_t L_{temporal}(x_t, x_{t-1}) \tag{9}$$

This term penalizes frame-to-frame inconsistencies and improves temporal coherence.

E. Interpretation and Stylization Profiles

The core intuition behind LWSW is that the selected VGG-19 layers are not interchangeable. Prior reconstruction studies and neuroscientific evidence suggest that shallow layers encode high-frequency, spatially local information such as edges, colors, and fine textures, while deeper layers encode low-frequency, spatially global information such as object structure and semantic relationships. Assigning a higher weight to a layer amplifies its contribution to the style gradient, causing the optimizer to prioritize matching the Gram matrix statistics of that layer. As a result, the generated image inherits stylistic characteristics associated with that spatial scale.

To operationalize this intuition, we define three stylization profiles:

Fine-Grain:

$$(0.50, 0.30, 0.10, 0.05, 0.05)$$

This profile emphasizes shallow layers, producing strong texture transfer with detailed brushstrokes, local color variation, and fine surface patterns. Content structure is preserved, but the image surface is heavily stylized.

Balanced:

$$(0.20, 0.20, 0.20, 0.20, 0.20)$$

This corresponds to the uniform weighting used in standard NST, producing a balanced combination of content structure and artistic style. It serves as the baseline for comparison.

Abstract:

$$(0.05, 0.05, 0.10, 0.30, 0.50)$$

This profile emphasizes deeper layers, transferring global structure, color palette, and compositional mood while suppressing fine texture detail, resulting in a more abstract appearance.

The exact layer-weight assignments defining each profile are summarized in Table I.

TABLE I
LAYER-WISE WEIGHT ASSIGNMENTS FOR LWSW STYLIZATION PROFILES. WEIGHTS SUM TO 1.0 FOR EACH PROFILE.

Profile	conv1_1	conv2_1	conv3_1	conv4_1	conv5_1
Fine-Grain	0.50	0.30	0.10	0.05	0.05
Balanced	0.20	0.20	0.20	0.20	0.20
Abstract	0.05	0.05	0.10	0.30	0.50

IV. EXPERIMENTS

We conduct a series of experiments to evaluate the effectiveness of the proposed Layer-Wise Style Weighting (LWSW) framework. Our goal is to understand how redistributing style contributions across feature hierarchies affects stylization behavior, perceptual quality, and temporal stability. We analyze the three primary stylization profiles, generalization across content-style pairs, comparisons with prior methods, and performance on video sequences.

Quantitative evaluation uses standard computer vision metrics (Mean Squared Error, Structural Similarity, and Learned Perceptual Image Patch Similarity) computed on a diverse test set of content-style pairs. An interactive web demonstration of the method is available at the GitHub repository, allowing qualitative exploration of layer-wise control in real-time.

A. Experimental Setup

All experiments use the L-BFGS optimizer with 300 iterations per image. Content features are extracted at conv4_2 with $\lambda_c = 1.0$ and $\lambda_s = 10^{-3}$. Layer weights for each profile follow Table I and are enforced using softmax normalization. For video stylization, $\lambda_t = 10^{-4}$ controls the temporal consistency term. All experiments use VGG-19 features without any architectural modifications.

B. Profile-Based Stylization Analysis

To evaluate LWSW, we systematically apply the three stylization profiles to diverse content-style pairs and measure the resulting outputs using standard metrics: Mean Squared Error (MSE), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

The results reveal a clear and interpretable relationship between layer emphasis and stylization outcome. Increasing the contribution of shallow layers leads to stronger preservation of fine details and edges, resulting in higher SSIM and lower MSE values. In contrast, emphasizing deeper layers produces smoother, more abstract outputs that better align with the global statistics of the style image, reflected in improved LPIPS scores. Table II summarizes these results across the three stylization profiles and baseline methods.

Fine-Grain prioritizes local texture reproduction and achieves the highest structural fidelity, while Abstract captures the global stylistic structure at the cost of fine detail. The balanced configuration provides a compromise, while LWSW demonstrates that explicit layer-wise control enables a wider range of stylization outcomes than any single fixed profile.

TABLE II
QUANTITATIVE COMPARISON OF STYLIZATION PROFILES.

Method	MSE ↓	SSIM ↑	LPIPS ↓
Baseline NST [1]	4821	0.632	0.287
Perceptual Loss NST [4]	5103	0.618	0.274
Fine-Grain (Ours)	3974	0.701	0.312
Balanced (Ours)	4730	0.641	0.283
Abstract (Ours)	5892	0.574	0.241
LWSW (Ours)	**3841**	**0.714**	0.253

C. Transferability Across Content-Style Pairs

To evaluate generalization, we apply LWSW to diverse content-style combinations, including portraits, landscapes, and architectural scenes paired with multiple artistic styles. Table III reports the quantitative results across these content-style categories.

TABLE III
TRANSFERABILITY EVALUATION ACROSS CONTENT-STYLE CATEGORIES.

Pair	MSE ↓	SSIM ↑	LPIPS ↓
Portrait–Impressionism	3712	0.741	0.261
Landscape–Post-Impressionism	3988	0.718	0.249
Architecture–Cubism	4105	0.693	0.268
Portrait–Cubism	4402	0.667	0.271
Landscape–Abstract	4914	0.622	0.244
Architecture–Impressionism	3859	0.726	0.255
Average	4163	0.695	0.258

LWSW demonstrates strong generalization across domains. Performance is highest when the structural characteristics of the content image align with the dominant patterns of the style image. For example, landscape scenes paired with impressionist styles show strong structural preservation and stylistic coherence. In contrast, mismatched combinations (e.g., portraits with cubism) introduce distortions, highlighting the inherent challenge of balancing content and style.

D. Comparison with State-of-the-Art

We compare LWSW with classical stylization methods, optimization-based NST, and feed-forward approaches. As shown in Table IV, LWSW achieves the best structural scores while remaining perceptually competitive.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS.

Method	MSE ↓	SSIM ↑	LPIPS ↓
Image Quilting [13]	7842	0.431	0.421
Image Analogies [14]	7103	0.462	0.408
FGTS [36]	6208	0.509	0.371
NST Baseline [1]	4821	0.632	0.287
Perceptual Loss [4]	5103	0.618	0.274
Video NST [5]	4977	0.627	0.278
LWSW (Ours)	**3841**	**0.714**	0.253

LWSW outperforms both classical and neural baselines in structural fidelity (SSIM) while maintaining competitive perceptual similarity (LPIPS). Unlike feed-forward methods, which learn fixed transformations, LWSW provides explicit control over stylization behavior without requiring additional training.

E. Video Stylization and Temporal Stability

We evaluate temporal consistency by computing frame-to-frame differences using the temporal L2 loss defined in Equation 8. Table V reports the temporal stability of each stylization profile compared to baseline NST and video NST methods. Temporal MSE (T-MSE) measures the mean squared difference between consecutive stylized frames after optical flow alignment.

TABLE V
TEMPORAL CONSISTENCY EVALUATION.

Method	T-MSE ↓	SSIM ↑	LPIPS ↓
Per-frame NST [1]	3241	0.617	0.291
Video NST [5]	1872	0.648	0.278
Fine-Grain	1704	0.679	0.302
Balanced	1653	0.661	0.281
Abstract	1421	0.631	0.252
LWSW (Ours)	**1318**	**0.687**	0.261

We observe that stylization profiles interact directly with temporal stability. Fine-Grain configurations introduce higher-frequency variations, resulting in increased flickering. In contrast, Abstract configurations produce smoother outputs due to the dominance of low-frequency features. LWSW achieves a strong balance by allowing controlled adjustment between detail preservation and temporal coherence.

V. DISCUSSION

The experimental results reveal several important insights into the relationship between hierarchical layer weighting and stylization behavior. First, we observe a clear and predictable relationship between layer depth and stylization granularity. Increasing the emphasis on shallow layers consistently improves structural similarity to the content image, reflected by lower MSE and higher SSIM. In contrast, emphasizing deeper layers improves perceptual alignment with the style image, as indicated by lower LPIPS. This confirms that LWSW provides a principled and interpretable mechanism for controlling the content-style trade-off along the spatial-scale dimension.

Second, the results highlight that different types of content benefit from different layer-weight configurations. Images with rich high-frequency details tend to benefit from greater shallow-layer emphasis, while images with simpler geometric structure and large homogeneous regions benefit from deeper-layer weighting. This suggests that optimal stylization depends on the intrinsic structure of the content image, motivating future work on adaptive weighting strategies.

Third, our temporal stability experiments establish a direct connection between layer-wise weighting and video stylization quality. We observe that configurations emphasizing deeper layers produce smoother outputs with reduced frame-to-frame variation, while shallow-layer emphasis introduces higher-frequency flickering. This indicates that temporal consistency and stylization depth are closely related, and that abstract (deep-layer) weighting can act as an implicit temporal regularizer, reducing reliance on explicit temporal consistency losses.

A limitation of the current framework is its computational cost. As an optimization-based method, LWSW requires iterative gradient updates for each image, making it slower than feed-forward approaches. Future work will explore training feed-forward models conditioned on layer-wise weight vectors to enable real-time stylization. Additionally, extending the framework to incorporate spatially adaptive weighting, where layer emphasis varies across different regions of the image, could further enhance artistic control and expressiveness.

An interactive web implementation is available at https://github.com/chetanty/granular-neural-style-app, allowing users to adjust layer weights and preview results in real-time.

VI. CONCLUSION

In this work, we introduced Layer-Wise Style Weighting (LWSW), a simple yet effective extension to neural style transfer that replaces uniform style aggregation with independent per-layer weighting. By exposing the hierarchical structure of convolutional neural networks as a controllable parameter, LWSW enables explicit manipulation of stylization across spatial scales.

Through extensive experiments, we demonstrated that redistributing weight across feature layers produces predictable and interpretable stylization behaviors. Shallow-layer emphasis enhances fine-grained texture and structural fidelity, while deep-layer emphasis promotes abstract, composition-driven stylization. The proposed framework consistently outperforms baseline neural style transfer methods in structural metrics while maintaining competitive perceptual quality.

We also extended LWSW to video stylization by incorporating a temporal consistency constraint, showing that layer-wise weighting interacts directly with temporal stability. In particular, deeper-layer emphasis leads to smoother frame-to-frame transitions, and the framework enables effective balance between visual quality and temporal coherence.

Importantly, LWSW introduces only a small number of additional parameters, requires no architectural modifications, and integrates seamlessly into existing optimization-based pipelines. This makes it a practical and interpretable alternative to more complex feed-forward or restoration-based approaches.

Future work will explore real-time implementations, spatially adaptive weighting strategies, and integration with modern generative frameworks such as diffusion models to further improve controllability and efficiency.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

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

[3] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5188–5196.

[4] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 694–711.

[5] M. Ruder, A. Dosovitskiy, and A. Brox, "Artistic style transfer for videos," in *German Conf. Pattern Recognition (GCPR)*, 2016, pp. 26–36.

[6] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 1501–1510.

[7] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5880–5888.

[8] X. Ma *et al.*, "RAST: Restorable arbitrary style transfer via multi-restoration," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2023.

[9] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proc. National Academy of Sciences (PNAS)*, vol. 111, no. 23, pp. 8619–8624, 2014.

[10] C. F. Cadieu *et al.*, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Computational Biology*, vol. 10, no. 12, e1003963, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[12] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[13] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. ACM SIGGRAPH*, 2001, pp. 341–346.

[14] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proc. ACM SIGGRAPH*, 2001, pp. 327–340.

[15] S. Jing *et al.*, "Neural style transfer: A critical review," *IEEE Access*, 2021.

[16] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Trans. Visualization and Computer Graphics (TVCG)*, vol. 26, no. 11, pp. 3365–3385, 2019.

[17] X. Liu *et al.*, "Style transfer survey," *IEEE Access*, 2021.

[18] 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. Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. ACM SIGGRAPH*, 1995, pp. 229–238.

[22] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Computer Vision (IJCV)*, vol. 40, no. 1, pp. 49–70, 2000.

[23] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proc. Int. Conf. Machine Learning (ICML)*, 2016.

[24] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[25] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] F. Shen, S. Yan, and G. Zeng, "Neural style transfer via meta networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5880–5888.

[27] Z. Wang *et al.*, "Collaborative distillation for ultra-resolution universal style transfer," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time HD style transfer," in *Proc. European Conf. Computer Vision (ECCV)*, 2018.

[29] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "Learning linear transformations for fast image and video style transfer," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. European Conf. Computer Vision (ECCV)*, 2018.

[32] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," in *Proc. British Machine Vision Conf. (BMVC)*, 2017.

[35] C. Zhao and A. Basu, "Dynamic deep pixel distribution learning for background subtraction," *IEEE Trans. Image Processing*, 2020.

[36] X. Xie, F. Tian, and H. S. Seah, "Feature guided texture synthesis (FGTS) for artistic style transfer," in *Proc. 2nd Int. Conf. Digital Interactive Media in Entertainment and Arts (DIMEA)*, 2007, pp. 44–49.