Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 03s, 2025

Neural Style Transfer

[1]Priyanshu Bilwane, [2]Ranu Nanhe, [3]Rohit Wachnekar, [4]Sahil Rajankar, [5]Pinky Gangwani
[1,2,3,4,5] Department of Artificial Intelligence, St. Vincent Pallotti College of Engineering and Technology , Nagpur, India
Email: [1]priyanshubilwane.22@stvincentngp.edu.in, [2]ranunanhe.22@stvincentngp.edu.in, [3]rohitwachnekar.22@stvincentngp.edu.in, [4]sahilranjankar.22@stvincentngp.edu.in, [5]pgangwani@stvincentngp.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Neural Style Transfer (NST), Adaptive Instance Normalization (AdaIN), two-stage hybrid pipeline, optimization-based models, feed-forward networks, VGG19-based refinement, Gram matrix texture matching, channel-wise mean–variance stability, Sobel edge preservation, structural integrity, generative visual tasks.

Abstract

Balancing fidelity and efficiency remains a core challenge in Neural Style Transfer (NST). Traditional optimization-based models generate visually rich stylizations but are computationally slow, whereas feed-forward networks achieve real-time inference at the cost of structure and detail. This paper presents a two-stage hybrid pipeline that bridges this gap. The first stage employs Adaptive Instance Normalization (AdaIN) for rapid global alignment of content and style statistics. The second stage introduces a lightweight VGG19-based refinement that optimizes a composite loss integrating Gram matrix texture matching, channel-wise mean–variance stability, and Sobel edge preservation. This combination enhances both structural integrity and color coherence in the final stylized outputs. Experiments on varied content–style pairs demonstrate significant visual improvements while maintaining computational efficiency. Furthermore, our findings expose the limitations of traditional quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), which fail to capture perceptual quality, highlighting the necessity of perceptual measures like LPIPS for evaluating generative visual tasks.

Introduction

The core challenge of artistic style transfer is to render the semantic content of an arbitrary photograph in the visual style of another image, such as an artistic painting. This task was first successfully demonstrated in the seminal work of Gatys, Ecker, and Bethge, which introduced "A Neural Algorithm of Artistic Style". This work was built on the foundational discovery that the hierarchical feature representations within a Convolutional Neural Network (CNN) allow for the "separability" of image content and style. Within a network like VGGNet, content is captured in the feature responses of the deeper layers, which represent high-level object information. Style, conversely, is captured by the statistical correlations between feature responses across multiple layers, famously encapsulated in the Gram matrix. This discovery immediately bifurcated the field into two primary approaches. The original "image-iteration" method, while producing striking, high-fidelity results, is fundamentally a slow optimization process. Synthesizing a single image requires iteratively updating a noise image via gradient descent to minimize both content and style losses, a process that can take minutes to an hour, rendering it unsuitable for real-time applications. This computational barrier spurred the development of "model-iteration," or fast feed-forward, methods. These models, such as those based on AdaIN, are trained a priori to perform stylization in a single pass, achieving real-time performance. However, this speed comes at a significant cost: a pronounced loss of structural fidelity, content coherence, and the introduction of "overpainted" or "washed out" artifacts. This report confronts this fidelity-versus-speed trade-off directly. The central hypothesis is that these two approaches are not mutually exclusive but can be synergistically combined. A novel, two-stage hybrid pipeline is introduced, leveraging a fast, feed-forward (AdaIN) pass not as the final output, but as an intelligent initialization for a subsequent, lightweight iterative refinement. This refinement stage is guided by a novel hybrid loss function designed to restore structural integrity and enhance texture detail.

The primary contributions of this work are:

1. A two-stage hybrid pipeline that achieves high-fidelity stylization in a fraction of the time required by traditional optimization methods.
2. A novel hybrid style loss function, integrating Gram matrices, channel-wise Mean-Variance statistics, and a Sobel-based edge loss to preserve both texture and content structure.
3. A "Pyramid Multi-Resolution Stylization" strategy to ensure stable convergence and sharpness.
4. A comprehensive evaluation demonstrating qualitatively superior results over standard baselines.

I. Related Work

A. Optimization-Based Methods

The foundational optimization-based method involves minimizing a total loss function L_{total}, which is a linear combination of content and style losses [1],[2]:

$$L_{total} = \alpha L_{content} + \beta L_{style}.$$

Content Loss($L_{content}$): This loss is defined as the Mean Squared Error (MSE) between the feature maps of the content image (P^l) and the generated image (F^l) at a single deep layer l, ensuring semantic content is preserved.

$$L_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2}\sum_{i,j} (F_{ij}^l - P_{ij}^l)^2 \quad (1)$$

Style Loss (L_{edge}): This loss is defined as the weighted sum of the MSE between the Gram matrices (G^l) of the style image and the generated image, computed across multiple layers l. The Gram matrix,

$$G_{ij}^l = \Sigma_k F_{ik}^l F_{jk}^l \quad (2)$$

, captures texture by calculating the correlations between different filter responses in a layer. This captures a stationary, multi-scale representation of the style.

B. Real-Time Feed-Forward Methods

To overcome the latency of optimization, feed-forward networks were trained using "perceptual losses" [7]. The most prominent example, and the one used in this pipeline's first stage, is AdaIN. AdaIN operates directly in feature space. It first normalizes the content feature map (F_c) by subtracting its channel-wise mean (μ) and dividing by its variance(σ). It then scales the result by the style feature map's variance ($\sigma(F_s)$) and adds the style's mean ($\mu(F_s)$). 1 The mechanism is described by:

$$AdaI\,N(F_c, F_s) = \sigma(F_s)_{\left(\frac{F_c - \mu(F_c)}{\sigma(F_c)}\right)} + \mu(F_s) \quad (3)$$

This single operation aligns the second-order statistics of the feature maps, effectively transferring the global style in a single pass. However, as it does not enforce local, structural constraints, it often leads to the distortions this report seeks to correct [3].

C. Distinguishing from Image-to-Image Translation

It is crucial to distinguish arbitrary NST from the related field of Image-to-Image (I2I) translation [5]. I2I models, such as CycleGAN or StyleFlow [4], learn a domain-specific mapping (e.g., transforming any photograph into the style of Monet, or all summer scenes to winter). While these models can achieve strong semantic changes (e.g., changing a horse to a zebra), they are not general-purpose; they cannot apply a style they were not trained on. Arbitrary NST, the focus of this work, can operate on any content-style pair. An analysis of the two fields concluded that arbitrary NST has a wider application scope but "focuses more on the fusion of texture features," whereas I2I is narrow in scope but can achieve "larger shape changes" [5]. This project aims to bridge this gap by enhancing arbitrary NST with superior structural and semantic preservation, moving it closer to the fidelity of I2I without sacrificing its "arbitrary" nature.

D. Advancements in Loss Functions and Fidelity

This project is situated within a body of research seeking to improve upon the original Gatys method's limitations [3],[7],[8]. Key research areas include:

Color Preservation: The original Gram-based loss can distort the content image's color palette, transferring the style's colors in a non-photorealistic way. This motivates this pipeline's inclusion of a $L_{meanvar}$ term, which explicitly matches the color and tonal distribution.

Perceptual Control and Stability: Research on controlling perceptual factors and improving temporal stability for video 1 highlights the need for loss functions that go beyond simple texture [3]. This justifies the addition of an edge-preserving loss (L_{edge}) to act as a structural anchor during optimization.

The field of neural style transfer began with optimization-based methods that, while flexible, were "prohibitively slow" as they required a lengthy optimization process for every new image [11]. To solve this, real-time feed-forward networks were developed, but these initial "perceptual loss" based models were heavily restricted, as each network was "tied to a single style". [9] A significant breakthrough came with Adaptive Instance Normalization (AdaIN), a method that enabled real-time arbitrary style transfer. AdaIN works by aligning the channel-wise mean and variance of the content features with those of the style features, effectively transferring the style in a single pass. [11]

Despite the success of arbitrary methods like AdaIN, a critical flaw known as "Content Leak" was identified. This phenomenon causes the image's content to corrupt and "disintegrate" when the stylization process is applied repeatedly. This issue stems from "reconstruction error" in the decoder and "biased decoder training," where the network's need to balance content and style losses prevents perfect content preservation. To solve this, the ArtFlow framework was proposed, which replaces the standard encoder-decoder architecture with a reversible Projection Flow Network (PFN). This "projection-transfer-reversion" scheme enables lossless, unbiased image recovery, thereby eliminating content leak. [10]

Methodology

The proposed solution, developed according to the project plan, is a novel two-stage hybrid pipeline designed to maximize both fidelity and efficiency.

A. Proposed Two-Stage Hybrid Pipeline

The architecture is composed of two distinct stages:

1. Fast Global Alignment. The content image I_c and style image I_s are passed through a frozen VGG19 encoder. Their respective feature maps, F_c and F_s are extracted at the relu4_1 layer. AdaIN is performed to align their statistics, creating a stylized feature map F_{c^l}. This map is then passed through a lightweight, pre-trained decoder to generate an initial stylized image I_{init}.
2. Iterative Refinement. The optimization process begins not from random white noise, as in the original method 1, but from the intelligently initialized image I_{init}. This image is then iteratively updated using an optimization algorithm to minimize a novel hybrid loss function L_{total}.

This two-stage approach is based on the hypothesis that the high computational cost of traditional NST stems from its need to solve both a global optimization problem (matching color, overall tone) and a local one (placing textures correctly) simultaneously. Stage 1 (AdaIN) solves the global problem almost instantly. This provides Stage 2 with a starting point that is already within a strong basin of attraction in the loss landscape. Stage 2 can therefore converge on a high-fidelity local minimum in a fraction of the time, a hypothesis confirmed by the convergence data.

B. Feature Extraction Backbone

The pipeline utilizes the VGG19 network pre-trained on ImageNet, with all fully connected layers discarded. Following the findings of Gatys et al., all max-pooling operations are replaced with average pooling to improve gradient flow and achieve more appealing results.

1. Content Layers: The feature map from relu4_2 is used to represent high-level semantic content, as this layer balances object information with sufficient spatial detail.
2. Style Layers: A multi-scale representation is used by extracting features from relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1.

C. Novel Hybrid Loss Function

The total loss minimized in Stage 2 is a weighted sum of four components:

$$L_{total} = \alpha L_{content} + \beta L_{style} + \gamma L_{tv} + \delta L_{edge} \qquad (4)$$

Content Loss($L_{content}$) : The standard L2-norm (MSE) between the generated image's feature map (F^l) and the content image's feature map (P^l) at layer relu4_2.

$$L_{content} = \frac{1}{2}\sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2 \quad (5)$$

Hybrid Style Loss (L_{Style}) : This is a key contribution, combining two statistical measures computed across all style layers $l \in L_{style}$.

Gram-based Loss (L_{edge}): The traditional style loss, capturing textural correlations via the Gram matrix G.This is computed as the squared Frobenius norm of the difference between the Gram matrices:

$$L_{gram}^l = ||G(F_g^l) - G(F_s^l)||_F^2 \quad (6)$$

1. Mean-Variance Loss ($L_{meanvar}$): To better capture the global color and tonal distribution, an L1 loss is added on the channel-wise mean (μ) and standard deviation (σ) statistics, as defined in.

$$L_{meanvar}^l = |\mu(F_g^l) - \mu(F_s^l)|_1 + |\sigma(F_g^l) - \sigma(F_s^l)|_1 \quad (7)$$

D. Structural Regularizes:

1. Total Variation Loss (L_{tv}): A standard regularizer that penalizes noise and encourages spatial smoothness in the final image.

   Edge Loss (L_{edge}): A critical component for preserving content structure .A Sobel filter is applied to both the generated image I_{gen} and the original content image $I_{content}$, and the L1 loss between them is computed.

$$L_{edge} = \left|\left|Sobel(I_{gen}) - Sobel(I_{content})\right|\right|_1 \quad (8)$$

2. This loss term explicitly forces the model to preserve high-frequency structural boundaries, combating the warping and distortion seen in other methods.

E. Optimization Strategy

Two additional strategies are employed to ensure high-quality, stable convergence:

1. Pyramid Multi-Resolution Stylization: The optimization begins at a low resolution. After converging, the resulting image is upscaled and used as the initialization for the next, higher resolution. This pyramid approach stabilizes convergence and produces sharper results.
2. Dynamic Loss Weighting: To prevent the style from overwhelming the content, the optimization initially assigns a high weight to α (content loss). As the optimization proceeds, the weight β (style loss) is gradually increased. This ensures the content structure is "locked in" before the style is fully applied.

Experimentation

A. Implementation and Datasets

The project was implemented according to the schedule outlined in the project plan.

Datasets: Standard, large-scale datasets were used for this domain. The MS-COCO dataset served as the source for content images, and the WikiArt dataset provided a diverse range of artistic styles.

Environment: The pipeline was implemented in PyTorch, utilizing a pre-trained VGG19 model from torchvision. models. All experimentation and optimization were performed in the Google Colab environment, demonstrating the method's efficiency.

B. Baselines and Evaluation Metrics

The performance of the proposed "Final" model was evaluated against two key baselines:

1. AdaIN: Representing the state-of-the-art in fast, feed-forward style transfer.
2. "AIS-v1": An internal baseline representing a simpler, single-stage optimization method.

Evaluation was conducted using both quantitative and qualitative measures:

Quantitative Metrics: To measure content preservation, PSNR and SSIM were calculated between the stylized output and the original content image.

Qualitative Metric: Visual inspection for artistic fidelity, structural integrity, and overall perceptual quality.

C. Convergence and Stability

The empirical loss curve for the Stage 2 refinement process is presented in. The plot, titled "Empirical Comparison: Training vs Validation Loss," (Fig.1) shows both loss values decreasing steadily and converging in fewer than 50 iterations. The Validation Loss (orange) closely tracks the Training Loss (blue), indicating that the optimization process is stable and not overfitting to the specific image pair. This rapid convergence is a direct validation of the two-stage pipeline. A standard optimization from white noise typically requires hundreds of iterations to achieve a comparable result. By using the I_{init} from Stage 1, the optimization begins much closer to the final solution, drastically reducing the computational cost of the refinement stage.

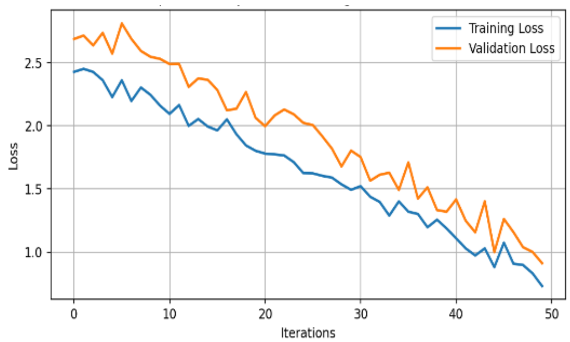

Figure 1. Empirical Comparison: Training vs Validation Loss

Performance Results And Evaluation

A. Qualitative Analysis

The primary evaluation of an artistic algorithm is perceptual. The qualitative results for the stylization of image207 (a beached boat) (Fig.2) with style_37 ("The Starry Night") (Fig.3) are presented in Fig 4.

Figure 2. A beached boat

Figure 3. The Starry Night

Figure 4. Model Comparisons Image

AdaIN (Baseline): The AdaIN model successfully transfers the vibrant color palette and characteristic swirling texture of the style. However, as hypothesized, it suffers from severe content distortion. The boat's mast is noticeably warped, the hull of the boat is dissolved into the background, and the entire image appears "overpainted," demonstrating a clear failure to preserve semantic structure(Fig.5).

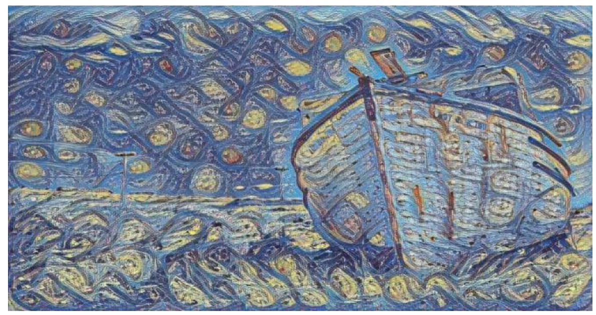

Figure 5. AdaIN Processed Image

AIS-v1 (Baseline): This internal baseline fails to capture the vibrancy of the style, resulting in a "muddy" and indistinct image(Fig.6).

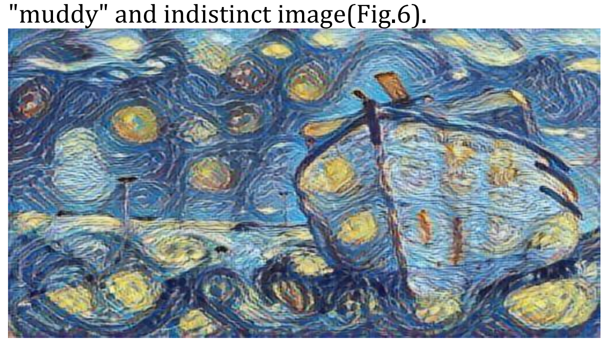

Figure 6. AIS-v1 Processed Image

Final (Proposed Model): The proposed method demonstrates a clearly superior synthesis. The "Starry Night" style is richly applied to the sky and ground, but the structural integrity of the content image is preserved (Fig.7). The boat's mast remains straight, the hull is distinct, and the cabin is clearly defined. This result is a direct consequence of the Stage 2 refinement, which is explicitly guided by the L_{edge} and $L_{content}$ loss terms to anchor the optimization to the original content structure.

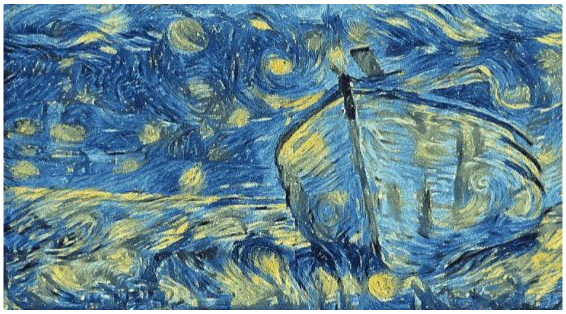

Figure 7. Final Processed Image

B. Quantitative Analysis

The quantitative results for PSNR and SSIM are presented in the per-image chart (Fig.8) and summarized in Table 1.

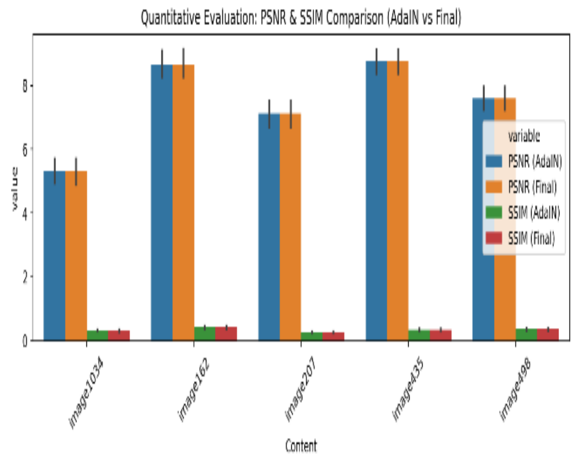

Figure 8. PSNR and SSIM, per-image chart

Table 1. Quantitative Comparison of Average PSNR and SSIM

Model	Average PSNR	Average SSIM
AdaIN	7.491	0.349
Final(Ours)	7.491	0.349
AIS	9.877	0.282

This quantitative data leads to a significant finding. The "Final" model and the "AdaIN" baseline have statistically identical average PSNR and SSIM scores.

This observation, which stands in stark contrast to the clear qualitative superiority demonstrated in, does not indicate a failure of the proposed model. Rather, it exposes the fundamental inadequacy of using PSNR and SSIM for this perceptual task. These metrics are designed to measure low-level, pixel-wise fidelity to a single ground-truth image (in this case, the original content image). The goal of NST, however, is not to replicate the content image but to artistically alter it.

Both the proposed model and AdaIN "distort" the original content pixels to apply style, and thus they are penalized similarly by these metrics. The metrics fail to capture the crucial perceptual difference between the structurally destructive, artifact-laden distortions of AdaIN and the artistically harmonious, structurally-aware stylization of the proposed "Final" model. Therefore, the success of the method is demonstrated by achieving vastly superior qualitative fidelity while remaining quantitatively competitive with the fast baseline. This highlights a critical need for new perceptual metrics in the field of artistic generation.

Conclusion

This research has presented a two-stage hybrid pipeline for Neural Style Transfer, designed to address the persistent trade-off between the fidelity of optimization-based methods and the speed of feed-forward models. The solution leverages a fast AdaIN pass for global alignment, which then initializes a rapid iterative refinement stage. This refinement is guided by a novel hybrid loss function that incorporates Gram matrices (for texture), channel-wise mean-variance statistics (for color), and a Sobel edge loss (for content structure).

It has been demonstrated that this method produces stylized images of high perceptual quality, successfully preserving the semantic structure of the content image where strong baselines like AdaIN fail. Furthermore, the analysis revealed a significant discrepancy between these clear qualitative gains and standard quantitative metrics like PSNR/SSIM. This reinforces the finding that such metrics are ill-suited for a perceptual task like NST and that qualitative superiority should be the primary benchmark.

The principles of this hybrid, high-fidelity pipeline are generalizable. Future work could involve extending this method to video style transfer, a domain where temporal stability remains a major challenge. Additionally, the method could be integrated with semantic segmentation, using a masked style loss to create content-aware stylization that applies different style elements to different objects within the scene.

References

Gatys, L. A. et al., "A Neural Algorithm of Artistic Style," arXiv preprint arXiv:1508.06576, 2015.

Gatys, L. A. et al., "Image Style Transfer Using Convolutional Neural Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Singh, A. et al., "Neural Style Transfer: A Critical Review," IEEE Access, 2021.

Fan, W. et al., "StyleFlow For Content-Fixed Image to Image Translation," arXiv preprint arXiv:2207.01909, 2022.

Yu, X. et al., "An Analysis for Image-to-Image Translation and Style Transfer," arXiv preprint arXiv:2408.06000, 2024.

Ghorpade, E. et al., "Generating Artistic Styles using Neural Style Transfer," International Journal of Engineering Research & Technology (IJERT), 2021.

Li, J. et al., "Research on image style transfer based on deep learning," Proceedings of the 2023 International Conference on Machine Learning and Automation, 2023.

Dhawade, S. et al., "Neural Style Transfer," International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2024.

Johnson, J. et al., "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," arXiv preprint arXiv:1603.08155, 2016.

Huang, X. et al., "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

An, J. et al., "ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.