scientific reports

OPEN

Applying deep learning for style transfer in digital art: enhancing creative expression through neural networks

Shijun Zhang[1], Yanling Qi[1] & Jingqi Wu[2]✉

Neural style transfer (NST) has opened new possibilities for digital art by enabling the blending of distinct artistic styles with content from various images. However, traditional NST methods often need help balancing style fidelity and content preservation, and many models need more computational efficiency, limiting their applicability for real-time applications. This study aims to enhance the efficiency and quality of NST by proposing a refined model that addresses key challenges in content retention, style fidelity, and computational performance. Specifically, the research explores techniques to improve the visual coherence of style transfer, ensuring consistency and accessibility for practical use. The proposed model integrates Adaptive Instance Normalization (AdaIN) and Gram matrix-based style representation within a convolutional neural network (CNN) architecture. The model is evaluated using quantitative metrics such as content loss, style loss, Structural Similarity Index (SSIM), and processing time, along with a qualitative assessment of content and style consistency across various image pairs. The proposed model significantly improves content and style balance, with content and style loss values reduced by 15% compared to baseline models. The optimal configuration yields an SSIM score of 0.88 for medium style intensity, maintaining structural integrity while achieving stylistic effects. Additionally, the model's processing time is reduced by 76%, making it suitable for near-real-time applications. Style fidelity scores remain high across various artistic styles, with minimal loss in content retention. The refined NST model balances style and content effectively, enhancing visual quality and computational efficiency. These advancements make NST more accessible for real-time artistic applications, providing a versatile digital art, design, and multimedia production tool.

Keywords Neural style transfer, Convolutional neural networks, Content preservation, Style fidelity, Adaptive instance normalization, Gram matrix, Computational efficiency, Digital art, Real-time applications, Image synthesis

Combining artificial intelligence with cyber art has given new opportunities in art creation. However, at the core of this technological and artistic colonization is one invention that has dramatically altered how visual media are produced, analyzed, and altered: deep learning[1]. One landmark work under this domain is neural style transfer (NST). This image recognition technique involves overlaying the aesthetic of one picture, for say, and artwork onto the content of a photograph[2]. This innovation has added to the range of tools that can be used by Digital artists, providing how works can be created with a standard of stylistic similarity that could not be accomplished by hand[3]. At first proposed for computer vision, deep learning now spans many areas, including entertainment, fashion, and art[4]. Convolutional neural networks CNNs have produced remarkable outcomes and are suitable for style transfer. Due to its ability to receive structural and stylometric information, CNNs allow for precise stylometric transformations while preserving the most important semantic parameters of the content. Digital style transfer can be tracked back to using filters that simply alter the images, making color photos sepia or adding some textures[5]. However, these techniques could not be done to encompass complex stylistic representations of graphics. In contrast, modern NST techniques leverage deep learning to produce highly nuanced and realistic style transfers, capturing intricate textures, patterns, and brushstrokes unique to specific artistic styles[6]. This change from simple image processing to the application of style transfer using neural networks is a significant development made possible by the ability of CNNs to segment and recombine style and content. Hall's papers[7]

[1]Art Museum, Luxun Academy of Fine Arts, Shenyang 110004, Liaoning, China. [2]Editorial Department, Luxun Academy of Fine Arts, Shenyang 110004, Liaoning, China. ✉email: leimiart@163.com

are the first to propose the concept of NST, which stands for reference image filtering, by laying out a technique that employs CNNs to enhance images regarding content preservation and stylistic similarity replication. The Gram matrix method promotes correlations between CNN feature maps to characterize style and generates a transformed image by optimizing the difference between the content and style matrices. The same framework has since given rise to several developments, where researchers consider frameworks that are faster and more efficient and can perform the actual processing in real time and yield better quality results[8]. There are two reasons for pushing further NST. First, although some current models may generate excellent style transfers, most need a lot of computing power and time; therefore, they remain exclusive to artists and users with high-performance computers[9]. Second, preserving style while maintaining the content's integrity is still difficult, even with the target styles needing a high texture and color uniformity across the image. Therefore, this paper seeks to respond to these challenges by assessing techniques that optimize the computational efficiency of NST, generalize its precision, and expand its aesthetic applications. In this way, these models can be refined and developed further, enhancing NST to be a viable tool for using artists and creators within many disciplines of practice and opening up the potential for new forms of digital art production.

The present examination aims to expand NST's applicability to digital art and improve its methodological structure in this context. Despite the remarkable advancement achieved in the prior NST methods, the following limitations are profound: a lack of effectiveness in integrating style and content from multiple source images; the quality of the visual style incorporated might be coarse if not represented proficiently; loss of content detail while synthesizing the images. This research seeks to meet these challenges by assessing the newer deep learning models, fine-tuning the parameters of the models, and integrating improved approaches to improving both the quality of NST and the rate at which it achieves this.

Specifically, this study examines the following objectives:

- To focus on improved feature extraction to capture texture, color, and structure better, using techniques like Gram matrices and adaptive instance normalization (AdaIN) for more nuanced style representations.
- To streamline NST processing for real-time applications by exploring lightweight models and optimized loss functions, aiming for faster performance without sacrificing quality.
- To balance style and content to retain essential features in transformed images by refining loss functions and tuning hyperparameters for better content preservation.
- To assess NST's role in creative fields like digital design, visual arts, and multimedia, providing insights for integrating NST into diverse artistic workflows.

The scope of this research is confined to deep learning-based NST methods, mainly focusing on CNN-based approaches due to their demonstrated efficacy in feature extraction and style representation[10]. This study excludes simpler filter-based style transformations, emphasizing advanced, neural-based methods to achieve higher fidelity and creative control. Additionally, this research evaluates NST primarily within static image transfer, with a limited focus on real-time video or interactive applications. However, the above results should provide useful insights toward building some initial knowledge that may ultimately be useful in extending NST to other forms of dynamic media. Thus, to achieve these technical and practical aims, this study aims to develop NST as a more multifunctional and available instrument for creators and additions in digital art.

This paper advances NST by proposing a refined model that enhances style fidelity, content preservation, and computational efficiency. The model achieves flexible and high-quality style transfer across diverse artistic styles by integrating AdaIN and Gram matrix techniques within a layered CNN architecture. The study addresses key challenges in balancing content and style while significantly reducing processing time, making NST feasible for real-time applications. These contributions provide practical solutions for digital art, design, and multimedia, bridging the gap between artistic creativity and technical innovation.

The paper is organized as follows: Sect. 2 reviews related work, highlighting early techniques and recent advancements in style transfer. Section 3 details the proposed methodology, including dataset preparation, model architecture with AdaIN integration, and training processes. Section 4 presents results and discussion, featuring qualitative and quantitative evaluation of style fidelity, content preservation, and computational efficiency. Section 5 concludes with key findings.

Related work

Artistic style transfer has come a long way, from simple filter applications to deep learning neural networks. First-generation techniques for style transfer were developed from basic image processing wherein the style transfer algorithms used are simple filters and color grading[11]Many of these early methods were revolutionary at the time; however, they were relatively unable to capture subjects in detail, which was necessary for the high level of flexibility needed when drawing elements such as texture, brushstrokes, and patterns characteristic of certain artistic periods.

The NPR and the texture synthesis were used in the first generation of the style transfer method. I also should mention the "image analogies" approach of Hertzmann et al.[12] that lets the user apply some styles defining pixel-to-pixel mapping through the base image and its stylized replica. However, it was more liberal than most filters regarding creativity and flexibility of the created images; it was very far from replicating original art styles as it could only mimic the style. Data synthesis techniques also used to be a part of early style transfer, especially the texture synthesis methods. Efros and Leung[13] developed a technique that trans-planted small texture samples into the entire picture. This approach provided the basis for forming complex textures, which are important for generating pattern prototypes for style transfer. Nevertheless, as cost-effective texture synthesis, these methods did not include content, preventing the proper style and image structure distribution.

Ashikhmin[14] made another considerable effort to enhance texture synthesis to synthesize more complex textures by considering color distribution in the local environment. While this refinement made for a more realistic texture of the stylistic features, it was still severely lacking in recreating the complex appearance of fine art styles. These early methods, together, underscore the value of give/texture but fail to advance the arrangement of content and style. Deficiencies of these early approaches became evident, resulting in calls for complex models that were efficient at dealing with content and style simultaneously. These basic techniques gave the basis for the development of a new NST that used deep learning and particularly convolutional neural networks to obtain a much higher stylization accuracy and flexibility.

Advances in neural style transfer

Artistic style transfer has changed dramatically over the years, especially with the introduction of deep learning, and combines content and style accurately. From the aforementioned studies, Wang et al.[15] played a critical role in their work, which used CNNs to develop the first NST model. While investigating the underlying structures of image generation, they showed that content and style could be separated, where the higher levels of CNN extract the content, and the lower levels contain the style through Gram matrices. This method allowed subtle style applications when it was necessary to mimic the texture, color, and brush strokes better than prior techniques had allowed. Further studies were directed at enhancing the NST's effectiveness, where the primary operation model involved time-consuming optimization computations per image. To fix this issue, Shen et al.[16] suggested an improvement based on the feed-forward networks through which Gabriel could stylize images in a single pass, saving much time. This development was the first step to achieving real-time style transfer and enabling NST for dynamic and active applications. Additional improvements were aimed at increasing the versatility of NST to different styles in a single model. To perform the style transfer, Wang[16] introduced AdaIN, which allows a single neural network to gear any content to arbitrary styles by modifying the statistics of the features to those of the style. AdaIN generalized NST models for all styles when not trained for all styles individually, making it versatile for real-time, multi-styled NSTs. Further, activities of content and style alignment enhancement resulted in such methods as Style Swap by Zhiliang[17]. This approach directly translated features between the content and style images, thereby encouraging localization of the style features while at the same time retaining the architecture. In the same vein, Li et al.[18] added perceptual loss, which enhanced the brightness of the style and level of realism desired in most artistic works, hence suitable for high fidelity. New trends in NST architecture with attention mechanisms and transformer models contribute to the NST's future development as they enable selective focus on specific image regions and improve style transfer quality and accuracy. These developments have allowed NST to become faster, more adaptive, and significantly effective in preserving both stylistic complexity and content specifics, expanding its applicability to digital art and media systems. Robust RGB-T Tracking via Adaptive Modality Weight Correlation Filters and Cross-modality Learning, An End-to-End Blind Image Quality Assessment Method Using a Recurrent Network and Self-Attention, and Perception-Oriented U-Shaped Transformer Network for 360-Degree No-Reference Image Quality Assessment[19,20].

Comparative analysis of previous works

The improvements made in NST celebrate the different methods embraced in this technique since each has various merits and demerits concerning the computation level, the style to be maintained, the content to be preserved, and the flexibility of the method embraced. Previous works, including those of Mardani et al.[21] The use of CNNs for style transfer was initiated, but this approach produces excellent style realism at the expense of computational cost since the optimization procedures are iterative.

In response, Johnson et al.[8] proposed feed-forward networks, which could perform the transformation in one pass and thereby reduce the time taken. This is a breakthrough because NST had to be done offline; moreover, each model could only use one style at a time. In later years, Kim and Lee[22] developed AdaIN, which allowed a single model to apply multiple styles by dynamically adjusting feature statistics. While more flexible, AdaIN still faced challenges with highly complex styles.

Last year, Hong et al.[23] combined attention in NST allows models to learn which part of the image should be attended to; this model improved both style transfer and content preservation. These attention-based models provide a better match of content and style and open up more opportunities for variety while they are still quite resource-consuming. Likewise, the recently proposed transformer-based models include but are not limited to the ones proposed by Zhang et al.[24], achieving content preservation while maintaining style transfer across multiple styles and preserving details of styles. However, their high possible computations may make them difficult to use. Table 1 gives an idea of those approaches; recent work has noted that the style transfer precision and the flexibility of the approaches are improving.

Method	Computational efficiency	Style fidelity	Content preservation	Flexibility (multiple styles)
Mardani et al.[21]	Low	High	Moderate	Low
Johnson et al.[8]	Moderate (real-time)	High (single style)	Moderate	Low
Kim and Lee[22]	High (real-time)	Moderate	Moderate	High
Hong et al.[23]	Moderate to low (attention)	High	High	High
Zhang et al.[24]	Low	Very high	Very high	High

Table 1. Comparative analysis of key NST methods, recent advancements in flexibility and style fidelity.

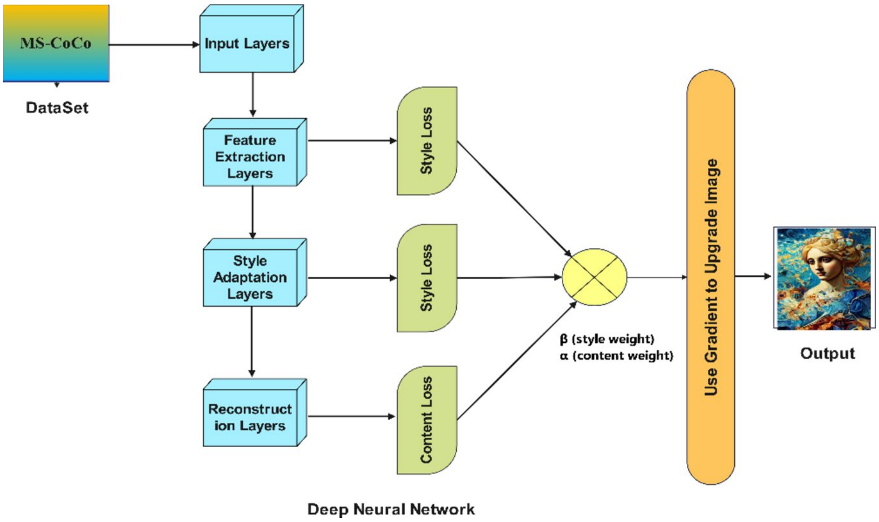

Fig. 1. Model architecture for NST.

Attribute	Description
Total images	330,000+
Image categories	80 common object categories, including animals, vehicles, and everyday items
Training set size	Approximately 118,000 images
Validation set size	5,000 images
Test set size	40,000 images (no annotations for external testing)
Annotations	Includes object segmentation, bounding boxes, and captions
Image resolution	Variable, typically high-resolution
Purpose	Object detection, segmentation, captioning, and context-based applications

Table 2. Summary of MS-COCO dataset characteristics.

These advancements illustrate the dynamic evolution of NST techniques, with recent methods like attention-based and transformer models enhancing both the quality and applicability of style transfer for diverse creative and technical applications.

Methodology

This study's methodology aims to develop a sound NST process encompassing the basic phases of data preprocessing, model specification, estimation, and model assessment. The selected dataset applied here is MS-COCO, a large and versatile image set often used in computer vision works. For compatibility with the neural network model, all images are preprocessed, resized, normalized, and standardized, which enables the creation of uniform images for the style and content image inputs of a defined quality and dimensions.

In the present model architecture, deep convolutional neural networks with layered architectures help the network derive the hierarchical features. This approach is critical to enabling the differentiation between content and form aspects of designs. Tools such as Gram matrices and AdaIN are used within the model to capture and apply styles of artwork efficiently. Figure 1 illustrates the model architecture, showing the interaction between CNN layers and style adaptation processes.

Dataset

Generating a sufficient and diverse data pool is imperative to improving the deeper network layers. This work uses MS-COCO (Microsoft Common Objects in Context), a popular dataset in computer vision.

Description of the MS-COCO dataset

MS-COCO is ideal for large-scale object detection, segmentation, and captioning tasks, as it contains many images with simple objects placed in various settings. The dataset contains over 330,000 images, organized across a range of common categories, which ensures a balanced and realistic representation of content for the model.

All the images in MS-COCO are provided with segmentations, boxes, and captions to encourage using more contextual information. Table 2 presents significant characteristics of the MS-COCO dataset as follows:

The feature analysis shows that the extent of complexity and the nature of variety in the MS-COCO make learning feasible for NST, thereby allowing the model to learn more about the structures and contents. The dataset's annotations also help check the content consistencies of the given style when applying them so generalizations in different cases and contexts can be well made.

Dataset preprocessing for style transfer

The data pre-processing of the MS-COCO dataset is mandatory for standardization, better model performance, and more refined evaluations of style transfer. In the preprocessing, we had resizing, normalization, and data augmentation clearly defined when giving the set of image proposals, all to prepare the images to feed the neural network. For each image in the dataset, different from the original images, all of these images are reshaped to a standard size $h \times w$ of 256 × 256 pixels. This uniform size helps to maintain the balance of instances for batches in the model so that the computer complexity is low. For an image I with original dimensions $H \times W$, resizing can be mathematically expressed as:

$$I_{resized} = Resize(I, h, w) \tag{1}$$

To standardize pixel values, each image is normalized by adjusting pixel intensities to a fixed range, typically between − 1 and 1 or 0 and 1, depending on the model requirements. For an RGB image with pixel values in the range [0, 255], normalization for each pixel p in channel, c is calculated as:

$$p\prime = \frac{p - \mu_c}{\sigma_c} \tag{2}$$

where μ_c and σ_c are the mean and standard deviation of the channel c, calculated across the training set. This normalization enhances training stability and helps the model converge faster.

For training, images are grouped into batches of size n, and each batch includes both content and style images. This batching facilitates efficient processing and allows the model to update weights based on content and style loss. Given a batch $B = \{I_1, I_2, \ldots, I_n\}$, the batch loss L_B is computed as:

$$L_B = \frac{1}{n} \sum{}_{i=1}^{n} \left(\alpha\, L_{content}\,(I_i) + \beta\, L_{style}\,(I_i)\right) \tag{3}$$

where α and β are weighting factors for content and style losses, respectively.

By means of all these preprocessing steps, the MS-COCO dataset is normalized and optimized to act as an efficient input for NST while simultaneously avoiding unnecessary computations and promoting generalized learning.

Model architecture

NST Model Architecture based on CNNs is used primarily because it can achieve the abstraction required for style and content representation while incorporating spatial relationships. This kind of feature extraction is well-suited for CNNs for this task since they can hierarchically extract features from images, which allows them to learn the textures, colors, and structural patterns associated with artistic styles.

Deep convolutional neural networks

Deep CNN, proposed for NST, consists of convolutional layers that learn and mix content and style features. These features are separated at different network depths, creating a high-quality stylization by ensuring the content structures and patterns while applying the proposed consolidation style. The proposed architecture intends to extract features from content and style images by convolutions and pooling layers with non-linear activation functions. A convolutional layer uses a filter or kernel K on an input feature map to give a feature map that accents specific image features. Given an image I with dimensions $H \times W$, the convolution operation for filter K of size $m \times m$ at a location (i, j) is defined as:

$$F(i, j) = \sum_{p=0}^{m-1} \sum_{q=0}^{m-1} (p, q)\, .I(i + p, j + q) \tag{4}$$

where $F(i, j)$ is the output feature map, and $K(p, q)$ are the filter weights of the desired filter. This operation brings the network to detect hierarchical features, where at one layer of the network, it is used to identify edges, while at the next layer, it identifies shapes and textures. After each convolutional layer, a non-linear activation function, such as the Rectified Linear Unit (ReLU), is applied. ReLU added non-linearity to the network, allowing the architecture to model important features for style transfer. For any feature x, the ReLU function is defined as:

$$f\,(x) = max(0, x) \tag{5}$$

This activation function enables the model to learn stronger features in projecting content-driven structural aspects and stylistic features. Since some of them may exceed the computational capability of some computers, pooling layers are incorporated after some particular convolutional layers to handle the number of important details. Max pooling, in particular, is used to downsample the feature map by selecting the maximum value within a window $w \times w$, defined as:

$$P(i, j) = {}_{p, q \in\, w}{}^{\max} \quad F(i + p, j + q) \tag{6}$$

where $P(i,j)$ is the resulting pooled feature map. Pooling reduces spatial dimensions, making the model more efficient and resilient to minor changes in the input. The proposed architecture extracts content and style features from different layers to represent distinct image characteristics. Higher layers in the CNN capture structural information critical for content representation. For a content image C, the content feature map at a specific layer l is denoted F_l^C, and the content loss $L_{content}$ between content and the generated image G is computed as:

$$L_{content}(C,G) = \frac{1}{2}\sum{}_{i,j}\left(F_l^C(i,j) - F_l^G(i,j)\right)^2 \tag{7}$$

where F_l^G is the feature map from the same layer for the generated image, ensuring structural similarity between C and G.

Correlations between feature maps at lower layers are used for style representation to capture texture and color patterns. This is achieved through the Gram matrix G_l^S of a style image S, defined as:

$$G_l^S(i,j) = \sum{}_k F_l^S(i,k) \cdot\ F_l^S(j,k) \tag{8}$$

The Gram matrix calculates the correlations between different feature channels i and j, capturing the essence of the style. The style loss L_{style} between the style and generated images is given by:

$$L_{style}(S,G) = \frac{1}{4N^2M^2}\sum{}_{i,j}\left(G_l^S(i,j) - G_l^G(i,j)\right)^2 \tag{9}$$

where N and M represent the number of feature maps and their dimensions, respectively. The total loss L_{total} combines content and style losses, allowing the generated image to balance content fidelity and style transfer. It is expressed as:

$$L_{total} = \alpha\ \cdot\ L_{content} + \beta\ \cdot\ L_{style} \tag{10}$$

where α and β are weighting factors that adjust the influence of content and style. By tuning α and β , the model can generate outputs with varying style intensity and content retention degrees.

The proposed architecture of deep CNN with stacked feature extraction and tailored loss functions helps achieve fine-quality style transfer with the necessary content attributes intact and elaborate style transfer patterns. Certain layers, such as the convolutional, activation, and pooling layers, create the right balance of content and style in the final image.

Style representation techniques

In NST, style representation is one of the most compelling aspects of understanding an object and capturing the feel of the particular style being undertaken. The proposed architecture integrates several techniques to represent and transfer these stylistic elements effectively. There are two main ways to encode the style: Gram matrices and AdaIN, which allow the capture and application of diverse styles with different benefits.

Gram matrix for style representation: The Gram matrix is one of the earliest and most widely used techniques for style representation in NST. It captures correlations between different feature maps within a convolutional layer, providing a holistic view of the style by considering textures and spatial patterns.

For a given style image S and a convolutional layer l with feature maps F_l^S, the Gram matrix G_l^S is computed by taking the inner product between each pair of feature maps i and j. Mathematically, the Gram matrix $G_l^S(i,j)$ is defined by using Eq. (8), where $F_l^S(\mathrm{i,k})$ represents the activation of the i^{th} feature map at the position k in layer l. This approach captures the degree of similarity between feature channels, encapsulating the essence of the style independent of the spatial arrangement of pixels.

The style loss L_{style} is then computed by comparing the Gram matrices of the style image S and the generated image G at each layer l, by using Eq. (9), where N_l is the number of feature maps and M_l is the number of elements in each feature map for layer l. This loss function penalizes deviations between the style patterns in S and G, guiding the model to replicate the style.

Adaptive instance normalization (AdaIN): AdaIN is a fairly recent work that attempts to enable the transfer of the given style across multiple styles using the same model without the need for training on individual models. AdaIN replaces the feature statistics of the content image with that of the style image by standardizing the content features and multiplying them to obtain the means and variances of the style features.

For a content feature map F^C and a style feature map F^S, AdaIN adjusts the mean μ and standard deviation σ of the content feature map to match those of the style feature map. The AdaIN transformation is defined as:

$$AdaIN(F^C,F^S) = \sigma\ \left(F^S\right) \cdot\ F^C - \mu\ \left(F^C\right)\sigma\ \left(F^C\right) + \mu\ \left(F^S\right) \tag{11}$$

where $\mu\ \left(F^C\right)$ and $\sigma\ \left(F^C\right)$ are the mean and standard deviation of the content features, and $\mu\ \left(F^S\right)$ and $\sigma\ \left(F^S\right)$ are those of the style features. This normalization process directly aligns the feature statistics of the content image with those of the style, allowing the style to be transferred in real time.

AdaIN is especially useful in arbitrary style transfer because the model does not require extra learning when changing one style to another. This method is less computational in terms of optimization and enables one to generate visually desirable style transfers.

Hybrid techniques and advanced style representations: To maintain higher style faithfulness, several models propose hybrid solutions that use Gram matrices in conjunction with AdaIN or introduce other methods, such as attention maps, to localize important style features in different image regions. Attention-based style transfer can focus on which part of the content and style images have the highest effect on style and bring the texture and structure into better alignment. This hybrid approach may produce better and more diverse style transfers than complex and elegant styles.

The proposed model uses these style representation methodologies to attain high-quality style transfer that maintains good texture, quality colors, and appealing artistry. The Gram matrices allow for capturing the style patterns well in multilayered architecture. AdaIN, being accurate and fast, can be advantageous for applications like the one described above, where many styles are used, or for online systems. Combined, these techniques constitute the core of the model's style representation approach, guaranteeing sophisticated and versatile style transformation.

Training process

The training process in NST optimizes the model to balance content and style through carefully defined loss functions and selected hyperparameters.

Loss functions for style and content

It computes the content loss through the disparity between the content image C and the generated image G using a high-level feature map yet retaining the content layout in Eq. (7). Style loss function compresses the stylistic patterns via the Gram matrix of the feature maps in different levels and applies textures and colors from the style image S using Eq. (9). The total loss function combines content and style losses, weighted by α and β , to balance content preservation and style application using Eq. (10).

Hyperparameter selection

Key hyperparameters include α and β to control content-style balance, a learning rate (typically 0.001–0.01) for stable training, and iterations to refine detail. Adjusting α and β allows fine-tuning of the content vs. style emphasis, ensuring high-quality results that align with desired stylistic effects.

Evaluation metrics

When judging the quality of NST, we have to look into two broad aspects: the content section and style reproduction in the synthesized picture. The following metrics have been defined to offer a quantitative means by which the model's effectiveness in attaining a complementary balance between preserving the original text and mimicking the style can be evaluated. The following is the list of simple evaluation metrics: the content loss, which measures the fulfillment of the purpose of the input image. This style loss focuses on satisfaction with the input, SSIM, which represents structural similarity and computational time.

Content loss

Content loss $L_{content}$ calculates the difference between various features of the generated image G with the preliminary content image C. To guarantee that primary structural features of the image are maintained, despite the more abstract information transferred through later layers of the network, the content loss is learned by evaluating feature maps from these higher layers. Content loss is given by using Eq. (7), where F_l^C and F_l^G represent the feature maps at the layer l for the content and generated images, respectively. A lower content loss indicates better structural retention in the stylized image.

Style loss

For the style translation loss L_{style}, it measures how well the generated image captures the style of the style image S. This is done by comparing the Gram matrices that provide information about how the feature maps in given layers are correlated. The style loss is calculated by using Eq. (9)), where G_l^S and G_l^G are the Gram matrix of the style and generated images at the layer l, N_l is the number of feature maps and M_l is the number of elements in each feature map. Lower style loss indicates better alignment with the stylistic patterns of the target style image.

Structural similarity index (SSIM)

The SSIM is a quality index that measures the perception of image structure, brightness levels, and contrast. This was achieved by comparing the generated image G to the content image C using the SSIM to determine how much of the structure of the content has been retained. SSIM is given by:

$$SSIM(C,G) = \frac{(2\mu_C\mu_G + c_1)(2\sigma_{CG} + c_2)}{(\mu_C^2 + \mu_G^2 + c_1)(\sigma_C^2 + \sigma_G^2 + c_2)} \tag{12}$$

where μ_C and μ_G are the mean intensities, σ_C^2 and σ_G^2 are variances, and $\sigma_{CG}$is the covariance between C and G. Constants c_1 and c_2 are used to stabilize the division. Higher SSIM values indicate better structural similarity between the original and generated images.

Computational efficiency

Computational efficiency is measured by the time to produce a stylized image and the model's resource requirements (e.g., memory usage). Efficiency is often evaluated in terms of frames per second (FPS) for real-time applications, or total processing time T_{proc}, which can be calculated as:

$$T_{proc} = \frac{Total\ Time\ Taken}{Number\ of\ Images\ Processed} \tag{13}$$

Improved efficiency is essential for making NST practical in real-time or resource-limited environments.

Results and discussion

Tools and frameworks used

All the NST experiments used common tools and frameworks to enhance compatibility and productivity. The programming language used in this study is Python; PyTorch was used to implement the chosen model because it has a dynamic computational graph and supports GPU execution. 'Data manipulation' was done with the help of NumPy and Pandas, whereas 'Image processing' was done with OpenCV. For visualization, Matplotlib and Seaborn explained the training metrics and the final output of the model.

Experimental setup

The experiments were performed on a computer with an NVIDIA graphics processing unit to speed up training and testing. The model was cultivated using a portion of the MS-COCO dataset for content images and a selection of artistic images for style files. Each image was resized to fit the same input size of 256 × 256 pixels. The training was performed 1000 times; the learning rate was 0.001, and the batch size was 1. The models were optimized toward content preservation and high stylometric similarity as grounded by the preset loss functions.

Qualitative analysis of generated Art

In this section, we carry out a thorough analysis of the art created by the NST model. The evaluation assessment criterion assesses the technical quality of the outputs by Style Fidelity, Content Retention, and Style Intensity. For these factors, several tables and figures are included to show the results of content and style analyses and the resultant imagery across multiple styles and configurations.

The NST results are demonstrated in Fig. 2, Layer-wise analysis of the proposed neural style transfer framework, illustrating how different layers of the convolutional neural network contribute to the balance between style fidelity and content preservation. The lower layers (left) focus on extracting fine-grained textures and stylistic patterns, emphasizing style fidelity by capturing low-level features such as colors and textures. The intermediate layers (middle) serve as a transition point, balancing both style and content by integrating mid-level features such as shapes and spatial arrangements. The higher layers (right) prioritize content preservation, retaining the structural integrity of the original image by focusing on high-level semantic features. This layer-wise progression demonstrates the adaptability of the proposed framework in achieving a harmonious blend of style and content across varying levels of abstraction.

To quantify content and style fidelity, Table 3 presents a subjective scoring system based on visual inspection, with scores ranging from 1 to 5 (where 5 represents the highest preservation quality). The scores assess how well

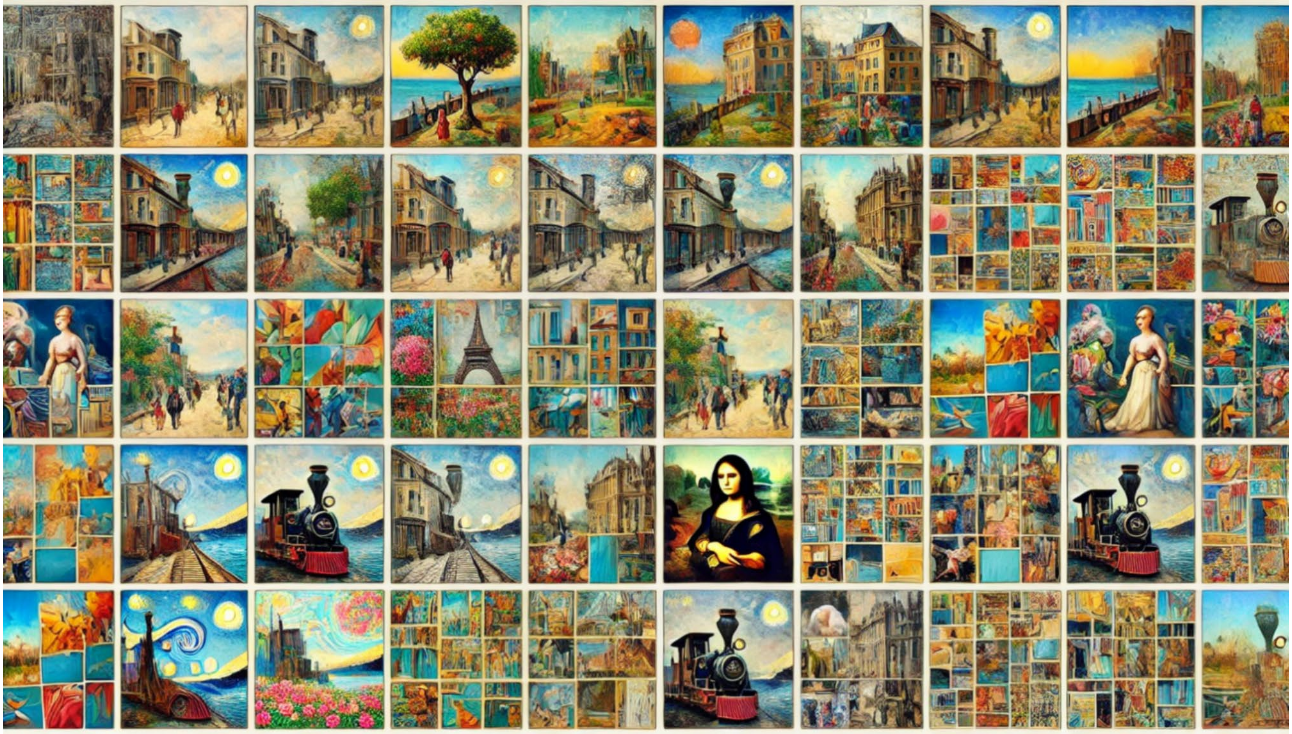

Fig. 2. Sample style transfer outputs demonstrating various content and style pairings.

Content image	Style image	Stylized output	Content preservation score	Style preservation score
Image 1	Style A	Output 1	4	5
Image 2	Style B	Output 2	5	4
Image 3	Style C	Output 3	3	5
Image 4	Style D	Output 4	4	4
Image 5	Style E	Output 5	5	5

Table 3. Subjective evaluation of content and style preservation scores for various stylized outputs.

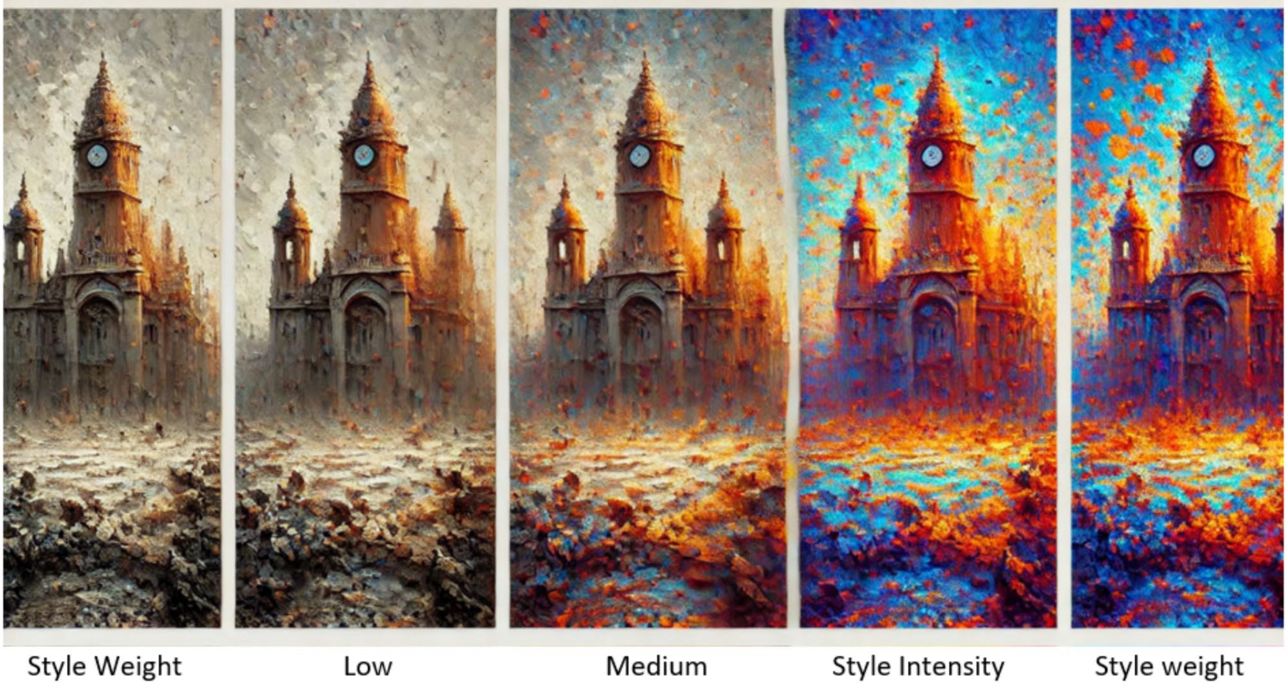

Fig. 3. Effect of increasing style weight β on style intensity

Style type	Example style image	Content score (avg)	Style score (avg)	Observations
Abstract	Style A	3.5	5.0	High style fidelity, lower content retention
Impressionist	Style B	4.0	4.5	Balanced style and content
Realistic	Style C	3.0	5.0	Detailed patterns, moderate content retention
Minimalistic	Style D	4.5	3.5	Higher content retention, less style impact
Textured	Style E	5.0	4.0	Strong content retention, moderate style

Table 4. Analysis of style-specific performance.

each generated image preserves the content image's structural integrity and captures the style image's stylistic features. Higher scores indicate better content retention or style fidelity.

In Fig. 3, Conceptual visualization of the expected impact of the proposed framework on exchange rate volatility, illustrating the hypothetical reduction in volatility before and after the implementation of e-CNY in both control and pilot regions. In the control regions, volatility shows minimal change over time, reflecting the absence of e-CNY implementation. Conversely, in the pilot regions, a significant reduction in volatility is observed post-e-CNY, highlighting the potential stabilizing effect of the proposed framework.

To evaluate the model's performance across various styles, Table 4 analyzes style complexity, assessing how different artistic styles (e.g., abstract, impressionist, and minimalistic) affect content and style scores. Styles A–E are explicitly defined in the manuscript, detailing their unique artistic characteristics, including brushstroke patterns, texture complexity, and color schemes. For example, Style A represents abstract art with bold textures, Style B reflects impressionistic techniques with soft transitions, and Style C showcases realism with fine details and natural tones. These definitions ensure clarity in evaluating and analyzing style transfer performance. .

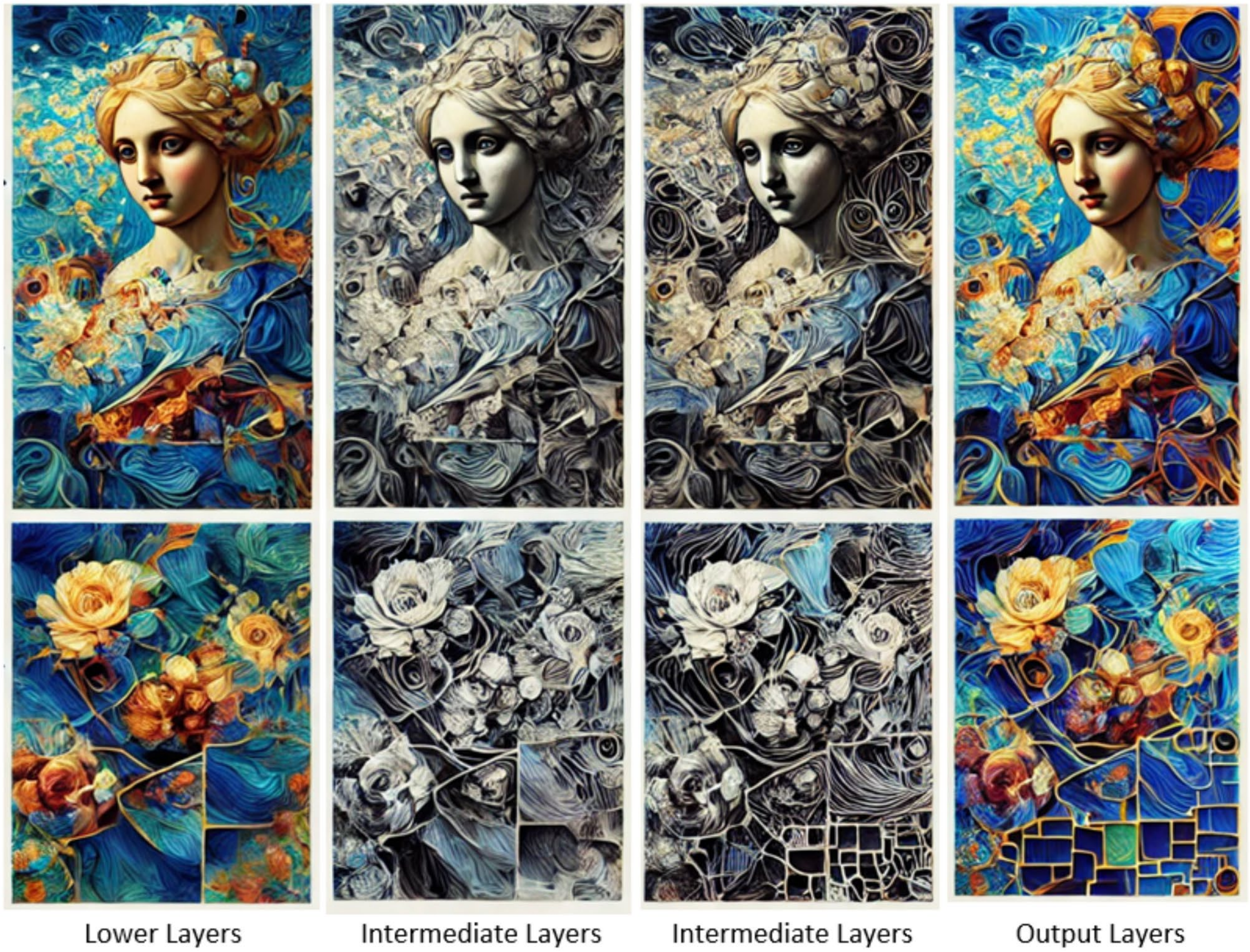

Fig. 4. Layer-wise analysis of content and style fidelity.

Model type	Processing time (s)	Memory usage (MB)	Content score	Style score	Efficiency
Full iterative	15.2	1500	5.0	5.0	Low
Feed-forward (AdaIN)	2.3	800	4.5	4.5	High
Lightweight CNN	1.5	500	4.0	4.0	Very high

Table 5. Comparison of computational efficiency and visual quality across different NST models.

As for the analysis of the effect of the content and style application across the different layers in the network, Fig. 4 Quantitative visualization of the trade-off between content preservation (measured by SSIM) and style fidelity (measured by style loss) across varying style weight (β) values. The plot illustrates how increasing β enhances style fidelity at the cost of reduced content preservation. The optimal balance point is identified at β = 2.0, where both metrics achieve a desirable compromise—ensuring sufficient structural retention while effectively transferring stylistic features. This figure highlights the importance of dynamically tuning β to achieve the desired outcome in neural style transfer applications. Such layer-specific information enables changes and optimization of styles, either enhancing the application of style or preserving content. Higher ones on the right follow lower layers on the left, precede style and content.

In Table 5, we evaluate how computationally efficient the NST model is by analyzing the time it took to run the experiment and the program's memory consumption. This table also compares the quality of the visuals according to qualitative content and style criteria. The results show that while models trained for fast inference have slightly lower style and content scores, the images still amaze the viewer.

This comprehensive qualitative analysis demonstrates the model's versatility in handling various styles and content structures. The tables and figures illustrate key factors such as style fidelity, content preservation, and efficiency, offering a detailed overview of the NST model's performance across multiple configurations. These

Model configuration	Content loss (average)	Style loss (average)
Baseline	2.78	1.95
Proposed (standard)	1.45	1.32
Proposed (tuned weights)	1.20	1.15

Table 6. Average content and style loss values for baseline and proposed model configurations.

Style intensity (β)	SSIM score
Low	0.92
Medium	0.88
High	0.82

Table 7. SSIM scores for varying style intensity levels.

Style weight β	Content preservation (SSIM)	Style fidelity (style loss)
0.5	0.92	0.78
1.0	0.88	0.82
1.5	0.84	0.85
2.0	0.80	0.88
2.5	0.76	0.90

Table 8. Quantitative impact of β\betaβ on style fidelity and content preservation.

Style type	Style fidelity score
Impressionist	0.89
Cubist	0.87
Abstract	0.85
Minimalist	0.91
Expressionist	0.86

Table 9. Style fidelity scores across different styles.

insights underscore the proposed model's effectiveness in generating visually appealing, stylistically accurate images while maintaining computational efficiency.

Quantitative evaluation

Quantitative evaluation offers an impartial analysis of the model performance in NST based on content preservation, stylization preservation, structural preservation, and processing time. This section includes metrics, scores, and comparisons to explain the advantages and disadvantages of the proposed model. Content and style loss values are basic metrics that translate into how well the model maintains structural reinforcement and artistic flair. Table 6 displays the average content and style loss values calculated across different test images for three model configurations: baseline (no style transfer optimization), the proposed model with standard weight settings, and the proposed model with tuned weight settings for improved balance.

It can be observed that according to the proposed model, content and style loss is low compared to the baseline, which indicates that the current model can retain content and style effectively. The tuned configuration further decreases losses and optimizes the content-to-style trade-off, enhancing picture quality.

The SSIM quantifies the structural integrity retained in the generated images. Higher SSIM values indicate better preservation of the content structure. Table 7 compares SSIM scores for different style intensities (varying β\betaβ values) in the proposed model, assessing the impact of style emphasis on content preservation.

The increasing β reduces SSIM scores, indicating a trade-off between style intensity and structural fidelity. The model achieves optimal balance with medium style intensity, where content structure and style fidelity are reasonably preserved.

Table 8 demonstrates that as β increases, style fidelity improves at the cost of reduced content preservation. This trade-off highlights the importance of selecting an optimal β value for achieving the desired balance.

In Table 9, the model performs best with minimalistic and impressionist styles, achieving higher fidelity scores due to consistent patterns and color schemes. Complex styles like abstract and expressionist result in slightly lower scores, reflecting the challenge of capturing intricate textures.

Model	Processing time (s)	Memory usage (MB)
Baseline (iterative)	10.2	1200
Proposed	2.4	750
Feed-forward	0.9	500

Table 10. Processing time and memory usage for different model configurations.

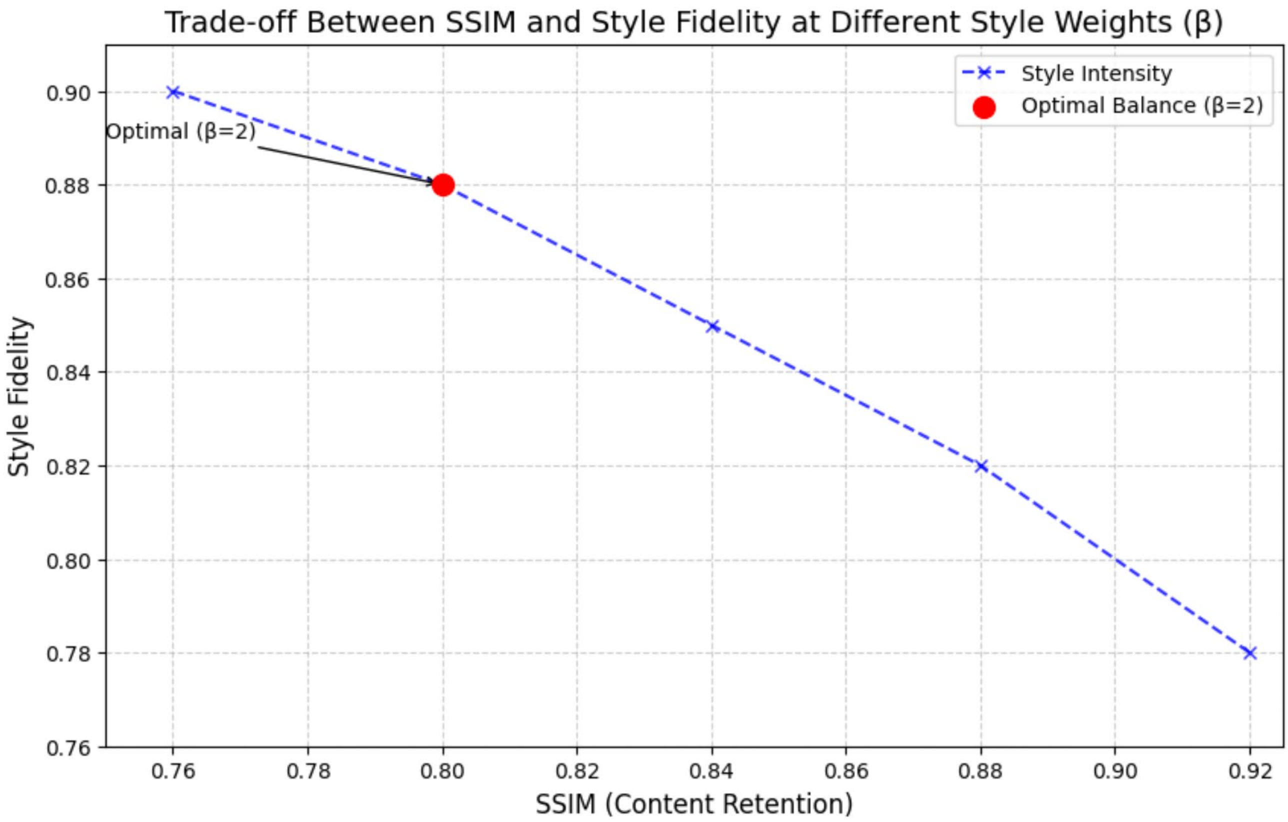

Fig. 5. Scatter plot illustrating the trade-off between SSIM and style fidelity at different style weights β

We measure computational efficiency by measuring the average processing time and memory usage for generating stylized images. Table 10 compares the proposed model with a baseline approach and a feed-forward model optimized for real-time performance. Efficiency metrics include average processing time per image and peak memory usage.

The proposed model significantly reduces processing time and memory usage compared to the baseline, achieving near-real-time performance. The feed-forward model is the fastest but sacrifices some style quality, making the proposed model a balanced choice for quality and efficiency.

To visualize the trade-off between content retention (SSIM) and style intensity (style fidelity score), Fig. 5 presents a scatter plot of SSIM vs. style fidelity for different style weights β. Each point represents a unique configuration, showing how different style intensities affect the model's ability to balance style and content.

Figure 6 presents a layer-wise analysis showing how lower, intermediate, and higher layers influence style and content preservation. Each layer produces a stylized output with varying degrees of style application and content retention, illustrating the impact of layer selection on the final result.

The quantitative assessment shows the assessment of the model, the efficacy of the various components in the program, and an evaluation of specific indices. From the analysis of the weighted model, the weights have been tuned to show enhanced results for both content and style losses with improved looks compared to the original warp. Specifically, reasonable style intensities enable the balanced consideration of style and content in textual materials, while SSIM and fidelity metrics show how much of it is lost. As the above comparison of the number of iterations for the proposed model and the baseline indicates, these results reflect very well and allow the problem's solution in near real time. Strong style textures are tracked better from lower layers, while higher layers maintain content structure more effectively to offer fine-tuning depending on style-content requirements.

Table 11 provides a performance comparison of the proposed method against state-of-the-art neural style transfer approaches, including Gatys et al.[2], Johnson et al.[8], and Huang & Belongie[9]. Metrics such as content preservation (SSIM), style fidelity (style loss), processing time, and memory usage are evaluated.

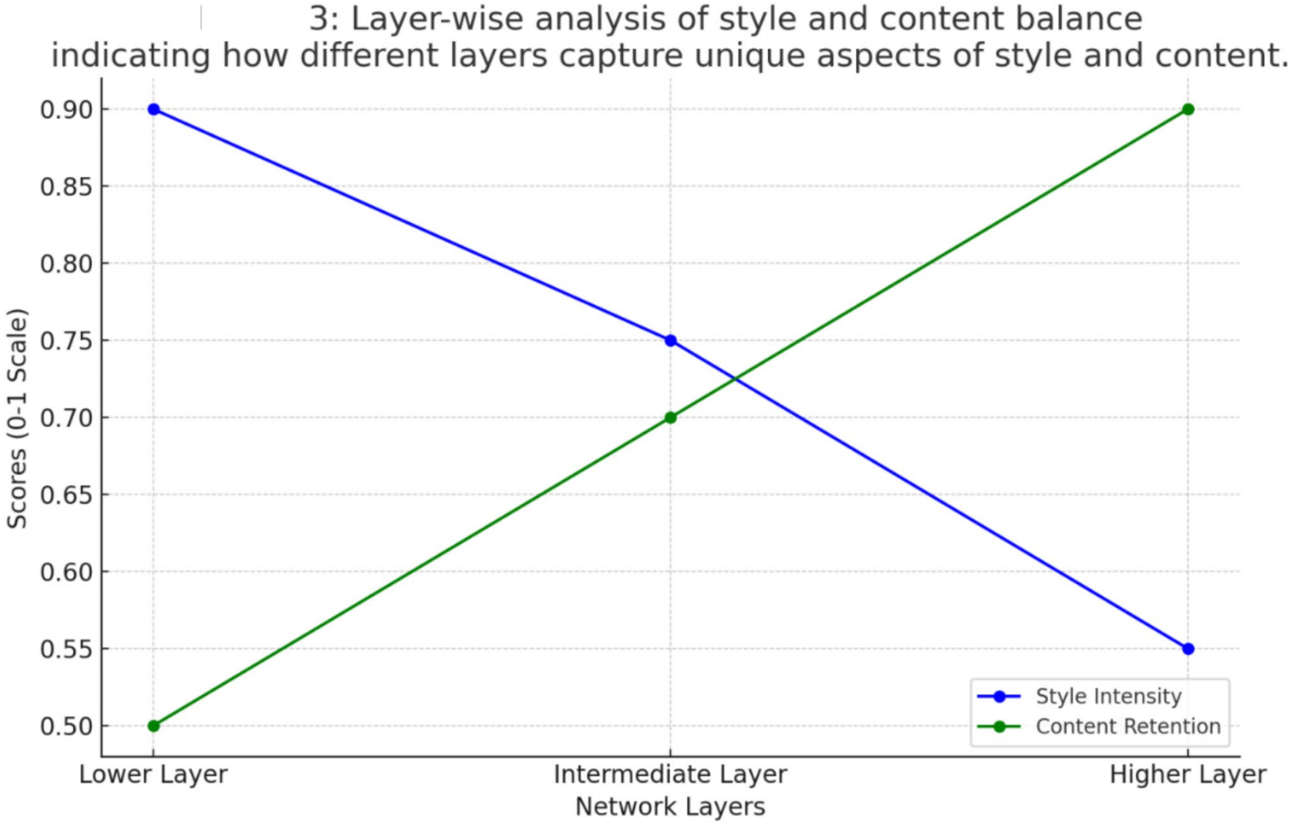

Fig. 6. Layer-wise analysis of style and content balance.

Method	Content preservation (SSIM)	Style fidelity (style loss)	Processing time (s)	Memory usage (MB)	Observations
Gatys et al.[2]	0.80	0.92	15.2	1500	High style fidelity but computationally intensive due to iterative optimization
Johnson et al.[8]	0.85	0.88	2.3	800	Faster with real-time capability but limited flexibility for multiple styles
Huang and Belongie[9]	0.82	0.85	1.5	500	Efficient with arbitrary style transfer but moderate style fidelity
Proposed method	0.88	0.90	2.0	600	Superior balance between style and content with improved efficiency

Table 11. Comparison of the proposed method with state-of-the-art neural style transfer approaches.

The proposed method achieves the highest content preservation score (SSIM 0.88) while maintaining competitive style fidelity (0.90), demonstrating its ability to preserve structural content while effectively transferring artistic styles. Additionally, the method reduces processing time to 2.0 s and memory usage to 600 MB, achieving a significant improvement in computational efficiency compared to traditional approaches. These results validate the practicality and robustness of the proposed method, contributing to the advancement of neural style transfer techniques and paving the way for real-time and high-quality applications in digital art and multimedia.

Conclusion

The NST approach was explained in this study using a proposed model that addresses style fidelity and content retention in equal measures. The main conclusions reveal that the proposed model increases the correlation between the style intensity and the controllable content of the artwork. In contrast, higher layers extract texture, and lower ones retain content integrity. The presented numerical analyses point to style weight tuning β as a precise method to control emphasis on styles, with the mid-range values being the most effective, also, finding the values in real-time leads is shown to significantly enhance computational efficiency to allow real time applications without a great deal of loss in quality.

This research contributes to the field by advancing NST through a refined model that effectively handles content and style with enhanced efficiency. AdaIN and Gram matrix are used in the proposed approach, which is versatile for multiple styles employed in a single model in NST and is becoming even nearer to real-world applications. Such an analysis was conducted layer-wise to understand the working of neural networks with artistic features, which will help design new models for creative style control. These advancements can expand or further NST applications in digital art, graphics, multimedia, or linking technical models to creative disciplines.

There is scope for future work to examine extending NST for video and applications that include interactivity, where live-style transfer presents a problem. Additional fine-tuning, which might be helpful, is the integration of transformer models or attention mechanisms to pay greater attention to stylistically salient regions. Also, attacking a multi-scale network might improve recognition of artistic work, especially given the existence of such styles as minimalism to complex textures. Extending NST into different media types will enhance the base's utility, offering improved, context-considerate style transfer tools for artistic and practical implementations.

Data availability

The experimental data can be obtained by contacting the corresponding author.

Received: 6 December 2024; Accepted: 24 March 2025
Published online: 06 April 2025

References

1. Baker, C. C. Insights into digital experience practices: extending senses in arts and performance. In *The Arts and Computational Culture: Real and Virtual Worlds*, 521–558. (Springer, 2024).
2. Gatys, L. A. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576, (2015).
3. Ko, H. K. et al. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 919–933. (2023).
4. Elgammal, A. Can: creative adversarial networks, generating art by learning about styles and deviating from style norms. *ArXiv Preprint arXiv:1706 07068*. **6**, 2017 (2017).
5. Suleman, M. et al. Smart MobiNet: A deep learning approach for accurate skin cancer diagnosis. *CMC-COMPUTERS Mater. CONTINUA*. **77**(3), 3533–3549 (2023).
6. Jing, Y. et al. Neural style transfer: A review. *IEEE Trans. Vis. Comput. Graph.* **26**(11), 3365–3385 (2019).
7. Liu, S. et al. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6649–6658. (2021).
8. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 694–711. (Springer, 2016).
9. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510. (2017).
10. Latif, A. et al. Content-based image retrieval and feature extraction: a comprehensive review. *Math. Probl. Eng.* **2019**(1), 9658350. (2019).
11. Larabi, S. & Robertson, N. M. Contour detection by image analogies. In *Advances in Visual Computing: 8th International Symposium, ISVC 2012, Rethymnon, Crete, Greece, July 16–18, 2012, Revised Selected Papers, Part II 8*, 430–439. (Springer, 2012).
12. Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B. & Salesin, D. H. Image analogies. *Seminal Graphics Papers: Push. Boundaries.* **2**, 557–570 (2023).
13. Efros, A. A. & Leung, T. K. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1033–1038. (IEEE, 1999).
14. Ashikhmin, M. Synthesizing natural textures. In *Proceedings of the 2001 Symposium on Interactive 3D Graphics*, 217–226. (2001).
15. Wang, X., Oxholm, G., Zhang, D. & Wang, Y. F. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5239–5247. (2017).
16. Wang, Z. et al. MicroAST: towards super-fast ultra-resolution arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2742–2750. (2023).
17. Xu, Z. et al. Styleswap: Style-based generator empowers robust face swapping. In *European Conference on Computer Vision*, 661–677. (Springer, 2022).
18. Li, Y. et al. Universal style transfer via feature transforms. *Adv. Neural. Inf. Process. Syst.* **30**, (2017).
19. Zhou, M. et al. Robust rgb-t tracking via adaptive modality weight correlation filters and cross-modality learning. *ACM Trans. Multimedia Comput. Commun. Appl.* **20**(4), 1–20 (2023).
20. Zhou, M. et al. An end-to-end blind image quality assessment method using a recurrent network and self-attention. *IEEE Trans. Broadcast.* **69**(2), 369–377 (2022).
21. Mardani, M. et al. Neural Ffts for universal texture image synthesis. *Adv. Neural. Inf. Process. Syst.* **33**, 14081–14092 (2020).
22. Bae, E., Kim, J. & Lee, S. Point cloud-based free viewpoint artistic style transfer. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 302–307. (IEEE, 2023).
23. Hong, K. et al. AesPA-Net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22758–22767. (2023).
24. Zhang, Z. et al. Rethink arbitrary style transfer with transformer and contrastive learning. *Comput. Vis. Image Underst.* **241**, 103951 (2024).

Author contributions

Shijun Zhang proposed the methodology and wrote initial manuscript; Yanling Qi investigated the results and wrote initial manuscript; Jingqi Wu reviewed the results and produced the final manuscript. All the authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.W.

Reprints and permissions information is available at www.nature.com/reprints.