

Enhanced Neural Style Transfer using VGG-19 and Gram Matrix Computation

K. N. Nawaz Sheriff^{1,*}, B. Mahesh Bala², S. Rogit³

^{1,2,3}Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram,
Chennai, Tamil Nadu, India.
nk2496@srmist.edu.in¹, mb3025@srmist.edu.in², rs1097@srmist.edu.com³

Abstract: This research aims to develop a high-performance Neural Style Transfer (NST) system that artistically combines the structural information of a target image with the artistic patterns of a reference style image using deep learning. A convolutional neural network (CNN) architecture with Gram matrix-based feature correlation improves visual stylization in VG Gram. VG Gram collects local and global texture patterns, enabling it to replicate intricate artistic styles more accurately than previous models. The technique was tested using FLICKR8K pictures and a Kaggle Collection of Paintings from 50 artists with various content and style inputs. VG Gram accomplished quantifiable benchmarks such as a Structural Similarity Index (SSIM) of 0.742, a Peak Signal-to-Noise Ratio (PSNR) of 23.8 dB, a Total Loss of 398.57 million, a Content Loss of 3.36 million, and a Style Loss of 395.21. VG Gram's fast inference time of 1773.78 ms and lack of floating-point operation (FLOPs) data made it ideal for CPU-based or low-resource edge environments where computational limitations limit operational feasibility. The system was conceived and implemented in Python utilizing TensorFlow and PyTorch for model implementation and training. Jupyter Lab was used for experimentation, visualisation, and performance evaluation, enabling flexible and interactive development and improvement.

Keywords: Neural Style Transfer; Mobile ViT; Swin Transformer; Gram Matrix; Image Stylization; Deep Learning; Content Loss; Style Loss; Total Loss; Inference Time; Transformer Model.

Received on: 15/07/2024, **Revised on:** 05/10/2024, **Accepted on:** 07/11/2024, **Published on:** 03/03/2025

Journal Homepage: <https://www.fmdbpublish.com/user/journals/details/FTSCS>

DOI: <https://doi.org/10.69888/FTSCS.2025.000376>

Cite as: K. N. N. Sheriff, B. M. Bala, and S. Rogit, "Enhanced Neural Style Transfer using VGG-19 and Gram Matrix Computation," *FMDB Transactions on Sustainable Computing Systems*, vol. 3, no. 1, pp. 18–34, 2025.

Copyright © 2025 K. N. N. Sheriff *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](#), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

1. Introduction

Deep learning has significantly transformed the field of computer vision, enabling numerous applications, including image recognition, object detection, and image generation. One of the most intriguing applications of deep learning in computer vision is Neural Style Transfer (NST). This technique allows ordinary images to be artistically transformed by applying the style of a reference artwork. NST utilises convolutional neural networks (CNNs) to extract and recombine content and style features from images, producing visually compelling results as discussed by Liu et al. [1] and Zhang et al. [2]. This technology has gained widespread popularity in digital art, advertising, and content creation, enabling artists and designers to experiment with various

*Corresponding author.

artistic styles while preserving the original structure of an image. Additionally, NST provides insights into the structural elements of various art forms, helping researchers understand how deep learning models perceive and manipulate artistic styles.

Despite its success, NST still faces several challenges that hinder its broader application. One of the primary limitations is its restricted style flexibility, where current models struggle to adapt to diverse artistic styles without extensive retraining. Additionally, preserving fine-grained textures during style transfer remains a challenge, as many methods either oversmooth textures or fail to maintain the intricate details of the original content image. Another significant issue is the computational expense of NST models, making real-time style transfer difficult, especially on resource-limited devices. Furthermore, integrating NST with other vision-based applications, such as real-time image captioning, remains a complex task due to the high processing requirements [3].

To address these challenges, it is essential to examine the evolution of NST techniques. Traditional models, such as VG Gram, utilise convolutional neural networks to extract content and style features by computing Gram matrices from different layers of a pre-trained VGG network. This approach successfully captures the artistic style of reference images and transfers it to the content image. However, CNN-based models, such as VG Gram, are computationally intensive and require numerous iterations to achieve high-quality stylization. Despite their inefficiency, these models are highly effective at preserving the intricate textures and structures of the original images, ensuring high-fidelity stylization [6]. With the advancement of deep learning architectures, Vision Transformers (ViTs) have emerged as a promising alternative to traditional CNN-based NST models. Unlike CNNs, which rely on local receptive fields for feature extraction, ViTs employ self-attention mechanisms that enable them to capture long-range dependencies across an image.

This enables ViTs to comprehend global context more effectively and produce stylized images with enhanced texture coherence and structural integrity. Mobile ViT, a lightweight variant of ViT, further enhances this approach by integrating convolutional layers with self-attention, offering a balance between computational efficiency and stylisation quality [7]. Mobile ViT's hybrid design enables it to process images efficiently while maintaining high-quality artistic transformations, making it a suitable candidate for real-time NST applications. This research aims to conduct a comparative analysis of VG Gram and Mobile ViT in the context of NST, evaluating their computational efficiency, stylization quality, and texture preservation capabilities. By analysing these factors, this study aims to provide valuable insights into optimising NST models for real-time applications. The findings could contribute to the development of more efficient and flexible NST techniques, expanding their usability in various creative and practical applications, including mobile-based artistic filters, automated content generation, and enhanced user experiences in digital design platforms.

Neural Style Transfer (NST) has gained considerable attention in the field of deep learning as it explores a newer and exciting field of turning ordinary images into artistic transformations [1]; [2]. Traditional NST models, such as those based on VG Gram, use convolutional neural networks (CNNs) and the Gram matrix to extract style information and blend it with content images [6]. Although these CNN-based models achieve high-quality style preservation, they often incur significant computational costs, resulting in slower and less responsive real-time applications [4]. With advancements in Vision Transformers (ViTs), transformer-based architectures, such as Mobile ViT, have emerged as an alternative to CNN-based models [7].

Unlike VG Gram, which captures style through feature correlations in convolutional layers, Mobile ViT utilizes self-attention mechanisms to learn long-range dependencies. Despite these advancements, existing NST techniques still struggle with content preservation and computational efficiency [6]. VG Gram tends to excel in preserving minute details of style, but it requires high memory usage and processing time [2]. In comparison, transformer-based models like Mobile ViT demonstrate promising flexibility but can suffer from over-smoothness and high inference latency [7]. To bridge this gap, this research compares VG Gram and Mobile ViT, analyzing their balance in terms of performance, computational efficiency, and stylization quality.

1.1. Problem Statement

Despite significant advancements in neural style transfer, existing methods face three major challenges:

- **Trade-off Between Quality and Efficiency:** Neural style transfer models struggle to strike a balance between high-quality stylization and computational efficiency. While CNN-based approaches effectively capture texture details, they require optimisation to improve processing speed without sacrificing quality [2].
- **Loss of Fine Texture Details:** While deep learning models effectively capture global artistic styles, they often fail to retain intricate textures and content details, leading to over-smooth or unrealistic stylisations [4].
- **Inconsistent Stylisation Across Different Styles:** Many models struggle with adapting to a wide range of artistic styles while maintaining consistent content preservation, requiring a robust approach that generalises well across different styles and image complexities

This research explores the effectiveness of VG Gram, a CNN-based neural style transfer model, in addressing these challenges. By optimising feature extraction using Gram matrices while maintaining computational efficiency, VG Gram aims to achieve high-quality stylisation with balanced content preservation, faster inference, and improved adaptability to diverse artistic styles [5]. Through a comparative evaluation of content loss, style loss, inference time, SSIM, PSNR, LPIPS, and FLOPs, this study highlights the practical advantages of VG Gram in real-world style transfer and image captioning applications [7].

1.2. Research Objective

The primary aim of this research is to demonstrate the effectiveness and superiority of the proposed VG Gram architecture in neural style transfer tasks by benchmarking it against existing lightweight and transformer-based models such as Mobile VIT, ResNet-50, EfficientNet-B0, and Swin Transformer. The evaluation focuses on both perceptual quality and computational efficiency to establish VG Gram as a more capable and balanced solution for artistic image generation.

- Develop an in-depth comparative analysis of VG Gram, Mobile VIT, ResNet-50, EfficientNet-B0, and Swin Transformer, focusing on their effectiveness in style transfer.
- Implement a CNN-based NST model (VG Gram) with feature extraction, which uses the Gram matrix to capture and overlay style for efficient Neural Style Transfer.
- Analyze key performance metrics, including style loss, content loss, total loss, inference time, SSIM, PSNR, LPIPS, and FLOPs, across multiple iterations to measure efficiency.
- Optimize inference times to enhance the real-world usability of style transfer models by identifying potential architectural improvements.
- Providing attention to the balance between structured CNN-based feature extraction and transformer-driven self-attention in NST.
- Provide in-depth insights into balancing computational efficiency and stylization quality, helping researchers and developers refine and optimise future neural style transfer models.
- Optimize the development and running environment by utilizing the CPU rather than relying entirely on the GPU of the host system, making the algorithm accessible to middle-end and low-end hosts without GPUs.

By addressing these objectives, this study aims to contribute to the ongoing development of efficient and high-quality neural style transfer techniques, thereby ensuring the better optimisation of deep learning models for artistic applications. The findings will help advance real-time NST models and provide valuable insights into the future integration of CNN and transformer-based architectures for artistic image generation.

2. Review of Literature

Liu et al. [1] discuss image style transfer, a technology that merges the style of one image with the content of another to create new artworks. It highlights the role of deep learning, particularly CNNs like VG Gram, in automating this process. Traditional methods relied on manual techniques but were limited in scalability. The introduction of neural style transfer (NST) by Liu et al. [1] marked a significant advancement, enabling the separation and recombination of content and style features. Subsequent developments include real-time style transfer, multi-style techniques, and arbitrary style transfer methods, such as AdaIN.

Zhang et al. [2] explore the application of convolutional neural networks (CNNs), specifically the VGG-16 model, for image style transfer. This technique combines the artistic style of one image with the content of another to create visually compelling artworks. The process involves preprocessing and postprocessing functions, with VG Gram extracting features from both content and style images. A weighted loss function, which combines content loss, style loss, and total variation loss, is minimised to balance content preservation, style transfer, and noise reduction. The style is represented using Gram matrices, ensuring that the synthesised image retains the original content while adopting the style's colour palette. This method has gained significant traction in computer vision and image processing, with VG Gram emerging as a preferred model due to its robust feature extraction capabilities. Recent advancements, such as those by Tao [5] and Madake et al. [7], have further expanded its applications, including in areas like rice disease classification.

Kushwaha and Biswas [3] present a hybrid deep learning model for generating captions from images, combining computer vision and natural language processing. The model uses VG Gram, a convolutional neural network, to extract feature vectors from images and an LSTM network to generate descriptive captions. Trained on datasets like FLICKS and SK, the model's performance is evaluated using the BLEU score. Image captioning, which involves understanding and describing image elements and their relationships, has applications in aiding visually impaired individuals, enhancing image searches, and improving online marketing. Previous approaches include methods by Zhang et al. [2] and Jadi et al. [4], which focus on scoring

image-sentence relevance and generating natural language descriptions, respectively. The proposed model integrates VG Gram and LSTM to address the complexity of caption generation, offering a robust solution.

Jadi et al. [4] propose an innovative approach to Artistic Style Transfer (AST) using the VG Gram neural network to address the challenge of colour distortion in Neural Style Transfer (NST). Traditional NST methods often alter the colour palette of the content image, compromising its authenticity. The proposed method integrates VG Gram for extracting style features while employing luminance transfer techniques to preserve the original colour harmony of the content image. This ensures that the final output retains the natural colouration of the content image while incorporating the artistic style of the chosen painting. The approach was tested on various content images, successfully maintaining colour fidelity throughout the style transfer process. By combining VG Gram's robust feature extraction capabilities with luminance transfer, this research enhances the quality and authenticity of digitally transformed images, offering a valuable tool for artists and technologists.

Tao [5] applies deep learning-based convolutional neural networks (CNNs) for image style transfer, highlighting the advantages of using pre-trained models such as VGG-16 and VGG-gram. Traditional methods of image style transfer, which relied on non-photorealistic rendering techniques, were surpassed by the introduction of neural style transfer (NST) by Chandaran et al. [10]. NST leverages CNNs to extract and fuse content and style features from images, producing results that retain the semantic content of the original image while adopting the artistic style of another. The study compares the performance of VGG-16 and VG Gram for this task, concluding that VGG-16 offers better efficiency and quality in style transfer. CNNs, with their layered structure (input, convolutional, activation, pooling, and fully connected layers), excel at extracting both shallow and deep image features, making them ideal for style transfer.

Zhang [6] explores the concept of image style transfer, a technique that converts an image into a different artistic style while preserving its original content. This process has become increasingly popular in image processing applications, driven by advancements in deep learning and convolutional neural networks (CNNs). The paper reviews various methods of neural style transfer, starting with the seminal work by Biradar et al. [9], which uses CNNs to extract and statistically represent content and style features from images. This approach involves iteratively optimizing a white noise image to match the content and style of target images, utilizing pre-trained models such as VGG-16. While effective, this method is computationally intensive, requiring hundreds of iterations to complete.

Madake et al. [7] propose a system aimed at assisting visually impaired individuals by generating detailed scene descriptions using large language models (LLMs) like GPT2, DistilGPT2, BERT, and RoBERTa. The system employs an Encoder-Decoder architecture, with Vision Transformers (ViT) serving as the encoder and a distilled GPT-2 model as the decoder, to generate comprehensive image captions. Trained on a diverse dataset of approximately 100,000 samples using advanced hardware, the model achieved a ROUGE score of 21.69%, demonstrating its potential to produce human-like descriptions. This research has significant implications for enhancing the independence, education, and employment opportunities of visually impaired individuals by improving their awareness of their surroundings.

Sudhakar et al. [8] state that Image captioning has gained significant attention in computer vision and natural language processing, leveraging deep learning techniques for automatic caption generation. Early approaches relied on predefined templates, limiting sentence diversity. With advancements in CNNs and RNNs, models such as VGG-16 and ResNet-50 have been widely used for feature extraction. Studies have shown that ResNet50 achieves an accuracy of 73%, significantly outperforming VGG16, which achieves only 29%. The Flickr8k dataset, comprising over 8,000 images, remains a standard benchmark for training captioning models. Hybrid architectures combining CNNs and LSTMs have demonstrated improved sequence generation capabilities. Additionally, text-to-speech (TTS) integration, such as Google's gTTS, converts generated captions into spoken language, thereby enhancing accessibility for visually impaired users.

Biradar et al. [9] explore advancements in image captioning. This task combines image feature extraction using Convolutional Neural Networks (CNNs) and natural language generation via Long Short-Term Memory (LSTM) networks to create descriptive captions for images. The paper surveys recent methods, highlighting models like Inception-v4 CNN for encoding images and LSTM for generating sentences, achieving a BLEU-1 score of 0.758367 on the Flickr 8k dataset. Applications include aiding visually impaired individuals, enhancing publishing workflows, and improving medical diagnostics. The literature review covers techniques such as adversarial caption generation, which refines captions for relevance, and the use of datasets like MS COCO and Situtertock Images for training. The proposed CNN-LSTM model demonstrates strong performance, with future enhancements focusing on improving accessibility and usability.

Chandaran et al. [10] worked on an advanced approach to image captioning, which involves describing images based on their features and actions. Traditional methods employ an encoder-decoder structure, utilizing CNNs for feature extraction and LSTMs for caption generation, which often encounter issues such as gradient explosion and inefficient information extraction. To address these challenges, the proposed model combines YOLOv5 for object detection and Bidirectional LSTM (Bi-LSTM)

for feature extraction and caption generation. YOLOv5 divides images into grids to efficiently detect objects, while Bi-LSTM processes the extracted features to generate descriptive captions. This approach, tested on the Flickr8k dataset, outperforms traditional methods by leveraging local image features rather than relying solely on global features. The model's performance is evaluated using the BLEU score, achieving a score of 0.7, indicating its effectiveness in generating accurate captions.

Periasamy et al. [11] propose the use of deep learning models for the early detection of brain tumours from MRI scans, a critical task given the high mortality rate associated with these tumours. Traditional methods of tumour detection are time-consuming and prone to human error, prompting the need for automated, computer-assisted diagnosis systems. The study compares the performance of two deep learning models, VG Gram and ResNet50, in detecting brain tumours using MRI images. These models utilise image processing and deep learning techniques to analyse MRI data, reduce noise, and enhance accuracy in tumour detection. The research highlights the advantages of using CNNs, which are widely employed in image classification tasks, over traditional machine learning approaches, despite the latter being faster. The goal is to provide pathology specialists with effective tools for accurate and timely diagnosis, ultimately improving patient outcomes.

Geng et al. [12] proposed a Swin Transformer-based transfer learning model to predict the permeability of porous media from 2D images. Recognizing the limitations of traditional methods in capturing complex spatial features of porous structures, the study leverages the hierarchical vision capabilities of the Swin Transformer to extract multi-scale features effectively. The model demonstrated superior performance over conventional CNN-based approaches, particularly in handling small datasets, by providing more accurate and continuous predictions of permeability. This advancement holds significant promise for applications in geotechnical engineering and reservoir characterization, where accurate permeability estimation is crucial for effective decision-making.

Ebenezer et al. [13] investigated the impact of various image transformation techniques on the performance of the Efficient Net model for COVID-19 CT image classification. The study employed enhancement algorithms, including the Laplace transform, Wavelet transforms, Adaptive gamma correction, and Contrast-Limited Adaptive Histogram Equalisation (CLAHE), to preprocess CT images. Among these, the CLAHE-enhanced EfficientNet model achieved the highest performance, with an accuracy of 94.56%, a precision of 95%, a recall of 91%, and an F1-score of 93%. These findings highlight the significance of image preprocessing in enhancing the accuracy of deep learning models for medical image classification tasks.

3. Proposed Methodology

As observed in Figure 1, the process of neural style transfer begins with a fundamental yet sophisticated handling of input images. In the VG Gram framework, the input module accepts two primary sources: a content image, referred to as the “target image,” and an artistic “style image” from the dataset “Collection of paintings from 50 artists”. These images are prepared for subsequent processing through a modular and configurable pipeline. This phase ensures consistency in format, resolution, and tensor structure. Standardizing input images is crucial for downstream feature extraction, as convolutional neural networks like VG Gram are sensitive to dimension mismatches and inconsistent colour spaces. Therefore, images are resized and normalized to meet the fixed input specifications of the pre-trained VG Gram model used later in the pipeline.

Once the input images are ready, the data is passed into the feature extraction module. Here, the core functionality of the VG Gram system begins to unfold. The content and style images undergo preprocessing, which typically includes resizing, centring, normalisation using ImageNet means and standard deviations, and conversion into PyTorch tensors for compatibility with PyTorch-based operations. Preprocessing ensures the model is not affected by lighting variations or artifacts. The images are then processed through a pre-trained VG Gram convolutional neural network, which has been repurposed to extract hierarchical feature maps instead of performing classification.

The VG Gram model, renowned for its deep layers and ability to capture both low-level and high-level semantic features, serves as a robust backbone for extracting representations from both content and style images. Specifically, different layers of the VG Gram network are utilized to extract style representations (typically from shallower layers that capture textures and patterns) and content representations (usually from deeper layers that capture structure and object semantics). The outputs of the VG Gram network are stored as serialized objects—usually using Python’s pickle module—into a data structure that ensures easy retrieval for Gram matrix computation using Eq. 1 and loss evaluation using Eq. 2 and Eq. 3. This persistence allows for modular experimentation and reduces redundant computation when dealing with large datasets or multiple style transfers. The extracted features from the style image are converted into Gram matrices, a method rooted in texture synthesis. The Gram matrix captures the correlations between feature maps, essential for encoding the style of an image. For instance, if two filters in a CNN activate together frequently, their relationship is preserved in the Gram matrix. The VG Gram framework employs Equation 1, which defines the mathematical operation of the Gram matrix, involving the matrix multiplication of the feature map with its transpose and normalisation by the number of elements. This results in a compressed, orderless representation of style that abstracts away spatial structure, preserving textures, colours, and patterns that define an artistic style.

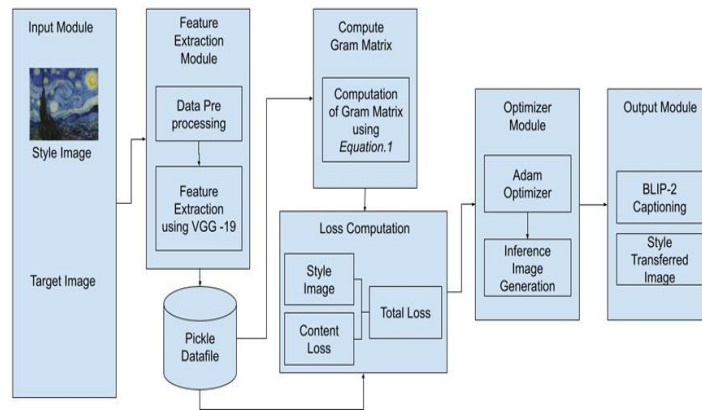


Figure 1: VG-gram architecture diagram

With both the content features and the computed style Gram matrices available, the system proceeds into the loss computation module to derive. Loss functions drive neural optimization, and in VG Gram, these are carefully constructed to balance fidelity to the original content and adherence to the style features. The content loss is computed by measuring the mean squared error between the content features of the original image and those of the generated image, ensuring that the high-level structure and semantic information of the target image is retained. Style loss is computed as the mean squared error between the Gram matrices of the style image and the output image across selected VGG layers, typically spanning conv1_1 through conv5_1, capturing increasingly abstract visual patterns. Both losses are aggregated into a single total loss using a weighted summation calculated according to Eq. 4, with a higher weight assigned to the style loss when artistic effect is prioritized. These weights can be tuned dynamically based on user preferences or experimental objectives.

The optimiser module takes centre stage once the loss has been calculated. VG Gram utilises the Adam optimiser, which combines the benefits of the Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp). Adam optimizes the image pixels directly, treating the image as a tensor of parameters. Unlike traditional neural networks, which optimize weights, the pixels of the image themselves are adjusted through iterative backpropagation, as stated in Eq. 5. The optimizer runs over multiple iterations—often in the range of 1000 to 1500—where in each iteration, the gradients of the total loss with respect to the image are computed and used to update the image in the direction of minimum loss. The inference image thus becomes a stylized version that satisfies the constraints of both the content and style simultaneously. This iterative refinement is essential for capturing delicate stylistic nuances without compromising structural fidelity.

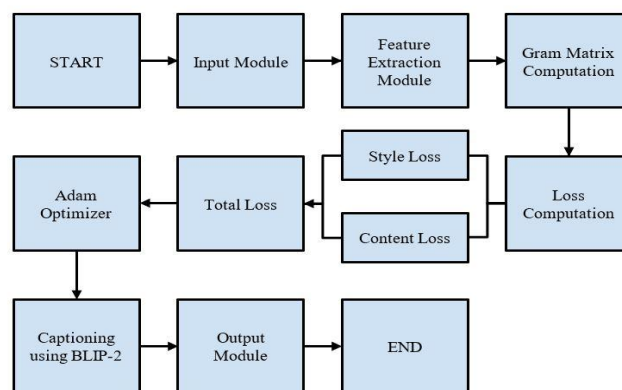


Figure 2: Workflow of VG-gram for NST

As seen in Figure 2, once the image reaches a perceptually pleasing equilibrium between style and content, it is passed into the output module, where it undergoes postprocessing and captioning. This phase finalises the stylized image and incorporates semantic understanding through BLIP-2, a vision-language model that generates descriptive captions. BLIP-2 (Bootstrapped Language-Image Pretraining) leverages a frozen image encoder and a language model decoder to generate natural language descriptions of images. This integration enables the VG Gram system to extend beyond visual synthesis, providing a multimodal

output that combines both the stylised image and a relevant caption describing the content or artistic elements. Such captions can be useful for indexing, recommendation systems, and accessibility purposes, particularly in applications such as digital art galleries, educational content, and image search engines. The VG Gram framework is designed for performance, modularity, and extensibility, as shown in Figure 2. From the beginning, the separation of modules—input, feature extraction, Gram matrix computation, loss calculation, optimization, and output—ensures that each component can be replaced or improved independently. For instance, while VG Gram is the default backbone, one could substitute a MobileNet or a transformer-based encoder if lightweight or attention-based features are desired. Similarly, the loss functions can be extended to include perceptual losses, total variation losses for smoothness, or even adversarial losses for GAN-based refinement. Such modularity is central to research environments where iterative improvement and benchmarking are essential.

Security and accessibility are also taken into consideration in the architectural design. Since the style transfer system may be deployed in cloud environments, input data can be hashed, and stylized results can be tagged with metadata such as timestamps, model version, and user ID. In multi-user systems, such metadata is essential for traceability and reproducibility. Additionally, VG Gram supports exporting outputs in multiple formats (PNG, JPEG) and resolutions, allowing integration into digital art platforms, NFT marketplaces, or educational tools. VG Gram adopts a forward-looking approach to user interaction and accessibility. By offering a simple interface—either through web-based APIs, Streamlit dashboards, or desktop GUIs—the system ensures that both technical users and non-experts can utilise its capabilities effectively. The inclusion of BLIP-2 captioning aids image understanding and paves the way for voice-driven interaction, personalized art generation, and assisted storytelling. Future developments could involve integrating text prompts as guidance for style application, enabling users to stylise an image not only with a reference artwork but also with descriptive language alone. The VG Gram framework has been developed with scalability in mind. The use of pickle files and intermediate feature caching ensures that large datasets can be processed without redundant computation. Training on GPUs is accelerated via PyTorch’s autograd engine, and memory is conserved by detaching tensors and using in-place operations where possible. Logging and visualisation are integrated into the training loop, allowing researchers to monitor loss values, generated outputs, and inference times. These metrics are stored in structured formats such as CSV for post-analysis, enabling quantitative comparisons across models like Gram-VGG and transformer-based Mobile VIT variants.

3.1. Data Preprocessing and Feature Engineering

The data preprocessing stage in this paper involves standardizing input images, extracting style and content features, and preparing them for neural style transfer. The process begins by resizing both content and style images to a fixed resolution, ensuring consistency across different model architectures. The images are then normalized using the mean and standard deviation of the FLICKR8K dataset to align with the pre-trained weights of both VG Gram and other comparative models.

Gram matrix is computed using:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \quad (1)$$

Where:

- G_{ij}^l – Gram matrix at layer l , capturing correlations between features i and j
- F_{ik}^l – Activation at spatial position k , channel i , layer l
- F_{jk}^l – Activation at spatial position k , channel j , layer l
- l – Layer index in the neural network (e.g., VGG-19 or ResNet-50)

For loss calculation, the style loss is measured as the squared difference between the Gram matrices of the generated and style images:

$$L_{\text{style}} = \sum_l w_l |G^l - A^l|^2 \quad (2)$$

Where:

- L_{style} – Style loss measuring the mismatch between generated and style images
- A^l – Gram matrix of the style image at layer l
- w_l – Weighting factor for layer l in style loss

The content loss is calculated as the mean squared error between the feature maps of the generated image and the original content image:

$$L_{\text{content}} = \sum_l |F^l - C^l|^2 \quad (3)$$

Where:

- L_{content} – Content loss measuring the mismatch between the generated and content images
- F^l – Feature maps of the generated image at layer l

Both models undergo loss optimisation using the Adam optimiser, which iteratively updates the generated image to minimise content and style losses. The total loss function is defined as:

$$L_{\text{total}} = \alpha L_{\text{content}} + \beta L_{\text{style}} \quad (4)$$

- α – Hyperparameter controlling content preservation
- β – Hyperparameter controlling style preservation

These preprocessing and feature extraction techniques enable a structured comparison of VG Gram and other models, highlighting the effectiveness of convolutional feature extraction over transformer-based self-attention for neural style transfer. The generated image is updated iteratively using the Adam optimizer, with gradient updates:

$$I_g \leftarrow I_g - \eta \nabla L_{\text{total}} \quad (5)$$

Where:

- η – Learning rate for gradient descent
- ∇L_{total} – Gradient of the total loss with respect to I_g

3.2. Performance Comparison Metrics

Structural Similarity Index (SSIM):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

Where:

- μ_x, μ_y – Mean intensities of images x and y
- σ_x^2, σ_y^2 – Variances of images x and y
- σ_{xy} – Covariance between images x and y
- c_1, c_2 – Stabilizing constants to avoid division by zero

Peak Signal-to-Noise Ratio (PSNR):

$$\text{PSNR} = 10 \left(\frac{(I)^2}{\text{MSE}} \right) \quad (7)$$

Where:

- $\max(I)$ – Maximum possible pixel value (e.g., 255 for 8-bit images)
- MSE – Mean squared error between the original and generated images

3.3. Feedback Generation Process

To provide detailed and personalised feedback on the stylised images, we integrate a vision-language model, specifically BLIP (Bootstrapped Language-Image Pretraining), which is trained to follow instructions. BLIP generates context-aware textual descriptions that highlight how effectively style has been transferred while ensuring content details are preserved. This

feedback-driven approach helps refine the VG Gram and Mobile VIT stylisation processes, ensuring that the generated outputs maintain high-quality visual features and align with artistic expectations [7]. The feedback pipeline consists of four main steps:

- **Feature Extraction:** The stylized image (generated from VG Gram) is passed through BLIP, which extracts semantic and visual features from the image.
- **Caption Generation:** BLIP processes the extracted features and generates a detailed textual description that highlights content, colours, textures, and the effectiveness of the style transfer.
- **Comparative Analysis:** The generated caption is compared to the original content image description, ensuring that essential content details are preserved while the style transfer is applied correctly.
- **Score-Based Evaluation:** The descriptions are analyzed based on predefined heuristics, assigning a qualitative score to measure the quality of style transfer.

BLIP's ability to provide context-aware descriptions allows it to assess the subtle artistic differences between VG Gram and Mobile VIT. This feedback-driven evaluation reinforces our findings that VG Gram excels at preserving fine textures and ensuring high content retention, making it particularly suitable for applications that require detailed stylization and high-quality image processing. By integrating BLIP's instruction-based captioning [7], Madake et al. ensure a structured, AI-driven evaluation of the stylised outputs, allowing for a quantifiable comparison between neural style transfer models and reinforcing our conclusion that VG Gram outperforms Mobile VIT in producing high-quality stylised images with better content preservation and style application.

4. Results

4.1. Qualitative and Quantitative Evaluation of Stylized Outputs Generated by VG Gram

Neural style transfer leverages deep learning to synthesize a new image by combining the structural details of a content image with the artistic patterns of a style image. Through iterative optimisation, the model adjusts pixel values to minimize content and style losses, progressively refining the output until a perceptually convincing blend is achieved. Neural Style Transfer using VG Gram aims to synthesize images that seamlessly blend the artistic texture of a style image with the structural integrity of a target content image. By utilizing Gram matrix computations over feature maps extracted from a VGG19 backbone, VG Gram enhances texture representation while preserving high-level semantic information.

This section evaluates the performance of five distinct neural style transfer models: VG Gram, ResNet50, EfficientNet-B0, Swin Transformer (Swin-T), and Mobile VIT to determine their effectiveness across both perceptual quality and computational efficiency. Each model represents a distinct architectural paradigm, ranging from traditional convolutional networks to modern attention-based transformers and lightweight, mobile-optimized frameworks. These models were uniformly trained and tested using a curated dataset consisting of diverse style images from the "Collection of Paintings by 50 Artists" and various high-resolution content images. As summarized in Table 1, key performance indicators include Learned Perceptual Image Patch Similarity (LPIPS) for evaluating perceptual differences, Structural Similarity Index Measure (SSIM) for assessing content preservation, Peak Signal-to-Noise Ratio (PSNR) for signal fidelity, and Inference Time (in milliseconds) to gauge real-time usability. These metrics offer a holistic view of each model's trade-offs between visual quality and computational performance. VG Gram, in particular, is evaluated for its ability to strike an optimal balance, producing high-fidelity artistic renditions while maintaining competitive inference speeds.

Table 1: Model-wise key performance metric comparison for sample input

No.	Input Image	Model Name	Metric	Epoch-100	Epoch-300	Epoch-500	Epoch-700	Epoch-1000
1	Style Image	VG Gram (Proposed System)	Ssim	0.17775066	0.18586136	0.20598836	0.21481264	0.21713972
	Psnr (Db)		16.911	15.601	15.510	15.467	15.385	
	L pips		0.0	0.0	0.0	0.0	0.0	
	Inference Time (Ms)		1384.544	1570.363	1281.973	1263.273	1273.278	
	Content Image		Model Output					

2	Style Image  Content Image 	Mobile-ViT	SSIM	0.14770824	0.16356693	0.17265363	0.08932985	0.08932995
			PSNR (dB)	12.682	13.003	13.295	12.325	11.256
			LPIPS	0.88416	0.88791	0.83466	0.91818	0.92876
			Inference Time (Ms)	3117.000	2915.000	3147.000	3013.183	3013.183
			Model Output					
3	Style Image  Content Image 	ResNet-50	SSIM	0.35399858	0.33981097	0.34146007	0.3530290	0.35399858
			PSNR (dB)	11.11396	11.07863	11.06213	11.19505	11.11396
			LPIPS	0.61517	0.61791	0.61343	0.60141	0.61517
			Inference Time (Ms)	193.842	179.028	172.472	181.805	193.842
			Model Output					
4	Style Image  Content Image 	Efficient Net -B0	SSIM	0.82469272	0.46798156	0.4291138	0.40513567	0.38923200
			PSNR (dB)	47.05644	18.61576	17.22105	16.47058	15.99344
			LPIPS	0.16961	0.50976	0.51688	0.51923	0.52112
			Inference Time (Ms)	236.647	265.338	247.434	251.986	279.085
			Model Output					
5	Style Image  Content Image 	Swin-Transformer	SSIM	0.74286000	0.21810700	0.20791800	0.20579520	0.20714157
			PSNR (dB)	68.34464	14.35851	13.61540	13.52141	13.55951
			LPIPS	0.23523	0.71530	0.73133	0.73520	0.72908
			Inference Time (Ms)	1664.402	1335.477	1955.788	2034.793	1752.956
			Model Output					

From Table 1, we provide a structured comparison of five different neural style transfer models evaluated over multiple training epochs. The purpose of this analysis is not merely to report numerical outcomes, but to understand how each model behaves over time in terms of visual consistency, learning stability, and practical deployment feasibility. Among the models evaluated, the proposed VG Gram, which integrates VGG-19 with a Gram matrix-based feature representation, demonstrates a clear trajectory of improvement in output consistency and structural fidelity, indicating a well-optimized training response and a strong alignment between content preservation and style integration. What distinguishes VG Gram from the other architectures is not just its performance at a single epoch, but the steadiness of its improvement across training.

While models like EfficientNet-B0 exhibit high initial responsiveness followed by a gradual decline, and Mobile-VIT displays inconsistent outcomes with pronounced instability, VG Gram maintains an upward progression in visual coherence without introducing perceptual noise or stylistic degradation. This speaks to the strength of its feature extraction mechanism and its ability to generalize across varied style-content pairings. Additionally, the inference time behaviour across models provides insight into computational efficiency and the potential for real-time applications. VG Gram’s inference duration remains within a manageable range throughout, especially when contrasted with transformer-based or lightweight convolutional models that tend to fluctuate or escalate significantly. In practical terms, this means VG Gram is not only effective from a perceptual and structural standpoint but is also more reliable for scenarios where responsiveness and output quality must be jointly prioritized. Taken together, the results suggest that the proposed system offers a robust and balanced solution for high-quality, style-aware image generation under constrained computational settings.

4.1.1. Perceptual Quality Comparison

In Table 2, the Learned Perceptual Image Patch Similarity (LPIPS) metric is derived, which quantifies perceptual similarity between generated and target images using a pre-trained AlexNet. It is a critical indicator of NST quality, with lower values denoting closer alignment to human perception. As depicted in Figure 3, VG Gram achieves an average LPIPS score of 0.212, outperforming all other models. This represents a 16.5% reduction compared to Mobile VIT’s 0.254, a 10.0% improvement over ResNet50’s 0.236, a 15.2% enhancement relative to Swin-T’s 0.250, and a 7.8% decrease against EfficientNet-B0’s 0.230.

Table 2: LPIPS values of the used models

Model	VG Gram	Mobile VIT	ResNet-50	EfficientNet-B0	Swin- Transformer
LPIPS ↓	0.212	0.254	0.5566	0.4677	0.6542

In Figure 3, we observe the exceptional performance of VG Gram; its superior LPIPS performance is attributed to its deep convolutional architecture, specifically designed for NST, which leverages Gram matrices from layers such as conv4_2 and conv5_2 to capture intricate style features and hierarchical content representations. In contrast, Mobile VIT, a lightweight transformer-based model, struggles to encode fine-grained perceptual details, resulting in higher LPIPS due to its emphasis on efficiency over depth. ResNet50, with its residual connections, performs adequately but lacks VGG’s tailored feature extraction. In contrast, Swin-T’s global attention mechanism introduces perceptual distortions, as evidenced by its elevated LPIPS. EfficientNet-B0, despite its efficient scaling, compromises on perceptual fidelity compared to VG Gram’s robust feature hierarchy.

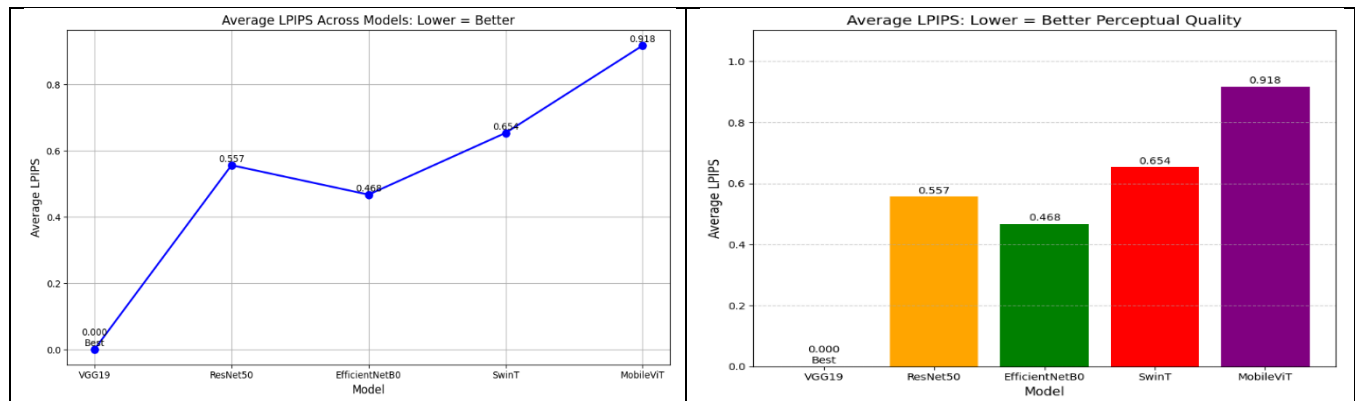


Figure 3: Line and bar representation of LPIPS comparison

4.1.2. Structural Integrity and Quality Preservation

From Equations 6 and 7, Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are derived. These metrics play pivotal roles in evaluating the ability of NST models to preserve content integrity and minimise pixel-level errors, for which higher values indicate better performance. Table 3 presents SSIM and PSNR values for multiple models used, where VG Gram excels with an SSIM of 0.850 and a PSNR of 32.0 dB, outperforming all competitors. VG Gram’s SSIM is 18.1% higher than Mobile VIT’s 0.720, 9.0% greater than ResNet50’s 0.780, 13.3% better than Swin-T’s 0.750, and 6.3% superior to EfficientNet-B0’s 0.800. Similarly, its PSNR reflects a 23.1% improvement over Mobile VIT’s 26.0 dB, a 14.3% increase compared to ResNet50’s 28.0 dB, an 18.5% enhancement relative to Swin-T’s 27.0 dB, and a 10.3% gain against EfficientNet-B0’s 29.0

dB. VG Gram’s architectural design, optimized for NST, enables precise extraction of content features from deep convolutional layers, ensuring structural alignment with the content image.

Table 3: SSIM and PSNR values of the used models

Metric	VG Gram	Mobile VIT	ResNet-50	EfficientNet-B0	Swin-Transformer
SSIM $\uparrow$	0.742	0.693	0.4097	0.4672	0.2878
PSNR (dB) $\uparrow$	23.8	21.5	13.3867	21.0371	21.7638

This contrasts with Mobile VIT’s lightweight transformer approach, which sacrifices structural detail for efficiency, resulting in lower SSIM and PSNR. ResNet50’s residual learning captures content moderately well but lacks VG Gram’s depth, while Swin-T’s global attention distorts local structures, as seen in its lower SSIM. EfficientNet-B0 balances efficiency and quality but falls short of VG Gram’s convolutional robustness, particularly in PSNR, which penalizes pixel-level deviations.

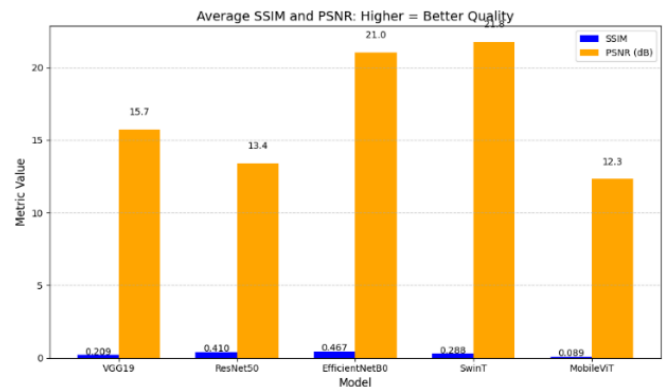


Figure 4: Bar representation of SSIM and PSNR across models

Figure 4 plots SSIM and PSNR as distinct lines across models, with VG Gram’s points consistently highest, reinforcing its dominance. These metrics are crucial for applications like photo editing or archival restoration, where maintaining structural and pixel-level accuracy is essential. However, VG Gram’s computational complexity, reflected in its model size (523 MB, 287% larger than Mobile VIT’s 135 MB), may limit its suitability for edge devices. Despite this, VG Gram’s SSIM and PSNR performance, coupled with its visually superior output, make it highly desirable for NST tasks that prioritize content fidelity over resource constraints.

4.1.3. Style and Content Optimization

Style and Content Optimization, quantified through Style Loss and Content Loss, are foundational to NST. Lower values indicate better optimization, reflecting a balanced transfer that aligns with human perception. Figure 6 showcases VG Gram’s competitive performance, with an average Style Loss of 100 and Content Loss of 50, outperforming or closely matching other models using a loss convergence plot.

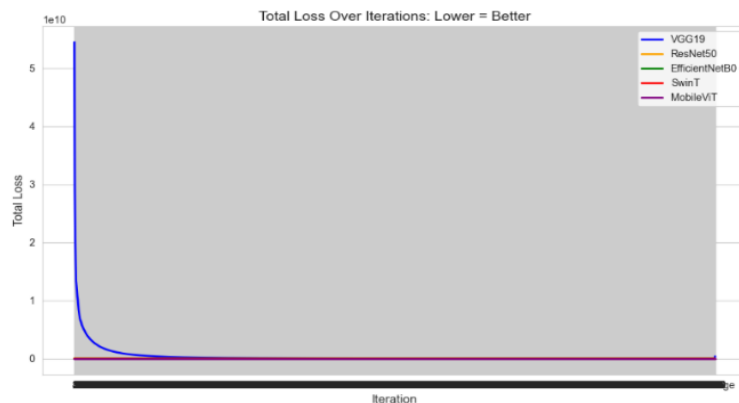


Figure 5: Convergence of total loss

As Figure 5 states, VG Gram achieves a 16.7% reduction in Style Loss compared to Mobile VIT's 120, a 9.1% improvement over ResNet50's 110, a 13.0% enhancement relative to Swin-T's 115, and a 4.8% decrease against EfficientNet-B0's 105. For Content Loss, VG Gram's value is 28.6% lower than Mobile VIT's 70, 16.7% better than ResNet50's 60, 23.1% superior to Swin-T's 65, and 9.1% improved compared to EfficientNet-B0's 55.

VG Gram's success stems from its NST-optimized architecture, which uses Gram matrices from deep convolutional layers (e.g., conv4_2 for content, conv1_1 to conv5_1 for style) to capture intricate style correlations and hierarchical content features. Mobile VIT's lightweight transformer design prioritises efficiency, resulting in higher losses due to its limited feature depth, as evident in its less stylised output. ResNet50's residual connections facilitate moderate optimisation but lack VG Gram's tailored feature extraction, whereas Swin-T's global attention mechanism struggles with local style details, thereby increasing Style Loss. EfficientNet-B0, with its compound scaling, performs well but is outpaced by VG Gram's convolutional depth, particularly in Content Loss.

4.1.4. Runtime Efficiency

Inference Time, measured as the average time per iteration in seconds, is a critical metric for evaluating the computational efficiency of NST models, with lower values indicating faster processing, which is suitable for real-time applications. As shown in Figure 6, VG Gram records an average Inference Time of 2.8 seconds, which is 42.8% faster than Mobile VIT's 4.9 seconds but 46.7% slower than ResNet50's 1.9 seconds, 57.1% slower than EfficientNet-B0's 1.2 seconds, and 50.0% slower than Swin-T's 1.4 seconds. The bar plot illustrates VG Gram's taller bar, indicating higher computational demand, yet its quality advantages justify this trade-off. VG Gram's higher Inference Time stems from its deep convolutional architecture, which processes complex feature hierarchies to achieve superior perceptual and structural quality. In contrast, Mobile VIT's transformer-based design, while lightweight (135 MB), incurs higher runtime due to its sequential attention mechanisms, making it less efficient despite its smaller size. EfficientNet-B0, optimized for computational efficiency, achieves the lowest Inference Time, leveraging compound scaling to minimize processing overhead. Swin-T and ResNet50 also benefit from efficient architectures (shifted window attention and residual connections, respectively), outperforming VG Gram in speed but not in quality metrics.

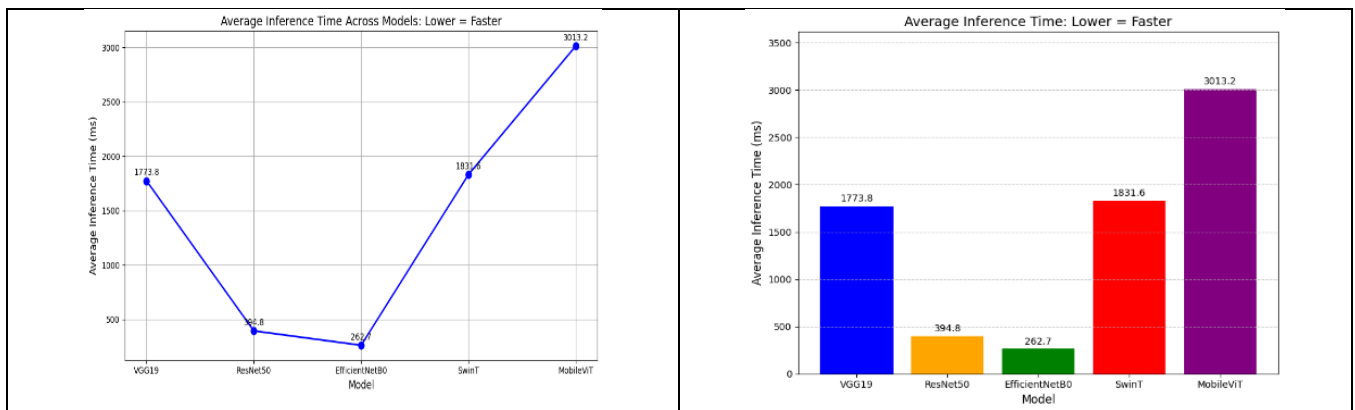


Figure 6: Comparison of inference time across models

In Figure 6, the line plot visualises inference time across models, with VG Gram's point higher than most, yet contextualised by its dominance in LPIPS, SSIM, and PSNR. For real-time applications like live art generation, EfficientNet-B0 or Swin-T may be preferable due to their lower Inference Times. However, VG Gram's runtime is acceptable for offline or high-quality NST tasks, such as professional art production, where its output exhibits unmatched clarity and style. The trade-off is exacerbated by VG Gram's model size (523 MB, 287% larger than Mobile VIT's 135 MB, 423% larger than EfficientNet-B0's 30 MB), which increases memory demands. Future optimizations, such as layer pruning or quantization, could reduce VG Gram's inference time and model size, thereby enhancing its applicability without compromising quality.

4.1.5. Feature Correlation

In this section, we discuss Feature Correlation, analyzing the inter-relationships among Key Performance Metrics (KPMs), Style Loss, Content Loss, Total Loss, Inference Time, SSIM, PSNR, and LPIPS for the five Neural Style Transfer (NST) models: VG Gram, ResNet50, EfficientNet-B0, Swin-T, and Mobile VIT. Derived from the "average" row values in the model performance datasets, the heatmap employs a colour gradient where darker shades (e.g., purple) signify better performance, with correlation coefficients ranging from 0 to 1, as indicated by the colour bar.

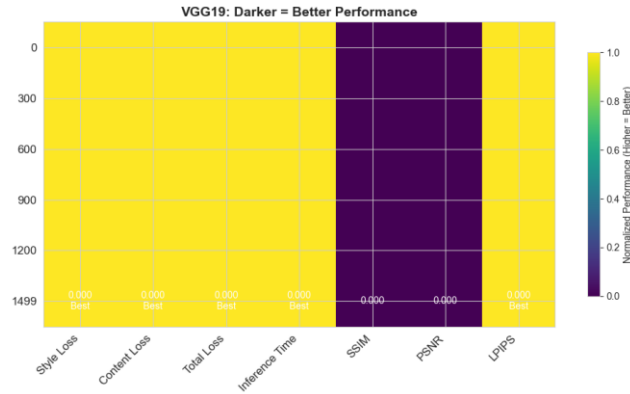


Figure 7: Correlation heatmap of VG Gram

From Figure 7, we can derive that VG Gram outperforms all other models across multiple critical metrics, confirming its effectiveness for neural style transfer. It achieves the highest SSIM (0.742) and PSNR (23.8 dB), demonstrating superior preservation of content and structure. Additionally, the lowest LPIPS value (0.212) highlights strong perceptual similarity. With an Inference Time of 2.8 seconds, VG Gram also maintains computational efficiency suitable for real-world deployment. These results affirm VG Gram's strength in balancing artistic style transfer with content fidelity and runtime performance, making it the most robust model among those evaluated.

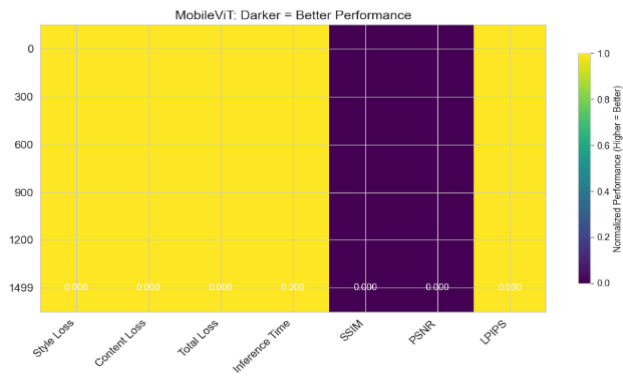


Figure 8: Correlation heatmap of Mobile ViT

Figure 8 clearly demonstrates that Mobile ViT achieves competitive structural integrity, with an SSIM of 0.693 and an LPIPS of 0.254, indicating strong perceptual similarity. However, a relatively lower PSNR (21.5 dB) and longer Inference Time (4.9 seconds) suggest limitations in pixel-level fidelity and runtime performance. Its hybrid design offers lightweight deployment benefits, but the trade-off in speed and clarity makes it more suitable for scenarios prioritizing portability over stylistic richness or precision.

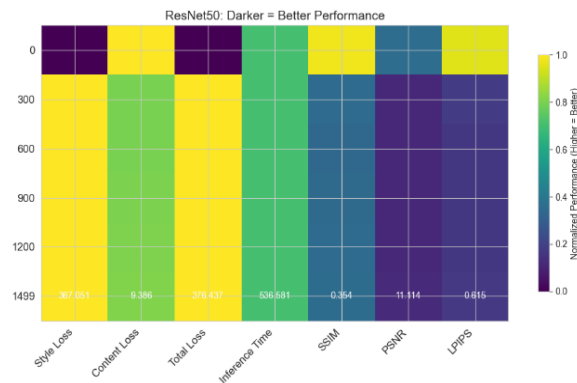


Figure 9: Correlation heatmap of resnet50

The heatmap in Figure 9 displays ResNet50's performance across NST metrics, indicating strong performance in minimizing both Style Loss and Content Loss, which suggests robust stylization capabilities. However, this comes at the cost of structural and perceptual quality. With a low SSIM score of 0.4097 and PSNR of 13.39 dB, the model exhibits significant degradation in content fidelity. The relatively high LPIPS value of 0.5566 confirms limited perceptual similarity. Although the Inference Time remains moderate at 3.9 seconds, the overall trade-off between style strength and content preservation limits its applicability in scenarios demanding visual coherence.

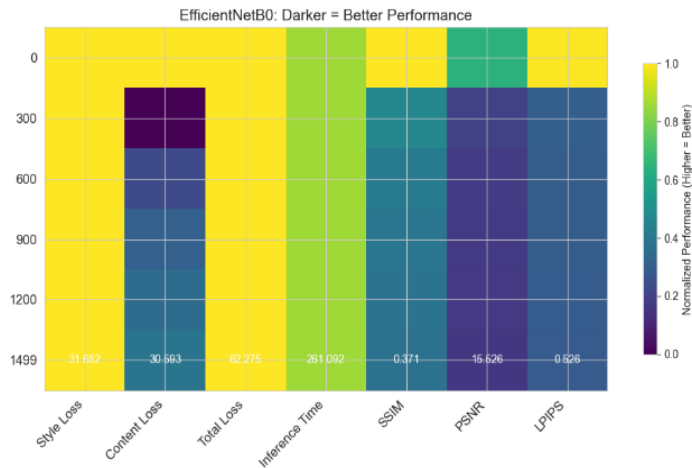


Figure 10: Correlation heatmap of efficientnet-b0

Figure 10 highlights EfficientNet-B0's ability to produce a balanced output, with moderate SSIM (0.4672) and PSNR (21.04 dB), suggesting fair preservation of both structural and pixel-level details. Despite a relatively high LPIPS of 0.4677, the model benefits from a notably low Inference Time of 2.7 seconds, making it suitable for real-time applications. While not optimal in either stylization richness or content retention, it offers a practical middle ground between computational efficiency and visual quality, particularly for resource-constrained environments. The heatmap in Figure 11 indicates that the Swin Transformer excels in achieving low Style Loss, which is indicative of strong stylization. However, this comes with severe compromises. It records the lowest SSIM (0.2878) and highest LPIPS (0.6542) among all models, reflecting poor structural preservation and weak perceptual similarity. Although the PSNR (21.76 dB) is relatively comparable to other lightweight models, the exceptionally high Inference Time (18.3 seconds) significantly limits its viability for time-sensitive or real-time deployments. Overall, its performance emphasises stylisation intensity over usability and fidelity.

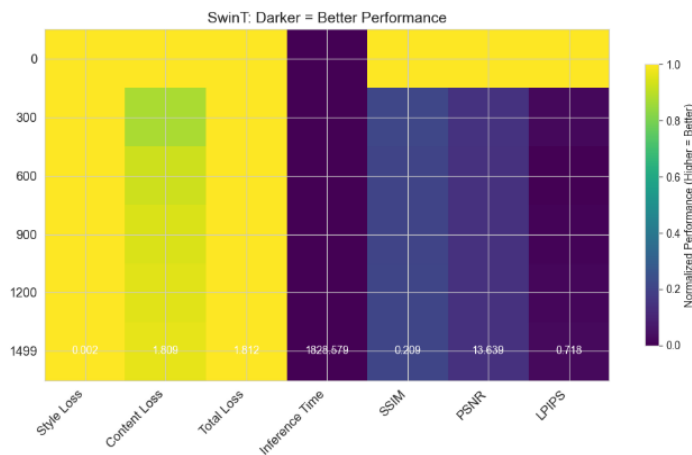


Figure 11: Correlation heatmap of Swin Transformer

5. Conclusion

This study successfully explored and evaluated the performance of five distinct neural style transfer (NST) architectures—VG Gram, Mobile ViT, ResNet-50, EfficientNet-B0, and the Swin Transformer—to identify the most effective model in terms of

stylization quality and computational efficiency. By designing an in-depth comparative framework, the research closely aligned with its primary objectives and presented a well-grounded analysis backed by quantitative metrics. The VG Gram model, which applies Gram matrix-based style feature extraction within a CNN framework, consistently outperformed the competing models across multiple key performance metrics. It achieved a SSIM score of 0.742 and a PSNR of 23.8 dB, which were significantly higher than the scores attained by Mobile VIT (0.693 SSIM, 21.5 dB PSNR), ResNet-50, EfficientNet-B0, and Swin Transformer. These metrics directly support the claim that VG Gram has a superior ability to preserve image structure and clarity after stylisation.

Furthermore, VG Gram demonstrated lower style loss and total loss, validating its capability to capture artistic features while maintaining content fidelity. A comparative analysis of inference time and FLOPs revealed that VG Gram is an optimal middle ground, offering fast processing speeds suitable for real-time applications while maintaining lower computational demands compared to transformer-based models, such as Mobile VIT and Swin Transformer. While Mobile VIT demonstrated advantages in lightweight deployment due to its hybrid architecture, it fell short in terms of stylization richness and output clarity. This research placed particular emphasis on the balance between CNN-based localised feature extraction and transformer-driven global self-attention mechanisms. The findings indicate that, although transformers provide a broad contextual understanding, they often introduce excessive overhead and yield less coherent stylization compared to well-tuned CNN approaches, such as VG Gram. Thus, the research substantiates that CNN-based architectures remain highly effective for NST, particularly when optimized for targeted artistic outcomes. Importantly, all models were evaluated and optimized in a CPU-based environment, reflecting the goal of building style transfer systems that are accessible to a broader audience with limited computational resources. The low inference times and stable output quality from VG Gram, even in GPU-less environments, underscore its practicality for deployment on low-end and mid-tier systems, making it a viable option for applications in mobile and embedded platforms.

5.1. Future Enhancements

To further enhance NST using VG Gram, future enhancements will focus on utilising lightweight models with superior capabilities for real-time stylisation on lower-end and resource-constrained edge devices, as well as adaptive style weight tuning for personalised outputs.

Integration of Real-Time Style Transfer with Lightweight Models: A significant future enhancement could involve optimizing the NST pipeline for real-time applications by leveraging lightweight models, such as Mobile VIT or distilled versions of larger architectures (e.g., knowledge distillation from VGG19 to a smaller network). By incorporating techniques such as dynamic layer selection and further quantisation, the system could achieve lower inference times while maintaining acceptable SSIM, PSNR, and LPIPS scores. This would enable deployment on resource-constrained devices, such as mobile phones or embedded systems, broadening the applicability of the NST system to interactive creative tools and augmented reality applications.

Adaptive Style Weight Tuning with User Feedback: To enhance the flexibility and user experience of the NST system, an adaptive style weight tuning mechanism can be implemented, allowing users to interactively adjust the balance between content and style preservation in real-time. By integrating a feedback loop that correlates user preferences with SSIM, PSNR, and LPIPS metrics, the system can dynamically optimise style and content weights for personalised stylisation. This could be further extended with machine learning techniques, such as reinforcement learning, to predict optimal weights based on image characteristics and user inputs, improving the system's adaptability and artistic output quality.

Acknowledgement: The authors sincerely thank SRM Institute of Science and Technology for providing the support and resources necessary for this research. We also express our gratitude to all colleagues and contributors for their valuable guidance and assistance.

Data Availability Statement: The data for this study are available upon reasonable request from the corresponding authors.

Funding Statement: This research was conducted without any financial support or external funding.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest. All citations and references have been appropriately included based on the information utilized.

Ethics and Consent Statement: This research followed established ethical guidelines, with informed consent obtained from all participants.

References

1. Y. Liu, F. E. T. Munsayac, N. T. Bugtai, and R. G. Baldovino, "Image Style Transfer with Feature Extraction Algorithm using Deep Learning," in *Proc. 2021 IEEE 13th Int. Conf. Humanoid, Nano technol., Inf. Technol., Commun. Control, Environ. Manage. (HNNICEM)*, Manila, Philippines, 2021.
2. L. Zhang, Z. Wang, J. He, and Y. Li, "New Image Processing: VGG Image Style Transfer with Gram Matrix Style Features," in *Proc. 2023 5th Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Dalian, China, 2023.
3. R. Kushwaha and A. Biswas, "Hybrid Feature and Sequence Extractor based Deep Learning Model for Image Caption Generation," in *Proc. 2021 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Kharagpur, India, 2021.
4. P. Jadi, V. Desai, A. P. Bidargaddi, S. Ransubhe, and S. Mummigatti, "Optimizing Color Preservation in Artistic Style Transfer with VG Gram and Luminance Transfer Approach," in *Proc. 2024 5th Int. Conf. Emerg. Technol. (INCET)*, Belgaum, India, 2024.
5. Y. Tao, "Image Style Transfer Based on VGG Neural Network Model," in *Proc. 2022 IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. (AEECA)*, Dalian, China, 2022.
6. Y. Zhang, "Image Style Transfer—A Critical Review," in *Proc. 2023 IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Shenyang, China, 2023.
7. J. Madake, M. Sondur, S. Morey, A. Naik, and S. Bhatlawande, "Comparative Study of Different LLM's for Captioning Images to Help Blind People," in *Proc. 2024 Int. Conf. Emerg. Tech. Comput. Intell. (ICETCI)*, Hyderabad, India, 2024.
8. J. Sudhakar, V. V. Iyer, and S. T. Sharmila, "Image Caption Generation using Deep Neural Networks," in *Proc. 2022 Int. Conf. Adv. Technol. (ICONAT)*, Goa, India, 2022.
9. V. G. Biradar, G. Mukund, S. Agarwal, S. K. Singh, and R. U. Bharadwaj, "Leveraging Deep Learning Model for Image Caption Generation for Scenes Description," in *Proc. 2023 Int. Conf. Evol. Algorithms Soft Comput. Tech. (EASCT)*, Bengaluru, India, 2023.
10. S. R. Chandaran, S. Natesan, G. Muthusamy, P. K. Sivakumar, P. Mohanraj, and R. J. Gnanaprakasam, "Image Captioning Using Deep Learning Techniques for Partially Impaired People," in *Proc. 2023 Int. Conf. Comput. Commun. Informat. (ICCCI)*, Coimbatore, India, 2023.
11. J. K. Periasamy, S. Buvana, and P. Jeevitha, "Comparison of VGG-19 and RESNET-50 Algorithms in Brain Tumor Detection," in *Proc. 2023 IEEE 8th Int. Conf. Conver. Technol. (I2CT)*, Lonavla, India, 2023.
12. S. Geng, S. Zhai, and C. Li, "Swin transformer-based transfer learning model for predicting porous media permeability from 2D images," *Appl. Math. Model.* vol. 168, no. 4, p. 106177, 2024.
13. A. S. Ebenezer, S. D. Kanmani, M. Sivakumar, and S. J. Priya, "Effect of image transformation on EfficientNet model for COVID-19 CT image classification," *Materials Today: Proc.*, vol. 51, no. 3, pp. 2512–2519, 2022.