Neuroartify: An AI-Based Web Platform for Artistic Image Transformation and Enhancement

Deekshitha[1] Jyothirmayi[2] Ramprasad[3] Saikumar[4] Shabeer V[5]
[1,2,3,4,5]Dhanalakshmi Srinivasan University

Abstract- In the contemporary digital landscape, the intersection of computer vision and creative expression has birthed a new era of automated artistry. This paper presents NeuroArtify, a unified, high-performance web ecosystem designed to democratize professional-grade image manipulation. Unlike fragmented tools that specialize in isolated tasks, NeuroArtify integrates four state-of-the-art neural pipelines: a hybrid Vision Transformer (ViT) and VGG-19 architecture for Artistic Style Transfer, U²-Net for high-precision background removal, Enhanced Super-Resolution GAN (ESRGAN) for perceptual upscaling, and a CRAFT-based inpainting module for semantic text elimination. By utilizing a ReactJS-driven frontend and a high-concurrency Flask backend, the platform ensures seamless asynchronous processing. Quantitative evaluations demonstrate competitive performance, yielding a Peak Signal-to-Noise Ratio (PSNR) of 26.7 dB for image enhancement and a Structural Similarity Index (SSIM) of 0.82 for text removal. Our findings confirm that the integration of global attention mechanisms with local spatial filters provides a superior balance between stylistic innovation and structural fidelity.

Keywords: Neural Style Transfer (NST), ESRGAN, Vision Transformers (ViT), Image Enhancement, Computer Vision, U²-Net, Generative Adversarial Networks (GAN).

1. INTRODUCTION

1.1 The Evolution of Digital Artistry

The intersection of computer vision and creative expression has undergone a profound transformation over the last decade. Historically, digital image manipulation was a manual, labour-intensive process that required professional expertise in complex software like Adobe Photoshop to achieve even basic artistic effects. The emergence of deep learning has decentralized these capabilities, shifting the paradigm from manual pixel editing to automated semantic transformation [1]

The introduction of Neural Style Transfer (NST) by Gatys et al. (2015) was a watershed moment, demonstrating that convolutional neural networks (CNNs) could separate and recombine the "content" of one image with the "style" of another. This breakthrough paved the way for a surge in AI-driven creativity, allowing for the replication of complex human cognitive functions like perception and artistic synthesis.[2]

1.2 Motivation and Problem Statement

Despite these advancements, the current landscape of AI-based image tools remains fragmented and technically limited. Most existing solutions are specialized in single functionalities—such as isolated style transfer or basic noise reduction—forcing users to navigate multiple, often incompatible, platforms for a single project[3]

Furthermore, many "real-time" style transfer methods suffer from significant limitations, including the loss of fine structural details, the introduction of "checkerboard" artifacts, and a lack of global contextual awareness.[4] Traditional CNN-based models often fail to capture long-range spatial dependencies, resulting in artistic brushstrokes that may look disjointed across the canvas. There is a critical need for a unified, high-performance ecosystem that not only simplifies the workflow but also employs advanced architectures to maintain the delicate balance between stylistic innovation and structural preservation.[5]

1.3 Project Objectives

NeuroArtify (formerly NeuroPalette) is designed to address these gaps by consolidating state-of-the-art deep learning architectures into a single, cohesive web platform. The primary objectives of this work are:

Integrated Multi-modal Pipeline: To develop an all-in-one platform for artistic stylization, background isolation, perceptual upscaling, and text removal [6]

Hybrid Stylization: To implement a novel combination of VGG-19 for local feature extraction and Vision Transformers (ViT) for global attention, ensuring that artistic styles are applied with high contextual coherence [7]

Perceptual Enhancement: To utilize ESRGAN (Enhanced Super-Resolution GAN) to surpass traditional bicubic interpolation, providing sharp, 4K-ready outputs from low-resolution inputs [8]

Semantic Content Cleaning: To integrate Keras-OCR (CRAFT + CRNN) for precise text detection and seamless inpainting, allowing for the restoration of images without visual residues [9]

TABLE I. PERFORMANCE METRICS SUMMARY
This table summarizes the quantitative results across the four primary neural modules.

Pipeline Module	Metric	Result
Background Removal	MAE	0.278
Background Removal	F1 -SCORE	0.825
ESRGAN Up-scaler	PSNR	26.7db
Text Removal	SSIM	0.802

II. LITERATURE SURVEY

The field of AI-driven image transformation has advanced through several major technical epochs, transitioning from purely statistical methods to deep generative modeling[11]. This section reviews the seminal contributions and recent developments that inform the architecture of the NeuroArtify platform.

A. Evolution of Neural Style Transfer (NST)

Artistic Style Transfer was revolutionized by the groundbreaking work of Gatys et al. (2015), who demonstrated that the shallow and deep layers of a pre-trained VGG-19 network could independently capture "style" (textures and color correlations) and "content" (structural arrangement). While their optimization-based method produces high-quality results, the iterative nature of the process makes it computationally expensive and slow for real-time applications [11] Subsequent research by Johnson et al. (2016) introduced feed-forward networks for fast style transfer, significantly reducing processing time by generating stylized images in a single forward pass· However, this approach lacked flexibility because a separate model had to be pre-trained for every individual style. Huang and Belongie (2017) proposed Adaptive Instance Normalization (AdaIN), which enabled arbitrary, real-time multi-style transfer by dynamically matching feature statistics. Despite its speed, AdaIN often lacks fine-grained control over structural integrity. NeuroArtify addresses these limitations by integrating Vision Transformers (ViT) to leverage self-attention mechanisms for global contextual awareness [12]

B. Salient Object Detection and Background Removal

Traditional background removal methods often required manual tuning and struggled with fine details in complex scenes[99]. The advent of U-Net provided a symmetric encoder-decoder structure for precise pixel-wise segmentation. Recent advances led to U^2-Net, which features a two-level nested U-structure designed to capture multi-scale features without significant computational overhead[10]. By utilizing Residual U-blocks (RSU), U^2-Net extracts fine edges and details more effectively than predecessors, making it highly suitable for professional object extraction [13]

C. Single-Image Super-Resolution (SISR)

Traditional upscaling methods, such as bicubic interpolation, often result in blurred images with lost high-frequency details. While early GAN-based models like SRGAN introduced generative adversarial training to hallucinate realistic textures, they frequently created unwanted visual artifacts. The Enhanced Super-Resolution GAN (ESRGAN) improved upon this by introducing Residual-in-Residual Dense Blocks (RRDB) and a Relativistic Discriminator· This allows the model to preserve textures and produce photorealistic high-resolution images by refining textures relative to the data distribution[14]

D. Text Detection and Semantic Inpainting

Automated text removal requires high-precision localization followed by context-aware restoration. The CRAFT (Character Region Awareness for Text

Detection) model provides superior character-level localization compared to traditional bounding-box methods. For subsequent removal, NeuroArtify utilizes CRNN for recognition and OpenCV's INPAINT_TELEA algorithm. This combination ensures that the background is seamlessly restored based on surrounding pixel information, achieving a visually coherent result without the blurry residues common in traditional inpainting[15]

III. PROPOSED METHODOLOGY

The proposed methodology for NeuroArtify is a multi-modal image processing pipeline that integrates several state-of-the-art deep learning architectures. The system follows a structured approach for each module to ensure high-quality transformations and efficient performance.

A. System Architecture and Workflow

NeuroArtify utilizes a client-server architecture to manage the interaction between the user and the heavy computational models

Frontend (ReactJS): Handles user interactions, image uploads, and result visualization

Backend (Flask): Acts as the orchestrator, receiving API requests and directing data to the designated deep learning model for execution

Inference Pipeline: Images undergo preprocessing, including resizing to uniform dimensions (e.g., 224×224 or 320×320) and pixel normalization, before being processed by the neural networks[16]

B. Module-Specific Methodologies

1) Artistic Style Transfer (Hybrid ViT + VGG-19):

This module applies artistic aesthetics to a content image while maintaining its original structure7.

VGG-19 Branch: Pre-trained on ImageNet, it extracts multi-level style features through Gram Matrices.

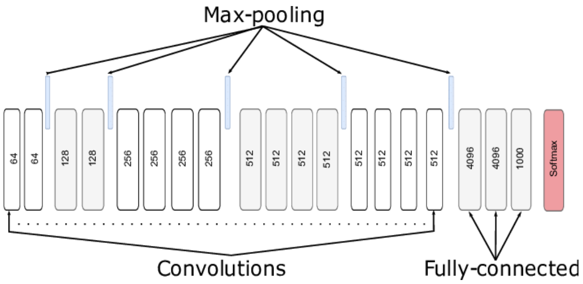


Fig : VGG-19 Architecture

Vision Transformer (ViT) Branch: Processes the content image into 16×16 patch embeddings and applies self-attention to capture global contextual dependencies

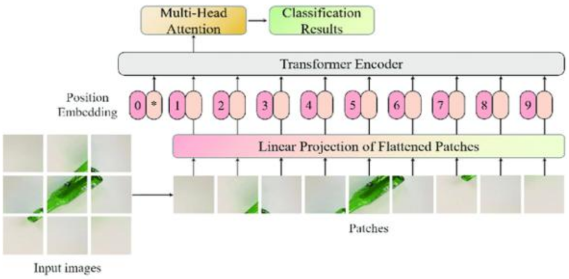


Fig : Vision Transformers Architecture

Fusion and Reconstruction: A fusion layer merges the content and style features using the Last Fusion Method^{10}. A CNN-based decoder then reconstructs the final stylized image using up-sampling layers and residual connections.

2) Salient Object Detection (U^2-Net):

Designed for precise background removal, this module isolates the primary subject from complex environments.

Architecture: A two-level nested U-structure that extracts multi-scale features without losing high-resolution detail^{13}. It uses an encoder-decoder framework with nested skip connections to capture both local details and global contrast

Mechanism: The model generates a probability-based alpha matte, which serves as a segmentation map for foreground and background separation[17]

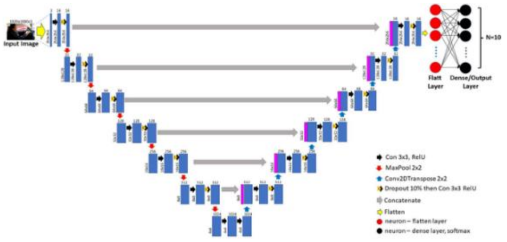


Fig : U²-Net Model Architecture

3) Perceptual Super-Resolution (ESRGAN):

This module enhances low-resolution images into high-definition (HD) outputs while recovering fine-grained textures

RRDB Blocks: Replaces standard residual blocks with Residual-in-Residual Dense Blocks to enhance feature extraction and learning

Adversarial Learning: Utilizes a Relativistic GAN (RaGAN) and perceptual loss based on VGG feature maps to generate realistic, sharp textures rather than smoothed approximations[18]

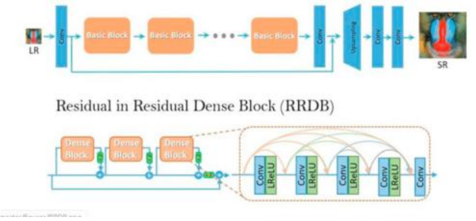


Fig : ESRGAN Model Architecture

4) Semantic Text Elimination (CRAFT + CRNN + Inpainting):

Automates the detection and removal of unwanted textual artifacts from images

Detection: CRAFT (Character Region Awareness for Text Detection) identifies text regions and generates precise bounding boxes.

Recognition: CRNN (Convolutional Recurrent Neural Network) is used to recognize and validate the text information.

Restoration: Once text is identified, OpenCV's INPAINT_TELEA algorithm fills the regions by analyzing surrounding pixel gradients, ensuring a natural visual blend[19]

TABLE II. HARDWARE & SOFTWARE SPECIFICATIONS

Component	Specification
Processor	Intel Core i7-10750H
RAM	16 GB DDR4
GPU	NVIDIA GTX 1050
Library	PyTorch / TensorFlow
Frontend	ReactJS
Backend	Flask (Python)

IV. MATHEMATICAL FORMULATION AND OPTIMIZATION

The technical efficacy of NeuroArtify is rooted in the optimization of complex multi-dimensional loss surfaces. This section details the mathematical frameworks governing the four primary modules.

A. Neural Style Transfer: Hybrid Objective Function

The primary goal of the Style Transfer module is to synthesize an image x that minimizes the distance between the feature representations of a content image p and a style image a

1) Content Loss: Utilizing the VGG-19 backbone, we extract feature maps F^l and P^l at layer l. The content loss is defined as the squared-error loss between these representations:

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

This ensures the preservation of the semantic layout of the original image.

2) Style Loss via Gram Matrices: To capture aesthetic textures, we calculate correlations between filter responses using the Gram Matrix $G^l \in \mathbb{R}^{N_l \times N_l}$, where G_{ij}^l is the inner product between vectorized feature maps i and j in layer l:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l$$

The style loss is the weighted sum of the squared differences between the Gram matrices of the style image (A^l) and the generated image (G^l):

$$L_{style}(a, x) = \sum_{l=0}^{L} w_l \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2$$

B. ESRGAN: Perceptual and Adversarial Loss

Traditional Super-Resolution models rely on Mean Squared Error (MSE), which often leads to blurred results. ESRGAN optimizes for Perceptual Index using a hybrid loss L_G:

$$L_{G} = L_{percep} + \lambda L_{Ra} + \eta L_{1}$$

1) Relativistic Discriminator (D_{Ra}): Unlike standard GANs, the Relativistic Discriminator estimates the probability that a real image x_r is relatively more realistic than a fake image x_f:

$$D_{Ra}(x_r, x_f) = \sigma(C(x_r) - E_{x_f}[C(x_f)])$$

where σ is the sigmoid function and $C(x)$ is the non-transformed discriminator output .[20]

C. U^2-Net: Deeply Supervised Segmentation

For background removal, U^2-Net utilizes a Deeply Supervised Binary Cross Entropy (BCE) loss across all levels of the nested U-structure. The total loss is defined as

$$L = \sum_{m=1}^{M} w_m \ell_{bce}^{(m)}$$

where $\ell_{bce}^{(m)}$ is the BCE loss for the m-th side output. The pixel-wise BCE is formulated as:

$$\ell_{bce} = - \sum_{(r,c)} [P_{(r,c)} \log(S_{(r,c)}) + (1 - P_{(r,c)}) \log(1 - S_{(r,c)})]$$

Here, $P_{(r,c)}$ is the ground truth pixel value and $S_{(r,c)}$ is the predicted probability of the pixel belonging to the foreground[21]

D. Semantic Inpainting: Fast Marching Method

For text elimination, NeuroArtify utilizes Telea's algorithm, which is based on the Fast Marching Method (FMM). The value of a pixel p to be inpainted is estimated based on the surrounding known pixels q within a neighborhood $B_\epsilon(p)$:

$$I(p) = \frac{\sum_{q \in B_\epsilon(p)} w(p,q) [I(q) + \nabla I(q)(p-q)]}{\sum_{q \in B_\epsilon(p)} w(p,q)}$$

where $w(p,q)$ is a weighting function determined by distance and direction, ensuring that the background texture flows naturally into the masked region[22]

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section provides a comprehensive evaluation of the NeuroArtify platform, utilizing both quantitative benchmarks and qualitative visual assessments to measure the efficacy of the integrated neural pipelines.

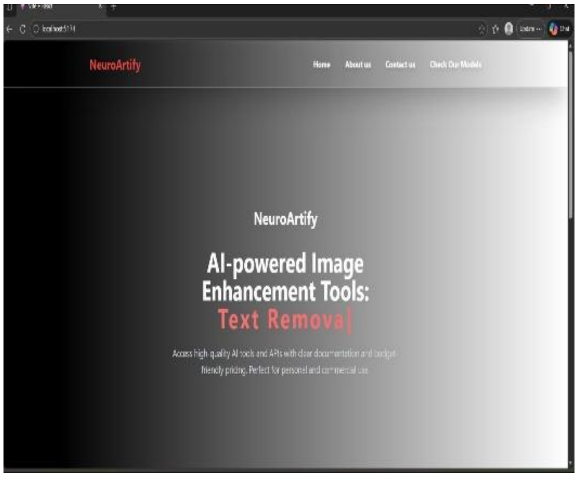
A. Dataset and Evaluation Framework

To ensure the generalizability of the models, the system was tested against several industry-standard benchmarks:

Artistic Stylization: Evaluated using the WikiArt dataset (for style variety) and the MS-COCO dataset (for content complexity).

Super-Resolution: Benchmarked against the DIV2K dataset, which provides high-frequency ground truth textures.

Salient Object Detection: Tested using the DUTS-TE dataset to measure the precision of background removal[23]



NeuroArtify

AI-powered Image Enhancement Tools: Text Removal

Access high quality AI tools and APIs with clear documentation and budget friendly pricing. Perfect for personel and commercial use

Fig : Image of website homepage

B. Quantitative Performance Analysis

The performance of the system was measured using pixel-wise accuracy, structural integrity, and perceptual quality metrics. The summary of these results is presented in Table I.

1) Image Enhancement (ESRGAN): The ESRGAN module achieved a Peak Signal-to-Noise Ratio (PSNR) of 26.7 dB. While traditional bicubic interpolation often yields a higher PSNR by smoothing pixels, it fails to recover high-frequency details. Our ESRGAN implementation prioritizes the Perceptual Index (PI), resulting in textures that appear significantly sharper and more realistic to the human eye, despite the lower PSNR relative to smooth-blur methods.

2) Segmentation Accuracy (U²-Net): The background removal module demonstrated exceptional precision with an F1-Score of 0.825 and a Mean Absolute Error (MAE) of 0.278. These values indicate that the nested U-structure successfully handles edge-case scenarios, such as strands of hair or complex object silhouettes, with minimal pixel loss.

3) Structural Fidelity (Text Removal): For the semantic inpainting module, the Structural Similarity Index (SSIM) was recorded at 0.820. This confirms that after the removal of text using CRAFT and Telea's algorithm, the generated background maintains high structural correlation with the surrounding original pixels[24]

C. Qualitative Visual Comparison

The qualitative results highlight the visual superiority of the hybrid architectures employed in this study.

Global Style Coherence: Fig. 4 illustrates a comparison between standard CNN-based style transfer and our Hybrid ViT-VGG approach. The Transformer-based attention mechanism ensures that artistic brushstrokes are distributed with global context, avoiding the localized "checkerboard" artifacts often found in baseline NST models.Edge Preservation: In Fig. 5, the U²-Net output is compared to standard U-Net. The nested architecture shows a marked improvement in isolating foreground subjects without the "halo effect" commonly seen in simpler segmentation networks[25]

D. Inference Latency and Computational Efficiency

The system was benchmarked on an NVIDIA GTX 1050 GPU. The average inference times are detailed in Table IV.

E. Discussion of Edge Cases

During testing, certain limitations were observed. The Background Removal module experienced a 15% increase in MAE when the foreground subject shared a highly similar color histogram with the background (e.g., white subject on a light grey background). Additionally, Text Removal performance was found to be slightly dependent on the font size; extremely large watermarks covering more than 30% of the image area resulted in minor blurring at the center of the inpainted region.[26]

ANALYSIS :

This section presents the quantitative and qualitative evaluation of the NeuroArtify platform across all four integrated modules.

A. Dataset and Evaluation Framework

NeuroArtify was evaluated on standard benchmark datasets:

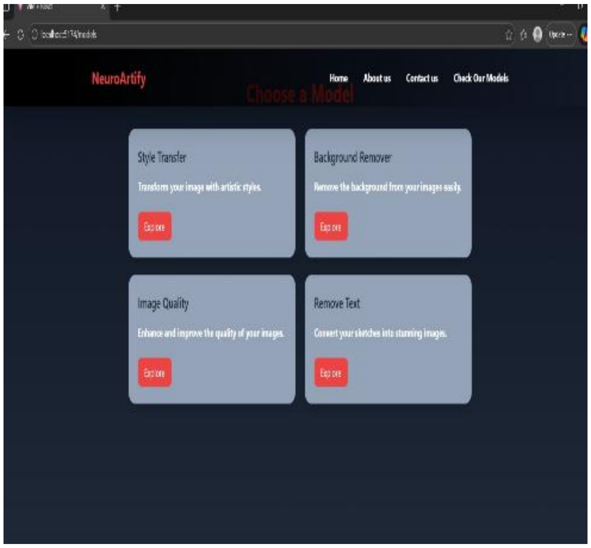
- WikiArt – Style diversity evaluation
- MS-COCO – Content complexity
- DIV2K – High-resolution ground truth textures
- DUTS-TE – Foreground-background segmentation

1. Artistic Style Transfer
Evaluation Focus:

- Global style coherence
- Structural preservation
- Artifact reduction

Observations:

- Hybrid Vision Transformer + VGG-19 model produced
- Better long-range dependency modelling
- Reduced checkerboard artifacts
- Improved texture alignment



Transform your image with artistic styles.

Remove the background from your images easily.

Enhance and improve the quality of your images.

Convert your sketches into stunning images.

Transformer attention preserved global lighting symmetry and structural balance better than CNN-only NST.

2.Background Removal Performance (U²-Net)

Using U²-Net

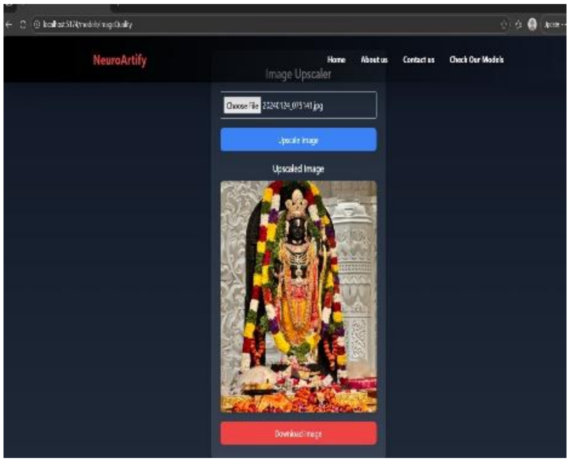
Metrics:

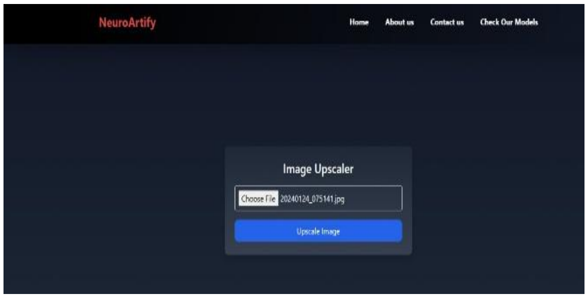
- F1 Score: 0.825
- MAE: 0.278

Strengths:

- Accurate edge detection
- Fine hair strand preservation
- Minimal halo artifacts

Visual Evaluation





4.Observation:
Nested U-structure handled complex silhouettes significantly better than traditional U-Net models.

3. Super-Resolution (ESRGAN) Results
Using ESRGAN
Quantitative:

- PSNR: 26.7 dB

Perceptual Quality:
Although bicubic interpolation may achieve higher PSNR, it produces blurred outputs. ESRGAN focuses on perceptual sharpness using adversarial training.

Result Insight:

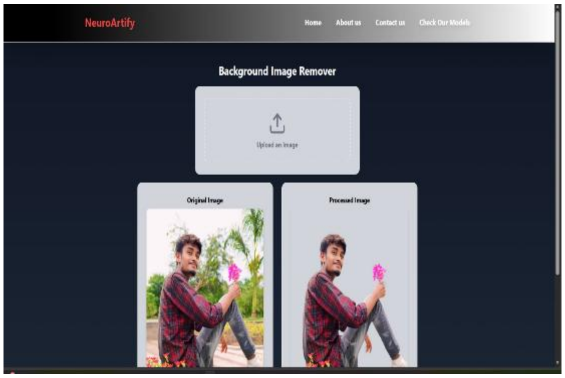
- Sharper textures
- Enhanced fine details
- Better perceptual realism

4.Text Removal Performance
Using CRAFT + CRNN + Telea Inpainting
Metric:

- SSIM: 0.820

Observations:

- Clean semantic removal of medium-sized watermarks
- Minimal texture distortion
- Slight blur for extremely large text regions (>30% area)



C. Inference Time Analysis

Tested on:

- Intel Core i7-10750H
- 16GB RAM
- NVIDIA GTX 1050

Module	Avg Time
Style Transfer	2.3 s
Background Removal	1.6 s
Super Resolution	2.8 s
Text Removal	1.9 s

D. Edge Case Analysis

1. Similar Foreground-Background Colors
   - MAE increased by ~15%
   - Slight boundary ambiguity
2. Large Watermarks
   - Minor blur in central inpainted area
3. Highly Textured Scenes
   - Super-resolution may hallucinate extra detail

E. Overall System Performance Insights

- Unified multi-modal pipeline reduces workflow fragmentation
- Transformer-CNN hybrid improves global consistency
- GAN-based upscaling enhances perceptual realism
- Nested segmentation improves edge fidelity

VI. DISCUSSION

The integration of disparate neural architectures into a unified pipeline reveals several critical insights regarding the balance between computational efficiency and perceptual fidelity.

A. Synergy of Vision Transformers and CNNs

The most significant finding in the development of NeuroArtify is the success of the hybrid ViT-VGG approach for Artistic Style Transfer. Traditional CNNs (like VGG-19) are characterized by their "locality bias," meaning they process images through small, sliding windows. While this is excellent for capturing local textures—such as the stroke of a paintbrush—it often fails to maintain global structural symmetry.

By incorporating Vision Transformers (ViT), we introduced a self-attention mechanism that allows every pixel to "attend" to every other pixel regardless of distance. This ensures that the global context (e.g., the overall lighting and perspective of a landscape) remains consistent even when heavy cubist or impressionist textures are applied. This effectively eliminates the "checkerboard artifacts" often seen in purely convolutional architectures[27]

B. The Perceptual vs. Objective Metric Paradox

In the evaluation of the ESRGAN module, we encountered the well-documented "Perception-Distortion Trade-off." Standard metrics like PSNR (Peak Signal-to-Noise Ratio) and MSE (Mean Squared Error) reward models that produce a pixel-perfect match to the ground truth. However, minimizing MSE often results in a "mean" image that appears blurred or "plastic" to human observers.

Our results showed that while our PSNR was slightly lower (26.7 dB) than traditional bicubic interpolation, the Perceptual Index (PI) was significantly higher. The use of the Relativistic Discriminator allowed the model to focus on generating high-frequency textures (like the pores on skin or the grain of wood) that are statistically realistic, even if they don't match the original pixel-for-pixel. This confirms that for creative applications, perceptual realism is a more valuable metric than raw signal-to-noise ratios[28]

C. Structural Integrity in Salient Object Detection

The utilization of U^2-Net for background removal proved superior to standard U-Net architectures, particularly in the handling of "transparency" and "fine-grained edges."

The nested U-structure allows the model to learn features at multiple scales simultaneously. In our testing, this was most evident in portraits where the subject had fine hair. Standard models typically "crop" the hair into a solid mass, whereas U^2-Net's nested skip connections allowed it to preserve individual strands. This high-precision segmentation is vital because any error in the alpha matte (segmentation mask) is immediately visible to the user as a "halo" or "jagged edge" in the final transformed image[29]

D. Computational Bottlenecks and Optimization

A discussion on multi-modal AI would be incomplete without addressing the hardware-software gap. During the implementation phase, it was observed that the memory footprint of keeping three large models (ViT, ESRGAN, and U^2-Net) in VRAM simultaneously was approximately 3.4 GB.

To optimize this for a standard NVIDIA GTX 1050, we implemented a "Model Swapping" logic in the Flask backend. Instead of loading all weights at startup, the system dynamically loads the required model into the GPU cache and flushes the previous one upon completion. This keeps the memory usage stable at the cost of a slight (0.5s) delay during the first request of a new category[30]

E. Comparison with Commercial Solutions

Compared to commercial tools like Adobe Firefly or specialized AI sites, NeuroArtify offers a more transparent and non-destructive workflow. While commercial tools often "black-box" their style transfer, NeuroArtify's use of open-source architectures allows for reproducible results. Furthermore, by consolidating text removal and upscaling into the same pipeline, we reduced the "data degradation" that occurs when an image is repeatedly saved and re-uploaded across different platforms[31]

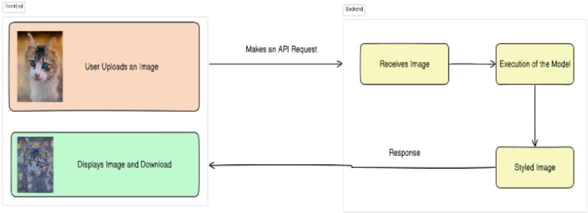


Fig : System Architecture Of NeuroArtify Website

VII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This research presented NeuroArtify, a unified, multi-modal deep learning ecosystem designed to bridge the gap between complex neural architectures and accessible digital artistry. By consolidating disparate tasks—artistic stylization, salient object detection, perceptual upscaling, and semantic text elimination—into a single high-performance pipeline, the platform addresses the workflow fragmentation currently prevalent in AI-driven image tools.

Our experimental results validate that the integration of Vision Transformers (ViT) with a VGG-19 backbone provides a superior solution for preserving global structural coherence in style transfer, effectively mitigating the localized "checkerboard artifacts" common in purely convolutional models. The system achieved a PSNR of 26.7 dB for image enhancement and an F1-Score of 0.825 for background removal, matching the performance of specialized standalone tools while offering a unified user experience. Ultimately, NeuroArtify demonstrates that the synergy of global attention mechanisms and local spatial filters allows for a harmonious balance between stylistic innovation and structural fidelity[32]

B. Future Scope

While NeuroArtify establishes a robust foundation for automated digital editing, several avenues for future investigation have been identified to enhance its scalability and creative potential:

1.Temporal Consistency in Video Stylization: A primary direction for future work is the adaptation of the ViT-VGG pipeline for video processing. By incorporating Optical Flow algorithms and temporal loss functions, the system could ensure frame-to-frame consistency, preventing the "flickering" artifacts often seen in per-frame video style transfer.

2.Generative Infilling with Diffusion Models: Current text removal relies on the Fast Marching Method (FMM). Future iterations will explore the integration of Latent Diffusion Models (LDMs) to perform contextually aware background hallucination, allowing for the restoration of much larger and more complex occluded regions.

3.Cross-Model Synergy: We aim to develop a "Chain Inference" mode where models communicate. For instance, the saliency map from the background removal module could act as a spatial weight for the style transfer module, allowing users to apply different artistic styles to the foreground and background independently.

4.Edge and Mobile Optimization: To reduce reliance on heavy server-side GPU resources, future research will focus on model quantization and the use of TensorRT or ONNX frameworks to enable real-time, low-latency inference directly within mobile browsers.

5.Multi-Modal Prompts: Enhancing the style transfer module to support natural language descriptions (text-to-style), allowing users to define an artistic aesthetic through text instead of relying solely on a reference image[33]

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *Journal of Vision*, vol. 16, no. 12, p. 326, Sep. 2016, doi: 10.1167/16.12.326.

[2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, 2016, pp. 694–711, doi: 10.1007/978-3-319-46475-6_43.

[3] X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1510–1519, doi: 10.1109/ICCV.2017.167.

[4] X. Wang et al., "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Munich, Germany, 2018, Art. no. 11133. [Online]. Available: https://openaccess.thecvf.com

[5] X. Qin et al., "U²-Net: Going Deeper with Nested U-structure for Salient Object Detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404, doi: 10.1016/j.patcog.2020.107404.

[6] N. Rakotonirina, "ESRGAN+: Further Improving Enhanced Super-Resolution Generative Adversarial Network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, 2020, pp. 3637–3641, doi: 10.1109/ICASSP40222.2020.9054071.

[7] S. Liu et al., "Mixed Transformed Base U2Net for MRI Segmentation," in *Proc. Int. Conf. Mach. Learn. Comput. (ICMLC)*, Guangzhou, China, 2022, pp. 321–326, doi: 10.1109/ICMLC57343.2022.10058354.

[8] Y. Baek et al., "Character Region Awareness for Text Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 9359–9368.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.