Image Style Transfer

Razan Alkharasani, Hanan Alshumrani, Maram Alnabrees, Fatima Alawami, Sajedah Alqudaihi , and Raghad Alamoudi

***Abstract*—Deep Learning has introduced style-transfer as an application of computer vision for creating artwork from the content of a source image, imitating the texture of a particular work of art. While traditional techniques will produce quality images when fed non-noisy high-resolution images, the quality diminishes significantly when the method is performed on images containing artifacts (artifacts) or low resolution. To address this limitation, this study proposes a method combining a Denoising Auto Encoder with a Convolutional Neural Network (CNN) built on top of VGG19. By using the auto-encoder for preprocessing and refining content images, the denoising process removed noise while keeping important details. After that, using the VGG19 model, both high-level and low-level content features and styles were extracted with the losses minimized to directly generate a stylised version of the original content image. The system was trained and tested on a collection of artistic and content images from Kaggle using Data Augmentation to improve the quality of the images. The results show that the pre-processing significantly increased the quality of the generated artistic images, as well as improving most input image sources' fidelity.**

***Index Terms*—Deep learning, Computer vision, Denoising autoencoders, Image style transfer, Neural networks, VGG19.**

I. INTRODUCTION

Deep learning (DL), which uses the structure of biological neural networks as a model, is an advanced and very powerful form of machine learning that is transforming the way computing works today, by allowing the model to learn hierarchical representations of data directly. This has led to tremendous progress in the development of many areas of computing including natural language processing and computer vision applications; one example is image style transfer. To facilitate research that requires large amounts of data, many researchers now use platforms such as Kaggle.com, which offer both hardware and access to huge amounts of unique data (such as the paintings and photographs used in this research project). The main goal of this project is to increase the ability of current methods of transferring style from one image to another when working with images that contain noise or low resolution. In this work, we are proposing a unified system that includes both a Denoising Autoencoder that scrubs and cleans the input image, and a VGG19-based network that processes the cleaned image with the artistic style. As a result, high-quality and detailed stylized images will be produced, bridging the gap between enhancing an image and making an image that is a work of art.

II. PROJECT GOALS AND OBJECTIVES

The goals associated with this project are focused primarily on data and performance metrics selected for Image Style Transfer. The first goal is to build an appropriate dataset by combining and normalizing three separate datasets from Kaggle together, adding in Data Augmentation to ensure sufficient variation in the dataset for the purpose of training. The project also will focus on the creation of a Denoising Auto Encoder that should be able to achieve a minimum of about 80% Test Pixel Accuracy for filtering out noise from the incoming images. Through these pre-processing techniques, the overall objective of the project is to build a VGG19-based neural network that is designed to minimize the amount of Content and Style Loss when used to generate the final output, thereby preserving the integrity and coherence of the original artwork being referenced.

III. RELATED WORKS (LITERATURE REVIEW)

Paper [1] follows in depth and structure how the entire image-matching pipeline, Feature-based image matching lies at the heart of many computer vision and photogrammetric tasks because it solves the fundamental correspondence problem: identifying the same physical points across different images so their geometric relationships can be estimated. This capability underpins everything from classical aerial image orientation to modern 3D reconstruction, SLAM, and Structure-from-Motion. The standard workflow usually follows a five-stage pipeline in which distinctive points are detected, their affine shape is estimated to compensate for viewpoint distortions, a dominant orientation is assigned to ensure rotational stability, a descriptor is computed from a normalized window around each feature, and finally, descriptors are compared to establish correspondences. While early methods such as SIFT, SURF, Harris, and FAST achieved invariance through carefully engineered mathematical designs, recent research has pushed toward learning-based models like LIFT and covariant detectors that try to learn directly from data. However, as this survey points out, the shift to deep learning has not solved all underlying challenges. Many widely used training datasets—for instance, the Brown/PhotoTourism patches—are generated from limited multi-view reconstructions, meaning that positive samples usually differ only slightly in viewpoint and artificially mined negatives do not reflect true wide-baseline variation. This causes learned features to perform impressively in controlled

tests yet degrade noticeably with oblique aerial imagery, severe geometric distortions, or when transferred across domains. The situation is even more constrained in photogrammetric datasets, where dense ground truth is scarce, especially for oblique aerial collections like Nex et al. (2015), making it difficult to evaluate matching quality beyond rough bundle-adjustment results. The review also notes that deep networks still fail to jointly learn affine shape and orientation reliably, and their ability to generalize from one imaging domain to another remains limited. Moreover, existing public benchmarks tend to emphasize consecutive views rather than the wide-baseline, off-nadir, or multi-sensor scenarios common in real-world applications. For these reasons, the authors argue that progress in the field depends on building richer and more representative datasets, designing evaluation protocols that consider full geometric consistency rather than isolated descriptor metrics, and developing feature-learning strategies that incorporate domain knowledge instead of relying solely on generic deep architectures.

Paper [2] offer one of the most practically grounded contributions to wide-baseline image matching by shifting the focus away from the usual patch-level metrics and toward what genuinely matters for real applications: how well a full matching pipeline can recover accurate camera poses. Instead of evaluating detectors or descriptors in isolation, they build a large, carefully curated benchmark the Image Matching Challenge constructed from 25 PhotoTourism scenes with COLMAP-derived camera poses and cleaned depth data, allowing them to assess pipelines end-to-end using mean Average Accuracy across both stereo and multi-view tasks. One of the most striking findings is that the performance of a system depends far more on how feature extractors interact with the robust estimator, especially RANSAC and its many variants, than on the specific detector or descriptor used. Their 6 experiments repeatedly show that common assumptions in the community do not always hold: a well-tuned classical configuration, such as SIFT paired with MAGSAC++, can outperform many modern deep learning-based features, even though deep methods tend to dominate on proxy metrics like repeatability or descriptor similarity. In fact, the authors illustrate how strongly these proxy evaluations can diverge from real geometric performance, with features that excel in controlled, small-baseline settings often faltering when confronted with the viewpoint and illumination changes typical of wide-baseline reconstruction. Their modular framework also exposes clear dataset-specific behaviors—thresholds and parameter settings that work for one scene or feature may completely fail on another—highlighting the sensitivity of SfM pipelines to seemingly minor choices. Overall, the paper underlines that real progress in wide-baseline matching depends not only on inventing new descriptors but on understanding and tuning the entire pipeline, and it offers rare, practical insight into what truly constitutes state-of-the-art performance when evaluated against downstream pose accuracy rather than simplified laboratory metrics.

Paper [3] introduce a diffusion-based stylization framework that tackles one of the core problems in image style transfer: how to apply an artistic style convincingly without destroying the structure or recognizability of the original image. Instead of relying on older approaches that summarize style with a single Gram matrix or global texture statistics, their method begins by building a much richer style distribution using Stable Diffusion equipped with Style-LoRA modules, allowing the system to capture the fine-grained statistical behavior of the reference artwork. The heart of their approach is the Style Matching Score, a KL-based objective that treats stylization as a distribution-matching problem and helps the generated image adopt the target style more faithfully. To keep the content intact, the authors add two mechanisms that work together: Progressive Spectrum Regularization, which steers the stylization process from low-frequency structure to high-frequency detail so that the overall layout of the original image remains stable while textures and brushstrokes appear later; and Semantic-Aware Gradient Refinement, which uses pixel-level relevance maps so that important regions like faces or key objects retain their shape while less important areas absorb more of the style. They validate the method on the 280-image PIE benchmark and a feed-forward version trained on the 60k-image LHQ dataset, comparing it against text-guided diffusion methods, exemplar-based transfer, and standard LoRA-based stylization. Across metrics such as LPIPS, CFSD, FID, and ArtFID, and in human preference studies, their method generally achieves the best balance between stylistic richness and content preservation. Although pure LoRA approaches can sometimes push the style more aggressively, SMS consistently produces results that look both artistic and structurally coherent, and the model avoids the instability often seen in GAN-based systems. The authors also highlight that the feed-forward variant can operate in real time, making the approach practical for interactive applications and not just academic demonstrations.

Paper [4] aimed to present a comprehensive overview of the progress in Neural Style Transfer (NST) using deep learning. The study's goal is to classify, analyze, and compare major approaches for transferring artistic onto natural images while preserving content structure. It utilized datasets such as MS-COCO, WikiArt, and ImageNet, which are commonly used to train and evaluate style transfer models. The algorithms that was conducted are CNN optimization methods, feed-forward networks, GAN-based architecture like CycleGAN and StyleGAB, and Transformer-based models such as StyTr2. Additionally, the study highlighted four feature selection techniques which are Adaptive Instance Normalization (AdaIN), Whitening and Coloring Transform (WCT), Conditional Instance Normalization (CIN), and attention-based normalization modules for enhancing semantic consistence between the content and the style.furthermore, for evaluation, they have used both qualitative and qualitative evaluations, employing metrics such as SSIM, PSNR, and FID to evaluate structural similarity, perceptual quality, and efficiency. The results showed that CAST and SANet achieved the best balance between content preservation and

style fidelity. However, the study showed some limitations which are high computational cost, instability while optimization, and limited generalization to unseen styles. Future research should focus on improving efficiency, adaptability, and semantic depth in deep learning-based image style transfer.

Paper [5] aimed to improve the CycleGan model for unpaired image style transfer to make it capable of producing a higher quality images. Their main goal was to enhance the image clarity and reduce the computational cost by modifying the generator and discriminator structures. The dataset that was used it Monet2Photo dataset which contains unpaired Monet paintings and real photographs, this dataset allows testing the model's ability to transfer the painting styles onto real photo without the need for paired examples. In this work, the paper replaced the traditional ResNet generator by a UNet generator, which skips connections to keep the global and the fine image details. Then, the PatchGAN discriminator is redesigned using depth separable convolutions to reduce the number of parameters and to improve the training efficiency. For testing the model's performance, the authors tested three loss functions which are L1, L2, and smooth L1. And two metrics have been used, the first metric is Peak Signal-to-Noise Ratio (PSNR) which measures how close the generated image is to the original, and the second metric is Structural Similarity Index Measure (SSIM) which measures the similarity in structure, brightness, and contrast. The results showed that L2 loss model achieved the best and most stable results, generating images with the highest PSNR and SSIM values. Also, the discriminator's parameters were reduced from 2.76M to 2.15M, which shows improving the model's efficiency without losing the image quality. However, the authors mentioned that the model still faces overfitting, high computational cost, and limited generalization to other datasets. These issues limit its scalability to broader and more diverse image translation tasks.

Paper [6] provided a comprehensive review of Neural Style Transfer (NST) methods, focusing on both image and video style transfer. The goal is to summaries the advancements in NST, such as highlighting current architectures like CNNs and GAN-based models, and identify the advantages, limitations, and gaps in real time NST systems. Also, the datasets that were used are MS-COCO and Cityscapes which are commonly used for object detection and segmentations tasks. Additionally, the study covered multiple deep learning algorithms applied in NST, such as CNNs, GANs, and CycleGANs. The key architectures discussed include DCGAN, CartoonGAN, Artsy-GAN, StyleBank, AdaIN, and ReCoNet. Each one was analyzed in terms of architecture, loss functions, and use sauce such as real-time video style transfer and artistic rendering. Furthermore, NST primarily relies on feature extraction from pre-trained CNNs rather than explicit feature selection. Then, the extracted features are used to represent content and style during the transfer process. The studies reviewed in the paper employ quantitative and qualitative evaluations. Quantitative metrics include FID, SSIM, and PSNR for measuring image quality and realism. Qualitative evaluations focused on visual perception, temporal consistency in videos, and user preference studies. For the results, the paper highlighted that GAN-based architectures, which are Artsy-GAN and CartoonGAN achieved faster training, better image quality, and higher diversity compared to traditional NST of CycleGAN models. Artsy-GAN was found to be up to 74.96% faster than CycleGAN at high resolutions. However, the paper pointed out major lineations such as high computational cost and difficulty maintaining color and temporal consistency.

Paper [7] presents an innovative use of neural style transfer for enhancing CAPTCHA security by blending textual content with artistic styles through a VGG 19-based model optimized using L-BFGS. The authors evaluate their approach on a dataset of 600 CAPTCHA samples collected from six websites, demonstrating a significant decrease in automated recognition success while maintaining human readability. Although conducted in a different context, the paper effectively demonstrates how deep style transfer can be adapted to preserve essential content features while introducing stylistic transformations. Its findings highlight the impact of controlled loss weighting between content and style, offering useful direction for achieving balanced and visually appealing transformations in image-based applications.

Paper [8] provides a comparative overview of deep learning techniques for image style transfer, focusing on the distinctions between CNN- and GAN-based frameworks. CNN-based models, such as Gatys' neural style transfer, are recognized for maintaining image structure but are computationally demanding, whereas GAN-based methods like CycleGAN and Cartoon-GAN achieve faster results with occasional loss of detail. The discussion emphasizes the importance of balancing speed and quality, an aspect critical for real-world systems that process multiple images in a consistent visual format. The review also notes the need for lightweight architectures capable of handling diverse styles efficiently, which can guide future efforts in optimizing model design and resource usage.

Paper [9] offers an extensive review of image and video style transfer methods, outlining the progression from early texture-based approaches to deep learning architectures such as CNNs, GANs, and attention-based networks. It highlights how diverse datasets like ImageNet, COCO, and WikiArt contribute to improving the generalization of stylization models across various domains. The study also points out persistent challenges, including limited detail preservation, high computational cost, and instability during training. These observations provide valuable insights into developing more efficient and stable models for consistent stylization tasks, particularly in applications requiring both accuracy and visual coherence. The paper further discusses evaluation metrics and loss functions that can serve as benchmarks for assessing new style transfer implementations.

Paper [10] provide a comprehensive survey of the neural style transfer (NST) literature up to their cut-off date. The review offers (1) a taxonomy of NST methods (image-optimization vs. model-optimization; parametric summary-statistics vs. non parametric patch/MRF methods; single-style vs. multi-style, photorealistic vs. artistic, etc.), (2) a detailed explanation of the foundational Gatys et al. formulation (Gram-based style loss + perceptual content loss), and (3) a systematic discussion of follow-up directions: fast feed forward nets (Johnson, Ulyanov), adaptive normalization (AdaIN), whitening/coloring transforms, patch-based methods, semantic/photorealistic approaches, GAN and collection based transfers, and extensions to video/3D. The survey also compares evaluation metrics (user studies, perceptual metrics, stylization fidelity vs. content preservation), highlights common failure modes (structure loss, instability, lack of photorealism), and outlines open challenges and promising research directions (controllable transfer, evaluation standards, semantics, and photorealistic constraints).

Paper [11] target the semantic mismatch problem in style transfer: when content and style images do not share the same object categories, naive transfer often produces semantically incoherent results (e.g., applying building strokes to a tree). They introduce semantic context matching, a strategy that leverages contextual co-occurrence statistics (computed from annotated datasets such as ADE20K) to relax strict category matching and find contextually similar region pairs between content and style images. Following matching, the method uses a hierarchical local-to-global network: a local context network transfers detailed stroke patterns between matched region pairs, producing intermediate local style transfer images; a global context network then fuses local results with global content/style cues to ensure image-wise coherence and reduce boundary artifacts. 10 Quantitative evaluations and a user study show that context-aware local transfer plus a global refinement yields stylizations more consistent with human perception than prior methods.

Paper [12] identify a recurring failure mode in neural style transfer methods: standard feature-based representations (e.g., VGG activations + Gram statistics) tend to scatter style textures across the whole image and thus disrupt important image structure, especially for inputs with complex spatial layouts or multiple depth layers. To address this, the paper proposes augmenting the usual content/style representation with an explicit structure representation composed of (1) a global structure signal (estimated depth map) and (2) a local structure signal (edge maps). The method trains a feed-forward generator under a composite loss that mixes perceptual (content & style) losses with depth and edge reconstruction losses. Architecturally the system uses two representation subnets (one for content/style features and one for structure features) together with a generator network; losses from both subnets guide stylization so that textures are applied while preserving global layout and local edges.

The article [13] Provides a thorough literature study related to deep learning-based style transfer between images, with emphasis on both image iterative and model iterative algorithms. The applications of diverse deep learning algorithms based on CNN, including VGG19, Markov Random Fields + CNN, CycleGAN, and fast feedback neural networks, are mentioned. The authors have mentioned the commonly cited datasets, including COCO and ImageNet, to compare algorithms with each other. The authors presented a solution related to the comparison between optimization algorithms and model algorithms, including the use of perceptual loss functions and conditional normalization to enhance multi-style transfer, resulting in images suitable for real-time transfer. The authors concluded that both model algorithms and deep learning algorithms are efficient, resulting in successful transfer and translation, respectively, but iterative algorithms are slow, unstable, or lack detailed texture, with subjective visual comparison as the primary metric of evaluation.

Paper [14] by Ruikun Wang. Describes a deep learning-based technique that incorporates cGANs with feedforward residual network functionality to automate the process of converting sketches to stylized images with the aim of greatly reducing manual labor in animation and art. The model includes a generator based on “U-Net” and a PatchGAN discriminator with a perceptual loss function computed with the help of a pre-trained VGG-16 network. Utilizing 80,000 resized images from the COCO dataset, the authors have trained this system to map sketches to fully colored, stylized images. It consists of creating conditional images followed by style transfer with an Adam-based optimization of losses representing content, style, and total variation. High-quality, realistic experimental results demonstrate performance comparable to the classic approach of Gatys et al. while running almost 1000 times faster, with a run time of about 10 seconds per image. However, this work also has some drawbacks: it requires a long model training time, sometimes loses fine details in the images, and needs better feature extraction for the replication of detailed artistic textures.

Paper [15] Proposes a novel framework that will renew image style transfer by incorporating Diffusion Models with AdaIN into high quality, stable, and flexible stylization. The research tries to overcome the common drawbacks of previous approaches, such as over-stylization, mode collapse, or poor content preservation. StyDiff incorporates a VDVAE encoder conducting multiscale feature extraction, AdaIN for accurately mixing the content and style features, and a multi-component loss function (content, style, element, and diffusion) to balance style fidelity with structural consistency. The network is trained with 100,000 images with content from the COCO dataset and 68,669 images** with style information from WikiArt, which are diverse images of various scenes and art styles. Once the reverse diffusion is accomplished, images are restored by StyDiff, which refines images by removing noise while following a desired style accurately. Experimental results showed that StyDiff surpassed current state-of-the-art approaches like StyleFlow, AdaAttN, and Stable Diffusion by

achieving the highest **SSIM**-0.7986-and the lowest LPIPS-0.05, hence presenting the best perceptual quality with well-preserved content. In the user study conducted with 100 participants, higher satisfaction with the outputs of StyDiff than the baseline further supported this result. This paper, however, has its own limitations, such as high computational cost and the use of a limited-scope dataset, which will affect the efficiency and generalization. Following that, future work will focus on the pursuit of better computational efficiency by optimizing models, the expansion of systems into video style transfer tasks with consideration of temporal consistency, and developing applications for interactive, real-time scenarios. On the whole, StyDiff does offer a strong, polished solution to image style transfer problems. The approach strikes a good merger between creativity and conservation. On the whole, StyDiff does offer a strong, polished solution to image style transfer problems. The approach strikes a good merger between creativity and conservation.

Paper [16] states that the Chinese painting requires advanced skills and years of training, and most of the existing transfer methods focus on Western painting, which differs from Chinese art. As a result, the author presents a novel method of Chinese printing style transform (CPST) that uses pre-trained VGG-19 to extract image features. The methodology focuses on transferring natural images into Chinese painting with special attention to the Chinese painting characteristics that include ink tone, brush stroke, space reservation, and yellowing. The methodology begins with classifying the images as freehand brushwork or fine brushwork using GoogLeNet V3 to guide the transfer strategy. The results show that the CNN model embedded the Chinese painting characteristics, and when compared with other existing methods, it outperformed the methods successfully and reproduced Chinese paintings using deep learning. Despite the success of this method, the author states that the method cannot handle figure painting yet, so future work can improve the CPST on this task.

Paper [17] discusses that most of the common transfer methods do not appropriately apply the color and the texture to the content image. This paper solves this issue by presenting a novel method that is a deep convolutional neural network (DCNN), which is based on the VGG19 model and a local and global style loss function. The style loss function is defined in several layers in the network to keep the detailed styles, while the global loss style is defined in several layers to preserve more global information. The proposed approach was compared with several methods, and the results clearly show that this model successfully preserved the color and the texture of the image while transferring the style of the image.

Paper [18] states that it’s a common issue in the field of image processing to match different images. This paper tries to solve this problem by introducing an intelligent method to match satellite images with aerial images. The process starts by using a generative adversarial network (GAN), specifically a smooth cycle consistent, to convert the satellite images into aerial images using style transfer. After that, both the D2-Net and the LoFTR models are used to match between the original aerial image and the generated aerial image. Finally, the transformation relationship is mapped to the serial-aerial images to get the final results. The results, after testing this method on several test sets, show that this method is successful and can improve the accuracy and the robustness of the matching algorithm. Since this method was applied only to these types of images, the author suggests future studies to apply this method to different image types, such as the infrared and optical images, and analyze the results.

IV. Data Acquisition and Preprocessing

In our project, three different datasets were used. The first dataset is the "Artistic Images for Neural Style Transfer" obtained from the Kaggle website that contains 49 images specially designed for style transfer projects [19]. The second dataset is the "Artistic styles 30k images dataset" that contains 68 high-quality artistic images, also designed to be used for style transfer tasks. It includes diverse styles such as Cartoon, Graphics, and many other categories [20]. The last dataset is the "Images for Style Transfer" which contains 37 content and style images for style transfer tasks [21]. These three datasets were combined and processed to ensure that no duplicate images were present resulting in a total of 130 unique images. After combining these three datasets, a normalization step was performed to scale the pixel values to a consistent range. The final dataset was then split into 80 % training set, 10 % validation set, and 10 % testing set.

Since the available dataset is still small, several augmentation steps were performed on the training and validation sets to increase the size of the dataset. First each original image was added to the augmented dataset without any modification. Next, we applied horizontal flipping that mirrors the image along the vertical axis. Followed by intensity scaling where each image was multiplied by predefined scaling factor from a list to improve the model's robustness to illumination variations. Moreover, brightness adjustment was performed similar to the intensity scaling where each image was multiplied by a brightness factor from a list. After that, a rotation step was applied where each image was rotated by a multiple of 90 degrees. Importantly, each image was augmented using one value from each augmentation list, resulting in five versions in the augmented dataset (original, horizontally flipped, intensity scaled, brightness adjusted, and rotated). All these augmentation steps do not rely on a random generator or function to ensure reproducibility. Together all these augmentation steps increased the size of the dataset to a total of 468 images and improved the model's generalization ability.

Since a denoising autoencoder was used in this project, Gaussian noise was added using a seeded random generator to allow the model to learn to reconstruct clean images from noisy inputs.

V. Model Development and Training

The modeling process in this project was designed as a two-stage pipeline, as shown in Figures 1 and 2. The former step is concerned with training a clean and stable image representation with the help of a denoising autoencoder, and the latter employs the neural style transfer with a pre-trained VGG19 network. This design choice was made because preliminary experiments indicated that style transfer is extremely sensitive to noise as well as minor visual artifacts in the input images. The textures in the stylized results were usually not stable, and unwanted distortions were also often found when using the raw images directly. The system adds an autoencoder as a preprocessing step to transform more structured and cleaner inputs, which results in more stable and visually appealing style transfer outputs.

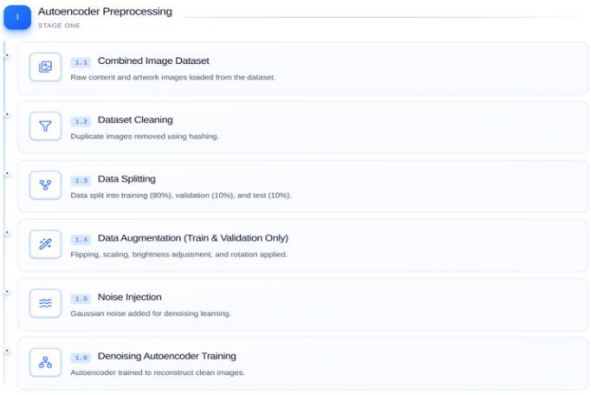

Fig. 1. Stage one: Autoencoder-based image preprocessing.

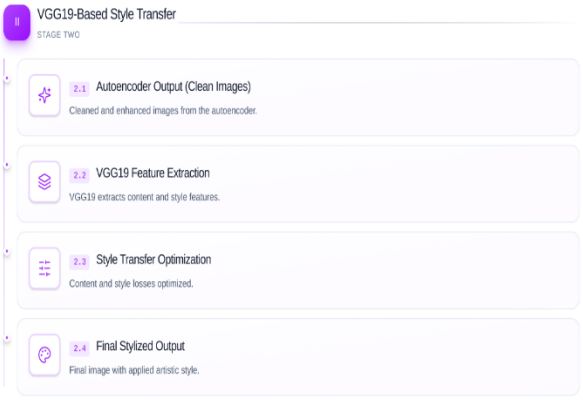

Fig. 2. Stage two: VGG19-based neural style transfer.

The autoencoder has a convolutional encoder-decoder design. The encoder is constructed in terms of stacked Conv2D layers with ReLU activation and thereafter, MaxPooling layers, which successively reduce the spatial resolution while capturing important visual features. The more depth the network has, the more filters it has to enable the model to learn richer and more abstract representations. UpSampling layers and convolutional layers are used by the decoder to reflect this structure to reconstruct the image in the original resolution. The last layer employs a sigmoid activation function to guarantee that the pixel values that are being reconstructed are within a useful range. Mean squared error was used as the reconstruction loss to train the model, as the goal is to reduce the difference between pixels of the input image and the output one. Also, the pixel accuracy measure is employed as a custom metric to evaluate the similarity of the reconstructed pixels to the original image within a low tolerance, which produces a more intuitive quality metric of reconstruction.

This training setup encouraged the autoencoder to learn a genuine denoising behavior, improving its ability to generalize to unseen images. After the dataset preparation and preprocessing steps described in the previous section, the resulting clean dataset was used for training the autoencoder. The data splits were used to support model training, validation-based tuning, and unbiased performance evaluation. During training, data augmentation was applied only to the training and validation sets to improve robustness while preventing data leakage into the test set. Gaussian noise was added to the training inputs to encourage the autoencoder to learn a denoising behavior rather than memorizing clean images. This training setup encouraged the autoencoder to learn a genuine denoising behavior, improving its ability to generalize to unseen images.

An autoencoder was trained using the Adam optimizer with an initial learning rate of 1×10^(-4) and a batch size of 4 for a maximum of 100 epochs. The validation loss was monitored with early stopping, and the training was terminated in case no further improvement was recorded, and the best-performing model weights are restored. The learning-rate scheduler was also used to slow the learning rate when the validation performance stopped increasing, which assisted in stabilizing the convergence in subsequent stages of training. Both reconstruction loss and pixel accuracy improved steadily during the training. The last validation pixel accuracy was around 88%, which means that the model could reconstruct clean images. The autoencoder, when tested on the unseen test set, reached an accuracy of approximately 81% on test pixels, confirming that the learned representations are highly generalizable even when compared to the training data.

The autoencoder became a preprocessing stage after training, which was used in the neural style transfer stage. The content and style images of high resolution were initially down sampled and sent through the trained autoencoder to create cleaned versions. This clean output was then mixed with the original image to conserve sharp edges, and at the same time use the advantages of the denoised and improved texture that the autoencoder had learnt. This measure meant that the style transfer procedure would be run on a refined image upon which the style transfer is carried out, instead of a noisy input. In the case of neural style transfer, the network employed was the VGG19 model pre-trained as a fixed feature extractor, with all weights frozen during optimization. Content representations were extracted from deeper convolutional layers to preserve the overall image structure, while style information was captured from earlier layers to represent textures and visual patterns at different scales. Gram matrices were used to model the style information, and the content loss

was computed as the difference between content feature representations. The cleaned content image was used to initialize the generated image, which was then optimized using the Adam optimizer for 1500 steps. The final objective function was defined as a weighted combination of content and style losses, allowing control over the balance between structural preservation and artistic appearance, based on ideas adapted from an existing Kaggle implementation [22]. The two-step process led to stylized results that preserve the content structure and clearly reveal the visual attributes of the chosen style image, which are more stable than the use of style transfer on original images.

VI. Evaluation and Analysis

The evaluation of the image processing pipeline has been performed on two different levels. In the first instance, the assessment of the Autoencoder was carried out based on pixel accuracy, reaching a result of 88% on the evaluation set and 80% on the test set.

The diagram below shows Validation Accuracy Curve: The accuracy curve reported a fast convergence to a stable point at 88% This clearly suggested that the network quickly optimized the performance on image reconstruction tasks and generalized the cleaning process on unseen samples, thus proving the robustness of the network.

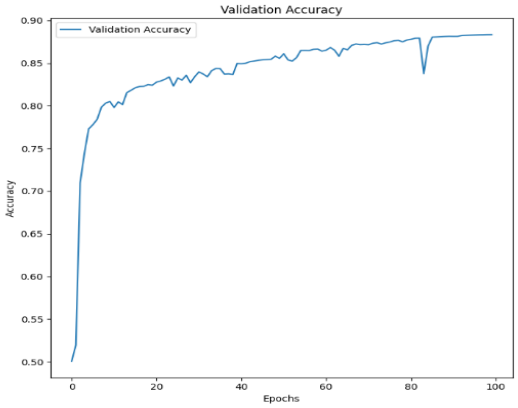

Fig. 3. Validation Accuracy Curve.

The below diagram shows Loss Curves: The training and validation loss curves were steadily parallel to each other and close to zero. This is a very essential observation that confirms the stable learning process with less overfitting, thus proving that the network's weights are optimally trained for generalizing the image cleaning task.

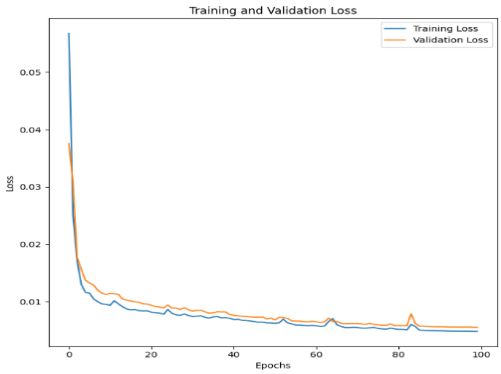

Fig. 4. Training amnd validation loss.

Secondly, the VGG19-based style transfer facilitated the stylization of the content, carrying it forward along with stylistic patterns. The style transfer is confirmed to be apt, with a need for improvement in terms of perceptual evaluation, loss function weighing, and variability in the dataset.

The below figure shows the VGG Neural Style Transfer image aptly combines the content with the style, which has a commendable balance between Content Loss and Style Loss. The generated image still lacks fidelity when it comes to the fine details (windows, which are blurred).

Fig. 5. VGG Neural Style Transfer image.

Analysis of Results

The VGG Neural Style Transfer system is very slow because it uses iterative optimization. Although pre-cleaning with an autoencoder improves visual quality, the down sampling and up sampling process generally removes details.

Potential Improvements

Speed-up: convert to feed-forward neural networks, or use GANs, to produce single-pass solutions instantaneously. Enhance the quality of the resulting image by applying super resolution, which uses total variation loss to maintain a smooth image.

VII. Conclusion(and Future Work)

The project has been a success in developing a high-quality deep learning system that combines two very essential components: Denoising Pre-processing, followed by Neural Style Transfer (NST). The Denoising Autoencoder worked exceptionally well (Test Pixel Accuracy of 80%), resulting in clean, refined inputs, which led to a balance in the resulting image generated from the VGG19 network style transfer, thereby capturing the structure of the content image while possessing the rich style of the reference image. Implications &Potential Application. The successful blend of these two models offers a variety of application domains:
Digital Art & Media: High-quality stylized asset creation for the graphic design and gaming sectors. Image Enhancement: This is pre-processing a noisy image, such as a low-light image, before applying creative filters. Future research may concentrate on enhancing the performance on the generalization set, handling multiple art styles, simplifying the learning process, and extending the technique to video-to-video style transfer tasks.

Appendix

The implementation code and pre-trained models are available at:https://github.com/Elnabrees/DL_Project_Code_Style_Transfet/blob/main/DL_Project_Code_Style_Transfet.ipynb

References

[1] L. Chen, F. Rottensteiner, and C. Heipke, "Feature detection and description for image matching: from hand-crafted design to deep learning," Geo-spatial Information Science, vol. 24, no. 1, pp. 58–74, 2021, doi: 10.1080/10095020.2020.1843376.

[2] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls, "Image Matching across Wide Baselines: From Paper to Practice," arXiv preprint arXiv:2003.01587, 2020. Avilable2003.01587v5.pdf.

[3] Y. Jiang, L. Jiang, S. Yang, J.-W. Liu, I. W. Tsang, and M. Z. Shou, "Balanced Image Stylization with Style Matching Score," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2025 (Open Access version by CVF; final on IEEE Xplore). Available : 1-s2.0 S0079642525001045-main.pdf.

[4] H. Li, L. Wang, and J. Liu, "A review of deep learning-based image style transfer research," The Imaging Science Journal, pp. 504–526, Oct. 2024, doi: 10.1080/13682199.2024.2418216.Available:https://doi.org/10.1080/13682199.2024.2418216.

[5] Y. Liao and Y. Huang, "Deep Learning-Based Application of Image Style Transfer," Mathematical Problems in Engineering, vol. 2022, pp. 1–10, Aug. 2022, doi: 10.1155/2022/1693892. Available: https://doi.org/10.1155/2022/1693892.

[6] A. Singh et al., "Neural Style Transfer: A Critical review," journal-article,Sep.2021,doi:https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9539183.

[7] H. Kwon, H. Yoon, and K.-W. Park, "CAPTCHA Image Generation Using Style Transfer Learning in Deep Neural Network," Lecture Notes in Computer Science, pp. 234–246, Jan. 2020, doi: https://doi.org/10.1007/978-3-030-39303-8_18.

[8] S. Ren and Y. Sheng, "Image Style Transfer Using Deep Learning Methods," IEEE Xplore, Feb. 01, 2022. https://ieeexplore.ieee.org/document/9745023

[9] Y. Xu, M. Xia, K. Hu, S. Zhou, and L. Weng, "Style Transfer Review: Traditional Machine Learning to Deep Learning," Information, vol. 16, no. 2, p. 157, Feb. 2025, doi: https://doi.org/10.3390/info16020157.

[10] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural Style Transfer: A Review," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 11, pp. 3365–3385, Nov. 2020, doi: https://arxiv.org/abs/1705.04058.

[11] Y.-S. Liao and C.-R. Huang, "Semantic Context-Aware Image Style Transfer," IEEE Transactions on Image Processing, vol. 31, pp. 1911–1923, 2022, doi: https://ieeexplore.ieee.org/document/9709704.

[12] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure Preserving Neural Style Transfer," IEEE Transactions on Image Processing, vol. 29, pp. 909 920, 2020, doi: https://ieeexplore.ieee.org/abstract/document/8816670 .

[13] Y. Shen, G. Tang, and Q. Xu , (PDF) studies advanced in image style transfer based on Deep Learning, ,https://www.researchgate.net/publication/369872081_Studies_Advanced_in_Image_Style_T ransfer_based_on_Deep_Learning .

[14] R. Wang, "Research on image generation and style transfer algorithm based on Deep Learning," SCIRP, https://www.scirp.org/journal/paperinformation?paperid=94650.

[15] Y. Sun and H. Meng, "Stydiff: A refined style transfer method based on Diffusion Models," Nature News, https://www.nature.com/articles/s41598-025-17899-x.

[16] J. Sheng, C. Song, J. Wang, and Y. Han, "Convolutional Neural Network Style Transfer Towards Chinese Paintings ",IEEE Access ,vol. 7, pp. 163719–163728, Jan. 2019 ,doi : https://doi.org/10.1109/access.2019.2952616.

[17] H.-H. Zhao, P. L. Rosin, Y.-K. Lai, M.-G. Lin, and Q.-Y. Liu, "Image Neural Style Transfer With Global and Local Optimization Fusion ",IEEE Access ,vol. 7, pp. 85573–85580, 2019 ,doi :https://doi.org/10.1109/access.2019.2922554.

[18] J. Zhao, D. Yang, Y. Li, P. Xiao, and J. Yang, "Intelligent Matching Method for Heterogeneous Remote Sensing Images Based on Style Transfer ",IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing ,vol. 15, pp. 6723–6731, Jan. 2022 , doi :https://doi.org/10.1109/jstars.2022.3197748 .

[19] S. Jha, "Artistic Images For Neural Style Transfer," *Kaggle.com*, 2024. https://www.kaggle.com/datasets/skjha69/artistic-images-for-neural-style-transfer

[20] simon graves, "Artistic styles 30k images dataset," *Kaggle.com*, 2025. https://www.kaggle.com/datasets/simongraves/artistic-styles-dataset (accessed Dec. 13, 2025).

[21] Soumik Rakshit, "Images for Style Transfer," *Kaggle.com*, 2018. https://www.kaggle.com/datasets/soumikrakshit/images-for-style-transfer

[22] alfredkondoro, "PyTorch - ImageStyle Transformers," Kaggle.com, Apr. 04,2024.https://www.kaggle.com/code/alfredkondoro/pytorch-imagestyle-transformers (accessed Dec. 13, 2025).