ArtifyAI: Text-Guided Artistic Image Enhancement and Stylization via Multi-Stage Diffusion and CLIP-Based Style Conditioning Proceedings

Ghaida Abdullah Alsabti
Princess Nourah Bint
Abdulrahman University

Ghaida Aziz Alanazi
Princess Nourah Bint
Abdulrahman University

Hessah Mohammed AlMujally
Princess Nourah Bint
Abdulrahman University

Wasan Mahdi Alomar
Princess Nourah Bint
Abdulrahman University

Abstract

*Today, images are everywhere—millions are taken and shared every day. Making these images look better has been done in many ways, from basic filters like the mean filter to more advanced AI models like CNNs. Recently, there's also been growing interest in changing the style of images to make them more creative. In this project, we created **ArtifyAI**, a system that uses the CodeFormer model to improve image quality while keeping important details. We also built a flexible setup that uses Canny Edge Detection, Stable Diffusion, and ControlNet to add animation and artistic styles. Our tool makes images clearer and more creative by combining enhancement with generative AI.*

1. Introduction

In today's digital culture, visual content is a primary means of connection and self-expression. People constantly share images to capture meaningful moments and preserve personal memories. However, when these images suffer from poor quality—due to low resolution, noise, or age—their emotional and aesthetic value can be lost. While numerous tools exist for either enhancing photos or applying filters, few offer a unified solution that intelligently restores image quality while also enabling creative transformation through artistic style.

To address this gap, we set out to build a system that uses generative AI not only to enhance degraded images but also to infuse them with stylized, artistic qualities aligned with popular trends on social media. Initially, our goal included animating still images, but due to limitations in model compatibility and computational resources, we refined our focus. This pivot allowed us to design a more lightweight and accessible solution that enhances photos and applies dynamic, visually compelling styles—bridging the gap between restoration and creativity.

This paper details the final version of our system, the underlying tools and models, and the design choices that shaped its development.

2. Related Work

ArtifyAI builds on advances in two primary research domains: **Image Enhancement** and **Text-Guided Image Generation using Diffusion Models**. The integration of methods from both areas allows ArtifyAI to intelligently enhance user-provided images and apply semantically relevant artistic styles with precision and control.

2.1. Image Enhancement

Image enhancement techniques have seen notable progress in recent years, particularly in face restoration. GFPGAN (Generative Facial Prior GAN) [1] emerged as a powerful solution for facial image restoration by leveraging generative priors. However, during the development of ArtifyAI, we encountered compatibility issues between GFPGAN and our broader system architecture. As a result, we adopted the **CodeFormer** model [2], which proved more suitable for our needs. CodeFormer offers robust face enhancement while maintaining fine details and delivering more consistent outputs across varied image inputs. Its compatibility and performance made it a more practical and effective component within our end-to-end enhancement pipeline.

2.2. Text Guided Image Generation with Diffusion Models

For artistic stylization, ArtifyAI leverages text-guided diffusion models to generate stylistic variations of images based on user-defined or AI-suggested prompts. This approach moves beyond traditional style transfer, which relies on reference style images, and instead uses natural language to control visual outputs.

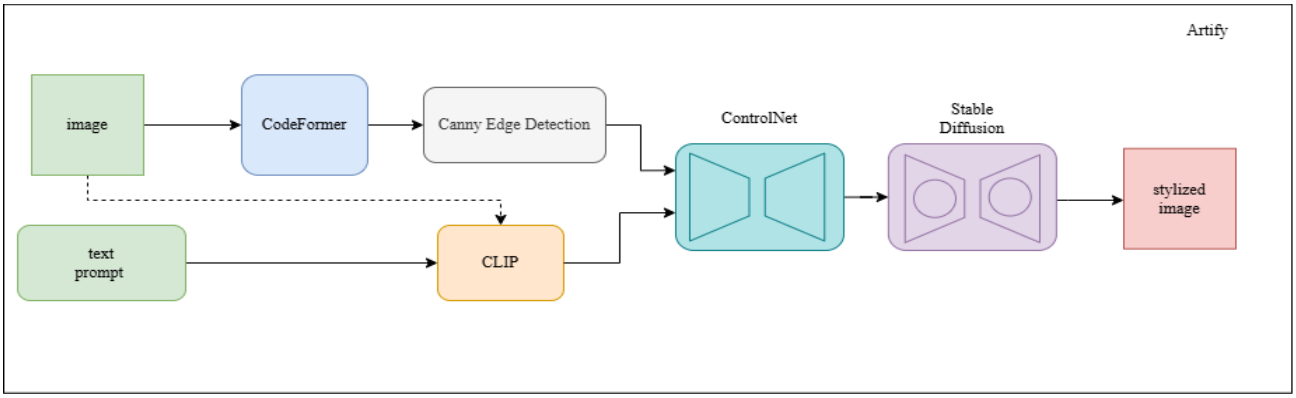

Figure 1: System flow of the Artify application. The dotted line denotes that CLIP's analysis of the image is optional.

The backbone of our system is Stable Diffusion [3], a powerful text-to-image generation model. To ensure spatial coherence and to preserve the structure of the input image during stylization, we incorporate a Canny edge detection module [4] in conjunction with ControlNet [5], which allows the diffusion model to follow the contours and layout of the original image. This combination enables precise stylization while retaining essential image content.

To intelligently guide the choice of artistic styles, we integrated CLIP (Contrastive Language-Image Pretraining) [6], which enables ArtifyAI to analyze the input image and suggest the most appropriate styles from a curated list. This CLIP-based suggestion mechanism ensures that the applied artistic modifications are not only aesthetically pleasing but also contextually relevant to the content of the image.

By combining these techniques, ArtifyAI delivers a seamless and interactive image transformation experience, merging state-of-the-art enhancement with text-driven creativity.

3. Data Description

ArtifyAI is designed to operate without dependence on a curated training dataset or the need for fine-tuning existing models. Instead, it utilizes a fully inference-based approach, processing multimodal inputs consisting of both images and textual style descriptions. This design choice ensures broad adaptability and user accessibility.

The system accepts user-uploaded images that can vary significantly in quality, resolution, and content. These inputs may include aged or degraded photographs—such as digitized family portraits, low-resolution snapshots taken with mobile devices, or even creative inputs like hand-drawn sketches or paintings intended to guide artistic transformations. Importantly, ArtifyAI does not require these images to conform to any fixed format or resolution, allowing the platform to support real-world use cases across both casual and professional domains.

Text input plays a central role in defining the artistic direction for image transformation. Users can specify stylistic preferences using natural language prompts (e.g., "Van Gogh style" or "cyberpunk cityscape"), which are then interpreted by the system to drive the stylization process via diffusion models. This enables a highly flexible interface where the user has creative control without the complexity of supplying reference images.

By relying entirely on pre-trained models such as CodeFormer, CLIP, and Stable Diffusion with ControlNet, ArtifyAI avoids the computational cost and rigidity associated with supervised model training. This architecture promotes scalability and enables immediate responsiveness to diverse user inputs, making it especially well-suited for dynamic web-based deployment.

4. Method

In this section, we detail the methodologies we adopted to implement ArtifyAI, an intelligent system for enhancing and artistically stylizing images that are older or of low-quality. Figure 1 illustrates the system flow of the Artify application.

This system is characterized by integrated models-and-methods usage in a single set of processing, and which aims to offer both technical high fidelity in resolution and creativity affordance - or exploratory potential. As such, it is an image pipeline with four principal elements in sequence: (1) enhancement of an image using a natural image restoration method using the CodeFormer application, (2) quantitative assessment of enhancement using PSNR, SSIM, and LPIPS metrics, (3) Stylizing an image using Stable Diffusion, with edge preservation and edge-based stylization using the Canny edge detector and

ControlNet and (4) recommendation of styles whereby the model determines the best artistic styles to apply based on the input image or an edge map of the input image.

To further increase the quality of intelligence in the system, we also implemented the CLIP model for assessing images contents semantically to aid choosing the most appropriate artistic style based on the context derived from the image. This integrated pipeline gives us the ability to take our images through a systematic process of improving the quality in an objective manner in the quantitative metric assessments of PSNR, SSIM, and LPIPS, and then through controlled high-quality artistic stylization on the content of the images.

4.1. Enhancement

The first phase of our pipeline is to improve the quality of degraded images. When we started this work, we first used the Generative Facial Prior GAN (GFPGAN). The advantages of using this model were that it has been shown to restore distinct features and details of a face from degraded images. In the end though, while we made our first advances with the original pipeline, we chose to not use GFPGAN for a couple of reasons: first, the demo we were able to access was made for public consumption, not for research; and second, the demo could not run in our software environment because of version incompatibilities. So, we switched over to the CodeFormer model, which has the advantage of being based on a well-documented research paper [2].

CodeFormer is a state-of-the-art face restoration model for gaining high-fidelity facial features from blurry or degraded images while maintaining face structure and natural appearance. Considering the robustness of the model and its reported performance, CodeFormer provided the best way to implement this as we knew we could gain consistent and realistic improvements.

4.2. Evaluation

To assess the quality of the improvement, we applied three quantitative metrics: Peak Signal-to-Noise Ratio (PSNR) to measure image fidelity, Structural Similarity Index (SSIM) to determine the structural correspondence between the input image and the output image, and the Learned Perceptual Image Patch Similarity (LPIPS) metric, which measures perceptual similarity based on representations in human vision. Together, these metrics provide a clear and reproducible assessment of the quality of improvements.

While these are all quantitative measures, we introduced the CLIP model for a semantic-level understanding of the images so we could assess the relevance between the visual content of the images and the artistic styles that were introduced. These factors added a qualitative assessment component into the evaluation which was distinct from pixel-based assessments.

4.3. Artistic Transformation

After refinement, we applied artistic style transfer using Stable Diffusion with ControlNet. ControlNet abstracted the basic structure of the original image using a Canny edge detector that uses edge extraction to develop a skeleton for the generation to proceed. It allowed the artistic transformations to still retain some identity.

To explain, the process begins by applying the Canny edge detector to the image, which determines the most prominent edges of the image while capturing the essence of the structure of the original image. When we identified these edges, we can create a guide or “blueprint” for artistic generation. When introduced into the generation process as ControlNet, this blueprint acts as a visual guide to ensure that while the artistic style is applied, and may alter the textures, colors, or finer details within the images, the core shapes, outlines, and layout of the original images remain intact.

This unique layering technique allows us to produce visual representations of the image that are still recognizable in terms of subject and context at the same time. This was especially valuable in use cases were retaining some aspect of the identity or meaning of the image was desirable, such as for personal portraits, historic images, or images with cultural value.

CLIP had another function at this stage - it helped the system match the image's semantic content to the most recent artistic style contextually relevant to each image, supporting stylistic choices that came close to their character's visual and thematic qualities.

4.4. Unique Contributions

One of the primary advantages of our system is its holistic design, which allows for incorporating image enhancement, objective evaluation, and artistic style modifications into one streamlined pipeline.

One unique thing to note about our system is that we used a CLIP model to analyze the semantic content of an image while also evaluating and recommending the best artistic style. Instead of picking a style arbitrarily or choosing styles based on personal preference, our system can use

Metric	Description	Ideal Value	Our Results
PSNR (Peak Signal-to-Noise Ratio)	Measures pixel-wise similarity between original and enhanced images.	Higher is better	25.59 dB
SSIM (Structural Similarity Index)	Measures structural similarity between images.	Closer to 1.0 is better	0.8445
LPIPS (Learned Perceptual Image Patch Similarity)	Measures perceptual similarity using deep features.	Closer to 0.0 is better	0.2856

Table 1: Evaluation Metrics

CLIP to make sound recommendations for styles that are metrics-aware and more suitably matched with the content of the image, e.g. if the subject is being expressed in cartoon, anime, classic, or realistic styles, the workflow can be enhanced and aligned to some degree of responsible standards.

Another benefit of the structure-guided generation that works well in the CG generated domain of abstraction is that it allows for a more stylized transformation of the input image while still retaining the identity of the subject and context of the input image. The objective evaluation effectiveness of methods like PSNR, SSIM, and LPIPS is also very advantageous, and it serves as a benchmark for us to verify the quality of the enhancement to the image, helping to reinforce the evaluation process. Overall, synthesizing these factors of systematic improvement, objective evaluation, semi-structured artistic transformational principles, and adaptive content sensitivity using image-wide references provides us with a fundamentally different way to manipulate images compared to a traditional image processing pipeline where only a small number of these factors are engaged.

5. Experiment

To evaluate the effectiveness of the ArtifyAI pipeline, we conducted experiments on images with varying levels of quality and diverse stylistic requirements. The pipeline consists of four key stages. First, low-quality input images were enhanced using *CodeFormer*, which restored fine details and improved the clarity of facial structures. Next, enhanced images were processed using *Canny edge detection* to extract structural information, which served as guidance for the *ControlNet* module during generation. In the third stage, a textual prompt describing the desired artistic style (e.g., *"realistic portrait"*, *"cartoon style"*) was analyzed using *CLIP*, allowing semantic alignment between the prompt and the visual content. Finally, the edge-guided and prompt-conditioned inputs were passed to *ControlNet* and *Stable Diffusion* to produce the stylized output, which preserved the subject's identity while applying the specified style.

5.1. Evaluation Metrics

To evaluate the performance of our model quantitatively, we employed standard metrics widely used in face restoration and generative art tasks: PSNR, SSIM, and LPIPS. A detailed breakdown of the results is provided in Table 1.

5.2. Results

We evaluated the ArtifyAI pipeline on a set of representative sample images to demonstrate its effectiveness across different stages. Figure 2 presents an example image as it progresses through the four key stages of the pipeline. The original image is a low-resolution input, which is then enhanced using CodeFormer to produce a more detailed and visually refined result. Next, structural features are extracted using Canny edge detection, providing guidance for the generation process. Finally, the stylized output is generated using ControlNet in conjunction with Stable Diffusion, conditioned on a prompt specifying a realistic style.

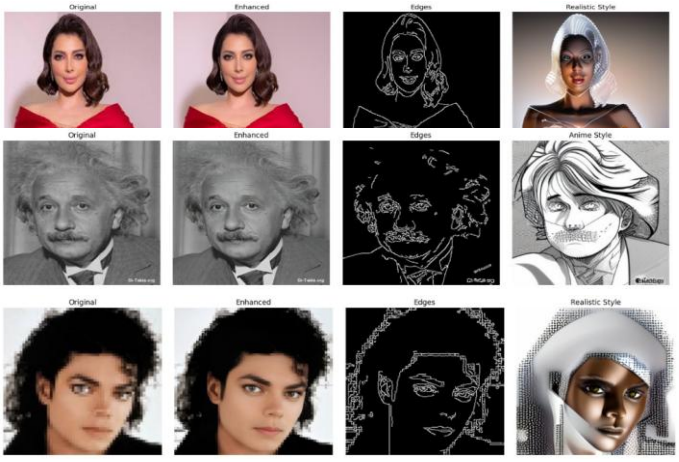

Figure 2: Left to right: the original low-quality input image, the enhanced version via CodeFormer, the detected structural edges using Canny filter, and the final stylized output using ControlNet.

This pipeline successfully maintained the core identity of the input image while producing visually compelling and stylistically relevant outputs.

5.3. Comparative Notes

While prior works such as GFPGAN and Codebook Lookup Transformers have primarily focused on specific aspects of image processing—namely, restoration or animation—our approach introduces a broader, more flexible pipeline that integrates multiple capabilities into a single system. GFPGAN, for instance, excels at facial restoration and detail recovery in portraits, but it is largely limited to enhancement tasks and lacks support for creative transformations. Similarly, approaches involving transformers for animation often require large datasets and heavy computational resources, making them less practical for general-purpose, lightweight applications.

In contrast, our method combines three powerful models—**CodeFormer**, **ControlNet**, and **CLIP**—to bridge the gap between technical enhancement and stylistic expression. CodeFormer enhances image quality by reconstructing structural details while preserving identity and texture, even in highly degraded inputs. ControlNet, on the other hand, enables edge-based structural conditioning, allowing for fine-grained control over how styles are applied without losing the original shape and composition of the image. CLIP adds another layer of intelligence by aligning style prompts with semantic understanding, ensuring that the generated artistic outputs are not only visually appealing but also conceptually relevant to the user's intent.

Together, these components form a cohesive system that offers greater control over both visual fidelity and creative transformation. This makes our approach particularly suitable for applications where both enhancement and personalization are essential—such as reviving family photos, transforming sketches into styled portraits, or creating social media–ready visuals with artistic flair. In summary, while earlier methods excel in isolated tasks, our integrated pipeline stands out by delivering a more expressive, controllable, and user-driven experience.

6. Conclusion

This project introduced ArtifyAI, a unified and intelligent system that enhances low-quality images and applies artistic styles using state-of-the-art generative models. By integrating CodeFormer for high-fidelity face restoration, Stable Diffusion with ControlNet for structure-preserving stylization, and CLIP for semantic analysis and style recommendation, our pipeline demonstrates a balanced approach between technical rigor and creative flexibility.

A key strength of ArtifyAI lies in its modular, inference-based design, which eliminates the need for task-specific training while remaining adaptable to diverse user inputs. The system not only restores degraded images but also offers users context-aware artistic transformation driven by language prompts—bridging the gap between visual restoration and creative generation.

Although our initial goal involved animating still images, constraints related to model compatibility and resource availability led us to pivot towards a more stable and scalable implementation. This decision ultimately allowed us to deliver a complete and well-integrated tool that supports practical use cases in both personal and professional domains.

Moving forward, there is potential to refine the CLIP-guided recommendation engine further, extend the system to support real-time transformations, or reintroduce motion-based outputs. Overall, ArtifyAI provides a solid foundation for future exploration at the intersection of image restoration, artistic stylization, and user-guided creativity.

References

[1] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration with Generative Facial Prior," *arXiv preprint arXiv:2101.04061*, Jan. 2021. [Online]. Available: https://arxiv.org/abs/2101.04061.

[2] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards Robust Blind Face Restoration with Codebook Lookup Transformer," Proc. NeurIPS 2022, Dec. 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c573258c38d0a3919d8c1364053c45df-Abstract-Conference.htmlA. Alpher. Frobnication. Journal of Foo, 12(1):234–778, 2002.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv preprint arXiv:2112.10752*, Dec. 2021. [Online]. Available: https://arxiv.org/abs/2112.10752.

[4] John Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[5] B. Zhang, H. Wang, Y. Zhang, and L. Zhang, "Adding Conditional Control to Text-to-Image Diffusion Models," arXiv preprint arXiv:2206.11253, Jun. 2022. [Online]. Available: https://arxiv.org/abs/2206.11253.

Appendix A: Code Implementation of ArtifyAI

This appendix provides a detailed description of the full

implementation of the ArtifyAI system. ArtifyAI utilizes advanced deep learning models such as CodeFormer for high-fidelity face restoration, Stable Diffusion with ControlNet for artistic style transfer, and CLIP for semantic analysis and style recommendations. The code provided facilitates image enhancement, style transformation, and quality evaluation (PSNR, SSIM, and LPIPS).

Please refer to the official GitHub repository: ArtifyAI GitHub Repository for the complete and executable code. (https://github.com/ghaidaaziz1/ArtifyAI-)

Key features of the implementation include:

Image Enhancement: Utilization of CodeFormer for high-quality face restoration.

Style Transformation: Leveraging Stable Diffusion and ControlNet to apply artistic styles while preserving image structure.

• Style Recommendation: Using CLIP to analyze and recommend appropriate artistic styles based on input images.

Quality Evaluation: Calculating image quality metrics such as PSNR, SSIM, and LPIPS to assess the effectiveness of transformations.