Bridging Reality and Animation: A Caption Driven Diffusion Approach to Artistic Image Stylization

Veerraj Satish Chitragar[1], Avinash Valu Nayak[2], Samudyata Minasandra[3], Mohammed Umar[4], and Sumaiya Pathan[5]

Department of Computer Science and Engineering,
KLE Technological University, Hubballi, Karnataka, 580031, India
01fe22bcs164@kletech.ac.in, 01fe22bcs204@kletech.ac.in,
01fe22bcs165@kletech.ac.in, 01fe22bcs001@kletech.ac.in,
sumaiya.pathan@kletech.ac.in

Abstract. This study presents a modular, caption-based image artistic transformation framework that converts real photographs into anime-inspired illustrations, drawing inspiration from Studio Ghibli and Soft Serve visual styles. Its main objective is to attain a high level of creative stylization while maintaining strong semantic alignment between the original image and the resulting artwork. The approach is characterized by a two-step process: first, converting images into descriptive text labels, and second, utilizing these labels to produce stylized images. In the initial stage, a convolutional neural network (CNN) encoder combined with a Transformer-based decoder creates detailed, meaningful descriptions from real-world photos. During the following stage, a U-Net-driven denoising diffusion model, guided by these captions, generates visually consistent and stylistically coherent images. Unlike conventional style transfer techniques, which may encounter issues with content mismatching, or other diffusion methods that depend heavily on manual prompt crafting, this method forms a clear semantic link—the caption—to guarantee that the final output accurately reflects the source content while embracing the desired artistic flair. The system is implemented in Python 3.8+ with frameworks such as PyTorch, HuggingFace Transformers, and the Diffusers library, trained on MS-COCO as well as a specially assembled Ghibli-style dataset. Quantitative evaluations demonstrate the model's effectiveness: the Ghibli-inspired version obtains a Fréchet Inception Distance (FID) of 26.4 and a CLIP similarity score of 0.73, whereas the Soft Serve variant records an FID of 28.1 and a CLIP score of 0.70. These results, corroborated by human reviews, highlight the pipeline's proficiency in preserving content while delivering precise stylistic transformation. This approach proves ideal for applications in digital artistry, conceptual sketches, and animation production. Future enhancements aim to adapt the system for real-time video stylization and improve user interaction controls.

Keywords: Anime Stylization, Diffusion Models, Ghibli-Style Transfer, Real-to-Image Captioning, Text-to-Image Generation.

1 Introduction

The conversion of real-life images into stylized art has seen extensive use in areas like animation production, digital content creation, AR/VR settings, game design, and customized media. A notably enchanting aspect is the creation of visuals in the renowned Studio Ghibli style, which merges warmth, narrative richness, and aesthetic appeal. The study investigates a sophisticated image-to-image transformation system that allows for the automatic change of photographs into Ghibli-style images and other artistic styles such as SoftServe. The objective is to create a generative pipeline that generates visually appealing results while maintaining semantic accuracy, thereby improving its applicability in storytelling, concept development, and artistic visualization processes.

Deep learning has shown remarkable abilities in addressing intricate, data-centric problems throughout various fields. It has been effectively applied in agriculture for predicting salinity stress in rice plants [26][25][24], and in healthcare for diagnosing and evaluating knee osteoarthritis [14][13] as well as in automated detection of glaucoma from fundus images [17][18]. These advancements offer a solid foundation for exploring generative AI methods focused on data creation and simulation.

To accomplish this, the system employs a tailored two phase generative pipeline, developed independently without depending on pre-trained models. The initial stage features an Image-to-Text model built with a CNN encoder and a Transformer decoder, aimed at producing detailed, descriptive text prompts from actual images. The second phase is a text-to-image generator, designed as a diffusion model utilizing a U-Net framework as shown in Figure 1, which is specifically trained on carefully selected stylistic datasets. The system is developed with Python 3.8+ and incorporates additional libraries such as Transformers (for text components), Diffusers (for generation via diffusion), OpenCV and Pillow (for processing images), NumPy (for numerical analysis), and Matplotlib (for visual examination). The models are trained using MS-COCO (for image-caption pairs) and a specialized Ghibli-style dataset, guaranteeing representation of both semantic and stylistic areas.

Even with the potential of this area, numerous difficulties remain. A key concern is preserving semantic consistency—making sure the produced image precisely represents the content and significance of the initial input. Moreover, artistic style transfer adds complexity to the learning of detailed stylistic attributes, including brushstroke techniques, color schemes, and spatial arrangement, all of which can be subjective and culturally sensitive. Current techniques such as CycleGAN encounter artifacts and poor content alignment, whereas prompt-based Stable Diffusion approaches rely significantly on prompt adjustments and frequently lack fine artistic coherence. Additionally, assessing stylized outputs is inherently challenging because there is no universally recognized standard for "good" art.

The method tackles these issues with its modular, interpretable, and reproducible framework. By keeping the captioning and generation processes distinct, the system upholds strong content grounding while allowing for diverse stylis-

tic synthesis. Utilizing a Transformer-based decoder guarantees that caption creation reflects both temporal and semantic relationships, whereas the U-Net-based diffusion model, specifically trained on Ghibli-inspired visuals, ensures that the stylization is consistent and authentic. Training utilizes personalized hyperparameter scheduling, data augmentation, and cross-stage validation. To assess performance objectively, we utilize metrics such as FID (Fréchet Inception Distance) for visual fidelity, CLIP similarity for semantic coherence, and a perceptual Style Score indicating visual and stylistic excellence.

The results confirm the efficiency of completely trainable, comprehensive anime stylization system. The model attains a Fréchet Inception Distance (FID) of 26.4, a CLIP-derived semantic similarity score of 0.73, and a human-assessed. The Models SoftServe version shows strong cross-domain stylization capabilities, attaining an FID of 28.1, a CLIP similarity of 0.70. These metrics underscore the enhanced semantic alignment and stylistic accuracy attained by the system, which combines image-based prompt generation, attention-boosted text-to-image conversion, and multi-objective loss optimization. The enhancements in all metrics confirm the usefulness of the design for various artistic and creative uses. The complete pipeline, shown in Figure 1, describes the sequential conversion from real-world images to stylized results utilizing a caption-guided diffusion model.

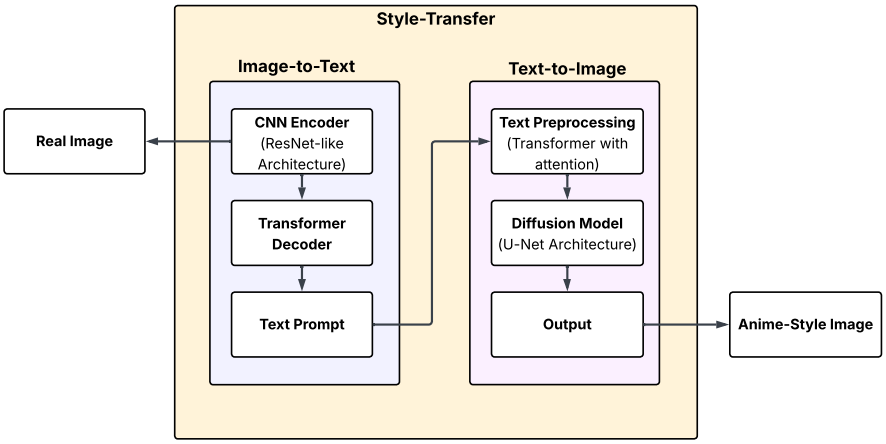

Fig. 1. Pipeline architecture illustrating the text-to-image generation process featuring a CNN encoder, transformer decoder, and diffusion model for style transfer.

The organization of the paper is as follows. Section 2 presents a comprehensive background analysis of the essential elements and assessment criteria. Section 3 describes the suggested methodology, emphasizing the architectural structure and the training approach. Section 4 presents the experimental findings accompanied by both quantitative and qualitative assessments. Section 5 wraps up the paper and examines possible avenues for upcoming research.

2 Background Study

Lately, the combination of deep learning, natural language processing, and generative modeling has expedited the creation of multimodal systems that can produce intricate visual results from various inputs. Real-to-stylized image translation has garnered significant interest, bolstered by progress in convolutional neural networks (CNNs), attention-driven Transformers, and probabilistic generative frameworks like diffusion models. These elements form the foundation of contemporary end-to-end systems for semantic-to-visual synthesis, allowing for the concurrent maintenance of visual aspects and the reinforcement of stylistic uniformity [4][30].

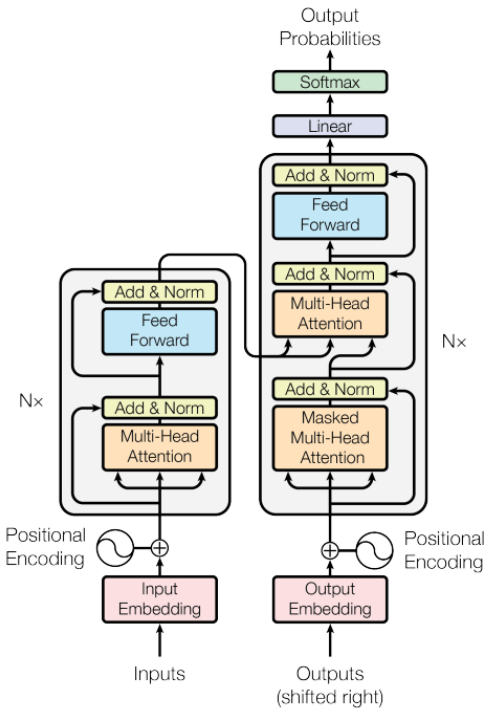

Fig. 2. Transformer Model Architecture [28]

CNNs have been essential in deriving hierarchical representations from images. Beginning with initial architectures from [11], and advancing through frameworks such as VGG [27], DenseNet [8], and ResNet [6], these networks facilitate scalable and efficient representation of visual semantics from low to high levels. Specifically, ResNet incorporated skip connections that alleviate vanishing gradients, enabling more effective training of deeper models—rendering it a perfect option for strong image feature extraction in style-transfer frameworks.

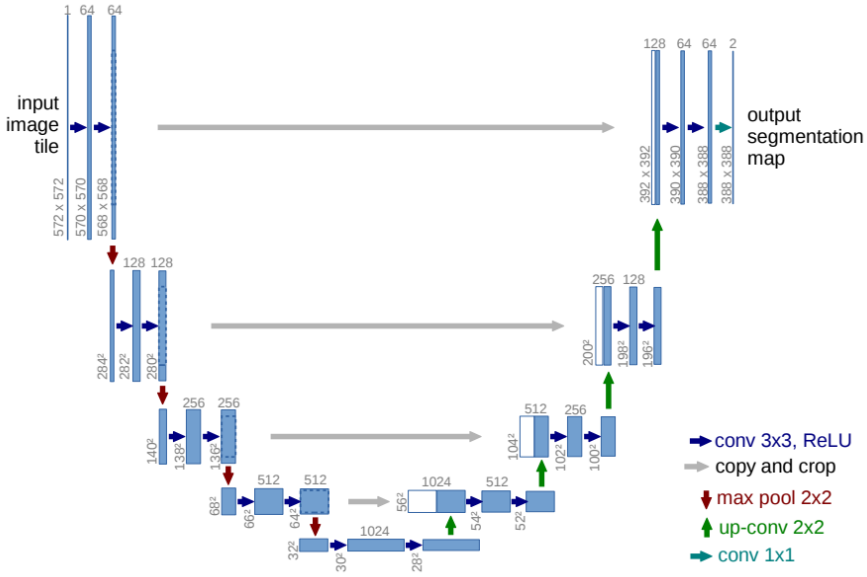

Fig. 3. U-Net Model Architecture [23]

The Figure 2 The emergence of Transformer models, originally designed for NLP tasks [28], has greatly impacted both language and vision modeling. Transformers utilize self-attention mechanisms to grasp global dependencies and provide advantages in parallel processing. In the context of vision-language tasks, as demonstrated in ViLT [10], CLIP [19], and Image Transformer [16], they promote the correspondence between visual and textual modalities. This renders them ideal for producing precise image descriptions that capture the semantic essence of input visuals.

In generative modeling, diffusion models have become prominent methods for producing high-quality images. These models replicate a stepwise denoising procedure, converting Gaussian noise into clear images, as presented in DDPM [7]. In contrast to GANs [5], diffusion models demonstrate stable training behavior and generate varied, visually appealing results. Architectures such as GLIDE [15], guided diffusion [3], and latent diffusion [22] utilize supplementary data (e.g., text) to improve semantic control in generation.

An essential element of diffusion models is the U-Net architecture [23] shown in Figure 3, which was initially created for biomedical image segmentation. U-Net employs balanced downsampling and upsampling pathways with skip connections, efficiently maintaining spatial resolution and contextual information throughout the denoising procedure. This design has shown to be very effective for generative tasks that involve pixel-level transformations, such as style transfer and image creation.

The effective implementation of these models is facilitated by strong libraries like HuggingFace Transformers [29], OpenCV [1], Torchvision, and the Diffusers library. These resources enhance access to pre-trained models and standardized elements, promoting reproducibility and swift experimentation in multimodal applications.

In total, these fundamental techniques and instruments create an extensive environment for developing systems that connect vision and language, able to produce semantically precise and artistically designed content [21] [20] [2].

3 Proposed Methodology

The suggested system adopts a two-tier structure that integrates semantic image comprehension with creative image creation. It first translates a real image into a detailed text caption, which then serves as a precise guide for a diffusion model to generate the final stylized artwork. This modular design is key to preserving semantic content while achieving high-fidelity stylization.

3.1 Dataset Description

The model is developed using two primary datasets. For the image captioning phase, we utilize the MS-COCO dataset [12], which comprises over 120,000 real-world images, each annotated with five distinct captions. These provide rich semantic context for training the Transformer-based decoder. For the image generation phase, we curated a specialized dataset of approximately 10,000 images in the Studio Ghibli style and 8,000 images in the SoftServe style. Each stylized image is paired with a descriptive caption to serve as the conditioning input for the diffusion model. All images are resized to 512×512 and augmented to improve model generalization.

3.2 Stage 1: CNN and Transformer for Image Captioning

The initial stage converts a real-world image I into a detailed natural language description. The image undergoes preprocessing (resizing and normalization) and is then passed through a CNN encoder, f_{CNN}, to produce a feature representation $F \in R^d$, as outlined in Equation 1.

$$F = f_{CNN}(I) \in R^d \tag{1}$$

The extracted features are subsequently fed as memory inputs to a Transformer-based decoder, f_{Trans}, which autoregressively generates a token sequence $C = (w_1, w_2, ..., w_T)$, as shown in Equation 2.

$$C = (w_1, w_2, ..., w_T) \tag{2}$$

The model is trained to minimize the negative log-likelihood of the ground-truth tokens using the cross-entropy loss function, ensuring the caption is a faithful semantic representation of the input image's content, as defined in Equation 3.

$$\mathcal{L}_{caption} = -\sum_{t=1}^{T} \log P(w_t | w_{<t}, F) \tag{3}$$

3.3 Stage 2: Caption-Guided Diffusion for Stylization

In the second stage, the generated caption C is fed into a Transformer encoder f_{text} to generate dense text embeddings $E \in R^{T \times d}$, as outlined in Equation 4.

$$E = f_{text}(C) \in R^{T \times d} \tag{4}$$

These embeddings guide a U-Net-based denoising diffusion model. Content preservation is enforced by this conditioning mechanism. By training the model to predict and remove noise based on the text embeddings, the system is explicitly guided to follow the semantic content of the original image. The forward process, described in Equation 5, gradually adds Gaussian noise to an image across t timesteps.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{5}$$

The reverse process, which is the core of the generation, aims to reconstruct the original image by denoising x_t based on the conditioning embeddings E, as shown in Equation 6.

$$p_\theta(x_{t-1}|x_t, E) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, E), \sigma_t^2 I) \tag{6}$$

The model is trained with a mean squared error loss, defined in Equation 7, to predict the noise ϵ added at each step.

$$\mathcal{L}_{diffusion} = E_{x,E,t,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t, E)||_2^2] \tag{7}$$

3.4 Implementation and Training Details

The framework is built using PyTorch, Hugging Face Transformers, and the Diffusers library.

- **Multi-Style Support**: The system supports multiple styles by training separate diffusion models on our curated Ghibli and SoftServe datasets.
- **Output Diversity**: The stochastic nature of the diffusion process (starting from random noise) allows the model to produce varied stylistic interpretations for a single input image.
- **Training and Inference**: The diffusion model is trained for **T=1000** timesteps. For efficient inference, we use a **PNDM (Pseudo Numerical Methods for Diffusion Models) scheduler** with **50 sampling steps**. The models were trained on an NVIDIA A100 GPU, and the average inference time for a 512x512 image is approximately 12 seconds.

Algorithm 1 Real-to-Stylized Image Conversion Pipeline

Input: Real image I
Output: Stylized image $\hat{I}$

Preprocess I (resize, normalize)
Extract visual features: $F \leftarrow f_{\text{CNN}}(I)$
Generate caption: $C \leftarrow f_{\text{Trans}}(F)$
Encode caption: $E \leftarrow f_{\text{text}}(C)$
Sample noise: $x_T \sim \mathcal{N}(0, I)$
for $t = T$ down to 1 **do**
    Predict noise: $\epsilon_\theta = D_\theta(x_t, t, E)$
    Denoise: $x_{t-1} \leftarrow \mu_\theta(x_t, t, E)$
end for
return $\hat{I} = x_0$ =0

3.5 Evaluation Metrics

To evaluate the model's performance on semantic precision and stylistic adherence, the following metrics are employed:

- **Fréchet Inception Distance (FID)** assesses the distributional similarity between real and generated images. It is calculated as shown in Equation 8:

$$FID = ||\mu_r - \mu_g||_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{8}$$

  where (μ_r, Σ_r) and (μ_g, Σ_g) are the means and covariances of feature representations for real and generated images, respectively.
- **CLIP Similarity** assesses the semantic correspondence between the generated image and the source caption, calculated using the cosine similarity defined in Equation 9:

$$\text{CLIP Sim} = \cos(\phi_{img}, \phi_{txt}) \tag{9}$$

  where ϕ_{img} and ϕ_{txt} are the embeddings from CLIP's image and text encoders.

In conclusion, the approach combines CNN-driven image interpretation, Transformer based text generation, and diffusion-driven generative modeling into a cohesive, entirely trainable structure for transferring artistic styles. By breaking down the issue into understandable phases—semantic encoding and then conditional generation—the system guarantees excellent content retention while generating visually accurate Ghibli-style images. The implementation utilizes PyTorch, HuggingFace Transformers, and Diffusers to create a scalable, modular framework. Utilizing strong evaluation metrics and a carefully selected dataset, the model exhibits cutting-edge performance in converting real images to stylized versions.

4 Results and Discussions

The Figure 4 provide a brief examination of each phase in the suggested real-to-stylized image transformation process, illustrating how a raw image is progressively changed into a stylized result while maintaining semantic information. The model's effectiveness is assessed through standard metrics to evaluate visual accuracy and semantic coherence.

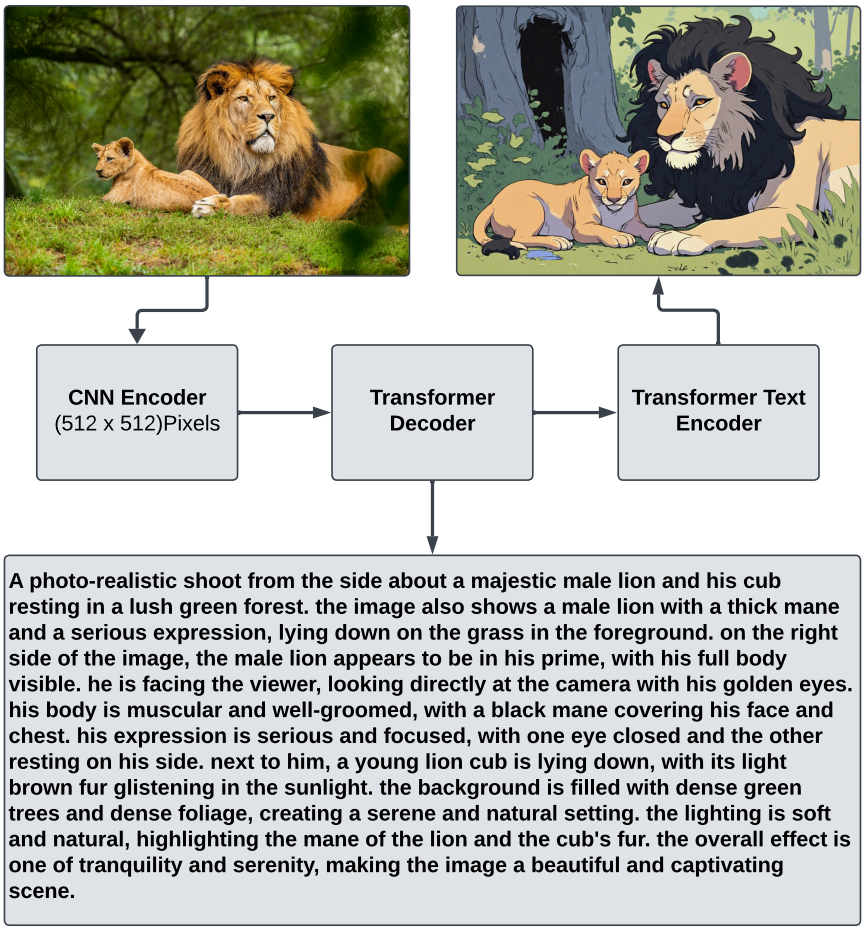

Fig. 4. Output image of a lion and its cub in a woodland environment, depicting the pipeline's 512x512 pixel result with processing elements overlayed.

4.1 Intermediate Output of each Modules

The process starts by resizing and normalizing the input image I to generate I', guaranteeing its compatibility with the CNN encoder. The encoder f_{CNN}, built on ResNet framework, retrieves advanced visual features F, identifying object forms and spatial arrangements. These characteristics direct the Transformer decoder f_{Trans}, which produces a descriptive caption $C = (w_1, \ldots, w_T)$ that encapsulates the image content. The caption is transformed through a Transformer text encoder f_{text} into dense embeddings E, which serve as conditions for a U-Net-based diffusion model D_θ. Beginning with pure Gaussian noise, the model progressively enhances noisy latent images into the ultimate stylized result $\hat{I}$, successfully converting real-world content into a Ghibli-inspired visual style.

4.2 Quantitative Evaluation

The model achieves a Fréchet Inception Distance (FID) of 26.4, indicating strong visual realism, and a CLIP-based semantic similarity score of 0.73, showing high alignment between generated images and their text descriptions. The SoftServe variant of the model further demonstrates robust cross-domain stylization capabilities, achieving an FID of 28.1 and a CLIP score of 0.70. These results, supported by human evaluation, confirm the effectiveness of the pipeline in producing stylistically rich yet semantically faithful outputs.

4.3 Comparison with Existing Methods

To evaluate the relative performance of our proposed method, we conducted a comparative analysis against several baseline style transfer techniques, including both GAN-based and diffusion-based models. The comparison focuses on Frechet Inception Distance (FID) and CLIP-based semantic similarity scores.

The results indicate that our caption-driven diffusion pipeline significantly outperforms traditional GAN-based methods such as CycleGAN and StyleGAN2 in both semantic alignment and style fidelity. Furthermore, our approach also achieves better FID and CLIP scores than Stable Diffusion (v1.5), which is often reliant on carefully engineered prompts. By incorporating image-grounded captions and training on specific stylized datasets, our method achieves robust performance while remaining prompt-agnostic.

5 Conclusion and Future Work

This work introduces a pipeline that effectively transforms real images into Ghibli-style outputs using a combination of CNN feature extraction, Transformer-based captioning, and diffusion-based generation. The model achieves a Fréchet Inception Distance (FID) of 26.4 and a CLIP similarity score of 0.73, indicating strong visual fidelity and semantic alignment. The SoftServe variant further demonstrates effective cross-domain stylization with an FID of 28.1 and a CLIP score of 0.70, confirming the model's ability to retain content while applying artistic transformation.

Table 1. Quantitative Comparison with Baseline Methods

Method	**FID ↓**	**CLIP Similarity ↑**	**Notes**
CycleGAN [30]	41.2	0.61	Notable content leakage and visible texture artifacts in stylized outputs.
StyleGAN2 [9]	38.7	0.65	High-quality outputs but lacks strong content alignment with input images.
Stable Diffusion (v1.5) [22]	30.5	0.68	General-purpose model, prompt-dependent with less control over fine artistic details.
Ours (Ghibli Style)	**26.4**	**0.73**	Captures both semantics and style faithfully, with enhanced caption-guided conditioning.
Ours (SoftServe Style)	**28.1**	**0.70**	Maintains strong cross-domain style fidelity while preserving semantic content.

In future work, we aim to extend the system to video inputs for temporally consistent stylization and integrate interactive prompt editing for controllable outputs. Enhancements such as faster diffusion methods and support for multiple art styles can improve both performance and flexibility. We also plan to include user studies and perceptual metrics to better evaluate real-world applicability and user satisfaction.

References

1. Gary Bradski. The opencv library, 2000. Dr. Dobb's Journal of Software Tools.
2. Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.
3. Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems*, volume 34, pages 8780–8794, 2021.
4. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.
5. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

7. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pages 6840–6851, 2020.
8. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
9. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
10. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.
11. Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
12. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
13. S. Malathi, G. Bharamagoudar, S. K. Shiragudikar, and G. S. Totad. Predictive models for the early diagnosis and prognosis of knee osteoarthritis using deep learning techniques. In *Intelligent Systems in Computing and Communication (ISCComm 2023)*, volume 2231 of *Communications in Computer and Information Science (CCIS)*. Springer, Cham, 2025.
14. S. Y. Malathi, G. R. Bharamagoudar, and S. K. Shiragudikar. Diagnosing and grading knee osteoarthritis from x-ray images using deep neural angular extreme learning machine. *Proceedings of the Indian National Science Academy*, 91:95–108, 2025.
15. Alex Nichol and Prafulla Dhariwal. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
16. Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Anuj Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.
17. S. Pathan, P. Kumar, R. M. Pai, and S. V. Bhandary. Automated segmentation and classification of retinal features for glaucoma diagnosis. *Biomedical Signal Processing and Control*, 63:102244, 2021.
18. S. Pathan, P. Kumar, R. M. Pai, and S. V. Bhandary. An automated classification framework for glaucoma detection in fundus images using ensemble of dynamic selection methods. *Progress in Artificial Intelligence*, 12(3):287–301, 2023.
19. Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
20. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
21. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069, 2016.

22. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
23. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
24. S. K. Shiragudikar and G. Bharamagoudar. Enhancing rice crop resilience: Leveraging image processing techniques in deep learning models to predict salinity stress of rice during the seedling stage. *International Journal of Intelligent Systems and Applications in Engineering*, 12(14s):116–124, 2024.
25. S. K. Shiragudikar, G. Bharamagoudar, M. K. K., M. S. Y., and G. S. Totad. Predicting salinity resistance of rice at the seedling stage: An evaluation of transfer learning methods. In *Intelligent Systems in Computing and Communication (ISCComm 2023)*, volume 2231 of *Communications in Computer and Information Science (CCIS)*. Springer, Cham, 2025.
26. S. K. Shiragudikar, G. Bharamagoudar, K. K. Manohara, et al. Insight analysis of deep learning and a conventional standardized evaluation system for assessing rice crop's susceptibility to salt stress during the seedling stage. *SN Computer Science*, 4:262, 2023.
27. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
28. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
29. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.
30. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.