Artistic Style Transfer with CycleGAN and Pre-trained Discriminators

Tienhsin (David) Chang *
University of Toronto
tien.chang@mail.utoronto.ca

Hyunseok (Peter) Jang *
University of Toronto
hyunseok.jang@mail.utoronto.ca

Aaron Xiaozhou Liu *
University of Toronto
aaronxiaozhou.liu@mail.utoronto.ca

Neil Mehta *
University of Toronto
neil.mehta@mail.utoronto.ca

Abstract

This project introduced an enhanced CycleGAN architecture designed for image-to-image translation, with a focus on translating real-world photographs into impressionist-style paintings. The primary objective was to assess the stylization quality and accuracy of the generated images produced by a CycleGAN model with pre-trained discriminators. Unlike standard CycleGANs, this model pre-trained its two discriminators as binary classifiers: one specialized in detecting impressionist style art, and the other focused on assessing image authenticity. This approach was expected to enable more stable training and precise control over both stylistic authenticity and image quality. The specialized sensitivity of each discriminator would complement the two generators: one to translate photos into impressionist art and the other to reverse the transformation. By incorporating cycle-consistency loss to retain structure and adversarial loss to encourage style adherence, it was expected that this architecture would demonstrate improvements in style transfer fidelity and content accuracy. By implementing this architecture, the project aimed to produce stylized yet content-accurate and realistic images, providing insights into the model's effectiveness in high-quality image-to-image translation tasks. The insights gained from this study could inform further advancements in artistic style transfer and general image-to-image translation applications.

1 Introduction

Image-to-image translation is a fascinating task within computer vision, attracting significant attention due to its applications in fields such as art, medical imaging, and data augmentation. At its core, image-to-image translation involves mapping an image from one domain to another while preserving its most essential content. This topic has notable implications for automated artistic stylization and improving medical data by translating scans into clearer versions, among many other uses. These potential uses position image-to-image translation as a valuable tool for advancing both practical applications and artistic creativity in an ever-evolving digital era.

The primary aim of this project was to explore whether pre-training discriminators before adversarial training was more effective in generating accurate images than allowing untrained discriminators to reach equilibrium with generators. As such, this project proposed a pre-trained discriminator CycleGAN architecture specifically designed to translate real-world images into specific art style paintings. In particular, this concept was demonstrated using the impressionist art style. Impressionist

*Equal Contribution (Alphabetical order of last names)

Final Project for CSC413 Deep Learning and Neural Networks (Fall 2024)

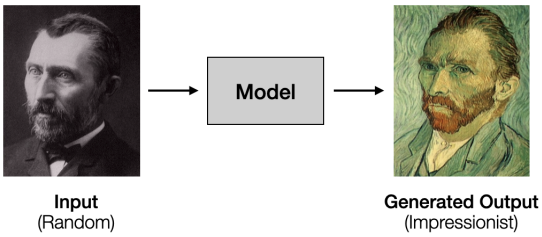

Figure 1: Visual Demonstration of the Proposed Model

art is characterized by vibrant colour palettes, small, visible brush strokes, and an emphasis on capturing light and movement over precise detail, making it a compelling style for challenging image-to-image translation [1]. By targeting impressionism, this project sought to address the complexity of transferring stylistic elements that are abstract and textural, exploring the capacity of GAN-based models to render stylistic features that go beyond traditional style boundaries.

Implementing impressionist stylization presents several challenges: the model had to accurately capture loose, textured brushstrokes, and color dynamics while preserving recognizable features from the original photo. This made it difficult to balance artistic abstraction with content fidelity, a balance critical to ensure that translated images remain stylistically true while retaining the content's essence.

The use of a Generative Adversarial Network (GAN) model was appropriate for this task, as GANs are well-suited to learning complex, non-linear transformations [2]. As seen in Figure 1, the input to the developed model is a photograph, such as a landscape or urban scene, while the output is an image of the same scene in an impressionist style. Both the input and output images were represented in standard image formats, such as JPEG or PNG, with a resolution of 224×224 pixels to maintain consistency. This format allowed the model to handle various image types and generate artwork suitable for digital applications. This approach leveraged the power of deep learning for extracting high-dimensional features, enabling the generation of visually compelling yet content-preserving images.

A link to the project's GitHub repository is provided in the footnote on this page. [2]

2 Background and Related Work

In recent years, GANs have become a popular method for style transfer, such as transforming real-life photos into artistic renditions [3]. An example of previous research in this field is the CycleGAN, work done by Zhu et al. in 2017. The introduction of the CycleGAN model enabled unpaired image-to-image translation by enforcing cycle consistency [4]. This helped maintain the contents of the artwork while allowing it to be flexible in style transformation between different domains, allowing it to be an effective tool for transforming photos into different art styles.

This project aimed to build on this by pre-training the two discriminators with a focus on transformations of photos into impressionist art. CycleGAN is used as the foundation for its ability to handle unpaired data, with the dataset of impressionist and non-impressionist photos.

3 Data

The datasets for this project consisted of four main types of images:

1. **Impressionist Art** (imp)**:** A collection of images of impressionist paintings in WikiArt, sourced from Kaggle [5]. These served as style references.
2. **Non-impressionist Art** (imp')**:** A collection images of non-impressionist paintings in WikiArt (e.g., realism, baroque, etc.).
3. **Real-life images** $(real)$**:** A collection of real-life photographs from ImageNet [6] and ArtiFact [7], both sourced from Kaggle.

[2] Link to the project's GitHub repository.

Impressionist Art (*imp*) | **Non-impressionist Art** (*imp′*) | **Real-life Photos** (*real*) | **Synthetic Photos** (*fake*)

Figure 2: Impressionist and Non-impressionist Datasets

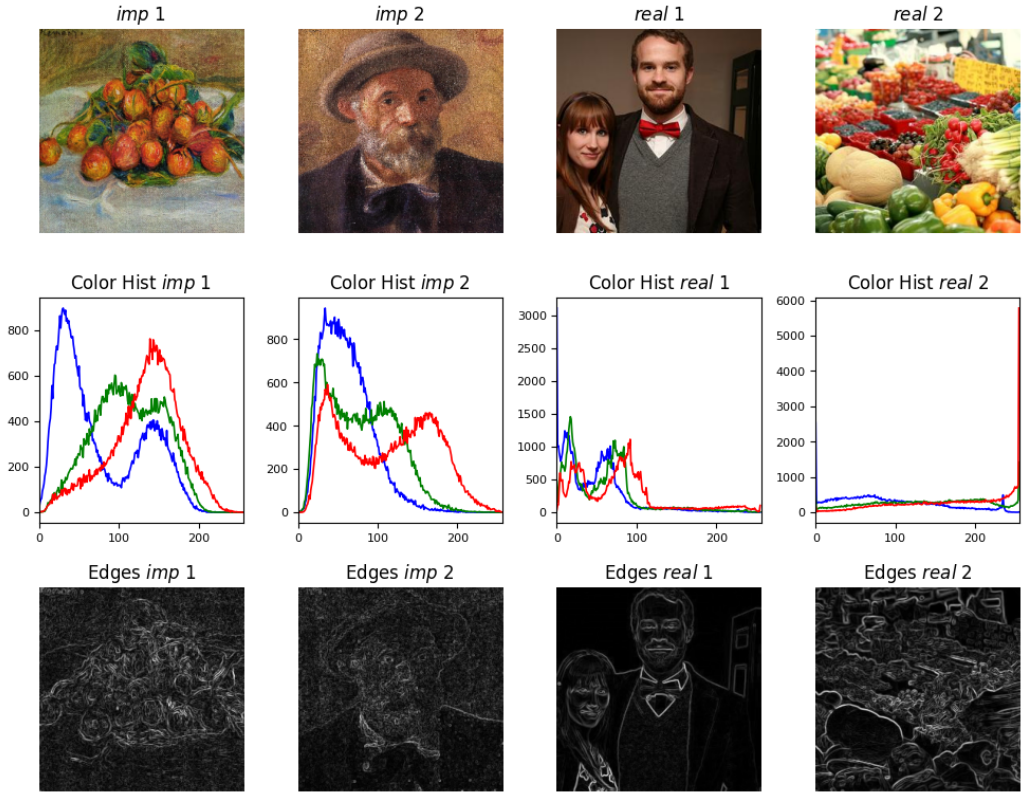

Figure 3: Comparative Analysis of Impressionist (imp) and Real-life ($real$) Images

4. **Fake real-life (synthetic) images** ($fake$)**:** A collection of synthetic, AI-generated images from ArtiFact.

As showcased in Figure 2, all four categories of images helped demonstrate the stylistic gap the model needed to bridge, with impressionist art showcasing abstract interpretations, vibrant colours, and varied brush work, non-impressionist art showcasing paintings with other properties, photographs offering realistic scenes, and synthetic images showing fake real-life images. All images were pre-processed to a standard resolution of 224×224 pixels, ensuring a uniform input size suitable for GAN training.

Figure 3 provides a comparative analysis of impressionist (imp) and real-life images ($real$), focusing on two key aspects: color histograms and edge-detected (texture) images. The analysis highlights significant differences between impressionist and real-life images in color distribution and texture. Impressionist images exhibit vibrant, saturated colors with concentrated peaks in blue and red

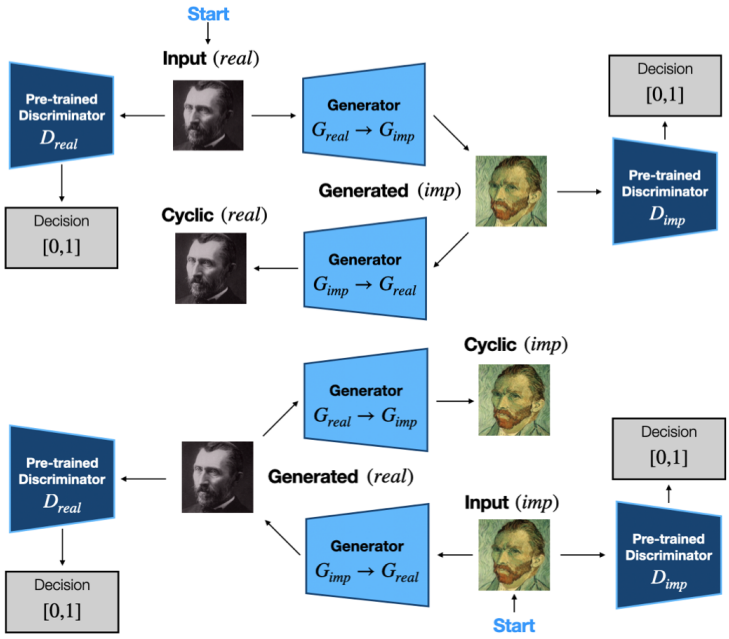

Figure 4: Simple View of the Proposed Model Architecture

channels, alongside soft, diffuse edges reflecting textured brushstrokes. In contrast, real-life images display balanced, natural color distributions and sharp, well-defined edges characteristic of realism. These differences underscore the need for a GAN to learn and apply vibrant color transformations and blended textures while preserving key content features from photographs, enabling authentic style transfer to impressionist art.

4 Model Architecture

The proposed model uses a CycleGAN-based architecture, which includes two pairs of generators and pre-trained discriminators. As seen in Figure 4, the two generators $G_{imp \rightarrow real}$ and $G_{real \rightarrow imp}$, perform bi-directional translation between the impressionist art and real-life domains. The pre-trained discriminators D_{imp} and D_{real} evaluate the quality of generated images via binary classification, ensuring that outputs from $G_{real \rightarrow imp}$ resemble impressionist art.

To expand on the structure in Figure 4, during adversarial training, each discriminator D_{imp}, D_{real} receives two types of inputs: one is a real image, and the other is a synthetic image generated by the corresponding generator. Each discriminator's job is to distinguish between the real and generated images, encouraging the generator to produce better images and punishing it for producing bad ones.

The key components of the model are:

- **Discriminators:** Pre-trained binary classifiers to ensure that generated images exhibit impressionist-style characteristics with realistic qualities
- **Generators:** Convolutional layers with residual blocks to maintain content features while transforming style features
- **Loss functions:** Adversarial (generator) loss to improve authenticity, cycle-consistency (discriminator) loss to preserve content, and identity loss to evaluate the color coherence
  - $\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$
  - $\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$
  - $\mathcal{L}_{\text{identity}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|G(x) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|F(y) - y\|_1]$

Table 1: Discriminator Architecture Comparison

Component	Custom	Pre-trained
Backbone	Conv Layers	ResNet50
Channels	32-256	2048
Reduction	4×4, s2	ResNet blocks
Pooling	AdaptiveAvg	Global avg
Dense	100→1	512→256→1
Norm	Instance	None
Dropout	0.5	0.3
LR	1e-4	3e-4

Table 2: Generator Models Comparison

Generator	1	2	3
Learning Rate	6e-4	2e-4	2e-4
Epochs	10	10	10
Batch Size	2	2	2
λ_{cycle}	10.0	20.0	20.0
$\lambda_{identity}$	5.0	10.0	10.0
Activation	LReLU	ReLU	ReLU
Architecture	Basic	Enhanced	Channel
Upsampling	Basic	Nearest	Nearest
Color Process	Std	Enhanced	Separate
Training	Single	Single	Split
Extra Loss	–	–	Balance

4.1 Discriminators

Both discriminators D_{real} and D_{imp} share the same architectures described below, with identical backbones and classification heads. While structurally identical, they were trained for different objectives: D_{real} learned to distinguish between natural photographs and synthetic images, while D_{imp} focused on differentiating impressionist artwork from generated artistic content. This sophisticated architecture, when paired with different sets of weights, was used by the discriminators to effectively specialize in their respective domain.

The initial attempt to train discriminators from scratch highlighted the challenges of building effective artistic classifiers. Despite careful architecture design, these models achieved limited success in distinguishing artistic styles and detecting synthetic images. The custom models struggled with feature extraction, evidenced by their need for high dropout rates (0.5) to prevent overfitting to superficial patterns.

Switching to pre-trained ResNet50-based discriminators significantly improved performance through transfer learning. The ImageNet-trained backbone provided robust feature hierarchies, while custom classification heads enabled fine-tuning for artistic classification. Using lower learning rates and dropout reflected stronger base features. However, even these improved discriminators occasionally misclassified images with subtle artistic elements, suggesting that purely discriminative approaches may have inherent limitations in capturing the full spectrum of artistic style. This experience demonstrates the value of transfer learning while highlighting the continuing challenge of computational art assessment.

4.1.1 From Scratch

Initial experimentation with the discriminator model led to the following: 4 convolutional layers, scaling a 224×224 pixel image with 3 input channels to a 12×12 image with 256 channels. Each convolution layer doubles the width, while decreasing the image size via a 4×4 kernel with a stride of 2. This decision was made to increase detail and texture learning with the multiple channels, while scaling down the image to decrease the amount of parameters for the fully-connected layers.

Average pooling was used to merge all the info learned from the output channels before passing the data to the MLP layers. Before each fully-connected layer, a dropout layer was used as a regularization technique to encourage exploration and prevent overfitting. Leaky ReLU was used as the activation function for the first fully-connected layer for its non-zero values for negative inputs, mitigating the dying ReLU problem.

Finally, the output of the second layer was interpreted directly as positive if it was greater than 0 and negative if it was less than or equal to zero. The loss of choice was the `BCEWithLogitsLoss` function in the PyTorch library. This was chosen as the discriminators were binary image classifiers and the function is a combination of a sigmoid layer with binary cross entropy log loss, which is more numerically stable.

Hyperparameters were tuned using grid search and resulted in the parameters in Table 1. Analyzing the parameters, it was initially anticipated a greater width to be more favorable as image classification

should require a lot of data, but larger widths did not seem to make any significant improvements, only serving to slow down training time. A batch size of 16 was favorable due to the amount of data in each picture. It seemed that a lower batch size allowed the model to pick out more specific patterns. The dropout rate of 0.5 seemed fairly high, but it did not seem to have a significant impact either in relation to the other parameters, while improving training time as less of the parameters were active. Lastly, a smaller learning rate was favored as it reduced oscillation.

4.1.2 Pre-trained

The final model architecture for the discriminators employs a pre-trained ResNet50 model, augmented with custom classification layers, departing from earlier attempts at fully custom architectures that proved challenging to optimize (described in Section 4.1.1). This design decision was motivated by the need to leverage robust feature extractors while maintaining the flexibility to adapt to the project's specific artistic style transfer task. The discriminator network (Figure **??**) consisted of two main components:

Feature Extraction: A ResNet50 backbone pre-trained on ImageNet, with weights initialized from the `IMAGENET1K_V1` dataset. This component leveraged the deep hierarchical feature representations learned from a diverse set of natural images, providing a strong foundation for distinguishing artistic styles.

Style Classification Head: A custom sequence of fully connected layers specifically designed for binary classification of artistic styles. This component included: an initial dense layer (512 units) to reduce the dimensionality of ResNet50's feature space, LeakyReLU activation ($\alpha = 0.2$) to introduce non-linearity while preventing dead neurons, dropout layers ($p = 0.3$) for regularization, a second dense layer (256 units) for further feature abstraction, and finally, a final output layer for binary classification.

4.2 Generators

The generator networks, $G_{real \to imp}$ and $G_{imp \to real}$, share base architectural elements but evolve across the three implementations. All models begin with a 7×7 convolutional layer with reflection padding, followed by two downsampling blocks that expand features from 64 to 256 channels. While they all utilize nine residual blocks, their activation functions and normalization strategies differ significantly. Generator 1 employs LeakyReLU activations throughout, whereas Generators 2 and 3 use ReLU activations.

The key architectural divergences appear in the design choices: Generator 1 uses basic upsampling, Generator 2 introduces nearest-neighbor interpolation, and Generator 3 implements separate channel processing with individual convolutional paths for RGB channels before merging. Additionally, Generator 3 adds channel-specific normalization steps. The final output layer maintains the 7×7 convolution with tanh activation across all models, but Generator 3 includes additional channel balancing mechanisms.

Generators 1 and 2 maintain a symmetric architecture focused on style transfer, while Generator 3 deviates with its channel-wise processing approach. These architectural variations, particularly in upsampling strategies and channel handling, led to different trade-offs between structural preservation and style transfer capabilities.

5 Results

The iterative model development revealed unexpected trade-offs in style transfer. Figure 5 presents a qualitative comparison between the original images, CycleGAN [4] outputs, and the team's three generator variations. While all models captured basic impressionist elements, they exhibited distinct trade-offs detailed in Table 2. Generator 1, employing conservative hyperparameters, maintained structural fidelity at the cost of artistic style, Generator 2, with increased loss weights and ReLU activations, achieved stronger brush stroke effects but compromised structural coherence, and Generator 3's attempt at color balance through channel-wise processing resulted in degraded overall performance.

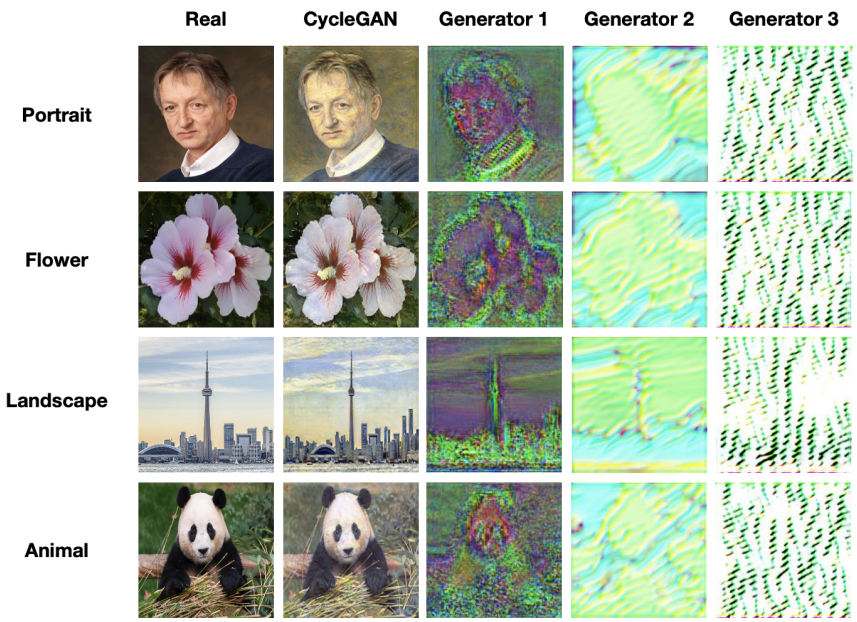

Figure 5: Comparison of Generated Images Across Models

The pre-trained CycleGAN `style_monet` model significantly outperformed the team's implementations, demonstrating superior brush stroke simulation and color adaptation while maintaining structural integrity. This performance gap, despite the project's architectural innovations, suggests that extensive training on curated artistic data may be more crucial than architectural complexity for style transfer tasks. This observation is supported by the team's discriminator performance, where both D_{real} and D_{imp} achieved strong training accuracy (96.8% and 91.1% respectively) but more modest validation performance ($\approx$86%), indicating the inherent challenge of generalizing artistic style assessment.

6 Discussion

The following subsections aim to break down and compare the pre-trained discriminator models (with custom layers vs. ResNet50) and the generator models (with custom layers vs. CycleGAN).

Discriminator Models: While the initial discriminator model was unable to achieve a higher validation accuracy, it did not overfit, with a training accuracy of similar value. In comparison, the discriminator model using the ResNet50 pre-trained weights was able to achieve greater validation accuracy, but overfitted to the training data. With some more tuning and compute resources, perhaps the initial discriminator model could have achieved greater results as it seemed to have more learning left to do.

Generator Models: The performance of the custom generator model reveals both strengths and limitations when compared to CycleGAN. While the team's model lacked significant detail and colour fidelity, as seen in Figure 5, it excelled in emphasizing the texture of brushstrokes, an essential characteristic of impressionist-style art. The generator captured basic shapes, such as the CN Tower and skyline, but struggled with high-fidelity details and lighting. In contrast, CycleGAN produced a more realistic and intricate output, retaining structural and colour accuracy, though its brushstroke texture was less pronounced.

7 Limitations

Despite the demonstrated effectiveness of the pre-trained discriminator approach, several key limitations warrant discussion. The model exhibits inconsistent performance on complex scenes with multiple subjects, often struggling to maintain coherent artistic style across all elements. This chal-

lenge is particularly evident in crowded urban scenes where the discriminator must simultaneously evaluate both global composition and local stylistic details. Additionally, the system shows reduced effectiveness in extreme lighting conditions and with unusual color palettes, suggesting potential overfitting to traditional impressionist color distributions in the pre-training phase.

A fundamental limitation lies in the trade-off between style transfer and semantic preservation. While cycle consistency loss helps maintain structural integrity, the model occasionally sacrifices important identifying features, particularly in faces and architectural details. The system also lacks fine-grained user control over stylization parameters such as brush stroke size or abstraction level, a common limitation in GAN-based approaches.

Hardware constraints further impacted performance, as the team's reliance on personal GPUs restricted the model's complexity and depth. Limited computational resources prevented the inclusion of additional layers in the generators, which are crucial for capturing finer details and complex stylistic transformations. With access to more powerful hardware, such as high-performance GPUs or distributed systems, the model could better handle intricate brush strokes, lighting variations, and abstract patterns that define impressionist art.

Future improvements could include multi-scale discriminators for complex scenes, decomposed style representations for better control, semantic-aware loss terms for feature preservation, and access to high-performance hardware to enable deeper architectures. These enhancements could significantly improve the system's practical utility and performance.

8 Ethical Considerations

Copyright: The issue with copyright regarding generative AI is that the dataset used may include data points that were illegally obtained or may be used for purposes outside of its outlined use cases. For example, many large tech companies in 2024 were buying datasets from 3rd-party sources to train their own LLMs and AI agents, some of which scraped data from the internet illegally [8]. In the case of this project, the WikiArt, ImageNet, and ArtiFact datasets was used. These are widely available, open-source datasets for training machine learning models to recognize, classify, and generate art [5]. In other words, they are well-established datasets that have no limits on their use cases in terms of legality.

Legitimacy of AI Generated Art: While creating a generative AI model for art may be legally permissible, it raises obvious ethical concerns. For instance, training a model on a digital artist's style to produce and sell similar art could devalue the artist's work. In this project, the generated art mimics impressionist painting, which is inherently tied to physical mediums like paint. This distinction preserves the unique details and imperfections of traditional art. Additionally, generative art can benefit artists by accelerating workflows, such as frame interpolation, or providing creative references. The team's model replicates the impressionist style, serving as a resource for artistic inspiration without replacing the physical qualities of traditional painting.

Privacy: Privacy is a serious issue. With that said, the privacy risks involved with this project's use case are reasonably minimal as images are only being converted into an impressionist-style digital painting. Additionally, the datasets this project uses are open-source, meaning they do not contain any sensitive data. Despite the low privacy risks, noise will still be introduced during training to limit the weight of any individual data point and reduce pattern recognition tendencies in the model.

9 Conclusion

The benefits of pre-training the discriminators are apparent: it allows for greater control over training generator models for style transfer in CycleGAN by providing a defined standard for the generator models to reach. Although these benefits were not fully realized in this project due to time and compute constraints, the work done in exploring pre-training discriminators for CycleGAN yielded promising results. Despite the short time span since this project's inception, it has demonstrated the feasibility of training generator models with pre-trained discriminators. Utilizing the CycleGAN architecture, the generator model trained on pre-trained discriminators was capable of learning and producing the textures of an impressionist art piece while maintaining some resemblance of the original image.

References

[1] G. Roque. Impressionism. In M. R. Luo, editor, *Encyclopedia of Color Science and Technology*. Springer, New York, NY, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 27, 2014.

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[4] 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 (ICCV)*, pages 2223–2232, 2017.

[5] Steubk. Wikiart dataset. Kaggle, 2023.

[6] dimensi0n. Imagenet 256x256 dataset. Kaggle, 2023.

[7] awsaf49. Artifact: Real and fake image dataset. Kaggle, 2022.

[8] Mia Sato. Apple, anthropic, nvidia, and salesforce reportedly used youtube videos for training data without copyright clearance, 2024. Accessed: 2024-11-09.