Deep Learning and Its Applications to Signal and Image Processing and Analysis

361.2.1120

Final Project

https://github.com/yuvalira/Monet_style_transfer

Yuval Ratzabi, Shahar Ain Kedem

Abstract

In this project, we addressed the challenge of unpaired image to image translation in the Kaggle competition *"I'm Something of a Painter Myself"*, which requires generating Monet style paintings from photographs.

Our baseline model was a CycleGAN with U-Net generators, which relies on adversarial, cycle-consistency, and identity losses to enable translation across unpaired domains. For the improved model, we developed a meaningful variation of CycleGAN by expanding the network capacity, integrating self-attention layers to capture long-range dependencies and adding perceptual loss and LSGAN loss to enhance stylistic fidelity and stability during training.

Our experimental evaluation, conducted over three independent runs, demonstrated that while the baseline CycleGAN effectively captured the overall style transfer, our proposed model consistently achieved superior MiFID scores, with improvements that exceeded the standard deviation of the baseline results. This provides strong evidence that the observed performance gains are both meaningful and statistically robust, confirming that architectural enhancements and refined loss functions yield more realistic and aesthetically coherent outputs.

Introduction, Objective, and Data Description

Introduction and Objective

This project addresses the "I'm Something of a Painter Myself" Kaggle challenge, which focuses on the problem of unpaired image-to-image translation. The specific task is to generate Monet-style paintings from real-world photographs without the availability of paired training data. This problem lies within the framework of Generative Adversarial Networks (GANs) and requires the model to learn a bidirectional mapping between two distinct image domains in the absence of explicit correspondences. The objective of this work is to develop and evaluate models capable of performing this translation with high perceptual quality and stylistic fidelity.

Dataset and Exploratory Data Analysis

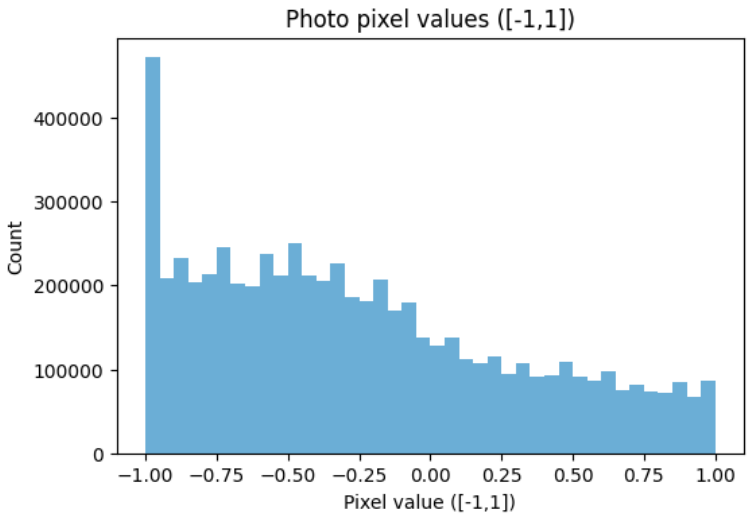
The dataset is provided by the competition consists of two domains:

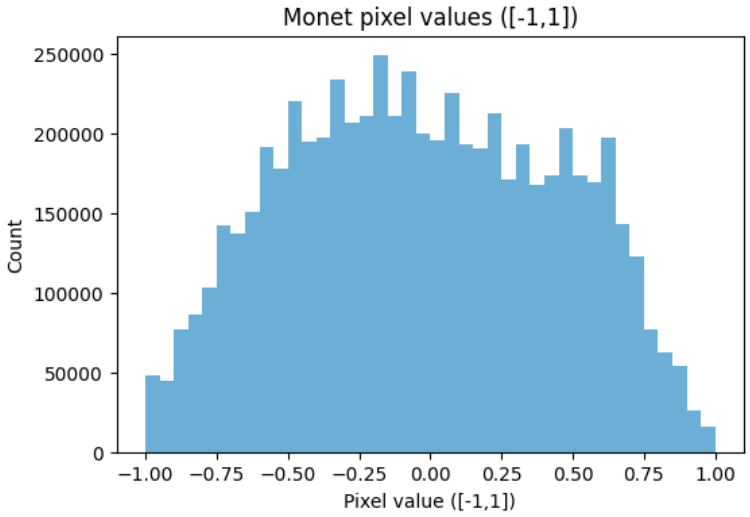
- Monet paintings: 300 images of artworks by Claude Monet, capturing his distinct style of brushwork, color palette, and composition.
- Photographs: 7,038 real-world landscape images, depicting scenes similar in subject to Monet's paintings (e.g., gardens, rivers, nature).

The dataset is unpaired, making the problem more challenging, requiring the model to learn style transfer without pixel-wise alignment.

Visual observation:

All images were resized and decoded into 256×256×3 tensors. Pixel values were scaled to the range [−1, 1]. Exploratory analysis of pixel-value distributions indicated a systematic domain shift, photographs tended to be darker and higher-contrast (mean 0.1887, std 0.5547), while Monet images clustered around mid-tones with softer contrasts (mean 0.0121, std 0.4741). This highlights the need for the model to learn not only stylistic transformations but also global differences in brightness and contrast.





Evaluation Metric

We used MiFID (*Memorization-informed Fréchet Inception Distance*), an extension of FID (*Fréchet Inception Distance*).

FID measures the distance between the feature distributions of real and generated images, extracted using a pre-trained Inception network. Lower values indicate higher similarity between distributions, and therefore, improved generation quality.

$$FID = \|\mu_r - \mu_g\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right)$$

MiFID adds a penalty for memorizing training images by computing the minimum cosine distance between generated and real images.

$$d_{ij} = 1 - \cos\left(f_{gi} - f_{rj}\right) = 1 - \frac{f_{gi} \cdot f_{rj}}{|f_{gi}||f_{rj}|}$$

$$d = \frac{1}{N}\sum_i \min_j d_{ij}$$

$$d_{thr} = \begin{cases} d & if\ d < \varepsilon \\ 1 & otherwise \end{cases}$$

$$MiFID = FID \cdot \frac{1}{d_{thr}}$$

Model 1

Model Architecture

Our baseline model is a CycleGAN Vanilla implementation using a U-Net generator and a PatchGAN discriminator.

Generator (U-Net):

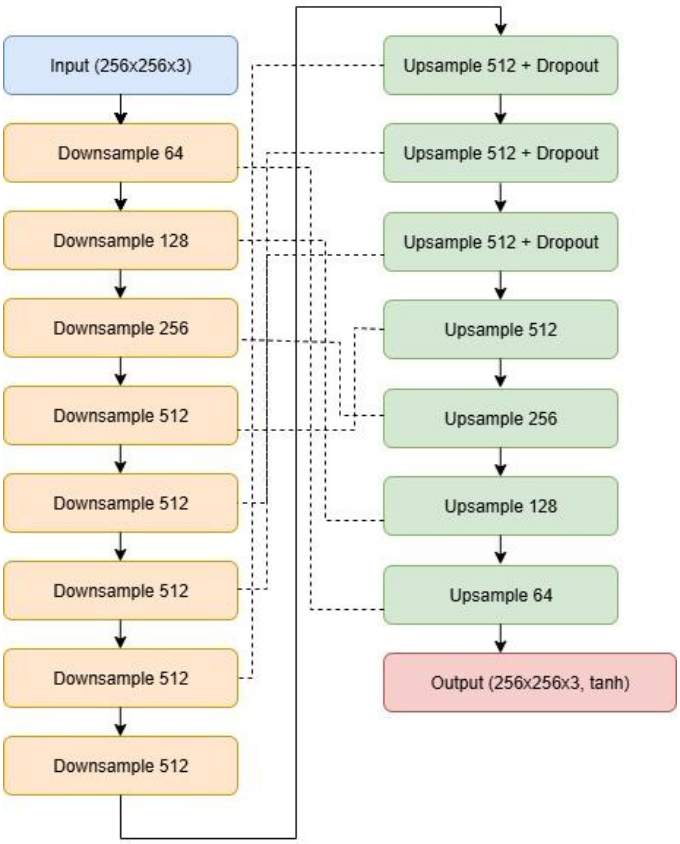
- Downsampling path: Eight convolutional blocks, each halving the spatial dimensions using a stride of 2. The method uses InstanceNormalization instead of BatchNormalization to improve style transfer stability.
- Upsampling path: Seven transposed convolutional blocks, each doubling the spatial dimensions. Dropout is applied in the first three layers of the upsampling path to reduce overfitting.
- Skip connections: Long skip connections link each downsampling layer to its corresponding upsampling layer to preserve spatial details and mitigate vanishing gradients.
- Output layer: A Conv2DTranspose layer with tanh activation produces the final 256×256×3 RGB image.

Discriminator (PatchGAN):

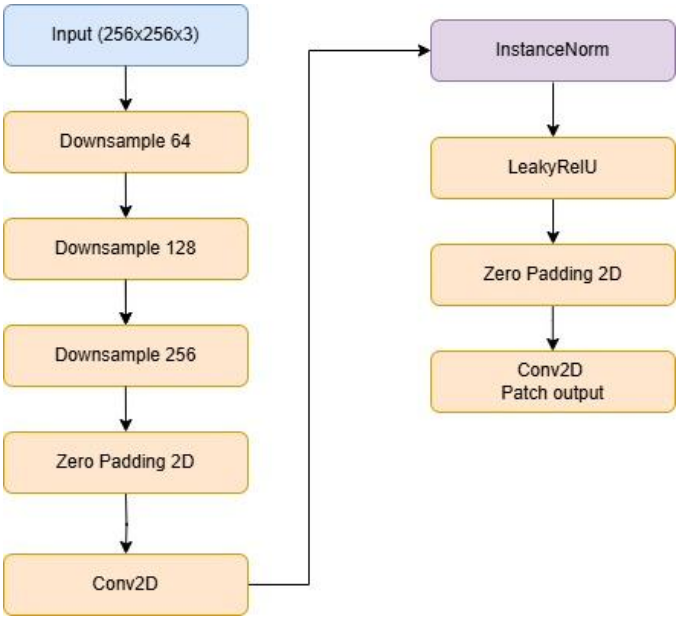
- Operates on image patches instead of full images, outputting a 30×30×1 matrix where each value represents the real/fake prediction for a local patch.
- Three downsampling blocks are followed by zero-padding and a convolutional layer, ending with a single output channel.
- Uses InstanceNormalization and LeakyReLU activations.

The CycleGAN model contains two generators (Photo→Monet and Monet→Photo) and two discriminators (Photo Discriminator and Monet Discriminator), trained jointly using adversarial, cycle-consistency, and identity losses.

Model Diagram – Generator:



Model Diagram – Discriminator:



Data Preprocessing

- Dataset: Monet paintings (300) and photographs (7,038) provided by the Kaggle competition.
- Preprocessing steps:
    1. Images were decoded from TFRecords, resized to 256×256, and normalized to $[-1,1]$.
    2. Mini-batches of 4 images were sampled with shuffling enabled to ensure varied pairings.
    3. Channel Matching: A per-channel normalization aligned photo statistics to Monet statistics.

Channel Matching method:

We created a deterministic, per-channel normalization that shifts Photo images so their RGB mean and standard deviation match those of the Monet domain. This step reduces the global tone/contrast gap between domains, allowing the generator to focus on style and texture rather than correcting brightness/contrast. It also stabilizes training and speed up convergence.

We computed per-channel statistics:

Photo: μ_p, σ_p

Monet: μ_m , σ_m

Then, for each pixel x and each channel we applied:

$$y = (x - \mu_p)\frac{\sigma_m}{\sigma_p} + \mu_m$$

And clip the result to $[-1, 1]$.

Hyperparameters and Loss Functions

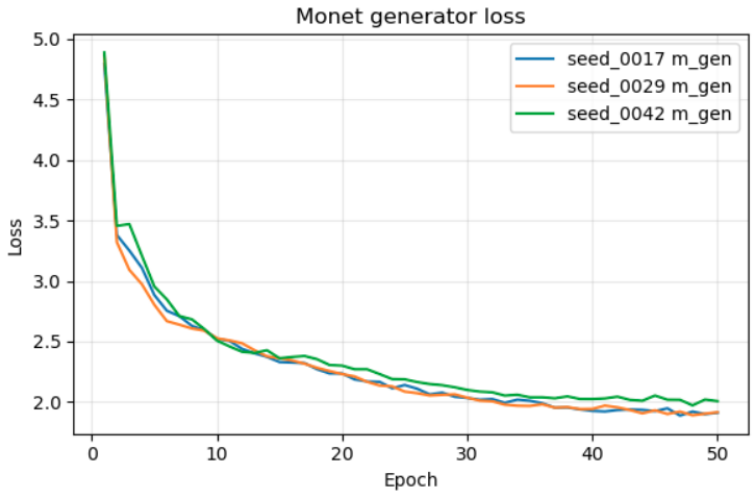
- Optimizers: Adam $lr = 2 \cdot 10^{-4},\ \beta_1 = 0.5$ for all four networks.
- Loss functions:

    1. Adversarial Loss: Binary Cross-Entropy (from logits) for discriminator and generator.

    2. Cycle Consistency Loss - L1 Loss between real and cycled images, scaled by λ.
    3. Identity Loss - L1 Loss between real and same-domain generated images, scaled by 0.5λ.

- Epochs: 5 (per seed) with seeds [42, 17, 29]
- Batch size: 4.

Convergence behavior



The figure above illustrates the Monet generator loss of the baseline CycleGAN across 50 epochs for three independent seeds. The loss decreases steadily throughout training and converges toward a stable low value near epoch 40. This monotonic decline indicates that the generator progressively improves at producing outputs that the discriminator cannot distinguish from real Monet paintings. While this reflects convergence, it also highlights a limitation: the discriminator in the baseline CycleGAN does not exert sufficient adaptive pressure, allowing the generator's loss to steadily fall without the oscillatory dynamics often expected in adversarial training.

Test results

Model	Average MiFID Score	Standard Deviation	Number of Epochs	Number of Seeds
Baseline CycleGAN	92.9	3.2	50	3

Qualitative examples:

Original: 0aec1f9701 (good)

Generated Monet

Original: 5ee0ea499b (good)

Generated Monet

Original: 7b4f952cda (bad)

Generated Monet

Original: 29c34c1abd (bad)

Generated Monet

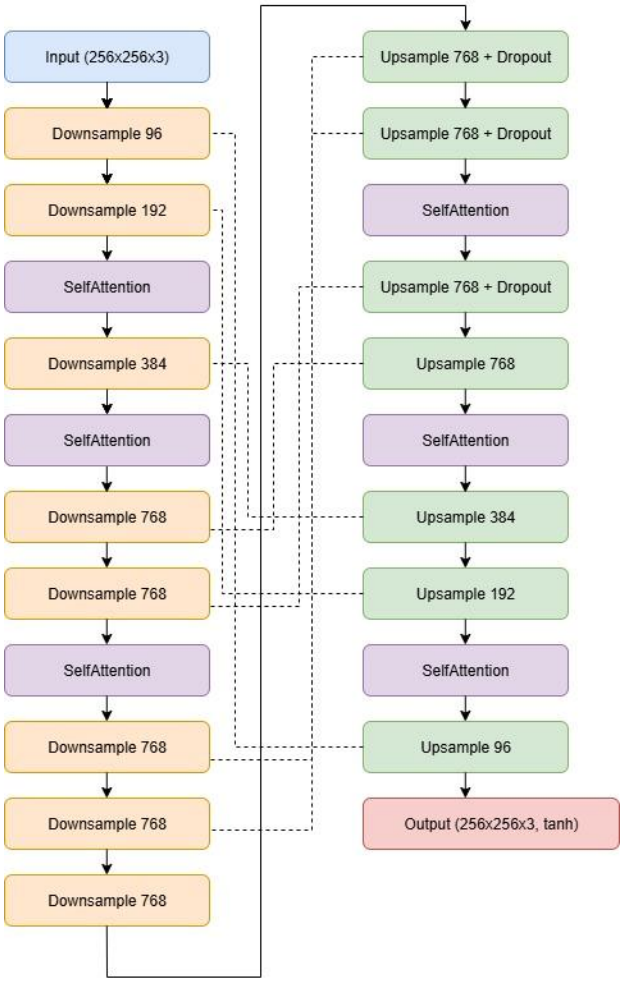
Model 2

The improved model is based on the CycleGAN framework for unpaired image-to-image translation, augmented with four key modifications :

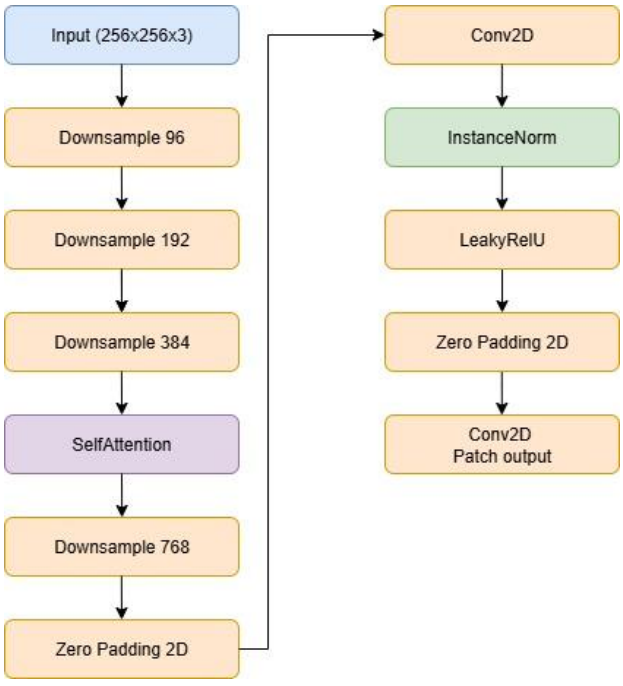
1. Self-Attention : Self-attention layers were introduced at intermediate stages of the generator, enabling the model to capture long-range spatial dependencies. This mechanism improves global structural coherence by allowing distant regions of the image to directly influence one another, which is critical for maintaining consistent artistic style across the entire scene.
2. Increased Network Capacity: Both the generator and discriminator architectures were widened by increasing the number of channels. This expansion provided greater representational capacity, allowing the networks to learn and reproduce more intricate stylistic features characteristic of Monet's paintings.
3. Perceptual Loss: In addition to the conventional adversarial, cycle-consistency, and identity losses, a perceptual loss was incorporated. This loss leverages feature representations extracted from a pretrained VGG19 network to encourage the generator to preserve high-level semantic content (e.g., object shapes) while accurately capturing fine-grained stylistic textures.
4. Hybrid Adversarial Objective: The adversarial loss was extended to combine the standard binary cross-entropy objective with the Least Squares GAN (LSGAN) formulation. This hybrid loss improves gradient flow and contributes to more stable training dynamics, thereby reducing mode collapse and improving convergence reliability.

Together, these modifications substantially improve the model's ability to generate realistic Monet-style images while preserving semantic fidelity and ensuring stable optimization.

Model Diagram – Generator:



Model Diagram – Discriminator:



Data Preprocessing

The preprocessing steps followed the same procedure as in the baseline model.

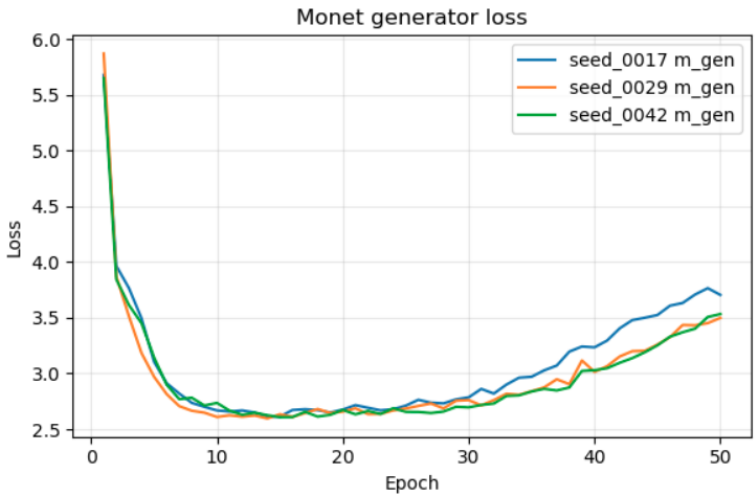
Hyperparameters and Loss Functions

The improved CycleGAN was trained for 50 epochs on a single GPU, with three independent runs using random seeds (42, 17, 29) to ensure reproducibility. The batch size was set to 1, which is standard for high-resolution GAN training. Optimization employed the Adam optimizer with the following settings $lr = 1.5 \cdot 10^{-4}$, $\beta_1 = 0.5$.

As we explained earlier the training objective integrated several complementary losses:

1. Adversarial Loss (BCE): Standard GAN formulation using binary cross-entropy.
2. Least Squares GAN (LSGAN) Loss: Applied to both generator and discriminator with a weight of 0.5 to improve gradient flow and training stability.
3. Cycle-Consistency Loss: Weighted by λ=10\lambda = 10λ=10 to enforce bidirectional mapping coherence.
4. Identity Loss: Weighted by 0.5 to encourage content preservation when the input already belongs to the target domain.
5. Perceptual Loss: Based on VGG19 feature representations, scaled by a coefficient of 0.002, to enforce semantic and textural fidelity.

convergence behavior:



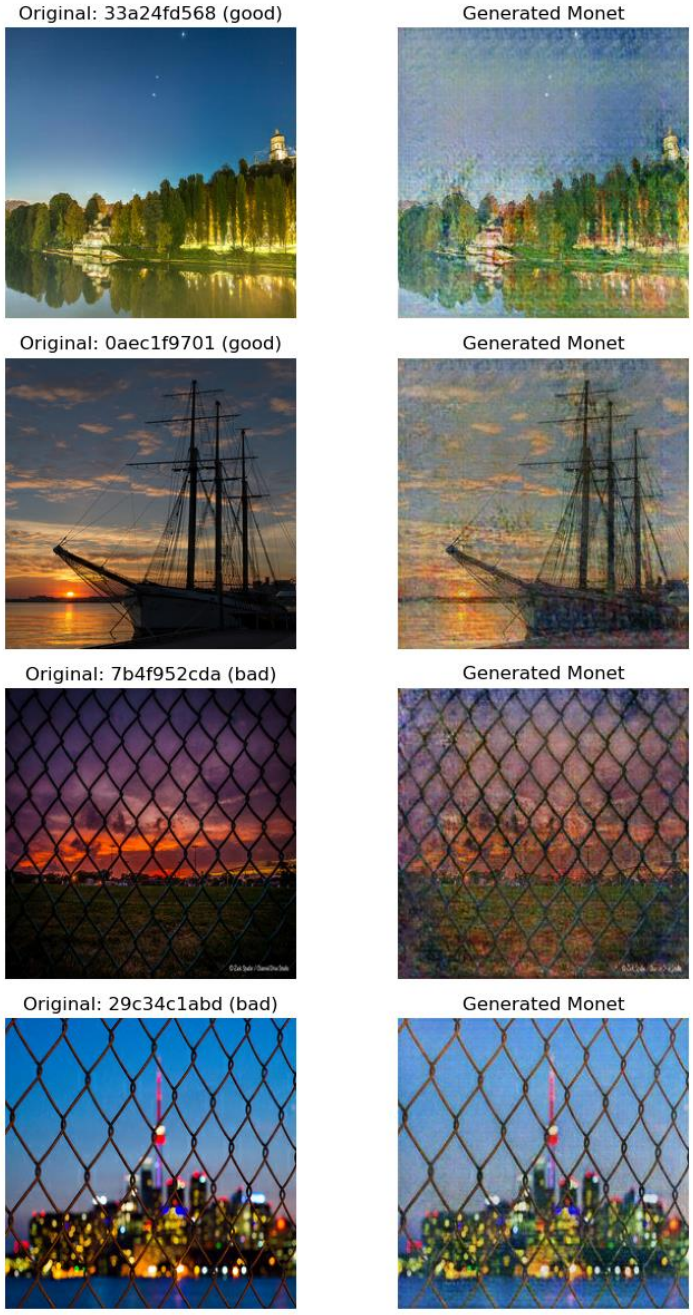
The figure above shows the Monet generator loss across 50 epochs for three training runs with different random seeds. The curves exhibit an initial sharp decrease followed by stabilization between epochs 10-20, indicating successful

learning of the photo to Monet mapping, with a later gradual rise reflecting the adversarial dynamics as the generator and discriminator adapt to each other. The similarity across seeds highlights the reproducibility of training.

Test results:

Model	Average MiFID Score	Standard Deviation	Number of Epochs	Number of Seeds
Proposed Model	87.1	1.3	50	3

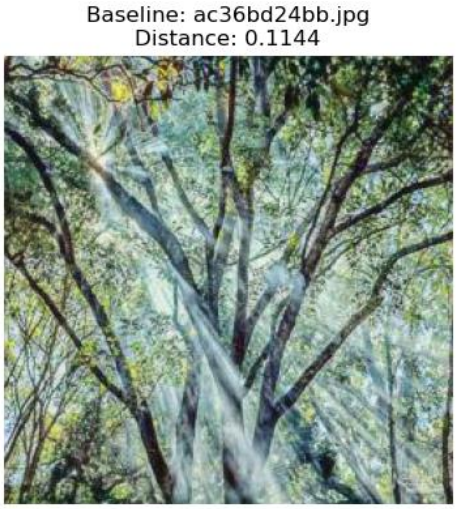
Qualitative examples:

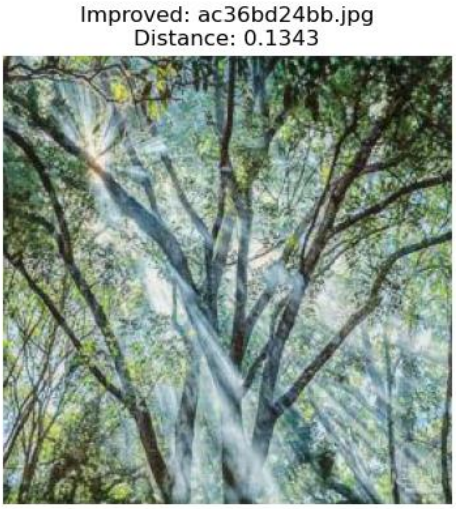


Results

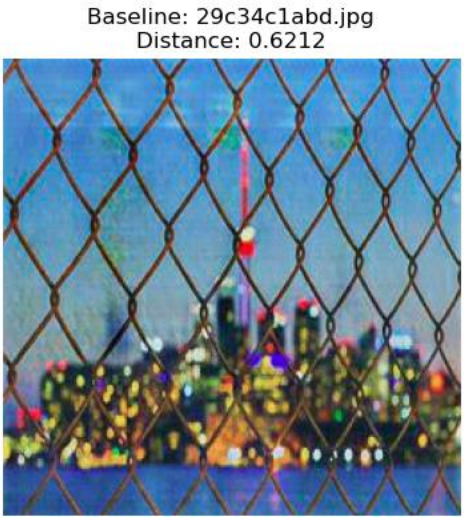
We compared the baseline and improved models by extracting InceptionV3 features from their generated images and computing cosine distances to the closest real Monet images. Per image distances quantify similarity, with lower values indicating better quality. Using these distances, we identified common images across both models and applied percentile thresholds (best 20% as "good," worst 20% as "bad"). This allowed us to systematically select representative cases in four categories:

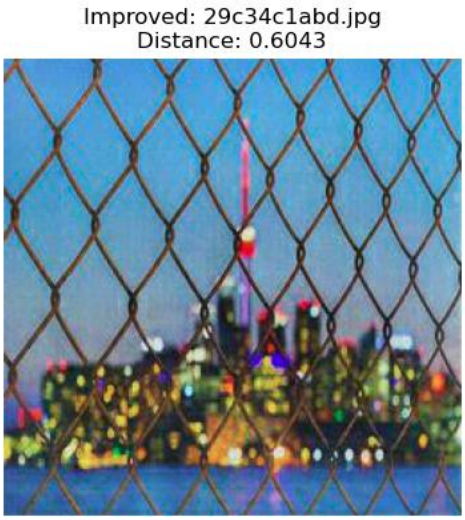
- Both models performed well:



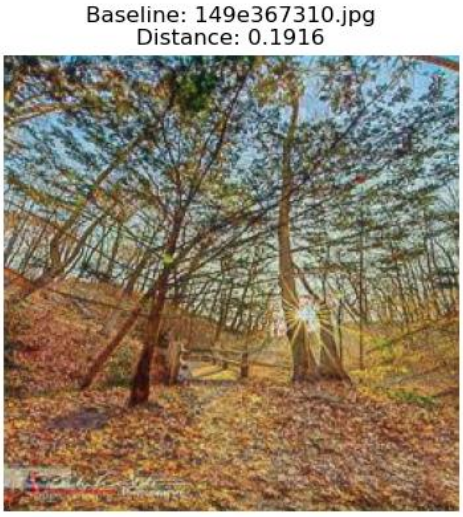


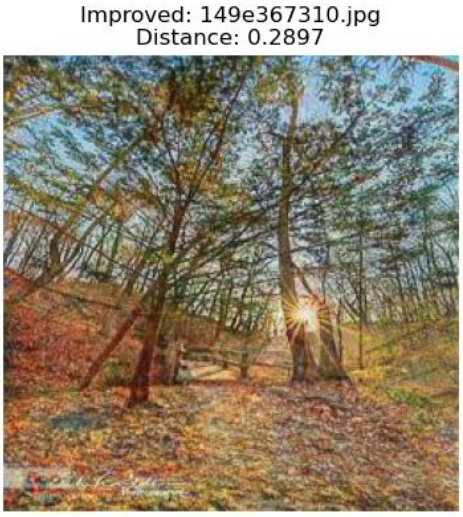
- Both models performed poorly:



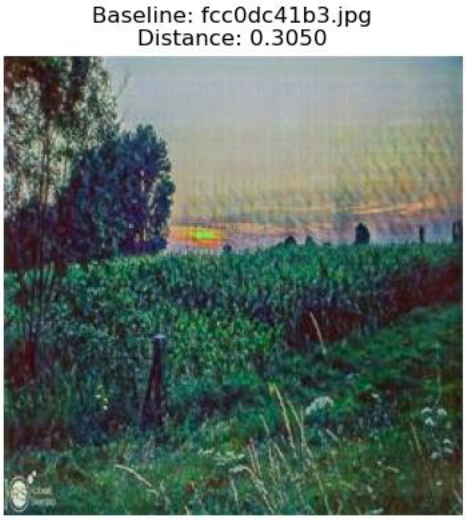


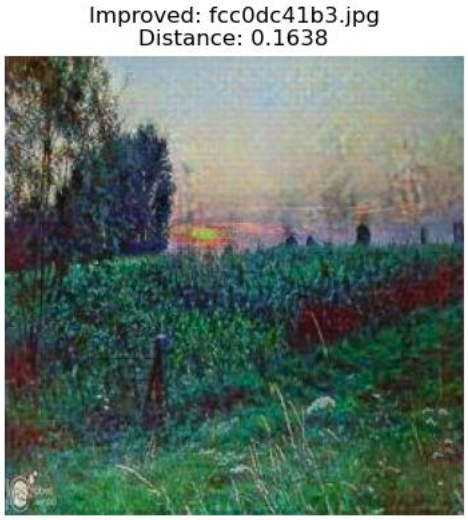
- Model 1 performed well and Model 2 did not:





- Model 2 performed well and Model 1 did not:





The MiFID results for both models:

Model	Average MiFID Score	Standard Deviation	Number of Epochs	Number of Seeds
Baseline CycleGAN	92.9	3.2	50	3
Proposed Model	87.1	1.3	50	3

Quantitative Results:

The baseline CycleGAN achieved an average MiFID score of **92.9** across 50 training epochs and three seeds, with a standard deviation of **3.2**. The improved model obtained a significantly lower average MiFID score of **87.1** and a reduced standard deviation of **1.3**. Since lower MiFID values indicate better perceptual fidelity, these results demonstrate that the improved model consistently surpasses the baseline. Moreover, the observed improvement exceeds the baseline's variability range (±3.2), providing strong evidence that the performance gain is statistically meaningful rather than attributable to randomness. The reduced variance further highlights the greater stability and reproducibility of the improved approach across runs.

Explanation for the reasons behind the outcome:

The superior performance of the improved model can be attributed to its architectural and training refinements. The incorporation of self-attention modules enabled the network to capture long-range spatial dependencies, thereby preserving global structural coherence. The perceptual loss, based on deep feature embeddings from a pretrained VGG19 network, guided the generators to maintain high-level semantic content while accurately reproducing fine textural details. The addition of the least-squares adversarial loss improved gradient flow, leading to more stable training dynamics and reducing the risk of mode collapse. Finally, widening both the generator and discriminator networks expanded their representational capacity, allowing for more precise modeling of complex patterns and stylistic nuances. Together, these enhancements explain not only the improved MiFID scores but also the reduced variance, as they provided stronger training signals and more expressive model components.

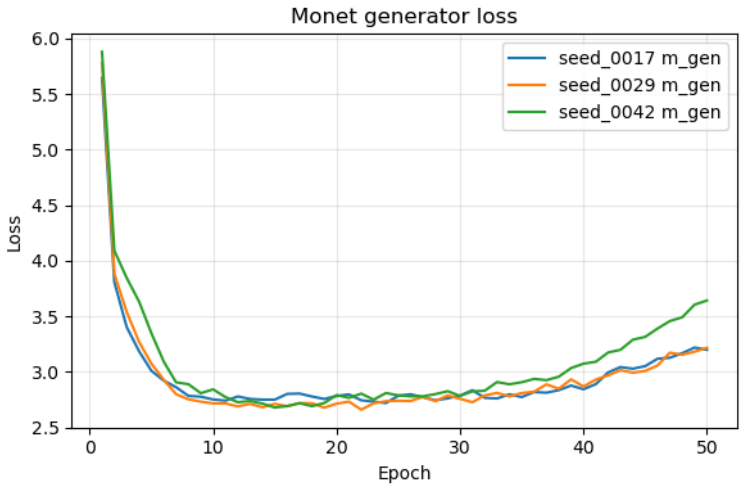
Ablation Studies

Ablation Setup and Motivation:

For the ablation study, we examined the contribution of self-attention mechanisms in the improved CycleGAN variant. In the full model, self-attention layers were integrated into both the generator and discriminator to capture long-range spatial dependencies and preserve global structural coherence. To isolate their effect, we trained an ablated model in which all self-attention modules were removed, while keeping every other architectural and training aspect identical.

The goal was to isolate and quantify the role of self attention in improving image translation quality and training stability. By comparing the performance of the attention enabled model against the attention free variant, we can determine whether self attention provides a meaningful benefit beyond standard convolutional processing.

Training Loss Comparison:



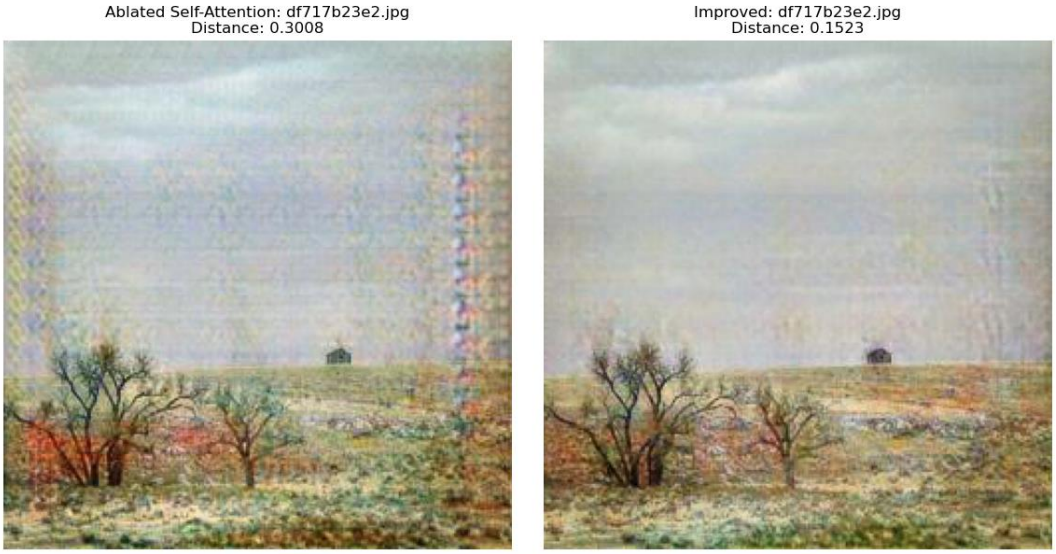
Including self-attention lowered the generator's minimum loss, reflecting improved fidelity. However, it also introduced greater variability across seeds after ~30 epochs, whereas models without self-attention, as we can see in the graph above, were more stable but slightly less accurate. This suggests a trade-off between higher quality and training stability.

Quantitative Results:

Model	Average MiFID Score	Standard Deviation	Number of Epochs	Number of Seeds
Proposed Model – No self attention	89.17	2.6	50	3
Proposed Model	87.1	1.3	50	3

Lower MiFID indicates superior perceptual quality. Removing self-attention degraded performance and approximately doubled the variability across seeds.

Qualitative Results:



The qualitative comparison clearly demonstrates the impact of removing self attention. The right hand image, produced by our full model, shows sharper textures, more coherent color blending, and improved global consistency, as evidenced by the lower distance score (0.1523). In contrast, the left hand image, generated by the ablated model without self attention, exhibits noticeable artifacts such as repetitive checkerboard like patterns in the sky and distorted textures around the edges. These artifacts arise because self attention helps the generator capture long range spatial dependencies, preventing local convolutional filters from producing redundant patterns.

Insights:

The ablation confirms that self-attention is a decisive contributor to both fidelity and stability. Removing it degraded MiFID from 87.1 to 89.17 while doubling variability (standard deviation from 1.3 to 2.6) across identical 50 epoch and 3 seed runs, indicating that the observed gain is robust and not seed dependent. Qualitative evidence further highlights that self attention suppressed grid like sky artifacts and edge distortions and yielded sharper textures with more coherent global color transitions, consistent with its role in modeling long range spatial dependencies that plain convolutions miss.

Conclusions and Summary

In this project, we addressed the challenge of translating photographs into Monet-style paintings, evaluating both a baseline vanilla CycleGAN model and an enhanced variant. The baseline model provided a strong starting point but exhibited notable variability across training runs. To improve upon this, we extended the model by incorporating self-attention layers, expanding its representational capacity, and introducing additional loss terms, including perceptual loss and the Least Squares GAN objective. These modifications enabled the model to better capture global structural dependencies, preserve semantic content, and stabilize adversarial training.

Quantitatively, the enhanced model achieved a mean MiFID score of **87.1** with a standard deviation of **1.3**, compared to the baseline CycleGAN's **92.9** with a standard deviation of **3.2**. Since lower MiFID values indicate higher perceptual quality, this demonstrates that our model not only surpasses the baseline in image generation quality but also outperforms it beyond the baseline's own variability range. The reduced variance further highlights the greater stability and reproducibility of our approach.

Overall, the proposed model enhances both the quality and consistency of photo-to-Monet translation, offering a more reliable framework for artistic image generation.

References

1. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017).
   Unpaired image-to-image translation using cycle-consistent adversarial networks.
   https://arxiv.org/abs/1703.10593
2. Ronneberger, O., Fischer, P., & Brox, T. (2015).
   U-Net: Convolutional networks for biomedical image segmentation.
   https://arxiv.org/abs/1505.04597
3. Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019).
   Self-attention generative adversarial networks.
   https://arxiv.org/abs/1805.08318
4. Johnson, J., Alahi, A., & Fei-Fei, L. (2016).
   Perceptual losses for real-time style transfer and super-resolution.
   https://arxiv.org/abs/1603.08155
5. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017).
   Least squares generative adversarial networks.
   https://arxiv.org/abs/1611.04076
6. Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., & Yoo, J. (2020).
   FID and improved precision and recall metric for generative models.
   https://arxiv.org/abs/2002.09797