Computer Vision Final Project 2025 Group 9

Sevde Yanik
4732565
Data and Computer Science

Sarp Tan Gecim
4175173
Physics

Abstract

Unpaired image-to-image translation has become a key technique in computer vision, with CycleGAN making a big impact, especially in artistic style transfer. In our project, we explore how CycleGAN can be used to transform real-world photographs into Monet-style paintings and back again.

We experimented with different training strategies to see how things like image resolution, cropping methods, and architecture tweaks affect the results. In particular, we looked at how downsampling, random cropping, and grid cropping influence translation quality, and what happens if we remove one of the discriminators.

To evaluate the outputs, we used Fréchet Inception Distance (FID) and Structural Similarity Index (SSIM) to measure both how realistic and structurally consistent the results are. We found that random cropping helped the model generalize better, while downsampling made training faster but at the cost of fine details. These observations give useful tips for anyone tuning CycleGANs for creative tasks like style transfer.

1. Introduction

Generative adversarial networks (GANs) have opened up exciting possibilities in image synthesis, powering applications like style transfer, image-to-image translation, and domain adaptation. One particularly well-known architecture is CycleGAN [4], which is great for unpaired image translation since it can learn to map between two domains without needing paired training examples. The trick lies in its cycle consistency loss, which helps the model reconstruct the original image after translation.

In this project, we used CycleGAN to translate between real-world photographs and Monet-style paintings. While the base CycleGAN model already works quite well, we were curious about how different training choices such as image resolution, cropping methods, and architectural tweaks might affect the results.

We chose the Monet2Photo dataset because it provides a visually distinct and widely used benchmark for artistic style transfer, making it ideal for evaluating stylization performance. The strong differences in color, texture, and structure between photos and Monet paintings make this a challenging and meaningful test case for CycleGAN-based models.

The central research question we explore is: *How do different training strategies and architectural simplifications affect the quality, efficiency, and stability of unpaired image translation with CycleGAN in an artistic context?*

So, we set up a series of experiments to test six different variations of CycleGAN, each focusing on one or more of the following:

- **Image resolution:** Does training on higher-resolution images help, or is it too computationally expensive?
- **Cropping strategies:** We tried both random cropping and fixed grid-based cropping to see which one helps generalization more.
- **Simplified architecture:** What happens if we remove one of the discriminators? Can we still get decent results?

To evaluate the models, we looked at both the visuals and some standard metrics such as Fréchet Inception Distance (FID) for perceptual quality and Structural Similarity Index (SSIM) for preserving structure. Overall, our experiments reveal some useful trade-offs between quality, efficiency, and complexity in training CycleGANs.

2. Related Work

CycleGAN, introduced by Zhu *et al*. [4], was a major breakthrough in unpaired image-to-image translation. By using a cycle consistency loss, it made it possible to train models without paired datasets, which opened the door for many creative applications, from artistic style transfer to domain adaptation.

Earlier work by Johnson *et al*. [2] showed how effective residual networks can be for style transfer. Their use of deep architectures with skip connections helped preserve

both content and texture during translation. This inspired the design of CycleGAN's generator, which uses residual blocks to learn rich features.

When it comes to evaluating generative models, the Fréchet Inception Distance (FID) and the Structural Similarity Index (SSIM) have become go-to metrics. FID [1] compares the distribution of generated images to real ones using features from a pretrained network, giving a sense of perceptual realism. SSIM [3], on the other hand, looks at similarities at the pixel level to assess how well the structure is preserved.

Our work builds on these foundations, but instead of proposing a new model, we take a closer look at how small changes, such as cropping strategy, image resolution, or removing a discriminator, affect performance. This lets us better understand what matters most when training CycleGANs for artistic style transfer.

3. Method

In this section, we describe the CycleGAN architecture we used, how we prepared the data, and the different model variations with which we experimented. Our goal was to understand how changes in training setup and architecture affect both visual quality and efficiency when translating between Monet-style paintings and real-world photos.

3.1. CycleGAN Architecture

CycleGAN [4] is built around two generators and two discriminators:

- G_{AB}: Translates real-world photos (**Domain A**) to Monet-style paintings (**Domain B**).
- G_{BA}: Translates Monet paintings back to photos.
- D_A: Tries to distinguish real photos from generated ones.
- D_B: Tries to distinguish real Monet paintings from generated ones.

Both generators are based on ResNet with six residual blocks [2]. These residual connections help the model keep fine details while applying stylistic changes. For the discriminators, CycleGAN uses a PatchGAN setup, which focuses on small image patches instead of entire images. This makes it better at capturing local texture patterns.

During training, two main losses guide the learning process:

- **Adversarial Loss:** Encourages generators to produce images that fool the discriminators.
- **Cycle Consistency Loss:** Makes sure that an image translated to the other domain and back comes out similar to the original.

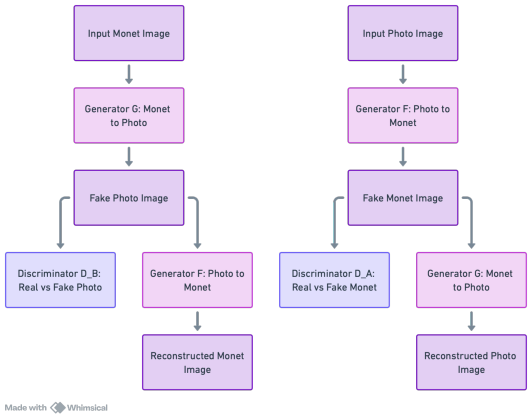

Figure 1. CycleGAN architecture with two generators and two discriminators for unpaired image translation.

The overall structure of CycleGAN is illustrated in Figure 1. It uses two generator-discriminator pairs to learn bidirectional mappings between two image domains. The cycle consistency loss ensures that translations preserve content by reconstructing the original image after a forward and backward pass.

3.2. Dataset and Preprocessing

We used the Monet2Photo dataset, which contains unpaired images from two domains: landscape photos and Monet-style paintings. Since CycleGAN doesn't need paired examples, we focused on how to prepare the data in a way that helps the model learn effectively.

The preprocessing steps we applied include:

- **Resizing:** Images were resized to either 128×128 or 256×256 depending on the experiment.
- **Cropping:** We tried both random cropping and fixed grid-based cropping to compare their impact on generalization.
- **Normalization:** All pixel values were scaled to the range $[-1, 1]$.

3.3. Model Variants and Rationale

To explore how training setup and architecture affect performance, we implemented six different variants of CycleGAN:

3.3.1 Model 1: Standard CycleGAN (256×256)

Our baseline model uses full-resolution (256×256) images with the original CycleGAN setup. It produces detailed results but takes longer to train.

3.3.2 Model 2: Downsampling to 128×128

Here we resized input images to 128×128. This speeds up training and reduces memory use, but can lead to loss of fine details.

3.3.3 Model 3: Random Cropping

Instead of resizing the whole image, we randomly cropped 128×128 patches from the originals. This adds variation and helps the model generalize better, but some outputs may look a bit inconsistent or blurry.

3.3.4 Model 4: Grid-Based Cropping

Each image is split into fixed 128×128 regions using a grid. This ensures that all areas of an image are seen during training, though this can sometimes lead to visual discontinuities if the model doesn't capture the connections between patches.

3.3.5 Model 5: Downsampling + Random Cropping

This hybrid approach uses a dynamic transformation using `RandomResizedCrop(128, scale=(0.8, 1.0))`, which randomly rescales and crops the image. It introduces more spatial and scale variation and proved to be a good middle ground between efficiency and quality.

3.3.6 Model 6: Removing One Discriminator

To test the role of adversarial supervision, we decided to remove one of the discriminators (either D_A or D_B). This helps us see whether both directions are really necessary for good results, and how much performance we lose (if any) with a simpler setup.

3.4. Training Details

All models were trained using the Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The key training settings were:

- **Batch size:** Between 4 and 8, depending on resolution and GPU memory.
- **Epochs:** 30–50, with checkpoints saved regularly.
- **Evaluation:** We used both visual inspection and metrics (FID and SSIM) to assess translation quality.

3.5. Expected Outcomes

Based on previous research and our own expectations, we figured that:

- Random and grid-based cropping could help the model generalize better and improve texture quality.
- Lower resolutions would probably lead to faster training but less detailed outputs.
- Removing one discriminator would reduce performance, but possibly not as much as expected.

4. Experiments

In this section, we present the experiments we conducted to evaluate different CycleGAN variants for translating between Monet-style paintings and real-world photos. We tested six model configurations, each designed to explore specific training strategies or architectural changes. Our goal was to understand how these choices affect image quality, training efficiency, and overall stability.

4.1. Training Setup

All models were trained on the Monet2Photo dataset. We applied different preprocessing strategies, such as downsampling and cropping, depending on the experiment. Each model was trained for 30 to 50 epochs using the Adam optimizer with a learning rate of 0.0002. We saved checkpoints regularly and evaluated the models both qualitatively and quantitatively.

To assess performance, we used the metrics FID and SSIM. FID helps evaluate the realism of generated images, while SSIM measures how well structural details are preserved.

4.2. Model Comparisons and Observations

4.2.1 Model 1: Standard CycleGAN (256×256)

This model was trained for 30 epochs using full-resolution images resized to 256×256 pixels. It adheres closely to the original CycleGAN setup and serves as our baseline. The results showed relatively detailed and stylistically strong outputs. However, due to the increased computational demand, training was slower. Visual artifacts, such as checkerboard patterns, occasionally appeared, particularly in flat-textured areas. Additionally, some translated Monet-style images retained too much of the original photographic content, which slightly weakened the stylization effect.

Observation: High-resolution training preserves finer details and texture but makes training harder and more expensive.

4.2.2 Model 2: Downsampling to 128×128

This variant reduces the input image resolution to 128×128 before training, also trained for 30 epochs. The model benefited from faster training and lower memory consumption, allowing more frequent updates. However, the resulting images lacked some of the sharpness and fine stylistic details present in higher-resolution outputs. Artistic textures

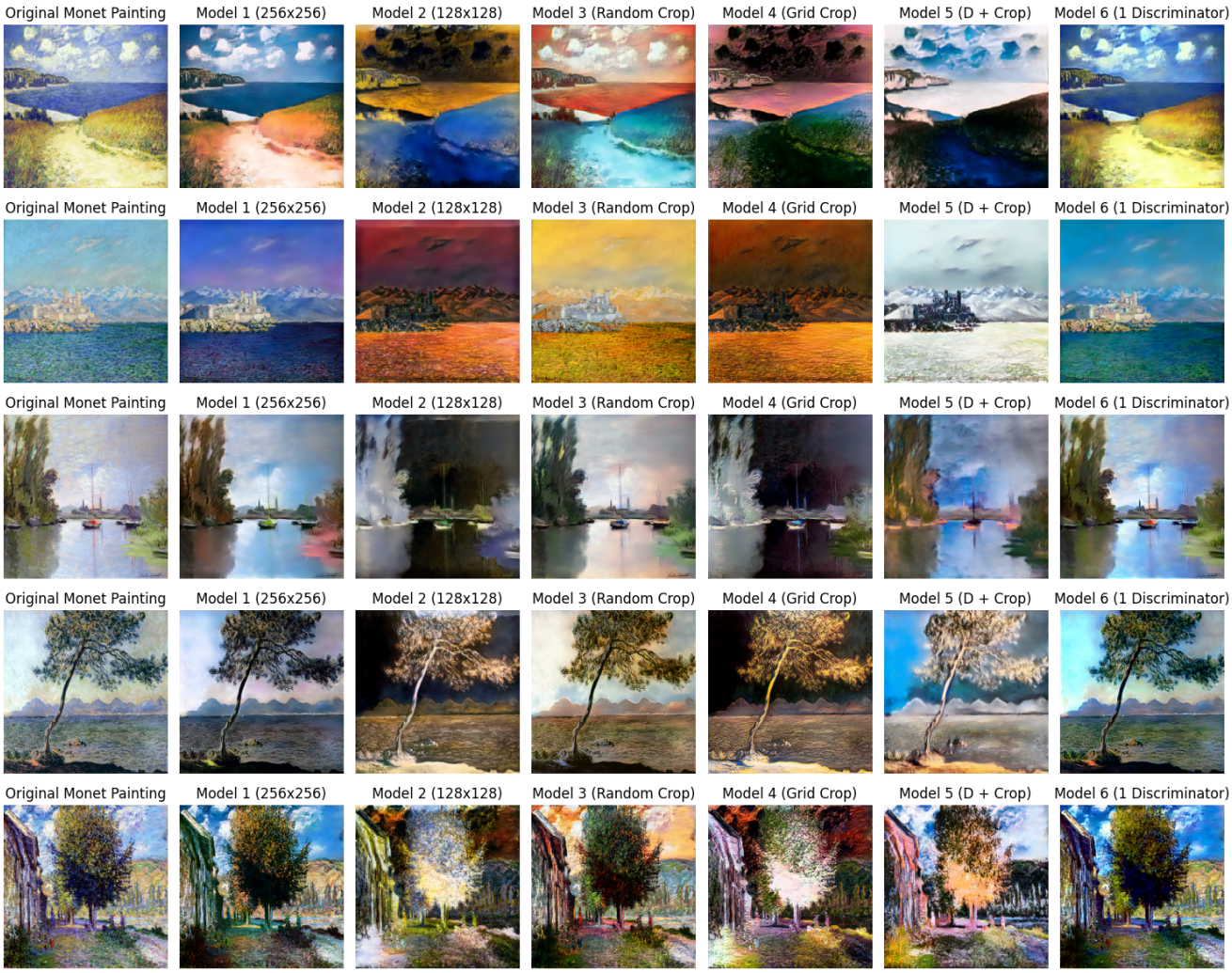

Figure 2. Monet → Photo translation results for all models (Model 1 to Model 6).

were often simplified, and structural details occasionally appeared blurred.

Observation: While computationally efficient, lower-resolution training sacrifices output sharpness and richness in style features. However, with longer training or additional regularization techniques, this model could potentially offer a viable trade-off between quality and efficiency.

4.2.3 Model 3: Random Cropping

In this experiment, each image was randomly cropped to 128×128. This method introduces spatial diversity by forcing the model to learn from different regions of an image at each iteration. It was trained for 50 epochs, allowing it to adapt to a wider range of input variations. It performed surprisingly well, the resulting outputs exhibited strong stylistic patterns and vibrant colors, often capturing the Monet aesthetic more distinctly than lower-resolution baselines. However, due to the stochastic nature of cropping, some outputs occasionally lost structural consistency, resulting in blurred or deformed shapes.

Observation: Random cropping performed well for generalization and diversity, but can be unstable without enough training.

4.2.4 Model 4: Grid-Based Cropping

In this model, each 256×256 image was systematically divided into four non-overlapping 128×128 patches using a fixed grid. This ensured that all spatial regions of the image were seen during training. Each crop was treated as an independent training sample, increasing dataset diversity without requiring additional images. The model was trained for 50 epochs on these grid-cropped patches.

This strategy led to more stable training and consistent learning of image structure. However, since the crops were

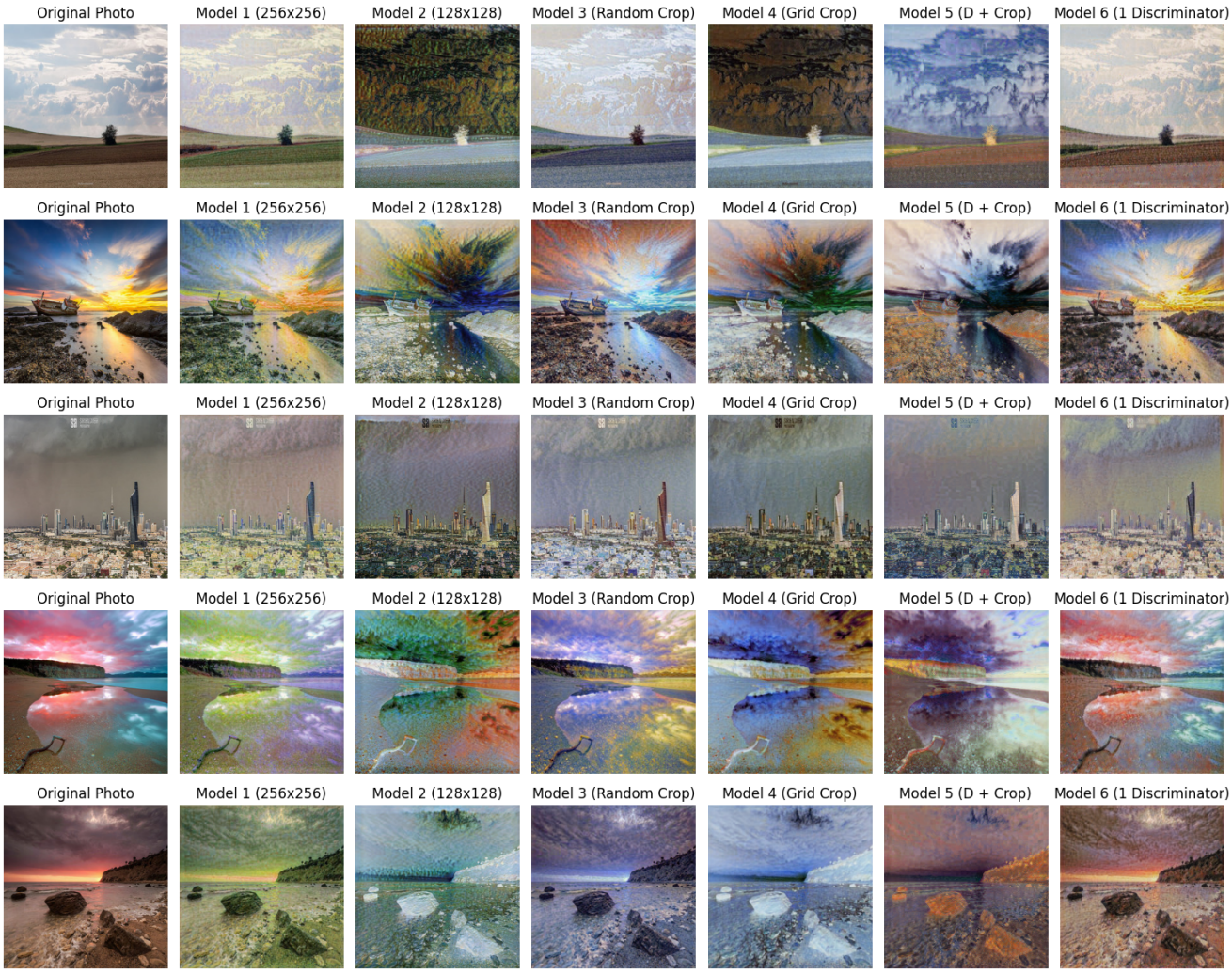

Figure 3. Photo → Monet translation results for all models (Model 1 to Model 6).

extracted independently, there was no guarantee of continuity between neighboring patches. As a result, the generated images sometimes exhibited visible artifacts at patch boundaries or lacked coherent global composition.

Observation: Better spatial coverage, but at the cost of global consistency.

4.2.5 Model 5: Downsampling + Random Cropping

This model applies a hybrid preprocessing strategy using `RandomResizedCrop`, which simultaneously downscales and randomly crops images to 128×128 patches. The scale range of (0.8, 1.0) ensures that cropped regions come from a downscaled area of the original image, introducing variability in both scale and position.

This approach improves generalization by exposing the model to different spatial contexts while reducing training cost compared to full-resolution models. Model 5 was trained for 50 epochs.

Observation: Balanced performance with solid results across metrics and visuals.

4.2.6 Model 6: Removing One Discriminator

This test simplified the architecture by dropping one of the discriminators. After training for 50 epochs, surprisingly, the results were still coherent and decently stylized. The textures weren't as rich, but the overall translation quality was still solid.

Observation: You can simplify the model and still get usable results which is useful especially in low-resource settings.

4.3. Quantitative and Qualitative Analysis

To better understand how each model performed, we computed FID and SSIM scores for both translation direc-

tions — Photo → Monet and Monet → Photo. These are shown in Table 1. Lower FID means the generated images look more realistic (closer to real samples in feature space), and higher SSIM means better preservation of structure.

For context, FID scores below around 150 are generally considered decent for stylization tasks, especially when working with artistic domains where exact realism is not expected. SSIM values closer to 1 indicate better structural preservation, though perfect scores are rare in style transfer.

A few things stood out:

- **Model 3 (Random Cropping)** produced visually compelling outputs with rich stylistic features and strong generalization.
- **Model 5 (Downsampling + Random Cropping)** generated consistently well-structured images and proved efficient in both training and inference.
- **Model 6 (1 Discriminator)** was a surprise, even with a simplified setup, it outperformed several full CycleGAN models.
- **Model 1** (256×256) maintained strong visual details but incurred higher computational cost and some instability.

Figure 2 and Figure 3 provide a qualitative comparison of the translations produced by each model for Monet → Photo and Photo → Monet tasks, respectively.

Overall, the models that included some form of cropping (especially random) tended to generalize better and produce more artistic outputs. But they also required longer training to stabilize.

4.4. Training Stability and Convergence

Training stability varied quite a bit across the models. The full-resolution model (Model 1) was the most unstable as its losses fluctuated more, and the outputs showed more variation in quality. On the other hand, the downsampled models (like Model 2 and 5) converged faster and more consistently.

Model 5 especially struck a good balance: it trained efficiently, stayed stable throughout, and produced strong visual results. Model 6 also trained quite smoothly, likely due to its reduced complexity, though its outputs weren't quite as sharp.

Summary: Stability and efficiency improved when using lower resolutions or fewer components, but high-resolution models still had an edge in fine detail if you have the resources to support them.

5. Ablation Study

To better understand the impact of our design choices, we ran an ablation study focusing on image resolution, cropping strategy, and architectural simplification. Each model tweak was isolated and analyzed to see how much it contributed to the overall performance.

5.1. Effect of Image Resolution

We compared models trained on 256×256 (Model 1) and 128×128 (Model 2) images.

- **256×256 (Model 1)** gave more detailed and textured outputs, but it also made training slower and more unstable.
- **128×128 (Model 2)** trained much faster and more smoothly but often produced blurrier results with less stylization.

Takeaway: Higher resolution improves image quality, but only if your hardware and training setup can handle the extra load.

5.2. Effect of Random Cropping

Random cropping (Model 3) helped the model generalize better by showing it different parts of the image during training.

- Outputs looked more varied and expressive.
- Occasionally, important structures got cut off, which led to blurry or inconsistent generations.

Takeaway: Random cropping adds valuable diversity but may need longer training to avoid instability.

5.3. Effect of Grid-Based Cropping

Model 4 used grid-based cropping to ensure the model sees all regions of the image.

- Coverage was more uniform.
- However, the model sometimes struggled with continuity across patches, resulting in edge artifacts.

Takeaway: Grid cropping is a good way to increase coverage, but it might hurt global consistency unless you handle transitions between patches well.

5.4. Impact of Downsampling + Random Cropping

Model 5 combined downsampling and random cropping using `RandomResizedCrop`. This added variation in both size and spatial content.

- It trained quickly and sometimes delivered visually pleasing results.
- Some fine details were lost due to reduced resolution, but overall structure and style were strong.

Takeaway: This combo offers one of the best trade-offs between training speed and quality a great choice if resources are limited.

Table 1. FID and SSIM scores for CycleGAN models. Lower FID is better; higher SSIM is better.

Model	FID (P → M)	FID (M → P)	SSIM (P → M)	SSIM (M → P)
Model 1 (256×256)	162.42	132.54	0.1747	0.1549
Model 2 (128×128)	189.68	162.57	0.1555	0.1213
Model 3 (Random Crop)	**149.43**	**132.01**	**0.1797**	0.1405
Model 4 (Grid Crop)	187.17	151.41	0.1348	0.1410
Model 5 (Downsampling + Random Crop)	181.07	139.02	0.1705	**0.1712**
Model 6 (1 Discriminator)	171.82	135.32	**0.1803**	0.1589

5.5. Effect of Removing One Discriminator

Model 6 only used one discriminator instead of two.

- We expected this to weaken the results, but it actually held up pretty well.
- The outputs were slightly smoother and less textured, but still coherent and visually reasonable.

Takeaway: A single discriminator setup might be a useful simplification for lightweight applications, especially if perfect detail isn't critical.

5.6. Evaluation Metrics and Performance Analysis

As shown in Table 1, each variation brought its own trade-offs in terms of FID and SSIM. For example:

- **Model 3** had the best FID scores and strong SSIM, showing how effective random cropping can be for improving realism and diversity.
- **Model 5** performed best on SSIM for Monet → Photo, making it a strong candidate for preserving structure.
- **Model 6**, even with its simplified architecture, outperformed several full versions, suggesting CycleGANs can be more flexible than expected.

Lessons Learned and Challenges: Small design choices, even ones that seem minor at first glance, can significantly impact model performance, training speed, and overall visual quality. Through our experiments, we realized that tweaks in image resolution, cropping strategies, or architecture can make or break the effectiveness of unpaired image translation, especially when working with artistic datasets like Monet2Photo. It's definitely worth exploring these simple modifications before diving into more complex model changes, particularly when working within limited computational environments.

One of the key challenges we encountered was the long training time when running models on CPU-only setups. Some models, especially the full-resolution 256×256 versions, took very long hours to train without access to GPU acceleration. This significantly limited our ability to iterate quickly or test multiple configurations. Although we used Google Colab for GPU access, we quickly ran into its session time limits, memory restrictions, and occasional disconnects.

To work around these constraints, we reduced image resolution in some experiments and designed lightweight variants (e.g., by removing a discriminator) to ensure our models could finish training within the time and hardware we had. These limitations ultimately influenced our choice of experiments and helped reinforce the importance of efficient design decisions in real-world research settings.

6. Conclusion

In this project, we explored how different training strategies and architectural choices affect CycleGAN's ability to translate between real-world photos and Monet-style paintings. By testing six different model variants, we gained a better understanding of how preprocessing, resolution, and network complexity impact the quality and efficiency of unpaired image translation.

Our findings show that:

- Higher resolutions (like 256×256) help preserve textures and fine details, but increase training time and instability.
- Random cropping improves generalization and stylization, but may need longer training to stabilize results.
- Grid-based cropping offers structured coverage but can create artifacts at patch boundaries.
- Combining downsampling with random cropping (Model 5) turned out to be a solid compromise, efficient to train, with consistently good results.
- Even with just one discriminator (Model 6), CycleGAN can still perform surprisingly well, hinting at possible ways to simplify the architecture.

Overall, we found that thoughtful choices in preprocessing and model design can make a big difference, sometimes more than large architectural changes. It's not always about making the model more complex; sometimes, simplifying the right way can go a long way.

6.1. Future Work

There are several directions we'd like to explore next:

- Adding perceptual loss functions, such as VGG-based feature loss, to improve high-level consistency and style.
- Using attention mechanisms to help the model focus on important regions and textures.
- Trying progressive training, starting with low resolution and gradually increasing it, to improve stability.
- Exploring transfer learning by initializing with pre-trained generators (like StyleGAN or BigGAN) to speed up convergence.
- Exploring alternative generator and discriminator architectures, such as varying the number of residual blocks or modifying the depth of the networks, to better understand how model capacity influences translation quality.

These extensions could help push the quality of unpaired image translation even further. But even without complex changes, our results show that careful experimentation with basic setup choices can already lead to meaningful improvements.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30. Curran Associates, Inc., 2017. 2

[2] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2

[3] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[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)*, 2017. 1, 2