CyclePix. Pix2Pix using CycleGAN

Apollinaria Chernikova
Innopolis University
Innopolis, Russia
a.chernikova@innopolis.university

Egor Machnev
Innopolis University
Innopolis, Russia
e.machnev@innopolis.university

I. Introduction

In recent years, the field of image-to-image translation has seen remarkable advancements, primarily driven by the development of generative adversarial networks (GANs). Among the most influential architectures are **CycleGAN** [1] and **Pix2Pix** [2], which have enabled a wide range of applications — from style transfer and image synthesis to domain adaptation and photo enhancement.

This project focusing on converting real-world photographs into stylized. To evaluate the impact of model architecture on style transfer quality, four experiments were conducted: two using Aivazovsky paintings and two using Ghibli-style images. Each domain was tested with both a ResNet-based generator and a U-Net-based generator. These experiments allow for a comparative analysis of how different architectures affect the translation fidelity, texture retention, and stylization strength.

II. Related Work

Image-to-image translation is a widely studied task in computer vision, with applications ranging from semantic segmentation to style transfer. Isola *et al.* [2] introduced a conditional GAN framework that learns a mapping from input to output images using paired datasets. Although effective, its reliance on aligned image pairs limits its applicability in domains like artistic style transfer, where such data is difficult to obtain.

In the context of artistic style transfer, many studies have focused on learning style representations from paintings. Gatys *et al.* [3] introduced a method for blending the content of one image with the style of another by optimizing a randomly initialized image to match the content features of the source and the style features of the target, as extracted from a pre-trained convolutional neural network (typically VGG). However, neural style transfer suffers from several limitations. First, it requires a separate optimization process for each new content-style pair, making it computationally expensive and unsuitable for real-time applications. Second, it often struggles to preserve the semantic structure of the original image, particularly when transferring complex or abstract styles. Third, the quality of the output heavily depends on hyperparameter tuning and the specific layers chosen for feature extraction.

To address these limitations, feed-forward models for style transfer were proposed, such as the one by Johnson *et al.* [4], which trained a single network to apply a specific style in real time. Similarly, *Huang et al*. [5] proposed Adaptive Instance Normalization (AdaIN), allowing arbitrary style transfer by aligning the mean and variance of content features to those of style features.

These models still rely on supervised training with aligned pairs or focus on global style features, which may not fully capture complex, domain-specific artistic styles. Unpaired image-to-image translation models like using CycleGAN. Zhu *et al.* [1] overcame this challenge by introducing a cycle consistency loss, enabling training without paired examples. This made it possible to translate between domains without the need for direct supervision, enabling a wide range of artistic and real-world applications.

III. Methodology

A. Model architecture

The architecture of CycleGAN is built upon a GAN framework, consisting of two generator networks and two discriminator networks, designed to handle unpaired image-to-image translation between two domains, A and B.

Each generator $G: A\rightarrow B$ and $F: B\rightarrow A$ is trained to translate images from one domain to another, while each discriminator D_A and D_B is trained to distinguish real images from fake images in its respective domain.

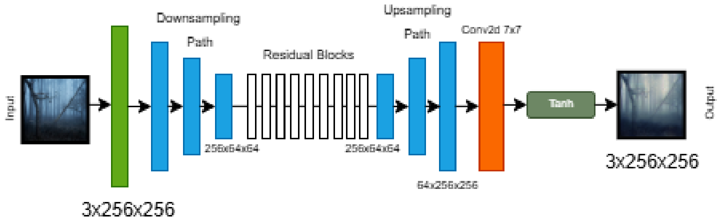

Fig 1. ResNet Generator Architecture

In the original CycleGAN paper, the authors propose using ***ResNet-based*** generators, which consist of a few convolutional layers, followed by several residual blocks, and then deconvolution layers. This architecture is effective at preserving the global structure of the input image while applying stylistic changes. In this project we consider two generator implementations:

- **ResNet generator**: closely follows the original CycleGAN design, using 9 residual blocks (for 256×256 images). This model excels at maintaining the content structure while translating textures and colors.
- **U-Net generator**: an alternative encoder-decoder architecture with skip connections between

corresponding layers in the encoder and decoder. U-Net is often used in Pix2Pix and similar tasks that require fine-grained alignment between input and output, making it a strong candidate for capturing more detailed spatial relationships.

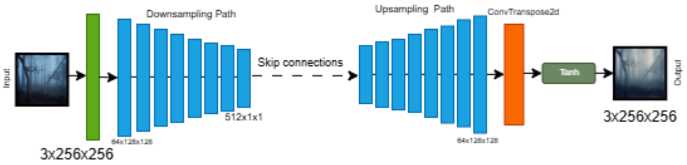

Fig 2. Unet Generator Architecture

By comparing both architectures, we aim to explore how the design of the generator affects the quality and fidelity of the translated images in different artistic domains.

B. Dataset

The dataset used in this project consists of three distinct parts: 500 paintings by Ivan Aivazovsky, 1,600 screenshots from Studio Ghibli films, and 6,287 real-world nature photographs.

The Aivazovsky paintings were selected to capture the dramatic lighting, rich color palettes, and sweeping seascapes typical of the artist's style. The goal was to train CycleGAN to transform real photographs into a style that reflects Aivazovsky's unique brushstrokes and atmospheric effects.

The Ghibli anime screenshots dataset was specifically chosen to test the versatility of CycleGAN in adapting to a completely different artistic style. While most CycleGAN implementations focus on artworks by famous painters, this dataset introduced a more dynamic animation style, characterized whimsical color schemes. The inclusion of such images allowed the experiment to assess how well the model can generalize to animated, non-photorealistic styles, expanding the boundaries of CycleGAN's applicability.

Finally, the real-world photographs were selected to serve as the base content for the image translations. These 6,287 provided the necessary ground truth for evaluating the effectiveness of the style transfer, ensuring that the model was able to produce realistic and meaningful results.

C. Image Buffer

An image buffer was implemented to store generated images and facilitate the training of the discriminator. The buffer helps maintain diversity in the images presented to the discriminator by providing a mixture of real images from the dataset and generated (fake) images from the generator. This approach is particularly important for improving the performance of the adversarial training process by avoiding issues such as mode collapse, where the generator might produce similar or identical outputs.

The buffer operates as follows:

1. During the training process, the generator produces a batch of fake images.
2. These fake images are added to the buffer, and if the buffer exceeds its maximum capacity, an old image is randomly replaced.
3. The discriminator is then trained with a mix of real images from the dataset and the fake images from the buffer.

D. Hardware Resources

All training experiments were conducted using the maximum computational resources available to us. The primary hardware used was the **NVIDIA T4 GPU**, which offers **15 GB of GPU memory** and **29 GB of system memory (RAM).**

IV. GitHub Link

Link to our GitHub repository with all experiments and evaluations: https://github.com/machnevegor/cyclepix.

V. Experiments and Evaluation

A. Training Pipeline

Most of hyperparameters were taken from original paper [1]. The models were trained using the Adam optimizer with a learning rate of 2e-4 and beta values of (0.5, 0.999). The cycle consistency loss weight was set to (10, 10) for both directions (MPM to PMP) and the identity loss weight was 0.5. Training was conducted for a maximum of 20 epochs.

Mixed-precision training was used for better efficiency on GPU, and training was conducted on a single GPU. The maximum training time was limited to 4 hours and 55 minutes, which resulted in shorter training durations for some of the experiments. Checkpoints were saved every 5 epochs, and the last checkpoint was always retained.

Due to limited hardware resources, the ResNet-based generator could only complete 13 epochs, preventing it from reaching the intended 20 epochs. However, the pipeline provided an efficient and reproducible framework for training the model under the given constraints.

B. Evaluation Pipeline

As in the original paper [1], we will employ the Discriminator loss for numerical comparison. This choice is motivated by the fact that Discriminator loss provides a direct measure of how well the model differentiates between real and generated images, which is essential for evaluating the performance of generative models like CycleGAN. However, it is important to note that finding a perfectly suitable metric to assess the quality of generation in such models is difficult. Common metrics are generally based on the pixel-wise similarity between real and generated images, such as Mean Squared Error (MSE) or Structural Similarity Index (SSIM). These metrics primarily focus on the preservation of low-level features and overall similarity.

In contrast, our task requires the model to maintain the structural integrity of the original image while transforming it into a different artistic style, making traditional metrics less effective. The challenge lies in evaluating the ability of the model to perform such transformations while preserving

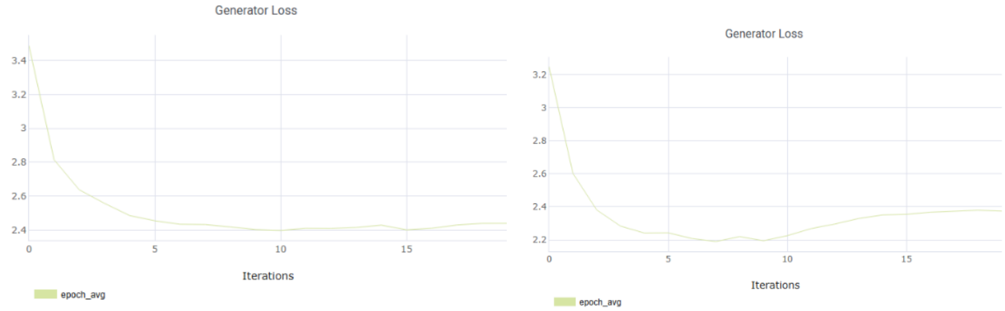

Fig 3. Unet Photo-to-Style Generator loss: Ghibli and Aivazovsky datasets

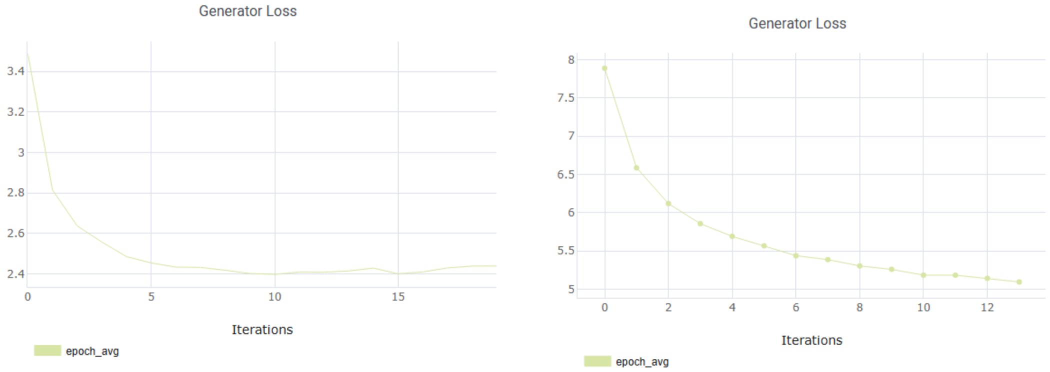

Fig 4. ResNet Photo-to-Style Generator loss: Ghibli and Aivazovsky datasets

important high-level content, which conventional metrics might not fully capture.

VI. Analysis and Observations

A. *Visual comparison*

When comparing the results on the Aivazovsky dataset, we observed notable differences in how the two model architectures perform. The ResNet-based model tends to preserve the color palette of the original images, whereas the U-Net architecture focuses more on transforming textures. However, U-Net struggles significantly with night-time scenes, often producing incoherent outputs. The ResNet model, on the other hand, applies a filter-like transformation but also fails to adequately handle images captured at night. This can be attributed to the dataset's bias — most of Aivazovsky’s paintings depict daytime seascapes, making it difficult for the models to generalize to other types of scenes. This limitation is inherent to CycleGAN itself, which is not well-suited for structural or shape transformations — a constraint explicitly mentioned in the original paper.

In contrast, the results with the Ghibli dataset were visually more appealing. Both models performed reasonably well, but the ResNet-based model yielded superior results. Unlike U-Net, which mainly overlays stylistic filters, ResNet made more deliberate changes to the structure and details of the images. This is particularly evident in the enhanced sharpness and clarity of lines. Given that the model achieved a final loss of only 5.09, the quality of the generated outputs is quite impressive. With access to more computational resources, we expect the performance could be improved even further.

B. *Value comparison*

Below is a table displaying the losses for all four models evaluated in this study.

Model name	Style Disc. Loss	Real Disc. Loss
Aivazovsky Unet	0.0764	0.1461
Aivazovsky ResNet	0.0307	0.1337
Ghibli Unet	0.0651	0.1398
Ghibli ResNet	0.1109	0.1361

Tab 1. Discriminator Losses comparison

For the Aivazovsky style, the ResNet-based model performs best overall. It achieves the lowest style loss (0.0307), meaning it transfers the painting style more accurately than the Unet model. Its real loss is also slightly better, suggesting its outputs look more realistic. The Unet model, while still effective, shows a higher style loss and a higher real loss, indicating it struggles more with both style transfer and realism—especially on night scenes, as noted earlier.

In the Ghibli experiments, both models perform more closely. The Unet shows a lower style loss, suggesting it better captures the unique visual look of Ghibli images. However, the ResNet has a similar real loss, which may indicate that it produces more visually coherent images despite a slightly higher style loss.

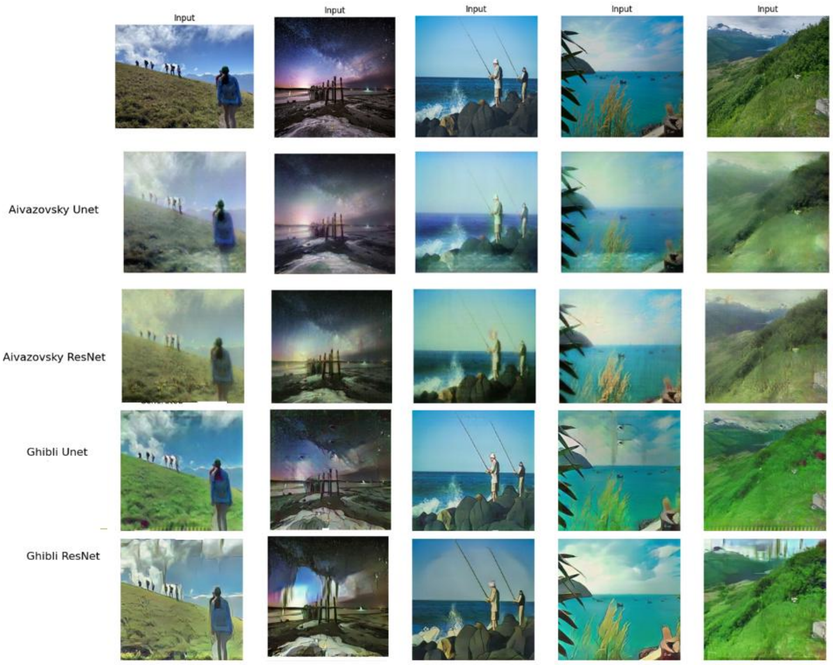

Fig 5. Images generation results

Overall, based on numbers, ResNet tends to produce more realistic results, while Unet can better adapt to specific stylistic features. The differences between the two are especially noticeable when the dataset contains scenes that are uncommon in the training data, such as night views in Aivazovsky paintings.

VII. Conclusion

Two generator architectures were evaluated—Unet and ResNet—on their ability to preserve image structure while applying stylistic transformations.

Our experiments demonstrate a clear trade-off between training efficiency and output quality. ResNet consistently produces more visually convincing results, especially when transferring the Aivazovsky style. It better captures the essence of the target domain and preserves structure, even in challenging conditions. However, this comes at the cost of significantly longer training times and greater demand on computational resources. Despite this, even with limited resources, ResNet was able to deliver high-quality results, suggesting that it generalizes well and benefits from architectural depth.

On the other hand, Unet trains much faster and requires less memory, making it a practical choice for situations where time and computational power are limited. While it may lack some of the subtlety and expressiveness of ResNet outputs, it still performs reasonably well—particularly for domains like Ghibli, where stylization is more forgiving.

For future work, it may be beneficial to explore hybrid architectures or improved loss functions that combine the efficiency of Unet with the expressiveness of ResNet. Additionally, incorporating attention mechanisms or training with perceptual loss (e.g., LPIPS) could help further enhance visual realism.

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, 2020, https://arxiv.org/abs/1703.10593

[2] Phillip Isola and Jun-Yan Zhu and Tinghui Zhou and Alexei A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks”, 2018, https://arxiv.org/abs/1611.07004

[3] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, “A Neural Algorithm of Artistic Style”, 2015, https://arxiv.org/abs/1508.06576

[4] Justin Johnson, Alexandre Alahi, Li Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”, 2016, https://arxiv.org/abs/1603.08155

[5] Xun Huang, Serge Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization”, 2017, https://arxiv.org/abs/1703.06868