Optimizing Conditional CycleGAN for Style Transfer

Sanjay Ravindran

STOR 512: Optimization for Machine Learning and Neural Networks

Spring 2024

Abstract

This project explores the development and optimization of a Generative Adversarial Network, specifically a Conditional CycleGAN, designed to transform human facial images into comic-style representations. The architecture comprises two generators and two discriminators, each responsible for transforming and validating images across domains while ensuring cycle consistency. This setup aims to not only map between domains but also to recover the original images post-transformation, thus maintaining a continuous transformation loop. The model employs adversarial and cycle consistency losses to ensure the generated images are both indistinguishable from real images and capable of reverting to their original forms, enhancing image quality and fidelity to the intended style. Optimization techniques such as gradient penalty and adaptive instance normalization were utilized to stabilize training and adapt to diverse inputs, while data augmentation broadened the model's exposure to varied facial features. Supported by high-performance GPU computing, these strategies expedited training and facilitated real-time adjustments. The efficacy of the Conditional CycleGAN was validated through qualitative assessments and quantitative metrics like the Inception Score and cycle-consistency loss, confirming the model's ability to effectively bridge distinct visual domains and showcasing the potential of CycleGANs in advanced image transformation tasks.

Contents

Contribution

I found a dataset of paired human and animated faces, and pre-processed it so it could be used to train the GAN. I used Google Colab's high-performance GPUs to speed up the training process. I then set up the model architecture and trained it. I evaluated the results by observing loss over training epochs, visual quality, and inception score.

1 Introduction

The rapid evolution of AI powered image generation has created many cutting-edge tools such as OpenAI's Sora and Generative Adversarial Networks. Generative Adversarial Networks (GANs), since their inception, have been at the forefront of this evolution, enabling the creation of highly realistic images from textual descriptions, noise patterns, or existing images. The aim of this project was to explore the capabilities of GANs and their variations further by focusing on the generation of comic book-style characters from real human face

images. This task not only challenges the model's ability to handle complex transformations between image domains but also tests creativity and adherence to a specific artistic style.
Given the unique adversarial nature of GANs and their success in various image transformation tasks, they presented an interesting framework for addressing this problem of guided style transfer. The project's goal was to develop a Conditional Cycle Generative Adversarial Network that could not only transform a given human face into a comic style but also revert any comic style face back to its original human form, thus maintaining cycle consistency. This approach ensures that the model learns a meaningful and reversible mapping between the two domains. It also learns to enhance its transformations through provided context, potentially making it more capable and adaptable.
Past literature on GANs provided foundational knowledge and insights into the potential and limitations of these networks. Studies such as Goodfellow et al.'s original paper on GANs and Lu et al.'s work on Attribute-Guided Face Generation Using Conditional CycleGAN were very influential and provided a good starting point for this problem. These papers highlighted the importance of using high-quality data, keeping track of Cycle Consistency, and providing useful conditional data to guide generation.
This project also draws heavily on related work in the field of style transfer, where neural networks adapt the style of one image to the content of another, and face recognition, where networks learn to identify and manipulate facial features accurately. By integrating concepts from these areas with GANs, the project aimed to create a model that learns to perform style transformation and retain and respect the unique characteristics of human faces.

2 Methodologies

The primary objective of this project was to explore and implement a Conditional Cycle Generative Adversarial Network (CycleGAN) to facilitate the transformation of human facial images into comic-style images and vice versa. This section outlines the model architecture, dataset preparation, training procedure, and the methodologies used to achieve these transformations.

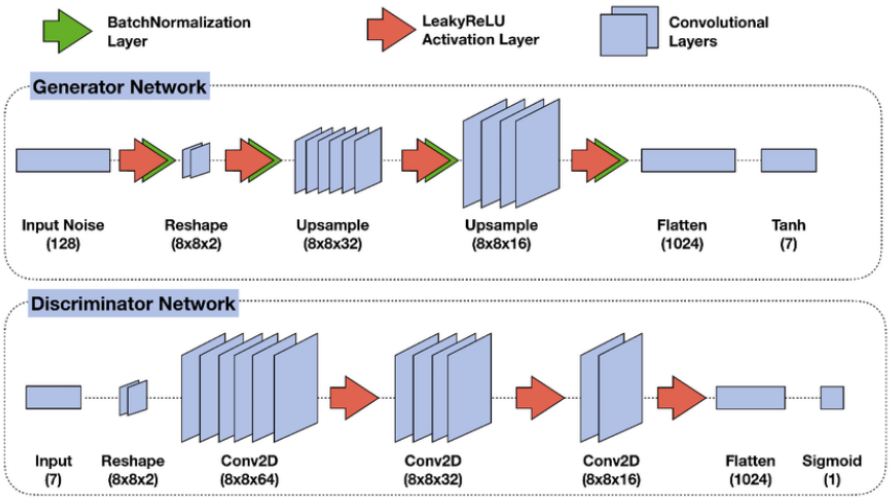

Figure 1: Generator and Discriminator Architecture

2.1 Model Architecture

The architecture of our Conditional CycleGAN consists of two main components: generators and discriminators. The model includes two generators (`gen_cond_comic` and `gen_cond_face`) and two discriminators (`disc_cond_face` and `disc_cond_comic`), structured to handle the forward and reverse transformations between the two domains.

- **Generators**: Each generator is tasked with translating images from one domain to the other. `gen_cond_comic` converts images from the human face domain to the comic style, while `gen_cond_face` performs the reverse. These networks are built using an encoder-decoder structure, where convolutional layers downsample the input image into a latent space, and subsequent deconvolutional layers upsample it to generate an image in the target domain. Residual blocks are included in the middle of the architecture to enhance feature transformation capabilities without adding excessive depth to the model.
- **Discriminators**: Each discriminator assesses the authenticity of the images generated by its corresponding generator, encouraging the generation of indistinguishable and high-quality images. The discriminators use a series of convolutional layers that gradually increase in depth and decrease in spatial dimensions, outputting a scalar prediction that signifies whether the input image is real or fake.
- Figure 2 shows how the Generators and Discriminators interact in a cycleGAN. Generator G translates between domain X and domain Y, while Generator F translates

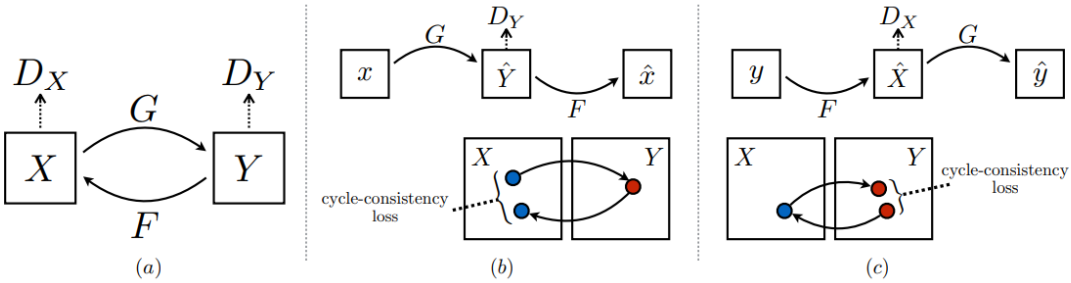

Figure 2: Training Flow and Cycle-Consistency Loss

between domain Y and domain X. Once Generator G translates X to Y, discriminator D_y judges this output translation $\hat{Y}$. After this, Generator F translates $\hat{Y}$ back to Domain X, resulting in $\hat{x}$. The difference between the original X and the reproduced $\hat{x}$ is the cycle-consistency loss. The same holds for the reverse of this process. In terms of the problem at hand, `gen_cond_comic` translates faces to comic faces, and `gen_cond_face` translates those comic faces back to normal faces. The difference between the original face and the reproduced face is the cycle-consistency loss.

2.2 Data Preparation

Figure 3: Human Face

Figure 4: Comic Face

The dataset consists of paired images from two distinct domains: human faces (Figure 3) and comic-style representations (Figure 4). Data collection involved curating a diverse set

of images that cover various facial expressions, angles, and lighting conditions to ensure robustness and generalizability of the model.

- **Preprocessing**: All images were resized to 256x256 pixels, a common dimension that balances detail with computational efficiency. Images were then normalized to have pixel values between -1 and 1, aligning with the tanh activation function used in the generator's output layer.
- **Augmentation**: To enhance the model's ability to generalize from limited data, we applied random horizontal flips and rotations as part of the data augmentation strategy. This approach helps in simulating different perspectives and variances, making the model less sensitive to exact image orientations and positions.
- **Normalization**: In the development of the model, Instance Normalization was employed within the convolutional layers of both the generators and discriminators. This choice was guided by the need for style normalization at the individual image level. Unlike Batch Normalization, which normalizes the input across the batch, Instance Normalization applies normalization for each sample independently. This method enhances the model's ability to adapt to various artistic styles without being influenced by the batch dynamics, crucial for achieving consistent style transfer.

2.3 Training Procedure

Training a CycleGAN involves alternating updates between the generators and discriminators. The loss functions used include:

- **Adversarial Loss**: Ensures generated images are indistinguishable from real images within each domain. The loss function for the generator G with respect to the discriminator D_Y is given by:

$$\mathcal{L}(G_{(X,Z)\to Y}, D_Y) = \min_{\theta_g} \max_{\theta_d} \left\{ \mathbb{E}_{y,z} \left[\log D_Y(y,z) \right] + \mathbb{E}_{x,z} \left[\log \left(1 - D_Y \left(G_{(X,Z)\to Y}(x,z), z \right) \right) \right] \right\}$$

- **Cycle Consistency Loss**: Minimizes the difference between original images and their cyclic reconstructions, crucial for retaining content while transforming style. The cycle consistency loss for the generators $G_{X\to Y}$ and $G_{Y\to X}$ is defined as:

$$\mathcal{L}_c(G_{X\to Y}, G_{Y\to X}) = \|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1 + \|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1$$

- **Identity Loss**: Sometimes used to preserve color composition and other details of the input image when translated back to its original domain.
- **Optimizer**: The Adam optimizer with a learning rate of 0.0002 and beta parameters of (0.5, 0.999). Learning rate schedulers were employed to decay the rate as training progressed, stabilizing learning as the model converged.

- **Batch Size and Epochs**: The model was trained with a batch size of 1 (due to the memory constraints and nature of image-to-image translation tasks) for 10 epochs. Training for more than 10 epochs did not significantly improve results.

2.4 Evaluation Metrics

Training loss, Inception Score, and visual assessments. These metrics helped gauge both the diversity, accuracy, and realism of the generated images, providing a comprehensive measure of model performance.

3 Experiments or results

3.1 Analysis of Training Losses

The primary objective of training was to minimize the adversarial losses associated with both the generator and discriminator networks in our conditional CycleGAN architecture. The progression of these losses throughout training is depicted in Figure 5.

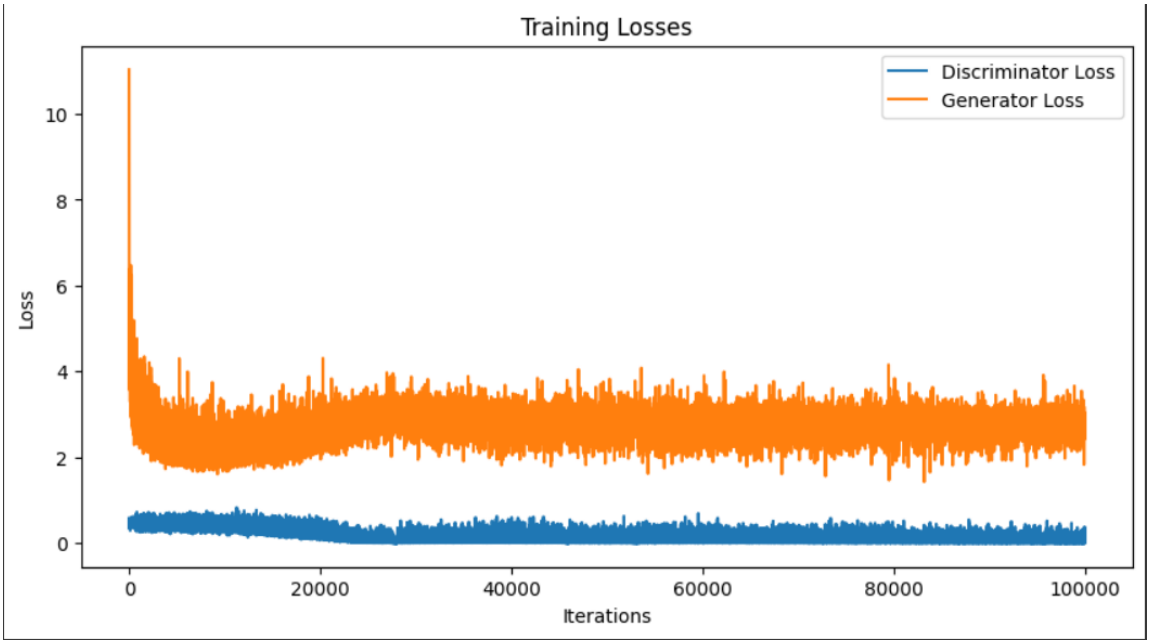

Figure 5: Training Losses Over 10 Epochs(100000 iterations)

As shown in Figure 5, the discriminator's loss starts high, indicating initial difficulty in distinguishing between real and generated images. However, it rapidly stabilizes, reflecting

the discriminator's improving capability to classify images correctly. This rapid improvement in the early stages is followed by a gradual decrease, settling into a stable loss as training progresses.
Conversely, the generator's loss begins slightly elevated and experiences significant variability initially, indicative of its struggle to fool the discriminator. Over time, as the generator improves through backpropagation and learns to produce more convincing images, this loss also stabilizes. Notably, the generator loss consistently remains higher than the discriminator loss throughout the training process, a common occurrence in adversarial training, which suggests that the generator always has room for improvement given the discriminator's effectiveness.
The dynamic between the generator and discriminator losses are crucial for understanding the adversarial training process. The adversarial push and pull dynamically balance each other, maintaining a learning environment conducive to generating high-quality outputs. This balance is pivotal in conditional adversarial networks where the generator not only learns to create plausible images but also ensures these creations adhere to conditional and style constraints.

3.2 Qualitative Analysis

- **Inception Score(IS)**: Inception score is used with GAN's to measure quality and diversity of generated images. It does this by computing the mean and standard deviation IS for all images. The higher the mean the more realistic according to the Inception V3 model, and the lower the standard deviation the more consistent. The mean and standard deviation for the Conditional cGAN was 1.00347 and 0.00013 respectively. The mean value is much lower than expected, which means the images were not very realistic. However, this should be taken with a grain of salt. The Inception V3 model decides whether an image is realistic or not based on the images it was trained on, which would have included human faces, but not animated faces. Thus, it would not regard the generated comic faces as being realistic, as it doesn't match what it was trained on. However, the standard deviation reveals some useful information. 0.00013 is very low, which means that the images had a very similar inception score. Furthermore, all the images were very similar. This means that they all had similar elements of style and quality, so the model translated similar elements of style and qualities into all the images.

3.3 Visual Analysis

One of the most insightful ways to gauge the performance of a GAN dedicated to image-to-image translation—such is through visual inspection of the output images. This is especially useful in this project, because the outputs are not standard or realistic, and thus are hard to evaluate. Qualitative analysis was conducted by comparing comic-styled

outputs generated by the model to input human face images and other comic faces.

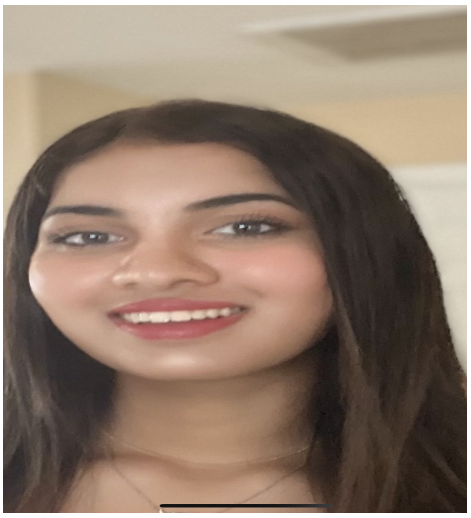

Figure 6: Human Face

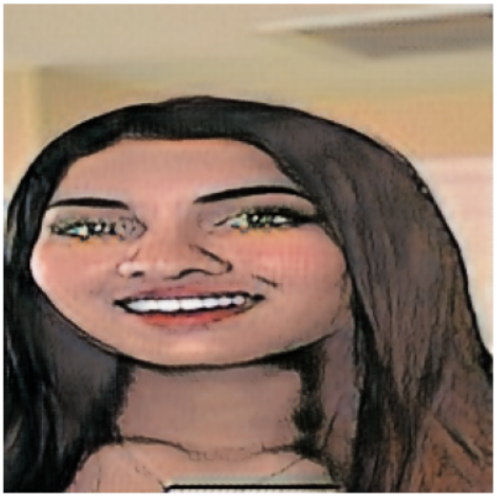

Figure 7: Comic Face- 5 Epochs

Figure 8: Comic Face- 10 Epochs

- **Face Quality**: As can be seen in figures 6 through 8, the model was fairly successful at translating human faces into animated faces. One important thing it seemed to learn over epochs was the resolution of the face. After 5 epochs, the comic elements were present, but the resolution of the comic face was not good. However, as seen in figure 8, after the full training cycle was completed, the clarity of the comic face became significantly better.

- **Adherence to Human Facial Features**: The model demonstrated a strong a capability in retaining the core features of the input human faces. The output images maintained identifiable characteristics such as eye placement, nose shape, and mouth expression. This adherence suggests that the network has effectively learned to map crucial facial features from the domain of human faces to the stylized comic domain, incorporating stylistic elements of comic art while preserving the original's identity.
- **Introduction of Comic Elements**: In terms of style transfer, the model successfully introduced several comic-like elements into the generated images. These elements include exaggerated facial expressions and enhanced contours that are typical in comic art, demonstrating the model's ability to blend two distinct artistic representations. However, the degree to which these comic features conform to traditional comic art varied based on the input, indicating room for refinement in achieving a consistent comic style.

Figure 9: Human Face

Figure 10: Comic Face: 5 Epochs

- **Sensitivity to Lighting and Image Quality**: A notable observation was the model's performance across different lighting conditions present in the input dataset. The outputs did not consistently handle variations in lighting, often retaining the lighting conditions of the input images rather than adjusting them to fit a standard comic illustration style, which typically features bold and uniform lighting. Also, in images where the lighting or quality was significantly different than the training images, the outputs were of lower quality. This limitation likely stems from the homogeneous nature of the training data, where images predominantly shared similar lighting and quality. This can be seen in Figures 9 and 10, as the translated comic

face is more similar to the input human face than a comic face. Further, it only changed the quality of the image, but did not introduce many comic characteristics.

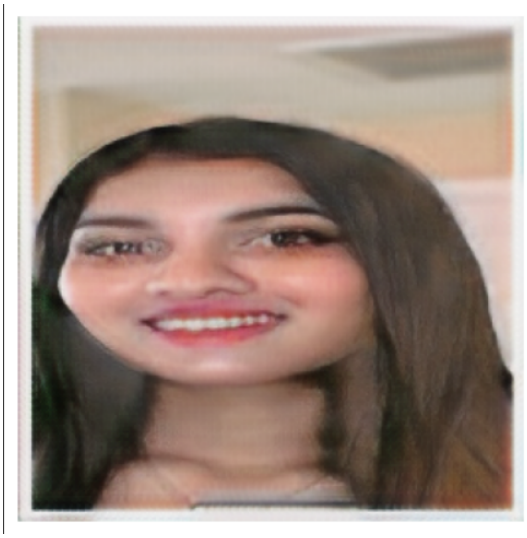

Figure 11: Recovered Human Face from Figure 8

- **Cycle Image Quality**: As can be seen in figure 11, the model was not as good at translating comic faces back to human faces. It was able to recover the most important characteristics from the original photo, but the quality is not as good as the original. However, the cycle did help produce better results for the human to comic translation, regardless of the quality of the recovered image.

Overall, the visual analysis of the generated images provided valuable insights into the model's current capabilities and limitations. While the model excels at maintaining facial integrity and introducing stylized elements, its performance across diverse lighting conditions highlights the importance of a varied training dataset. Future work could explore the incorporation of a more diverse set of images in training or the implementation of specialized layers or loss functions in the network architecture to better handle variations in lighting and image quality.

4 Discussions and findings, conclusions

4.1 Discussion

The Conditional CycleGAN developed in this project has shown promising capabilities in translating human facial images into comic styles under well-controlled settings. This

model harnesses the unique properties of cycle consistency and conditional inputs to ensure that transformations not only adhere to the artistic domain of comic imagery but also retain the essential features of the original images.
Despite some challenges with image diversity and lighting variations, largely due to the homogeneous nature of the training dataset, the model adeptly handles images that resemble the training data in terms of quality and lighting. This limitation, while notable, does not detract from the model's ability to produce high-quality transformations where conditions align with those of the training data. The performance in these cases highlights the effectiveness of the cycle consistency and adversarial training methods employed, which have been optimized to balance fidelity and stylization accurately.

4.2 Findings

he Conditional CycleGAN not only demonstrated its effectiveness in style transformation but also provided insights into the adaptability of neural networks to highly specific artistic tasks. Key findings from the project include:

- **Effectiveness of Cycle Consistency:** The implementation of cycle consistency played a crucial role in reinforcing the model's ability to preserve content while facilitating style transformations. This finding underscores the importance of cycle consistency not just as a tool to measure loss, but in enhancing the quality of generative models.
- **Impact of Conditional Inputs:** The use of conditional inputs allowed for targeted style transformations, demonstrating that the model could be effectively guided through additional contextual information. This adaptability suggests potential applications beyond comic style generation, such as in personalized content creation and other forms of artistic style transfer.
- **Model Sensitivity to Training Data Characteristics:** The model's performance highlighted its sensitivity to the characteristics of the training data, particularly in terms of lighting and image quality. This sensitivity, while limiting in some respects, provides valuable lessons on the critical role of dataset composition and preprocessing in the training of generative models.
- **Optimization Techniques:** The application of advanced optimization techniques, including adaptive learning rates and gradient penalty, was found to significantly enhance training stability. These techniques not only improved the convergence rate but also reduced the occurrence of common issues such as mode collapse, pointing to their vital role in the successful training of complex models like GANs.

These findings highlight the nuanced interplay between model architecture, training dynamics, and data characteristics. They suggest that while Conditional CycleGANs are

powerful tools for artistic style transformation, their performance is intricately linked to the conditions and constraints set by their training environment.

4.3 Conclusion

In conclusion, this project has successfully demonstrated the potential of Conditional CycleGANs in bridging the gap between human facial imagery and comic-style representations. The model's performance in a controlled setting not only underscores its potential in digital art but also highlights the robustness of the optimization techniques used during its training.

Future work will focus on expanding the diversity of the training data and exploring more sophisticated data augmentation techniques to improve the model's performance under a broader range of conditions. Such efforts aim to enhance the generalizability and utility of the model, making it a more effective tool for artists and creators in various fields. Additionally, integrating feedback mechanisms to adapt the transformation process to user preferences could further personalize the output, aligning it more closely with specific artistic goals.

Overall, the findings from this project contribute valuable insights to the field of machine learning and neural networks, particularly in the optimization of generative models for complex image transformation tasks. The successful application of cycle consistency and conditional inputs within a GAN framework sets a foundation for future research and development in this area.

Bibliography

[1] Lu, Y., & Tang, C.-K. (n.d.). *Attribute-Guided Face Generation Using Conditional CycleGAN*. Retrieved May 8, 2024, from `https://openaccess.thecvf.com/content_ECCV_2018/papers/Yongyi_Lu_Attribute-Guided_Face_Generation_ECCV_2018_paper.pdf`

[2] Zhu, J.-Y., Park, T., Isola, P., Efros, A., & Research, B. (n.d.). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. `https://arxiv.org/pdf/1703.10593`

[3] ResearchGate — Find and share research. (n.d.). ResearchGate. https://www.researchgate.net/