ISSN 2007-9737

A CycleGAN Framework for Anime Style Image Synthesis based on U-Net and Self-Attention

Harshad Sharma, Smita Das*

National Institute of Technology Agartala,
Department of Computer Science & Engineering,
India

{hs1989ts, smitadas.nita}@gmail.com

Abstract. In the contemporary realm of digital imagery, Image Synthesis (IS) is a technique employed to craft artificial images that encapsulate specific content tailored to user preferences. Owing to the intricate and time-intensive nature of this process, researchers have turned to Generative Adversarial Networks (GANs). These networks based on back-propagation signals, alleviate the need for extensive annotated training data. This paper introduces an Animation Style Image Synthesis from natural images by utilizing the CycleGAN framework. Initially, the exploration of CycleGAN's capabilities focused on style transfer to integrate it with U-Net architecture for the creation of anime-style images. Gradually, enhancements in feature extraction and the improvement in overall image quality has been carried out by incorporating self-attention layers and ResNets. The experimental outcomes for proposed architecture have been verified against established evaluation criteria that indicates a promising direction for research in the field of IS.

Keywords. Image synthesis, computer animation, generative adversarial network, cycleGAN, neural style transfer, self-attention layer.

1 Introduction

Animated images or Anime are now-a-days a part of modernistic art style depicting creativity, story-lines, and vivid expression of emotions. Anime is not only limited to art or science, but its commercial usage includes movie production, advertising media, educational sectors and many more. It has the potential to pique interest in the most abstract notion of ideas, subject and what not. No matter how intriguing that might appeal, creating anime is not an easy task, due to its laborious human involvements.

In recent times, advancement in deep learning techniques like Neural Style Transfer or GAN[6] has shown prominent results in order to generate high dynamic range Image Synthesis [21]. For the anime production houses, it is quite important to learn applying style of one domain to another and to create artistic effects on the images. Embellishing image in a most aesthetic way and to create a unique signature style is one of the reasons why Deep Generative Models are hot research topics in today's world.

Image synthesis is referred to as the technique for producing such realistic anime style images from various natural image dataset. Researchers are increasingly applying IS technique to various field of research [22] such as: motion deblurring, image decomposition, face sketch synthesis, Focus Fusion Attention Mechanism etc. In the field of Computer vision research such as: image recognition, object detection, segmentation and tracking, IS enables computers to derive meaningful information from digital images and videos. For the specific task of Image Synthesis, StackGAN can create realistic text-to-image synthesis and image-to-image Translation is possible using pix2pix GAN that is based on a conditional adversarial network to transform a source image into a target image. Moreover, SRGAN can create high-to-low resolution images and even Cross Domain Image Generation, image blending[16] and generation of anime characters [27] are also possible in various GAN approaches .

1.1 Motivation

Generative Adversarial Networks (GANs) are used in many exciting ways, and one of the most popular is turning regular pictures into anime-style images. This trend has become very popular on social media, where people love using anime versions of their photos as profile pictures. But most editing tools aren't good at making these transformations look smooth and natural, and doing it by hand is hard for most people. That's why researchers are still working on better ways to create high-quality anime-style images quickly and easily.

1.2 Problem Statement

In this study, a CycleGAN-based framework for the creation of anime images from natural images has been introduced. This particular variant of GAN is ideally suitable for this work due to its straightforward structure and impressive outcomes. At first, source images have been collected from diverse datasets and style references have been taken from various animation films, compensating for the lack of paired source and style images. Thereafter, the CycleGAN model is refined by incorporating self-attention layers to minimize the computational load on GPUs. In the final stage of model development, we have employed U-Net as the generator within the CycleGAN framework for enhancing it further by embedding self-attention layers in both the up-sampling and down-sampling processes.

1.3 Research Gaps and Possible Contributions

Image Synthesis is revolutionizing the field of Computer Vision with its progressive and significant advancements. The success of this technique is largely due to its adaptable architecture, tailored optimization goals, and specialized loss functions for particular tasks. Despite of all these advantages, few research gaps have been pinpointed that warrant further investigation. In the context of image synthesis, researchers often utilize datasets that might exhibit class imbalance which can lead to biasness towards a specific class of images. Again, there is a lack of reliable evaluation metrics in the research of IS that might yield inconsistent results with specific dataset when evaluated from a human perspective. Apart from that, increased computational complexity in case of high resolution images often forces researchers to depend on GPU rental services or to adopt the Transfer Learning approach. For that reason, there arise a significant opportunity for research into optimizing generative models like GANs for better performance in environments with restricted hardware capabilities. Following are some of the key contributions that our research offers:

— Proposed
a CycleGAN framework for transforming a real-world image into an anime-style depiction. The CycleGAN framework retains the essence of the original natural image and provides a seamless blend between reality and the distinctive characteristics of anime artistry.

— The CycleGAN architecture is further modified to include Self-Attention layers, which focus on capturing global dependencies within the data. The strategic placement of self-attention layers in the up-sampling and down-sampling stages has been designed to refine the network's ability to focus on pertinent features across different scales of the image.

— Rigorous analysis has been carried out to determine how the self-attention layers influence the generator's capability to accurately reconstruct images.

Therefore, in continuation of this paper, Section 2 describes the prevailing research in the field of GAN and different GAN variants for image synthesis along with details study of CycleGAN. In Section 3, details discussion has been carried out for proposed methodology along with dataset and Cycle GAN architecture. Section 4 conducts experimental analysis based on various training parameters and details discussion. Finally the paper is concluded with future direction of work in section 5.

2 Related Work

In this section, a brief study of background of GAN and an overview of the recent literature related to both GAN and CycleGAN has been conducted.

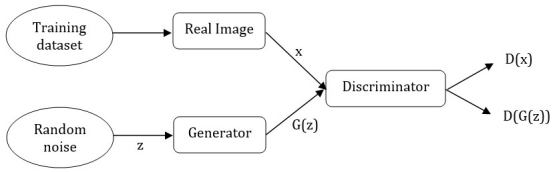


Fig. 1. Typical GAN Architecture

2.1 Background of Generative Adversarial Network

GANs fall under the category of unsupervised learning in machine learning framework but are trained in a self-supervised fashion. GAN involves the learning of the intricate underlying data patterns in the input data distributions and uses it to generate near similar data distributions that are possibly identical or represent the same semantic notion as that of the input data. A typical GAN architecture is shown in Figure 1. GAN uses two neural networks namely Discriminator(D) and Generator(G) to generate quality data. Discriminator differentiates between the real and the fake data distribution, while generator creates data distribution similar as training data. Several advancements took place in GAN such as: critical change in architecture, loss function or inclusion of other frameworks. Conditional generation in GAN have been proposed to leverage the process of image generation.

2.2 Neural Style Transfer

Neural Style Transfer(NST)[2] is a popular image blending technique that takes two images namely, Source Image and Style Image, and tries to modify the source image to look like in a way of Style image. It is done with the help of CNNs that transfer style to the desired image or content. Behind the scene it reaps benefit of VGG[13] pre-trained model to encode the style and embellish the source image. This transferring is done in such a way that it retains the feature of the source image and at the same time matches the texture of the style image. Although NST can generate desired artistic style in no time, but that doesn't makes it perfect technique for the job, since it may still transfer style to semantically wrong region. Also it may not generate images devoid of artifacts like blunt edges, inconsistent shading etc. Apart from NST, a Neural Architecture Search (NAS) is also useful to find the optimal descent directions [4] for generating feature maps for a specific optimal knowledge distillation process.

2.3 Different GAN Variants for Image Synthesis

For Image synthesis, various GAN variants have been used extensively. CycleGAN variant has been initially designed for Image-to-Image Translation without paired images, however, [11] has proposed a similar method with paired training examples in a supervised fashion. Another GAN variant, CartoonGAN [3] has proposed a method that takes a real-world image and converts it into a cartoon or anime-style image. The authors emphasized the improvement of the image quality and also the usage of novel losses for the similarity and semantic retention of the image. AnimeGAN [28] has proposed a lightweight implementation of the GAN that turns a natural realistic image into an Anime rich style image. This approach fuses the neural style transfer technique with GAN in order to create an Anime like texture. MontageGAN is yet another variant that trains multiple parts or layers of the image and then places it together like a puzzle in order to form a larger and more meaningful image. The authors [23] have proposed a method that comprises two steps. The first step trains GAN to generate different parts of the image, followed by a global GAN that learns to put various parts generated in the previous step, so as to form a complete image. SC-FEGGAN [12] has proposed a novel method for image editing by making use of an end-to-end trainable CNN, that synthesizes image. Moreover, the authors formulate an additional style loss, the strategy can also provide outcomes that are realistic. EditGAN [17] is a first of its kind GAN driven image editing framework, that has allowed high-level and high-precision image editing and that too in a semantic level, by making use of segmentation mask. In order to get the intended output image with only the required features modified, manipulating particular features typically requires enormous datasets and specialists to know which characteristics to change inside the model. Authors in [14]

have discussed an Exemplar-Based Conditional Broad-GAN to incorporate sub-net architecture for feature matching and reconstruction by using the matching information between the target and reference image. A feedback spatial attention de-hazing network based on the recurrent structure has been proposed in [33]. The attention-based estimation on the residual block is used to adapt the value of pixels with different distributions to restore haze-free image generation. In summary, their method combines semantic understanding, exemplar-based learning, and conditional GANs to produce realistic color images.

2.4 CycleGAN

CycleGAN utilises an aided cycle consistencies loss for unpaired image translation. This idea has its origin back to language translation, where if a sentence is in language X, and had been translated to Language Y, then doing the reverse translation, the original or near original sentence must be obtained. As discussed in [34], CycleGAN has managed to surpass CoGAN[18], BiGAN[20], SimGAN[24] etc. Out of the two generators, one can be used to input photographs from domain X to create false images that resemble domain Y, while the second generator can be used to input images from domain Y to create fake images that resemble domain X. The quality of the image generations is then somewhat improved by using discriminators to assess the realism of creating false pictures. Generators then utilise the discriminator to identify what has to be changed to deceive the discriminator.

In [8], CycleGAN has been used to convert MRI to CT Image of head and pelvic region.

While the MRI to CT synthesis was successful in head images but the results of pelvic images do not have much variations due to the presence of joints and muscles. In another research work [15] of cross-modality medical image processing, MR to CT image synthesis has been done based on U-Net and CycleGAN. Authors claimed better performance based on mean absolute error (MAE), higher structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) in both directions of CT/MR synthesis. Authors in [25] has used CycleGAN with double U-Net to get spatial consistencies of 3D medical images by avoiding memory-heavy 3D convolutions. Another research work [32] has developed an advancement in CycleGAN concept known as switchable CycleGAN model that has been used for multi-contrast brain MRI image synthesis of pediatric structural brain MRI images. Authors of [29] has proposed a cycleGAN variation DicycleGAN which is based on the deformation-invariant CycleGAN architecture.

Spatial transformation network based on thin plate spline has been used to train the unpaired and unaligned data and to generate synthesised images aligned with the source data. The authors in [30] employ a multi-adversarial mechanism based on CycleGAN for blind motion deblurring and generates high-resolution images iteratively. The hidden layers of the generator are gradually supervised, leading to implicit refinement and continuous production of high-resolution images. The authors introduce a structure-aware mechanism that enhances the algorithm using the edge map as guidance information. Additionally, they incorporate multi-scale edge constraint functions.

3 Proposed Method

In this section the proposed ANIMIS architecture based on CycleGAN has been depicted in details along with dataset.

3.1 Dataset

In this paper, two distinct categories of images have been utilized: content images and style images. Content images primarily represent the main subjects or objects within the image, while style images capture the ambient aesthetic or background elements. For the content images, widely accessible datasets such as Monet2Photo[19], which includes 1193 paintings by Monet and 7038 photographs depicting natural scenes, as well as the Flickr8K[9] dataset, which comprises 8092 images, have been used. These datasets provide a rich variety of foreground elements for the experiment. On the other hand, the style images have been sourced from various anime films, including titles like “Your Name,” “Weathering with You” and “Suzume.” Additionally, a diverse

collection of digital artworks has been curated and compiled from social media platforms [31, 26, 1] to serve as a comprehensive reference for the desired style. This amalgamation of style sources aims to create a unified benchmark for the overall visual style that the experiment seeks to achieve. The derived dataset used for the experiment purpose is available online from the corresponding author on request.

3.2 Data Pre-processing

During the initial processing phase, the images have been uniformly resized to a 256x256 pixel resolution.

This resizing has been performed using the LANCZOS filter to adjust the image dimensions and to execute random cropping. Furthermore, the PIL Library has been utilised for image enhancement techniques such as sharpening, smoothing, and blurring to refine the style images.

To capture dynamic sequences, the frame capture rate varies from 2 to 10 frames, depending on the specific requirements. Additionally, the RandomHorizontalFlip function[10] is applied to randomly flip the images horizontally, a technique that typically has minimal impact on the perception of both natural and anime-style images. The ColorJitter operation[35] is also employed to introduce variability in the images' coloration. After completing all the requisite pre-processing steps, the resulting training images are prepared as exemplified by the sample content images provided in Figure 2(a) and style images as shown in Figure 2(b).

3.3 Purpose of Self-Attention Block

Self-attention blocks are special parts of neural networks that help the model focus on different sections of its input. In language tasks like translation or understanding emotions in text, they've been very useful. In image generation using GANs (Generative Adversarial Networks), self-attention helps the model understand how different areas of an image relate to each other.

This leads to images that look more realistic and varied. These blocks allow GANs to notice connections between far-apart parts of an image, which regular layers might miss. They also improve image quality by letting the model pay attention to important regions. Self-attention works well alongside convolutional layers, filling in gaps where those layers fall short in capturing the full picture. Compared to older methods like RNNs and LSTMs, self-attention is faster and more flexible, making it better for both training and generating images efficiently.

3.4 Why U-Net?

U-Net is a neural network originally made for medical image segmentation, now widely used in computer vision and language tasks. It has a U-shaped structure with two parts: an encoder that extracts key features from an image, and a decoder that rebuilds it into a detailed map. A special feature of U-Net is its skip connections, which link matching layers in the encoder and decoder. These help preserve important spatial details. The decoder also uses transposed convolutions to improve image resolution. By combining fine and broad features, U-Net produces accurate results, making it ideal for tasks needing precise image analysis.

In this paper, the U-Net architecture is employed as a generator within the CycleGAN framework.

The U-Net's adeptness at grasping both the minute and overarching elements of the input image is pivotal for producing translations of superior quality, which justifies its integration into CycleGAN as a generator. Other than down-sampling and up-sampling, U-Net's capability to perform style transfer across different domains is harnessed, facilitated by the CycleGAN structure. As a generator in GANs, U-Net's proficiency in rendering high-resolution images replete with intricate details, crucial for achieving realism. Furthermore, the skip connections empower the decoder network to tap into multi-scale information from the encoder, thereby preserving delicate details through the up-sampling process. This feature is especially beneficial in image-to-image translation tasks where fine nuances are prone to being overlooked.

Fig. 2. (a) Content Image and (b) Style Image after pre-processing

3.5 Proposed CycleGAN Architecture

In this paper, the image synthesis process is founded upon a combination of CycleGAN, U-Net, and self-attention mechanisms. The implementation harnesses the PyTorch deep learning framework, renowned for its efficiency and rapid execution.

The CycleGAN architecture has been refined by incorporating a self-attention layer that is carefully calibrated to avoid imposing additional computational burdens on GPU resources. The inclusion of the attention layer in both the Generator and Discriminator is aimed at enhancing performance and capturing more accurate spatial relationships within the images. During the initial phase of implementation, we have conducted a series of experiments with various modifications to the standard CycleGAN architecture. While the original CycleGAN employs a Resnet[7] backbone, we experimented with varying the number of residual blocks to observe their impact. These modifications and their effects were evaluated based on the average loss recorded during training sessions. For the final version of our implementation, we opted for a U-Net-based generator within the CycleGAN framework, augmented by the integration of self-attention layers along both the Down-Sampling and Up-Sampling pathways of the U-Net. This approach is detailed in the subsequent steps of the proposed CycleGAN architecture.

3.5.1 Generator

As outlined earlier, the generator has been crafted utilizing the PyTorch framework, featuring a U-Net supported structure. It is composed of 8 Down-Sampling/Encoding layers succeeded by 8 Up-Sampling/Decoding layers. A significant feature of this design is the incorporation of a long skip connection which effectively mitigates the issue of gradient vanishing. Within the network, each Encoding layer block is constructed with a convolution layer, followed by Instance Normalization, LeakyReLU activation[5] and strategically placed self-attention layers. This configuration is consistently replicated across the Up-Sampling/Decoding layers as well.

3.5.2 Discriminator

In the discriminator component of the architecture, we have implemented PatchGAN[11]. It is structured with a series of Down-Sampling layers, maintaining the same composition as the generator's Encoding layers. The generator in this architecture is designed to faithfully recreate the original anime-style image. When an anime image

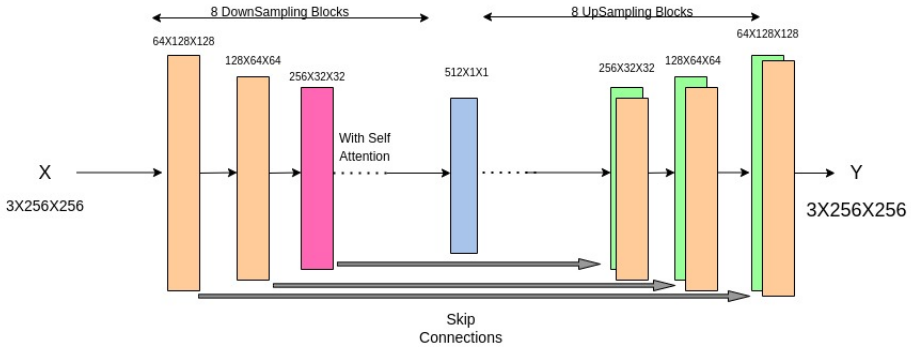


Fig. 3. Proposed CycleGAN Architecture

is input into the anime-to-photo generator and the resultant image is subsequently reintroduced into the system, the generator is expected to reproduce the initial anime image. This principle also applies to images processed by both generators, aiming to retrieve the original images. To quantify the fidelity of the image recreation, the L1 loss is employed, serving as a metric to preserve information throughout this cycle. The schematic diagram describing the overall network architecture is shown in Figure 3.

3.5.3 Loss Objective

The training of CycleGAN involves a suite of loss functions that facilitate the learning process for mapping between the generator and discriminator networks. The loss objectives encompass three main types: Cycle Consistency Loss, Adversarial Loss and Identity Loss. These loss objectives collectively guide the networks towards producing accurate and reliable image translations. The loss function is described as:

$$\begin{aligned}\mathcal{L}\left(G,F,D_x,D_y\right) = \mathcal{L}_{GAN}\left(G,D_y,X,Y\right) + \\ \mathcal{L}\left(F,D_x,Y,X\right) + \\ \lambda\mathcal{L}_{cyc}\left(G,F\right) + \\ \mathcal{L}_{identity}\left(G,F\right).\end{aligned} \tag{1}$$

1. **Cycle Consistency Loss** The application of cycle consistency loss is crucial to ensure that images generated by the model can be accurately reverted to their original domain. Essentially, this means that if an image is transformed from domain A to domain B, and subsequently from domain B back to domain A, the final image should closely resemble the initial one. This loss function acts as a safeguard to maintain the integrity of the image throughout the translation process:

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}\left[\|F(G(x))-x\|_1\right] + \\ \mathbb{E}_{y\sim p_{data}(y)}\left[\|G(F(y))-y\|_1\right].\end{aligned} \tag{2}$$

2. **Adversarial Loss** Within the CycleGAN framework, adversarial loss serves as the cornerstone of the training process. It motivates the generator to create images that are indistinguishable from real ones, effectively deceiving the discriminator. Conversely, the discriminator is trained to discern between genuine and generated images. For each domain in CycleGAN, there are two distinct adversarial losses that guide the network towards generating convincing and authentic-looking images:

$$\begin{aligned}\mathcal{L}_{GAN}\left(G,D_Y,X,Y\right) = \mathbb{E}_{y\sim p_{data}(y)}\left[\log D_y(y)\right] + \\ \mathbb{E}_{x\sim p_{data}(x)}\left[\log(1-D_Y(G(X)))\right].\end{aligned} \tag{3}$$

3. **Identity Loss** The purpose of identity loss in CycleGAN is to ensure that the generator preserves the style of an image that already pertains to the target domain. In simpler terms, when an image from domain A is translated to domain B and then translated back to domain A, the outcome should be the original, unchanged image. This loss function helps maintain the

authenticity of the image's style throughout the translation cycle.

$$\mathcal{L}_{identity}(G,F) = \mathbb{E}_{y\sim p_{data}(y))}\left[\|G(Y)-Y\|_1\right] + \mathbb{E}_{x\sim p_{data}(x))}\left[\|F(X)-X\|_1\right]. \quad (4)$$

4 Results and Analysis

Prior to delving into the outcomes of the experiment and extracting meaningful conclusions, it is imperative to consider the intricate details and compromises inherent to the experimental setup.

4.1 Experiment Details

The conducted experiments were carried out using minimal hardware resources and accessible online platforms such as Google's Colab and Kaggle.

Challenges arose due to the extensive volume of training data coupled with limited VRAM and storage capacities, leading to memory shortages. To circumvent these issues and efficiently compute the attention scores, we strategically reduced the batch size during training and implemented attention layers in areas where computational demands needed to be minimized. This approach allowed us to manage resource constraints while maintaining the integrity of the experiment.

4.2 Training Parameters

For training of model we have utilizes the support of GPUs from kaggle. We have used Nvidia's GPU P100 and T4, which is based on pascal and Tesla architecture and known to have high throughput and performance. To address the issue of high VRAM issue leading to out of memory error, we have set a batch size of **8** and **16** for the experiment.

We also have utilised PyTorch Lightning module. We have set *learning rate* to be **0.00020**, have used Adam optimizer with the beta values of **0.5** and **0.999**, invoked a learning rate decay after **140** epochs, out of a total of **250** epochs for whole training session. The initialization of weight of the neural network is done from the standard normal distribution from the range of 0 and 0.2 respectively.

The reason behind using the learning rate decay is that a falling learning rate enhances the learning of complicated patterns in the training dataset and an initially high learning rate prevents the network from memorising noisy material. The training parameters are mentioned in Table 1.

Table 1. Training parameters

Parameter	Value
GPU	Nvidia's GPU P100 and T4
Batch size	8, 16
Learning rate	0.00020
Total epoc	250
Adam optimizer values	0.5, 0.999
Standard normal distribution	0 - 0.2

Despite the experiments being conducted under stringent limitations, the outcome of our proposed model is deemed to be significant. It is anticipated that alleviating the hardware restrictions would lead to enhanced performance and more robust results.

The experimental procedure is divided into two stages dictated by the training parameters. Initially, a ResNet-based generator is employed within the CycleGAN architecture, which was subsequently succeeded by a U-Net-based generator. To thoroughly evaluate the experimental findings, both qualitative and quantitative analytical methods have been employed.

4.3 Qualitative Analysis

Figure 4 presents a comparative analysis of four distinct U-Net configurations. The source images have undergone synthesis through a U-Net-based generator. The resulting image columns represent variations of U-Net, each modified with self-attention layers. In (a) actual images are shown followed by attention in down-sampling in (b) and attention in up-sampling in (c) respectively. The fourth column (d) showcases a standard U-Net without any self-attention mechanisms, and finally (e) includes self-attention in both the up-sampling and down-sampling blocks.

Visually, the impact of self-attention layers on color and contrast is evident, as they actively engage with different segments of the sample images. Notably, the dual incorporation of self-attention in both the up-sampling and down-sampling stages appears to yield a more authentic anime style transformation of the source images.

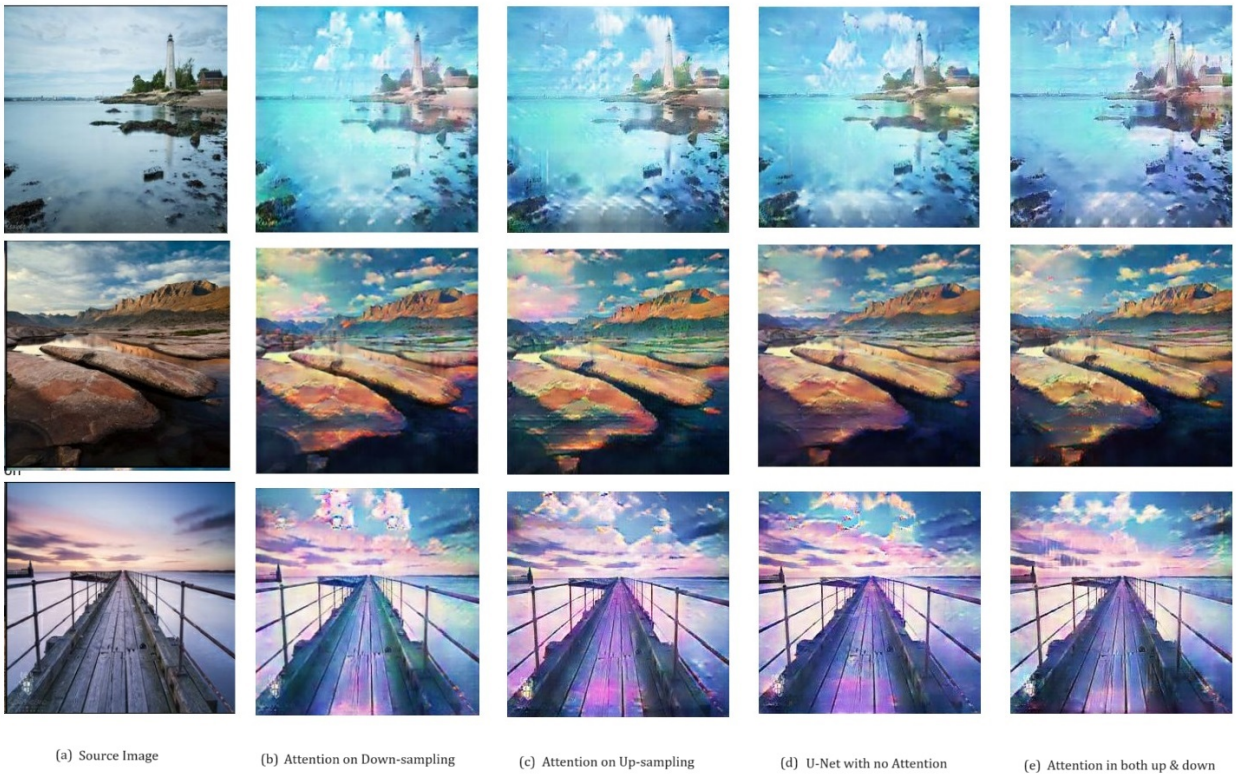


Fig. 4. Qualitative assessment of the proposed model and its variations

4.4 Quantitative Analysis

Quantitative analysis in image synthesis task involves assessing the performance based on various loss curve plots and evaluation metrics.

4.4.1 Analysis based on Loss Curve Plots

The loss curve plots serve as a testament to the stability and consistency of loss saturation. The loss curve analysis of the generator and discriminator for the different ResNet variations are shown in figure 5 and figure 6 for up-sampling and down-sampling respectively.

The loss curve analysis of the generator for simple U-Net without attention, self-attention on up-sampling block, self-attention on down-sampling block and self-attention on both up-sampling & down-sampling block are shown in figure 7. As depicted in Figure 7, the integration of self-attention within both the up-sampling and down-sampling blocks of the U-Net based generator leads to a more stable loss saturation rate compared to other variants. The U-Net architecture excels at extracting lower-level features, which are then relayed to the corresponding up-sampling block. The addition of self-attention mechanisms enhances the generator's ability to discern dependencies among the features, thereby elevating the quality of image generation. Conversely, the discriminator's reliance on basic down-sampling blocks hinders the generator's development, manifesting as fluctuations or irregularities in the loss curve.

The reason behind the fluctuation in the performance traces back to the GANs stability issues, the number of epochs it runs with, and the diversity of the dataset used in the training which is also supported by the loss curves depicted in

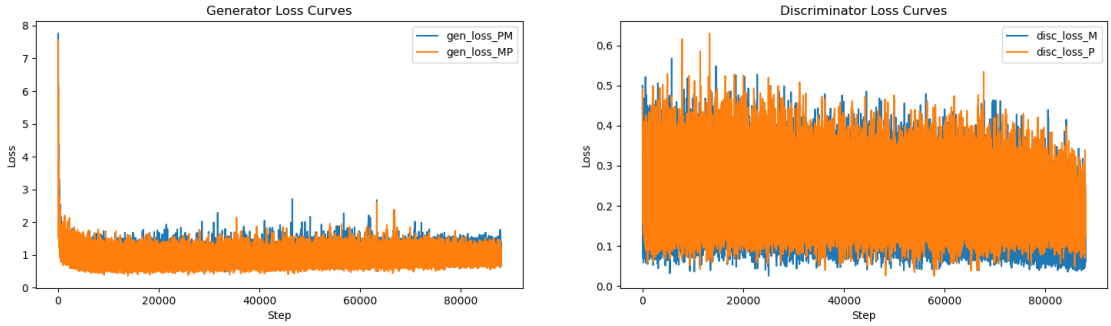


Fig. 5. Loss Curves for (a) Generator and (b) Discriminator in ResNet for Up-sampling

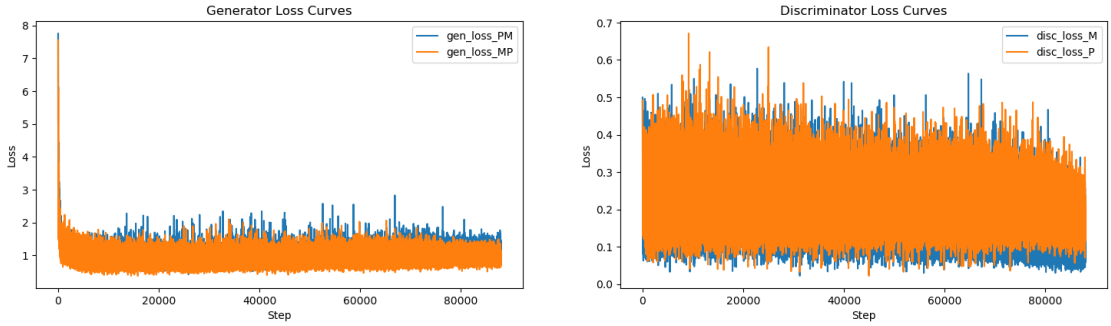


Fig. 6. Loss Curves for Generator and Discriminator in ResNet for Down-sampling

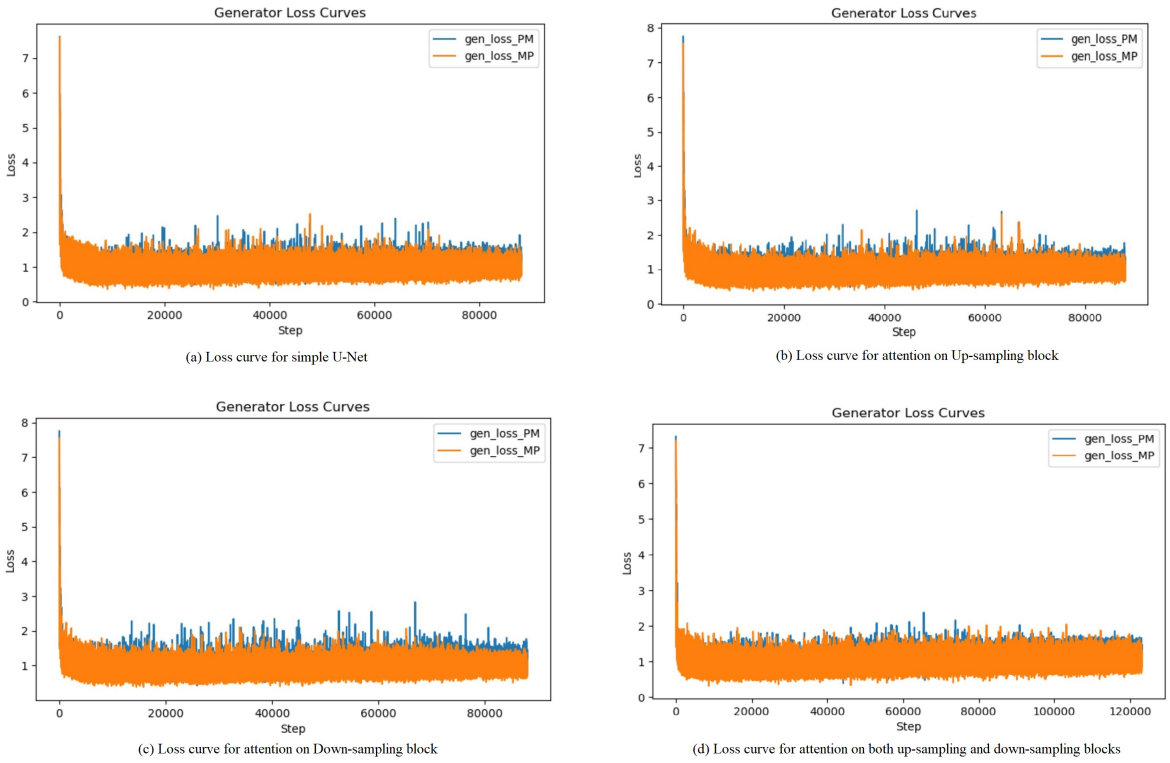
figure 7. Although this model has been trained with the premises of hardware constraint, limited window time and inadequate choices for hyper-parameter settings, the model has the potential scope for improvement. The improvement may include running the model on several settings, with a larger dataset and better processing power with an extended training routine.

4.4.2 Analysis based on Evaluation Metrics

Assessing the quantity of synthesized images presents a challenge due to the subtle nuances involved and the scarcity of reliable metrics.

Nevertheless, several metrics such as Frechet Inception Distance (FID), Inception Score (IS), and Kernel Inception Distance (KID) are widely recognized for their straightforwardness and expedited assessment capabilities. FID measures the distance between feature representations from real and generated images, considering both their distribution and covariance, making it a dependable gauge for image quality in various applications like image production, style transfer, and domain adaptation. A lower FID score indicates superior image quality. The IS metric evaluates the entropy of predicted class probabilities and the distinction between the class probabilities of real and synthesized images.An increased IS value signifies an enhancement in the quality of the synthesized images, with higher values being preferable. KID serves as an alternative to FID by relaxing the Gaussian assumption and computes the squared maximum mean discrepancy between the inception features of real and generated images using a polynomial kernel. Here, a lower KID value denotes better performance. Additionally, precision and recall values are also utilized for a more comprehensive quantitative analysis. Precision assesses the correctness of positive predictions, while recall gauges the comprehensiveness of positive predictions. Consequently, achieving both

ISSN 2007-9737



(a) Loss curve for simple U-Net

(b) Loss curve for attention on Up-sampling block

(c) Loss curve for attention on Down-sampling block

(d) Loss curve for attention on both up-sampling and down-sampling blocks

Fig. 7. Loss Curves of Generator for U-Net variations

high precision and high recall values are desirable. Table 2 shows the evaluation metrics for ResNet based variations.

From Table 2, it has been found that ResNet-8 with attention shows lowest FID value, while ResNet-9 with attention has highest IS value and lowest KID value. Again, with respect to Precision and Recall, ResNet-8 with attention outperforms other variants. Therefore, the evaluation metrics of ResNet variants in this experiment does not yield any definitive conclusions. There may be several reason behind this indefinite culmination, such as: ResNet architecture increases the computational cost during training and inference that may suffer from vanishing gradients or over-fitting issue. While residual connections help mitigate the vanishing gradient problem but they can also propagate noise or irrelevant features that can impact the quality of synthesized images. Again, ResNet blocks operate on local patches of the images whereas in image synthesis task global context is crucial. ResNet focuses on low-level image features such as: edges, textures etc. but may not capture high-level semantic information effectively which is essential for image synthesis.

The evaluation metrics for the U-Net-based variations are consolidated in Table 3. According to the data presented in Table 3, it is evident that the simple U-Net architecture without self-attention block outperforms its counterparts. This variant of U-Net boasts the lowest FID score, highest IS value, lowest KID value relative to the other variants. Also precision and recall values are higher as compared with other U-Net variants.

The introduction of self-attention mechanism can add complexity to the network requiring higher computational power however, this issue may be exacerbated with fine tuning of the self-attention mechanism. Additionally, the interaction between the self-attention mechanism and the discriminator

Table 2. Summarization of evaluation matrices with ResNet based variations

Matrices	ResNet based variations			
	ResNet-9	ResNet-9 with attention	ResNet-8	ResNet-8 with attention
FID	64.49056	65.8078	67.86121	61.2728
IS	5.62384 +/- 0.11134	5.88081 +/- 0.16586	5.80632 +/- 0.1641	5.33647 +/- 0.1363
KID	0.05738	0.04726	0.05815	0.05245
Precision	0.71228	0.72634	0.67988	0.72923
Recall	0.36971	0.4058	0.25931	0.4291

Table 3. Summarization of evaluation matrices with U-Net based variations

Matrices	U-Net based variations			
	Simple U-Net	Self-attention in up-sampling block	Self-attention in down-sampling block	Self-attention in both blocks
FID Score	34.334	36.635	37.218	37.279
Inception score	4.989+/-0.241	4.908+/-0.213	4.955+/-0.175	4.977+/-0.181
Kernel inception distance	0.02347	0.0255	0.0267	0.0263
Precision	0.66596	0.6486	0.6206	0.6227
Recall	0.23075	0.2074	0.2173	0.2002

network in a GAN could influence performance. While the discriminator network is adept at identifying genuine versus generated images by detecting specific image features, the integration of self-attention may interfere with this ability, leading to sub-optimal performance. The comparison of evaluation metrics across tables 2 and 3 indicates variability in performance among U-Net and ResNet variants with U-Net variants demonstrating superior results as compared to those based on ResNet.

5 Conclusion and Future Work

The paper presents a CycleGAN framework tailored for the synthesis of anime-style images from natural images. To optimize computational efficiency, the architecture incorporates a U-Net-based generator enhanced with self-attention layers in both the up-sampling and down-sampling stages.

Empirical findings suggest that the straightforward U-Net variant surpasses the ResNet-based alternatives across various evaluative metrics. Although U-Net is renowned for its efficacy in image segmentation, its application to style transfer may encounter limitations. Nevertheless, by formulating precise loss objectives and leveraging U-Net's proficiency in assimilating both style and content from images, it is feasible to synthesize new images that embody these elements.

While the focus of this paper is anime-style image generation, the principles could potentially be expanded to encompass the creation of 3D shapes, figures, and avatars, particularly relevant to the burgeoning field of procedural generation and digital avatars within the context of the Metaverse.

References

1. **Anonymous, Danbooru community, Branwen, G. (2021).** Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset. `https://gwern.net/Danbooru2020`.
2. **Cai, Q., Ma, M., Wang, C., Li, H. (2023).** Image neural style transfer: A review. Computers and Electrical Engineering, Vol. 108, pp. 108723.

3. **Chen, Y., Lai, Y.-K., Liu, Y.-J. (2018).** Cartoongan: Generative adversarial networks for photo cartoonization. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9465–9474.
4. **Chen, Z., Qiu, G., Li, P., Zhu, L., Yang, X., Sheng, B. (2023).** Mngnas: Distilling adaptive combination of multiple searched networks for one-shot neural architecture search. IEEE Transactions on Pattern Analysis and Machine Intelligence.
5. **Dubey, A. K., Jain, V. (2019).** Comparative study of convolution neural network's relu and leaky-relu activation functions. Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018, Springer, pp. 873–880.
6. **Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014).** Generative adversarial networks. arXiv preprint arXiv:1406.2661.
7. **He, K., Zhang, X., Ren, S., Sun, J. (2016).** Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
8. **Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J. L., Sugano, N., Sato, Y. (2018).** Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size. Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, pp. 31–41.
9. **Hodosh, M., Young, P., Hockenmaier, J. (2013).** Flickr8k dataset. Journal of Artificial Intelligence Research, Vol. 47, pp. 853–899.
10. **Hussain, Z., Gimenez, F., Yi, D., Rubin, D. (2017).** Differential data augmentation techniques for medical imaging classification tasks. AMIA annual symposium proceedings, American Medical Informatics Association, Vol. 2017, pp. 979.
11. **Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A. (2017).** Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
12. **Jo, Y., Park, J. (2019).** Sc-fegan: Face editing generative adversarial network with user's sketch and color. Proceedings of the IEEE/CVF international conference on computer vision, pp. 1745–1753.
13. **Koonce, B., Koonce, B. (2021).** Vgg network. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, pp. 35–50.
14. **Li, H., Sheng, B., Li, P., Ali, R., Chen, C. P. (2021).** Globally and locally semantic colorization via exemplar-based broad-gan. IEEE Transactions on Image Processing, Vol. 30, pp. 8526–8539.
15. **Li, Y., Li, W., Xiong, J., Xia, J., Xie, Y., et al. (2020).** Comparison of supervised and unsupervised deep learning methods for medical image synthesis between computed tomography and magnetic resonance images. BioMed Research International, Vol. 2020.
16. **Liang, W., Dong, L., Wang, R., Yan, D., Li, Y. (2022).** Robust document image forgery localization against image blending. 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE, pp. 810–817.
17. **Ling, H., Kreis, K., Li, D., Kim, S. W., Torralba, A., Fidler, S. (2021).** Editgan: High-precision semantic image editing. Advances in Neural Information Processing Systems, Vol. 34, pp. 16331–16345.
18. **Liu, L., Xie, Z., Chen, Y., Deng, Q. (2023).** Co-gan: A text-to-image synthesis model with local and integral features. International Conference on Neural Information Processing, Springer, pp. 243–255.
19. **Park, T. (2017).** Monet2Photo dataset.

20. **Sarv Ahrabi, S., Momenzadeh, A., Baccarelli, E., Scarpiniti, M., Piazzo, L. (2023).** How much bigan and cyclegan-learned hidden features are effective for covid-19 detection from ct images? a comparative study. The Journal of Supercomputing, Vol. 79, No. 3, pp. 2850–2881.

21. **Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R. (2024).** Fast high-resolution image synthesis with latent adversarial diffusion distillation. SIGGRAPH Asia 2024 Conference Papers, pp. 1–11.

22. **Sharma, H., Das, S. (2024).** A brief study of generative adversarial networks and their applications in image synthesis. Multimedia Tools and Applications, Vol. 83, No. 7, pp. 21551–21581.

23. **Shee, C. F., Uchida, S. (2022).** MontageGAN: Generation and assembly of multiple components by GANs. arXiv preprint arXiv:2205.15577.

24. **Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R. (2017).** Learning from simulated and unsupervised images through adversarial training. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2107–2116.

25. **Sun, B., Jia, S., Jiang, X., Jia, F. (2023).** Double U-Net CycleGAN for 3D MR to CT image synthesis. International Journal of Computer Assisted Radiology and Surgery, Vol. 18, No. 1, pp. 149–156.

26. **Tan, W. R., Chan, C. S., Aguirre, H., Tanaka, K. (2019).** Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing, Vol. 28, No. 1, pp. 394–409. DOI: 10.1109/TIP.2018.2866698.

27. **Thedwichienchai, N., Siriborvornratanakul, T. (2024).** 2d virtual youtuber character generation using generative adversarial networks. Fifth International Conference on Computer Vision and Computational Intelligence (CVCI 2024), SPIE, Vol. 13169, pp. 82–90.

28. **Tu, S. (2023).** Improving the effect of low-resolution face images output in AnimeGAN. Eighth International Conference on Electronic Technology and Information Science (ICETIS 2023), Vol. 12715, pp. 449–455.

29. **Wang, C., Papanastasiou, G., Tsaftaris, S., Yang, G., Gray, C., Newby, D., Macnaught, G., MacGillivray, T. (2019).** Tpsdicyc: Improved deformation invariant cross-domain medical image synthesis. Machine Learning for Medical Image Reconstruction: Second International Workshop, MLMIR 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 2, Springer, pp. 245–254.

30. **Wen, Y., Chen, J., Sheng, B., Chen, Z., Li, P., Tan, P., Lee, T.-Y. (2021).** Structure-aware motion deblurring using multi-adversarial optimized cyclegan. IEEE Transactions on Image Processing, Vol. 30, pp. 6142–6155.

31. **Westlake, N., Cai, H., Hall, P. (2016).** Detecting people in artwork with cnns. European Conference on Computer Vision, Springer, pp. 825–841.

32. **Zhang, H., Li, H., Dillman, J. R., Parikh, N. A., He, L. (2022).** Multi-contrast MRI image synthesis using switchable cycle-consistent generative adversarial networks. Diagnostics, Vol. 12, No. 4, pp. 816.

33. **Zhou, Y., Chen, Z., Li, P., Song, H., Chen, C. P., Sheng, B. (2022).** Fsad-net: Feedback spatial attention dehazing network. IEEE transactions on neural networks and learning systems.

34. **Zhu, J.-Y., Park, T., Isola, P., Efros, A. A. (2017).** Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

35. **Zini, S., Gomez-Villa, A., Buzzelli, M., Twardowski, B., Bagdanov, A. D., van de**

Weijer, J. (2022). Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training. arXiv preprint arXiv:2202.07993.

Article received on 05/09/2024; accepted on 21/08/2025.
**Corresponding author is Smita Das.*