

# Ink-To-Tint: Manga Artisan

Anannya Popat, Tarit Kandpal, and Govind Sudarshan

**Abstract**—The manga industry often grapples with the intensive labor involved in the creation process, commonly resulting in overworked artists. The meticulous nature of manga drawing typically results in publications featuring black-and-white illustrations. This choice, while traditional, can diminish the reader’s engagement, particularly when an anime adaptation is not yet available. Moreover, while certain art styles are favored, it can be excessively demanding and redundant for artists to reinterpret a series into another style. The implementation of automation through conditional Generative Adversarial Networks (cGANs) for colorization and style transfer via Stable Diffusion model offers a solution. This innovation promises to alleviate the artists’ burden while simultaneously broadening the appeal to diverse audiences, potentially increasing readership and expanding the market. Our initiative aims to foster a greater appreciation of manga as an art form across a wider audience.

**Index Terms**—Manga Colorization, Style Transfer, Pix2Pix conditional Generative Adversarial Networks (cGANs), Stable Diffusion, Image Processing

## 1 INTRODUCTION

IN the realm of computer vision and image processing, generative models, notably Generative Adversarial Networks (GANs) and Diffusion models, have risen to prominence. Their application has extended across a multitude of domains, ranging from the colorization of monochrome imagery to the synthesis of visual content from textual descriptions. These models have substantially automated and streamlined various tasks, thereby augmenting the efficiency of numerous workflows.

Despite their widespread use, an exploration into the application of GANs and diffusion models within the manga industry remains relatively nascent. The dwindling engagement with manga, exacerbated by the surge in anime’s popularity and the conventional monochromatic presentation of manga, has precipitated a discernible decline in this cultural medium. The labor-intensive nature of manga creation, coupled with stringent time and budget constraints, further precludes artists from producing colored editions on a large scale.

Motivated by our passion for manga, we propose the adoption of generative methodologies to address these challenges. In particular, we posit that conditional GANs could be instrumental in the automated colorization of manga pages, provided they are trained on a suitably diverse colored dataset (as shown in Fig. 1)

Additionally, we intend to experiment with the stable diffusion v1.5 model for the transformation of manga art styles. Given the distinct artistic styles characteristic of individual manga series and varying reader preferences, an automated style conversion could potentially revitalize reader engagement.

This paper investigates the utility of Pix2Pix conditional GANs for the colorization of manga pages and assesses the

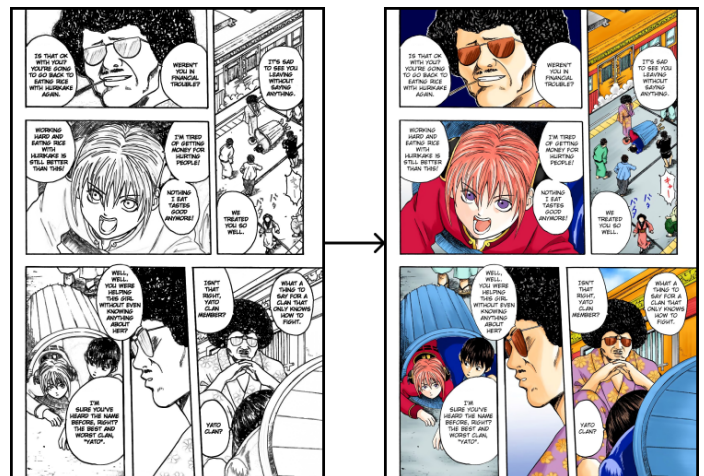


Fig. 1: Illustration of manga page colorization

efficacy of fine-tuning a pre-trained stable diffusion v1.5 model for the transfer of manga styles. By venturing into this uncharted application of generative models, we aim to contribute to the evolution and accessibility of manga as an art form.

## 2 RELATED WORK

A plethora of research has been conducted on the colorization of black-and-white images utilizing variants of these models. Notably, [1] presents the restoration of black and white photographs through conditional Generative Adversarial Networks (cGANs), offering a generalization of the colorization process beyond the confines of basic Convolutional Neural Network (CNN) architectures.

Moreover, the application of diffusion models for style transfer has witnessed a surge in popularity. References [2] and [3] highlight advancements where base diffusion models have been fine-tuned to specific requirements. This surge in innovative algorithms presents an opportunity to reinvigorate the waning appreciation of manga as an art

- Anannya Popat is with the Department of Computer Science at the University of Toronto.
- Tarit Kandpal is with the Department of Computer Science at the University of Toronto.
- Govind Sudarshan is with the Department of Computer Science at the University of Toronto.

form. Our project aligns with this trajectory, aiming to revitalize manga through the colorization of manga pages using Pix2Pix conditional Generative Adversarial Network (cGAN) — an image-to-image translation model [4]. Additionally, we endeavor to transform one manga art style to another using stable diffusion using prompts to guide the training process of the diffusion model as done in InstructPix2Pix [5]

While automatic colorization using cGANs has been successfully implemented in previous studies [6], these implementations often fall short in authentically replicating the actual manga dataset. Typically, textured and grayscale inputs were employed, resulting in idealistic outputs and an ease of achievement that does not accurately reflect authentic manga styles. To address this gap, our report demonstrates the conversion of a colored manga dataset to its sketched version using image processing techniques such as dodging and dilation.

### 3 DATASETS

For the dataset required for our colorization task, we procured a collection of 58,642 colored manga images from Kaggle, which served as the labeled training data for our cGAN model. Regarding the style transfer component of our research, we obtained a dataset of 1500 manga pages pertaining to four distinct manga art styles—Naruto, Jojo, Mob Psycho, and One Piece—legally, in the form of electronic mangas (Fig. 3).

#### 3.1 Preprocessing for Manga Colorization

The input for the Generative Adversarial Networks (GANs) consists of decolorized renditions of the images obtained from Kaggle. In the process of decolorization, we explored two distinct approaches.

In the initial approach, we straightforwardly converted the colored manga images to grayscale. However, this simplistic grayscale transformation preserved all the textures and shading in the decolorized outcome. Although our model rapidly adapted to the training data comprising these shaded mangas, it struggled when presented with unseen true manga images. This challenge arose due to its over-dependence on textures and shading. A noteworthy observation was made that authentic mangas from the unseen dataset scarcely exhibited any textures or shading, explaining the limitations of our initial model on such images. [4] also uses a similar shaded input for their models.

For the second approach, our objective was to eliminate these textures and shading from the grayscale images. To achieve this, we employed an image processing technique known as dodging. In the dodging process, we initially obtained an inversely filtered version of the grayscale image through bilateral filtering. Subsequently, we performed a division operation, dividing the grayscale image by the inversion of the inversely filtered image. This procedure resulted in a sketched effect on the original image, effectively erasing textures and shaded regions while preserving outlines.

However, post-application of this technique, we observed that the edges became excessively thin and required

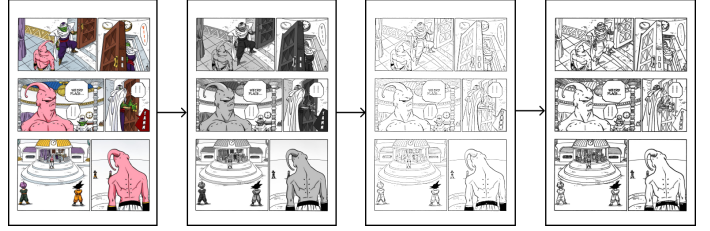


Fig. 2: Conversion of colored manga image to authentic manga style using grayscale, dodging and dilation.

reinforcement. To address this, we inverted the sketched image and applied morphological dilation, increasing the thickness of the white portions (edges) of the inverted sketched image. Consequently, after inverting this dilated image, we successfully obtained images that closely resemble authentic black-and-white mangas. The process flow is visually demonstrated in Fig. 2

In the concluding step of our process, we employ a filtering mechanism to exclude undesired manga pages. This is achieved through the calculation of a metric termed “white-ratio.” This metric serves as an effective indicator, representing the ratio of pure white pixels to the total number of pixels within an image. Subsequently, we set a threshold for filtering, removing images that have a white ratio falling below 0.1 or exceeding 0.9. This strategic filtration allows us to eliminate images that are predominantly white or overly colored. Following this rigorous filtering process, we are left with a refined collection of 49,217 images, which are deemed suitable for utilization by our models.

#### 3.2 Preprocessing for Style Transfer

In this dataset, the images exhibited variations in form and were standardized to a uniform aspect ratio of 2:3. Initially, each image was resized to dimensions of 400x600 pixels. For the purpose of consistency across different manga art styles, we calculated distributions of white and black pixel ratios for each style. Subsequently, we employed quantile analysis, selecting the 0.25 and 0.75 quantile values to establish thresholds for image filtering. Images with white and black ratios falling outside these thresholds were excluded from the dataset. This process resulted in a refined dataset comprising approximately 1,472 datapoints, encompassing all labeled styles.



Fig. 3: Dataset for diffusion (a) Naruto, (b) Jojo, (c) Mob Psycho (d) One Piece

Further, to facilitate the style transfer task, we developed a standardized prompt template instructing the model to convert manga pages from one art style to another while

preserving the narrative and text integrity. The template prompt is phrased as follows: ‘Convert the given manga page to enter manga name art style without altering the story and the text.’ Using this dataset, we paired images of different art styles to create each final datapoint, ensuring that the input and output styles were distinct.

The training and testing data distribution for the pix2pix cGAN model and stable diffusion model is shown in Table. 1:

TABLE 1: Training and testing data distribution for both the models

Type	cGAN	Stable Diffusion
Training Data	49,200	1440
Testing Data	17	32

## 4 PROPOSED METHODS

### 4.1 Conditional Generative Adversarial Networks for Colorization

In 2014, Goodfellow et al. [7] introduced Generative Adversarial Networks (GANs), a novel generative model capable of producing images from noise, highly regarded in the machine learning community. A GAN consists of two parts: a generator (G) and a discriminator (D). The generator aims to replicate the data distribution, minimizing the loss function  $\log(1 - D(G(z)))$ , with  $z$  as noise input. The discriminator’s role is to distinguish between real and generated samples.

For our manga colorization project, we utilize a Pix2Pix conditional GAN (cGAN), which differs from traditional GANs by generating images from existing input rather than noise [4]. The visual depiction of its system architecture can be seen in Fig. 4. Our cGANs generator translates black-and-white manga pages into colorized versions, while the discriminator assesses their authenticity.

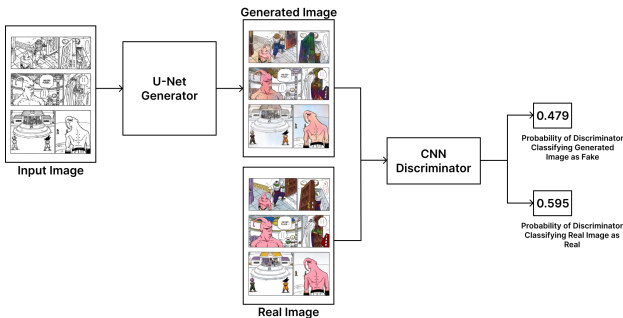


Fig. 4: cGAN System Architecture

#### 4.1.1 Generator Architecture

The Generator in our setup uses a U-Net structure [8], as illustrated in Fig. 5 to keep detailed features during image translation. Unlike usual Pix2Pix methods, we didn’t add noise to the input. This choice, made to suit the specific task of image colorization, helps the model learn more effectively, focusing on replicating the input image with color without unnecessary variations in outlines. Additionally,

we’ve enhanced our U-Net with Residual blocks between layers, improving its ability to maintain detailed features, like text in manga images. These Residual blocks are crucial for keeping key elements like text and edges intact, significantly improving our colorization results.

#### 4.1.2 Discriminator Architecture

The Discriminator, integral to the conditional GAN (cGAN) architecture, operates as a convolution-based classifier. In the framework shown in Fig. 6, its input is a 4-channel image, comprising a black and white (monochromatic) image (1 channel) and a colored image (3 channels). The introduction of the black and white input with the colored image is what characterizes our model as ‘conditional.’ This concatenated input allows the Discriminator to determine whether the colored output is an authentic (real) image or a synthetic (generated) one.

#### 4.1.3 Loss Functions

The loss function employed for the discriminator is formulated as the average binary cross-entropy loss. This calculation encompasses two components: firstly, the loss for real image being predicted as real; secondly, the loss for the generated image being predicted as fake. This is mathematically represented in Equation. 1.

$$L_D = \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \frac{1}{2} \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

The loss function used for the generator, as seen in Equation. 4 consists of the weighted average of binary cross-entropy loss (Equation. 2) with L1 loss (Equation. 3). The generator tries to maximize the probability that the discriminator is fooled into thinking the generated images are real.

$$L_{BCE} = \mathbb{E}_{z \sim p_z(z)} [\log D(G(z))] \quad (2)$$

$$L_{L1}(G) = \mathbb{E}_{x,y,z} [||y - G(x; z)||_1] \quad (3)$$

$$L_G = L_{BCE}(G; D) + L_{L1}(G) \quad (4)$$

## 4.2 Stable Diffusion for Style Transfer

In this study, we employed the InstructPix2Pix training methodology [5] for our diffusion model. This technique enables the model to process an input image and modify it according to a given instruction prompt. Specifically, our objective was to configure the model to accept a manga page and adapt its art style in response to the prompt.

### 4.2.1 Architecture

For the underlying architecture, we selected the Stable Diffusion v1.5 model [9] (as shown in Fig. 7, utilizing pre-trained weights from MeinaMix v10. This version of the stable diffusion model is adept at generating animated images and operates effectively without the need for complex prompts.

Due to constraints in time and computational resources, the model training was limited to one epoch, involving 1,440 datapoints, and validation was conducted on a set of 32 datapoints. In the post-training image generation phase,

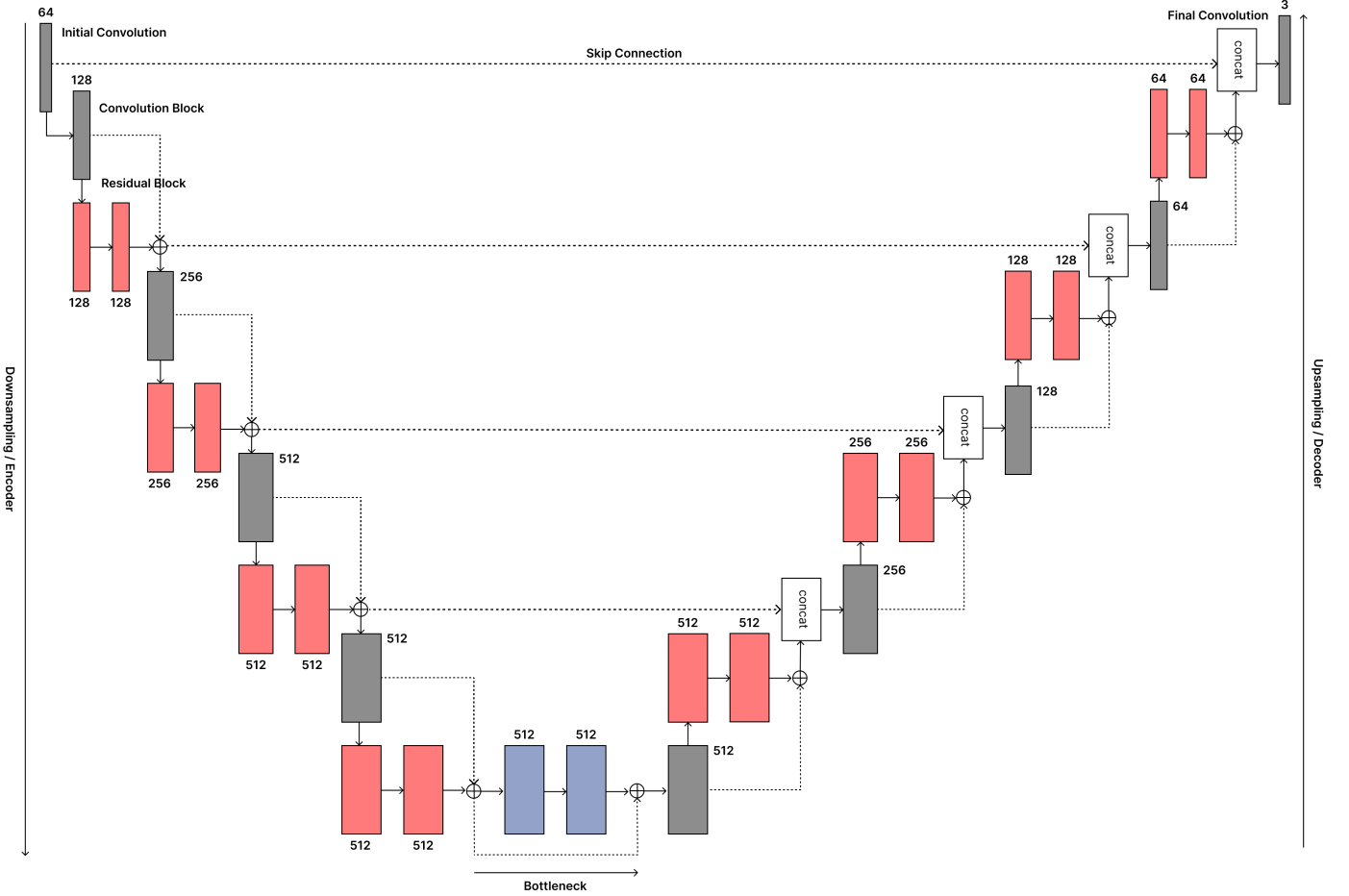


Fig. 5: U-net architecture for the generator

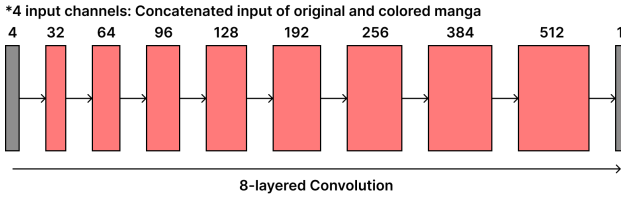


Fig. 6: Simple CNN architecture for the discriminator

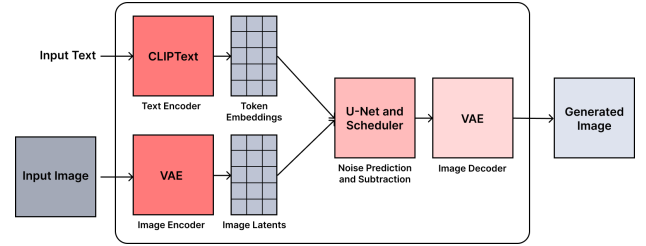


Fig. 7: Stable Diffusion System Architecture

we implemented 150 denoising/inference steps (against a default of 100), a guidance scale of 6 (default being 7, where a higher value promotes greater adherence to the textual input), and an image guidance scale set at 30 (default is 1.5, with higher values leading the generated image closer to the initial image). Furthermore, a negative prompt, 'Color', was utilized to guide the model towards producing black-and-white images.

#### 4.2.2 Loss Function

Equation 1 presents the Mean Squared Error (MSE) loss, also known as L2 loss, utilized for the Stable Diffusion model. This loss function facilitates a pixel-by-pixel comparison, essential for assessing the accuracy of the generated image against the actual image.

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (5)$$

### 4.3 Optimization

The table below quantifies the compute and time taken to train the cGAN model (49,200 training datapoints) and to fine-tune the stable diffusion model (1440 training datapoints).

## 5 RESULTS

### 5.1 For Manga Colorization

Fig. 8 illustrates a marked progression in the quality of colorization with each successive interval of 20 epochs for

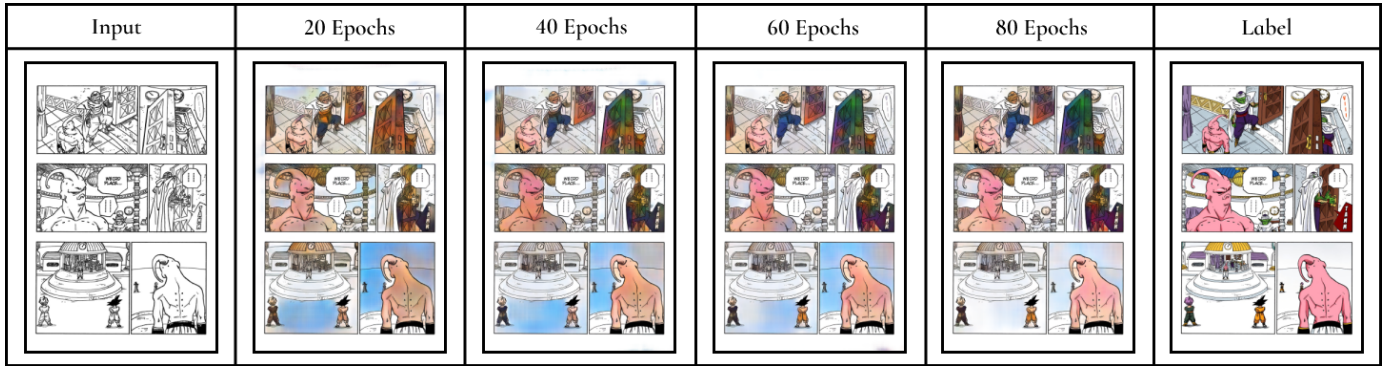


Fig. 8: Colorization improvement of manga page (not in training data) shown progressively

TABLE 2: Compute and duration for cGAN and Stable Diffusion model

Type	cGAN	Stable Diffusion
Runtime per epoch w/o GradScaler	19 hrs	N/A
Runtime per epoch w GradScaler	23 mins	9 hrs
VRAM utilized	7.1GB	22.7GB

a total of 80 epochs. Notably, the phenomenon of color spillage observed in the initial stages diminishes over time, and the hues associated with the characters exhibit a discernible deepening in intensity. Furthermore, the application of our model to unseen data yields commendable results, as evidenced by the vivid and accurate colorization. This outcome is indicative of the model’s robust ability to generalize effectively to novel manga pages (as can be seen in Fig. 9). Table 3 presents the discriminator’s final classification probabilities, demonstrating the model’s effective performance with balanced scores near 0.5 for correctly identifying fake images as generated and real images as authentic.

TABLE 3: Discriminator classification probability

Probability	Fake	Real
Generated Image	0.479	0.541
Actual Image	0.405	0.595

To facilitate the precise colorization of novel manga pages, the proposed model can undergo a fine-tuning process using a select set of colored pages. This targeted refinement enables the model to effectively apply its learned colorization strategies to subsequent, uncolored pages of the manga series. Such an approach ensures consistency and accuracy in the colorization of new manga content, leveraging the insights gained from the limited, yet representative, colored dataset.

The table below (Table. 4) quantifies the decrease in generator loss progressively after sets of 20 epochs. The generator and discriminator loss is also illustrated by the loss graph below in Fig. 10 and Fig. 11.

## 5.2 For Style Transfer

The experimental phase of our study on style transfer was undertaken utilizing a constrained dataset over a finite



Fig. 9: Prediction on unseen data (not present in Kaggle dataset)

TABLE 4: Generator loss shown epoch-wise

Loss	20 Epochs	40 Epochs	60 Epochs	80 Epochs
Generator Loss	10.76	8.817	7.513	7.446

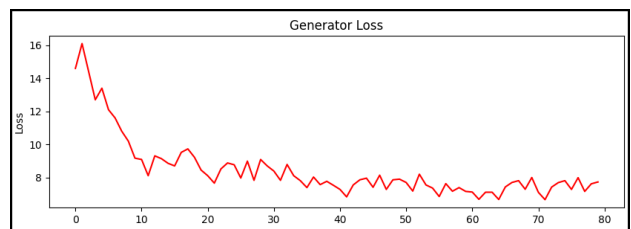


Fig. 10: Loss graph for generator

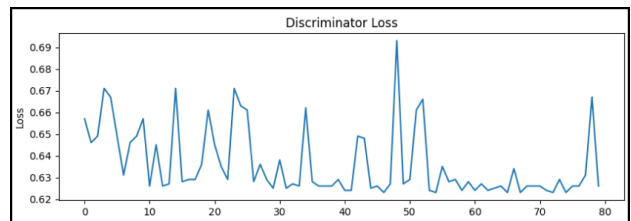


Fig. 11: Loss graph for discriminator

duration. Preliminary outcomes, as depicted in Figure X,

