APS360: ANIMEMORPH - GAN ANIME GENERATION

Apurva Agrawal
Student# 1009327845
apurva.agrawal@mail.utoronto.ca

Justin Li
Student# 1007731297
jstn.li@mail.utoronto.ca

Ryan Neil Alumkal
Student# 1009960632
ryan.alumkal@mail.utoronto.ca

Shirley Li
Student# 1007636569
shir.li@mail.utoronto.ca

1 INTRODUCTION

The purpose of our project is to generate anime-style images from an image of a person. Anime refers to a style of animation originating from Japan, which has garnered a large audience over the years throughout the world, and with the popularization of Generative Adversarial Networks (GANs), there has been more demand with people interested in resembling their favorite anime characters. Our motivation for this project stems from the fact that commercial image editing software falls short of achieving the results that we and many others are hoping for. Additionally, creating anime images manually in specific styles requires professional artistic skills that may not be available to everyone. Our goal is to take in an image of a person and show an illustration of what they'd look like in their favorite anime style.

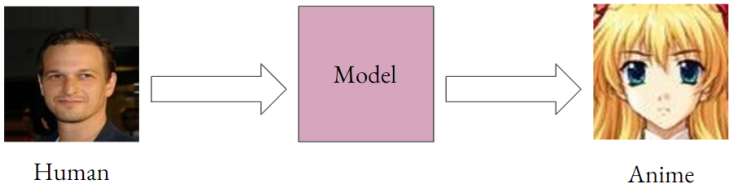


Figure 1: Diagram of the expected result of the project.

To do this, we will develop a deep learning model, specifically, a GAN, that would allow the translation of the features seen in a person's face to one in an anime-style. We aim to preserve the unique characteristics of a person, while applying elements unique to anime characters, such as big eyes, colorful hair, etc. CycleGANs would be one GAN model that is better at learning the mapping between human faces and anime characters. Additionally, because of the nature of CycleGANs, they do much better than say our baseline Deep Convolutional GAN (DCGAN). This is because CycleGANS not only learn mapping functions between two domains X and Y given training samples, but they also have two discriminators that are used for matching the distribution of generated images to the data distribution in the target domain Zhu et al. (2017). Lastly, CycleGANs use cycle consistency losses to prevent the learned mappings from contradicting each other and can try to reconstruct the fake generated images back to the original input.

2 ILLUSTRATION

An illustration of our final human-to-anime face generator is presented in Figure 2.

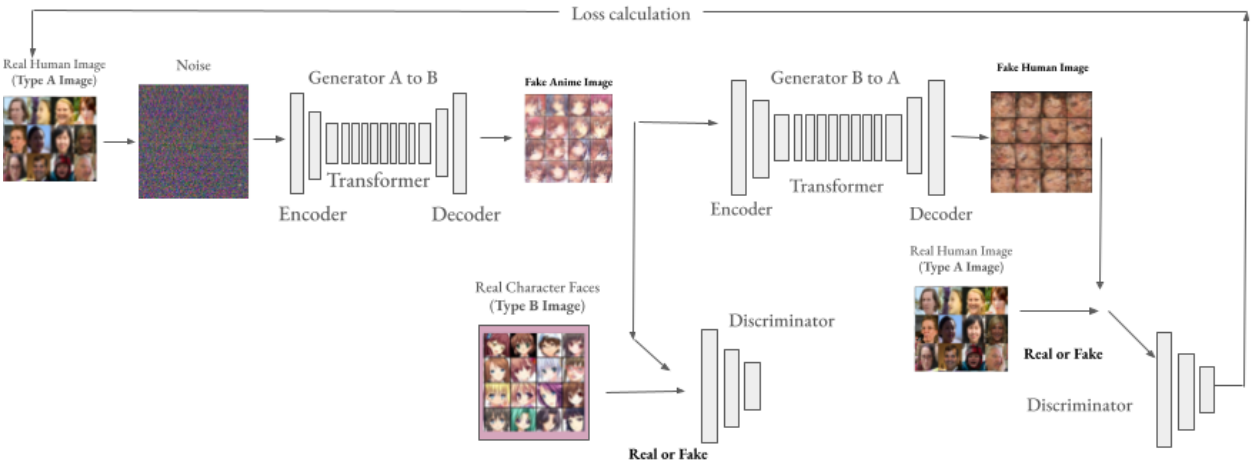


Figure 2: Primary model CycleGAN.

3 Background & Related Work

There has been a variety of research and newly developed applications around anime face generation. The following five related works are the most relevant to our topic.

The Medium article **"Style Transfer: Human to Anime Faces"** by Jake Valenciano explores converting human faces to anime-style faces using GANs. It details the use of CycleGAN and other techniques for style transfer. Then, they also address the challenges in training models to capture the distinctive features of anime art from human faces, considering both technical and artistic aspects (OkayBooster, 2019).

In the paper, **"AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation,"** a new method for applying visual styles from one headshot photo to another is introduced. The technique uses a multi-scale approach to align local statistics, such as contrast and lighting, between reference and target images. By applying Convolutional Neural Networks (CNNs), the method effectively replicates stylistic elements used by professional photographers. It adapts to different facial features and includes an automatic selection mechanism to find suitable reference images from a dataset. This unique method has been tested thoroughly and has proven itself by producing high-quality, stylistic transformations of head-shot photos (Li et al., 2022).

Reface is a mobile app that uses artificial intelligence to create animated versions of faces from photos, enabling realistic face swaps and animations. Leveraging deep learning algorithms, like GANs, Reface maps facial expressions and movements from source videos onto user photos, producing high-quality, lifelike animations. The app is user-friendly, requiring only a photo upload to generate animations, and integrates with social media for easy sharing. Reface exemplifies significant advancements in face animation and manipulation, making sophisticated AI technology accessible to the public and fostering new forms of creative expression and engagement in entertainment and social media (Ref).

The paper **"Landmark Assisted CycleGAN for Cartoon Face Generation"** presents a method to enhance CycleGAN by integrating facial "landmarks" to generate high-quality animated faces, using human faces as the input. The approach includes landmark-guided and landmark-matched local discriminators to accurately maintain facial structure consistency across all output data. The training is done in a two-stage process. Initially, the model is trained to achieve coarse results by not using local discriminators, then, specific facial features are refined using the discriminators. This method improves the visual quality of the input faces by maintaining its structural integrity and also enhancing prominent, noticeable features Wu et al. (2019).

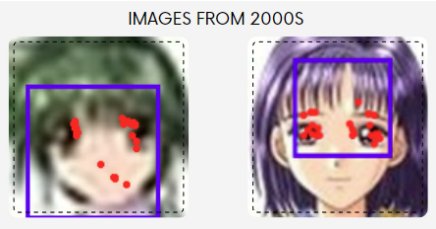
Aside from the above mentioned sources, our group most widely used the **"Face2Anime"** project as a reference to develop our finalized primary model. This project uses CycleGAN, like what our group had in mind, and optimizes Cycle Consistency Loss, Generator Loss, and Discriminator Loss to ensure accurate transformations. The project trains on a dataset of 3,500 real female faces and 3,500 anime faces, experimenting with different hyper-parameters like batch size and a "Random Pool" method to prevent mode collapse, an issue we faced while developing our model. From taking inspiration from the methods suggested by the developers of this project, we tackled the issue of mode collapse and settled on certain hyper-parameters such as the batch size and number of epochs. (Imtri1998)

4 Data Processing

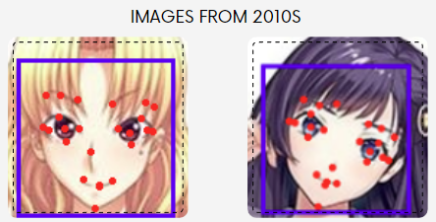
We require two data sources for our model. The data that we compiled consisted of one containing anime faces and the other containing human faces (CelebA) (Liu et al., 2015).

Our first dataset is the anime-face dataset from Kaggle (Churchill, 2019) containing 63,632 anime-face photos. From this dataset, we selected 10,000 images that contain anime characters with various styles, genders, and ages. We preprocessed the dataset by resizing all 63,632 images to 128x128 to reduce input size and keep all input images uniform. This dataset originally contained anime photos spanning the years 2000-2019. Our group decided to pick 10,000 from the newer style images (2011-2019), as they had better quality and had the anime photo style that we were interested in generating.

Additionally, we aligned the images with facial landmarks. This helped us remove unnecessary noise and invalid images in our dataset. After running anime landmarking software (Hysts, 2023) we removed duplicate anime characters, anime faces with unclear features that the landmarking software did not pick up, and characters with side profiles only. In addition, by Figure 3 we can see that on average the anime photos from the 2000s compared to the 2010s were lower quality and supported our decision in using a dataset with the newer images. While the landmarked photos were not directly used in the primary model inputs, using the landmarking software helped us pick out anime faces with reduced noise.



(a) Anime photos from 2000s

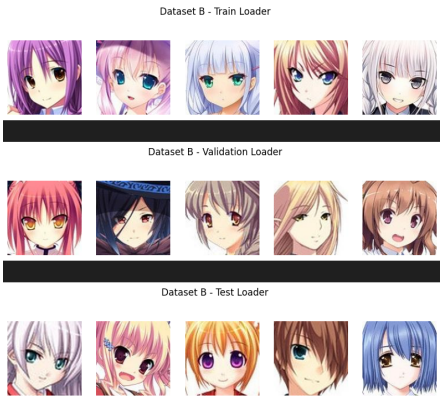


(b) Anime photos from 2010s

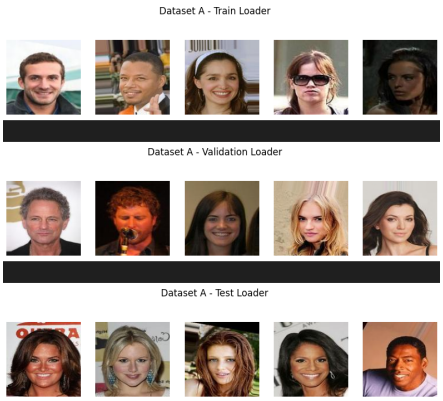
Figure 3: Anime photos landmarked.

Next, the CelebA dataset contains ∼200,000 face images of celebrities. Again we resized all the images to 128x128 to ensure all the images were the same size and then selected 10,000 human face images with 5,000 male and 5,000 female faces and combined them into one large dataset. Again we aligned the images with facial landmarks, which helped us remove unnecessary noise in our dataset. The reason we used the same amount of males and females is to make sure our model tries not to develop any bias towards one specific gender.

Finally, after preprocessing all of our data we used a data loader which applies a transformation on our data and splits our data in training, validation, and test sets with a (0.7, 0.15, 0.15) ratio. The following photos in Figure 4, display a few examples of what our two classes of data look like.



(a) Anime data loaders



(b) Human data loaders

Figure 4: Comparison of data.

5 ARCHITECTURE

The main architecture used for this project will be a CycleGAN, Zhu et al. (2017). The CycleGAN consists of two GANs in which will contain a generator and a discriminator. The role of the first generator is to create photos of anime characters (Type B) from a human's portrait photo (Type A), while the role of the second generator is to regenerate the portrait photo from the anime photo. Each generator is made up of an encoder, transformer, and decoder. The encoder will be a convolutional block with normalization layers and activation functions to extract higher-level features and compress the image size. The transformer will contain residual blocks that transform the image from Type A to Type B, Dubey & Singh (2023). Lastly, the decoder is made up of transpose convolution layers, which will build the lower-level features back. The second generator follows the same structure but instead generates image Type A from Type B. When translating Type A to Type B, noise is added to the original human image before it is fed into the generator to provide better results.

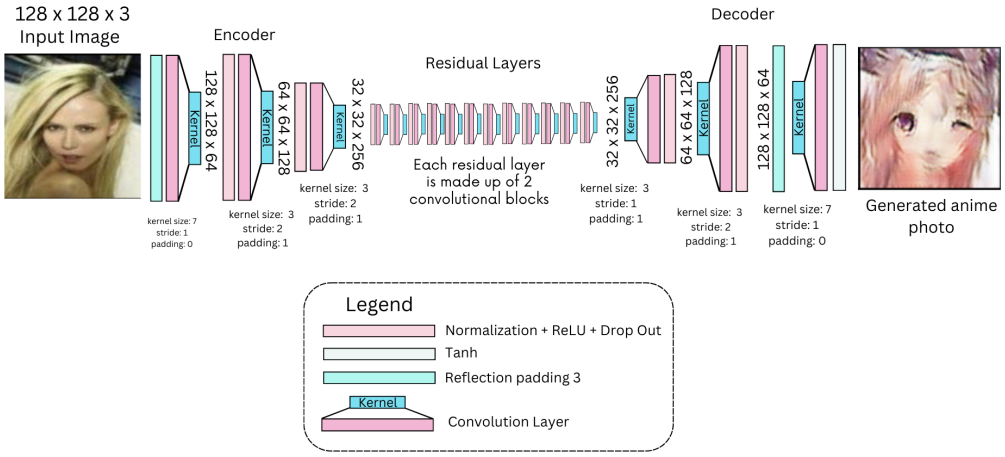


Figure 5: Primary model generator architecture.

The discriminator is made up of CNNs with multiple convolutional layers and outputs a single scalar output to check whether or not the input image(s) is a real photo or a generated one.

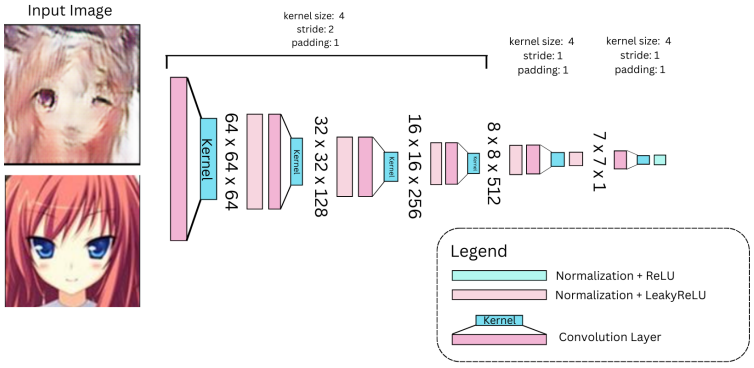


Figure 6: Primary model discriminator architecture.

6 BASELINE MODEL

Our baseline model will be a version of a DCGAN. In comparison to our main model, DCGAN is a more simplified version of a CycleGAN. Unlike traditional DCGANs that generate images from noise, the baseline model will generate images from a provided image (human faces).

The GAN consists of two components, a generator and a discriminator, a description of each can be found in previous sections. Both networks are implemented using CNNs to process the images. The layers include Conv2D, BatchNorm2D, and LeakyReLU blocks in sequence, for both the generator and discriminator(Radford et al., 2016). Finally,

the model will be updated using adversarial training, a generator aiming to fool the discriminator, and the discriminator aiming to better distinguish between real and fake images.

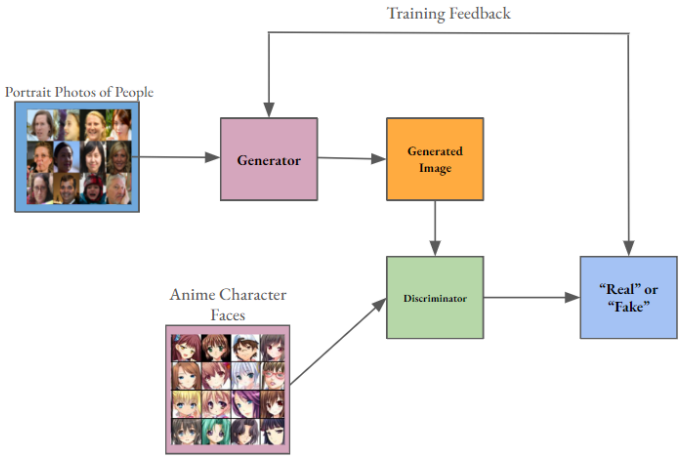


Figure 7: Illustration of DCGAN mode.

7 Quantitative Results

First, to best understand the model's performance, we measured and graphed the relationship between the training and validation loss of our primary model. This would provide us with a concrete result as to how well the model is generalizing. To minimize our training loss, we tried a variety of hyper-parameters, and although initially facing many issues where the discriminator was not distinguishing between real and generated images adequately, resulting in a high discriminator loss, we fixed it by increasing the number of epochs the model is trained for. This was the final step to achieving the training loss, as shown in Figure 8.

From analyzing the graphs for both anime and human generation, it is clear that the training and validation losses decrease over time, indicating that the model is learning. The gap between the two lines is also small, suggesting good generalization. This pattern continues until around epoch 80 where the training loss seems to stabilize. This is most likely due to the relatively low amount of data used, at 10,000 per dataset.

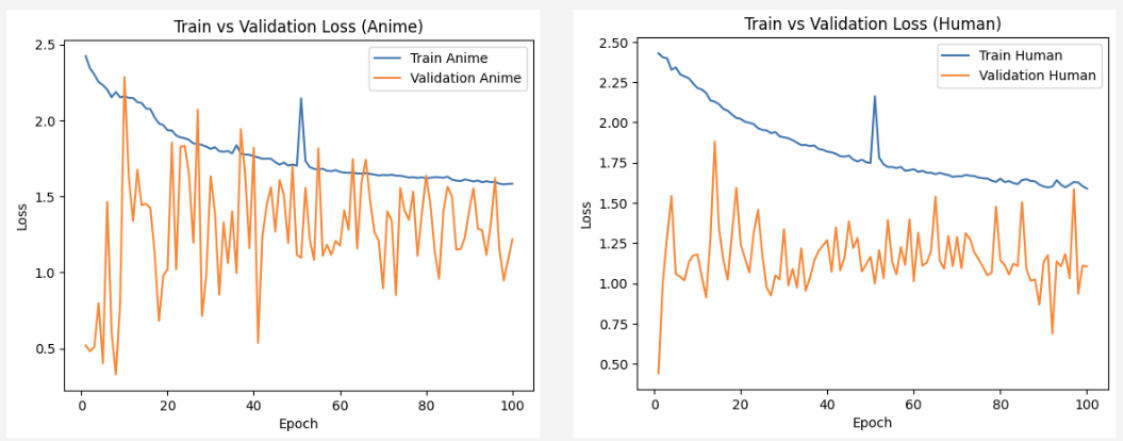


Figure 8: Train vs validation loss of primary model.

Furthermore, we used the Frechet Inception Distance (FID) metric to evaluate the quality of the generated images for the primary model, primary model with noise, and baseline model, and compare their results numerically (Brownlee,

2019). In essence, the FID metric simply measures the similarity between two sets of images by comparing the distributions of feature vectors extracted from these images using a pre-trained Inception network. As CycleGAN aims to preserve content while changing style, FID helps ensure that while the style is transferred effectively, the content remains coherent and similar to the real-world counterpart. So, when calculating the respective FID scores for the models, and comparing their results we were pleasantly surprised to discover that our primary model with noise performs better than the rest when translating from human-to-anime character and our primary model without noise also performs significantly better in terms of accuracy of image reconstruction, compared to the other models - shown in Table 1. As both models perform better than the baseline model and our primary model with noise results in a better FID score for anime image generation (our project goal), we concluded that our primary model performs within our group's expectations.

Table 1: Performance metrics comparison.

	Primary Model	**Primary Model with Noise**	**Baseline Model**
FID Score for Anime Images	2071.5011	**1974.66**	2445.22
FID Score for Human Images	**1014.40**	1223.27	1570.28
Accuracy for Anime Face	**90.73%**	89.27%	50.51%
Accuracy for Human Face	**95.80%**	49.80%	49.23%

8 Qualitative Results

We used test images, which were randomly selected from the whole dataset, making up 15% of the whole dataset, to visualize the model's performance. It is split into two different results, anime-to-human (Figure 9) and human-to-anime (Figure 10). The model was trained for 100 epochs.

Figure 9: Anime-to-human-to-anime translation.

Figure 10: Human-to-anime to human translation (noise added to figure for context).

As seen above, the model does sufficiently well given the computation power available to us. It is successfully able to capture the intricate details of humans (ex: facial structure, features, proportions) and anime characters (ex: large eyes, small mouths, colourful hair) in both the generated human and anime faces respectively.

Noise was added to human-to-anime translation as this approach yielded more visually appealing results compared to without noise. However, the reconstructed human photos look less accurate due to the noisy input, which did affect the CycleGAN training process.

Mode collapse also occurred around 60 epochs, which caused certain inputs to result in pixelated/distorted outputs. This may be due to the relatively low amount of data, at 10 thousand images per dataset.

Figure 11: Mode collapse occurring for certain input images (left) and output (right).

Additionally, the model did not perform the best for humans who were vastly different in terms of diversity compared to the trained anime characters. This mainly occurred for male humans due to the anime dataset having mostly feminine-looking characters. Additionally, humans with darker skin, accessories (ex: glasses, hats), and old age did not perform as well due to the lack of variety in the anime dataset in these categories. Most of the anime characters from the original dataset used for training were young and female with light skin.

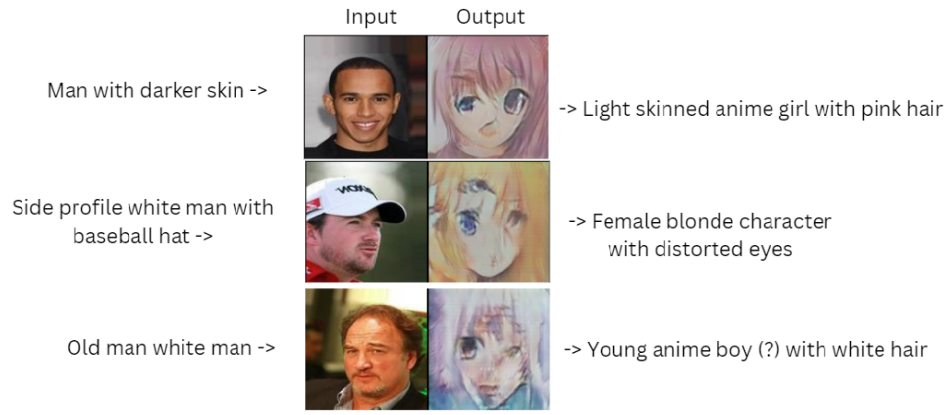


Figure 12: Limits of the model.

9 Evaluate model on new data

To evaluate the model on unseen data, we collected a new dataset from Kaggle of ∼2700 female and ∼2700 male photos Gupta (2020). This is a separate dataset from the CelebA dataset and is used to test the final performance of the generator human-to-anime model as well as to calculate the overall accuracy of the reconstructed photos. We wanted to choose a dataset that has a broad spectrum of genders, races, and ages similar to the CelebA dataset so that we can test how our model reacts to new faces. To pre-process the new dataset we cropped and resized the images to 128x128 (Figure 13), randomly picked out 750 of both female and male faces with a total of 1500 testing images, and fed it into the pre-trained model of the generator human-to-anime.

Upon inspecting the training and validation loss graph as well as the output images from each epoch, the team decided to use the model generated on the 67th epoch as the final model for evaluation. Figure 14 shows some samples of the output from the testing dataset. Since the primary objective of this project is to generate human photos into anime ones, only one generated needs to be used for evaluation. One major observation is that around 40 % of the generated testing images were prone to mode collapse, where the discriminator would constantly get fooled no matter what input image is being fed in. Compared to the training images, the testing images had less accurate positioning of facial features

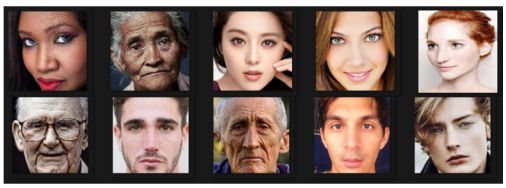


Figure 13: Sample of testing data from Kaggle.

and were less accurate at transferring style and colour. Many male human photos turned out with noisy outputs of a light blue colour, this suggests that the discriminator architecture can be improved further.

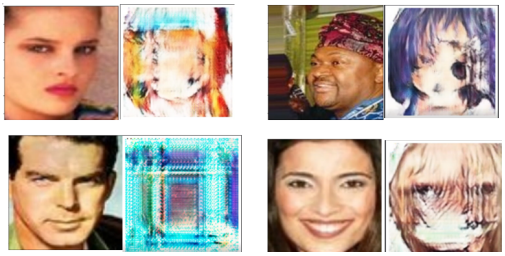


Figure 14: Sample of testing data output.

Other than the male and female datasets from Kaggle, we also inputted noise, animals, and plants to test the limits of our model as shown in Figure 15. We can see that the cactus produced the results closest to an anime figure, which we hypothesize could be due to the similar shaping between the cactus and human faces. The model identified it as a bowl-cut-like hairstyle. This testing also further confirms our assumption that our model cannot transfer style and colour with 100 % accuracy.

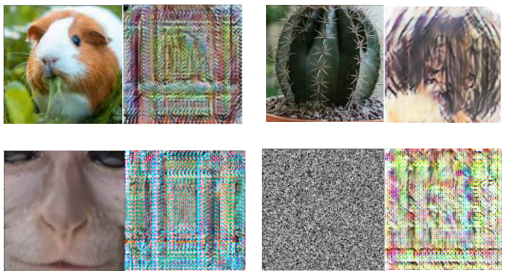


Figure 15: Testing output of guinea pig, cactus, noise, and monkey.

10 Discussion

We believe that our model of the CycleGAN is performing quite well given the input data and parameters. This is primarily because the model can successfully catch minute details within the anime/human faces and use them to generate a more accurate result. To be precise, the model accomplishes this by rigorously training and optimizing the cycle-consistency loss, which is a part of the training loss that ensures that if an image is translated to the other domain (real to anime, or vice versa) and back, it should closely resemble the original image. Furthermore, we were able to test for the image reconstruction more accurately thanks to the FID metric as mentioned in the quantitative results section above.

However, we found it unusual how the results for certain test images were not as well defined as others. This might have been caused by various factors, such as mode collapse, which occurred around 60 epochs, insufficient data, given that our model only trained on 10 thousand images per dataset (anime and human), and the volatility of the validation loss (Figure 8) which indicates inconsistencies, i.e. overfitting and lack of generalization within the model.

Additionally, we attempted to use our landmarked images as inputs to facilitate shape transformations between human and anime domains. The idea was to enhance feature extraction by developing multiple discriminators, each focused

on specific facial regions, using the landmarked features as inputs. In Figure 16 you can see how we can crop local patches (eyes, nose, and mouth) for local discriminators as input. This approach aimed to capture more nuanced details and improve the model's ability to perform accurate shape transformations.

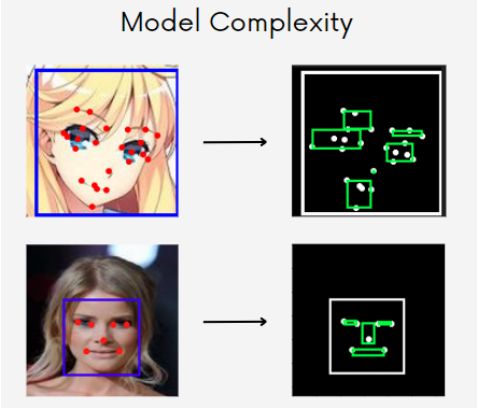


Figure 16: Landmarking facial patches.

However, the computational demands of this method proved too demanding to continue beyond the initial testing phase. While we could not create a refined implementation due to these challenges, we believe that revisiting and refining this approach in future work could lead to significant improvements in the model's performance and the quality of the generated images.

From this experience, we learned how to implement a CNN model with PyTorch and develop a functional animated image generator, given the image of a person. We also discovered how simple changes in parameters to the model can majorly impact the final result. For instance, something as simple as applying ReLu in the generator compared to LeakyReLu, and a batch size of 1 helped us generate better results. Finally, by working together as a team, we also learned how to distribute the tasks amongst ourselves, based on their difficulty and individual strengths. Certain tasks, such as training the model, took longer than expected, and some tasks, like adjusting primary model parameters. and using landmarking, were harder than anticipated. Thus, by adjusting and adapting to the challenges, our team managed to complete the project as efficiently and fruitfully as possible.

11 Ethical Considerations

First, altering a person's face to an animated version poses various ethical concerns, especially regarding privacy and consent. This can lead to unwanted consequences, negatively impacting a person's self-esteem and body image if the image is not generated mindfully. Thus, to mitigate this issue, we have decided to be selective of the images on which we train our model. To specify, our dataset does not include images of animated characters that are not non-human (ex, elves, monsters, robots), rendering it unlikely for the model to generate images that do not reflect a person's likeliness.

Furthermore, anime and artwork data collection poses legal risks, such as intellectual property infringement. Acknowledging these ethical issues when creating AI models is essential as it ensures that when these AI technologies are developed and deployed, they do not harm others and adhere to legal standards. Therefore, we have ensured that all research materials used to develop our model are for public use and will not cause copyright infringement. This ensures that the images we use have been gathered with the subject or artist's explicit consent.

Inclusivity is another issue. To avoid representation bias, the model must be trained on a diverse dataset. This includes considering various demographic characteristics such as skin tone, body type, age, and gender when collecting data to train the model. While this has not been the case for human data, this has been an issue for the team when it came to anime data, as there is no publicly available data for anime characters with diverse characteristics.

REFERENCES

Reface. URL https://reface.ai/.

Jason Brownlee. How to implement the frechet inception distance (fid) for evaluating gans, Oct 2019. URL https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/.

Spencer Churchill. Anime face dataset, Oct 2019. URL https://www.kaggle.com/datasets/splcher/animefacedataset/data.

Shiv Ram Dubey and Satish Kumar Singh. Transformer-based generative adversarial networks in computer vision: A comprehensive survey, Feb 2023. URL https://arxiv.org/abs/2302.08641.

Ashwin Gupta. Male and female faces dataset, Sep 2020. URL https://www.kaggle.com/datasets/ashwingupta3012/male-and-female-faces-dataset.

Hysts. Hysts/anime-face-detector: Anime face detector using mmdet and mmpose, 2023. URL https://github.com/hysts/anime-face-detector. GitHub repository.

Imtri1998. Imtri1998/face2anime-using-cyclegan. URL https://github.com/lmtri1998/Face2Anime-using-CycleGAN. GitHub repository.

Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, 24: 4077–4091, 2022. doi: 10.1109/tmm.2021.3113786.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

OkayBooster. Style transfer: Human to anime faces, Dec 2019. URL https://medium.com/@jakethevalencian/style-transfer-human-to-anime-faces-5464ec3ab8e.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, Jan 2016. URL https://arxiv.org/abs/1511.06434.

Ruizheng Wu, Xiaodong Gu, Xin Tao, Xiaoyong Shen, Yu-Wing Tai, and Jiaya Jia. Landmark assisted cyclegan for cartoon face generation, 2019. URL https://ar5iv.labs.arxiv.org/html/1907.01424.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.244.