Face Aging using CycleGANs

Vagelis Anagnostopoulos Orestis Vaggelis Andreas Kouridakis

24/6/2024

Abstract

Face aging, the process of making a person's face look older or younger in photos, is a challenging task in the field of computer vision. The complexity of this task stems from the various factors influencing the aging process, such as facial expressions, lighting, and environmental conditions. Additionally, collecting paired images—photos of the same individual taken many years apart—is both expensive and difficult, limiting the performance of many conventional methods. The project's goals are two: first to explore the use of CycleGANs for face aging, as CycleGANs can learn the mapping between two domains using unpaired image data, and second, to integrate attention mechanisms in the generator and the discriminator to determine if we can enhance the results by capturing the intricate details of the aging process more effectively.

1 Introduction

Transforming a face to appear older or younger, known as face age transformation, is a significant and challenging task in the field of computer vision. This process allows us to set any target age for an input face image and expect an output that accurately reflects the desired age characteristics.

The ideal age transformation model should satisfy two primary criteria: identity preservation and natural appearance. Identity preservation ensures that the transformed image maintains the original identity of the person, while natural appearance guarantees that the generated faces look realistic across the age transformation.

In this project, we aim to exploit the strengths of CycleGANs for facial aging. Our first goal is to explore how CycleGANs can be used to learn the mapping between young and old faces using unpaired image data. This will allow us to avoid the challenges associated with collecting large paired datasets. Our second goal is to extend the CycleGAN model by integrating attention mechanisms into both the generator and the discriminator.

Starting with a review of related work on GANs, CycleGANs and attention mechanisms, we detail our methodology. Then, we describe our dataset preparation, including the selection and preprocessing of images from the UTKFace dataset. We then present our approach to integrating attention mechanisms into the CycleGAN framework. Finally, we evaluate the performance of our model using both qualitative and quantitative metrics, demonstrating the effectiveness of our approach in producing realistic age-progressed images.

2 Related work

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [9], have been widely used on tasks of image regeneration. GANS consist of two adversarial models, the Generator G and the Discriminator D. The generator creates images that aim to resemble real images, while the discriminator evaluates the authenticity of these images, distinguishing between real and generated ones. In other words D and G play the following two-player minimax game with value function V (G, D):

$$\min_{G} \max_{D} V(D, G) = E_{x \sim p_{\text{data}}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{1}$$

z is a vector , sampled from the prior p(z), to the data space.

2.2 Cycle-GANs

CycleGANs [1], a variant of the traditional GAN architecture, are particularly effective for addressing image-to-image translation tasks, commonly known as style transfer. This involves transforming an image from one category to another, such as changing a summer landscape into a winter one or converting a photograph into a painting. The key advantage of CycleGANs is their ability to perform these transformations without the need for paired images, making use of unsupervised learning to derive relevant features from two distinct categories of data. Each translation requires two generators and two discriminators, thus training two GANs concurrently. A fundamental aspect of CycleGANs is cycle-consistency, which ensures that if an image is transformed from category A to B and then back to A, the final image closely matches the original. This is enforced by a cycle-consistency loss, which measures the similarity between the original and the twice-transformed images.

$$\mathcal{L}_{\text{cyc}}(G, F) = E_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$$

Additionally, an identity loss term ensures that when a generator processes an image from its target category, the output remains close to the input.

$$\mathcal{L}_{\text{identity}}(G, F) = E_{y \sim p_{\text{data}}(y)}[\|G(y) - y\|_1] + E_{x \sim p_{\text{data}}(x)}[\|F(x) - x\|_1]$$

In the context of face aging, CycleGANs learn to extract distinguishing features of young and old faces, enabling the model to apply aging characteristics to a young face or rejuvenate an old face. The final loss of the CycleGAN can be expressed as follows:

$$L_{\text{total}} = L_{\text{GAN}} + \lambda_{\text{cycle}} \cdot L_{\text{cycle}} + \lambda_{\text{identity}} \cdot L_{\text{identity}}$$

For our CycleGAN experiments, we set $\lambda_{\text{cycle}} = 10$ and $\lambda_{\text{identity}} = 5$. As suggested by the authors of the CycleGAN paper, we applied learning rate decay after epoch 100. The initial learning rate was set to 0.0002. To mitigate oscillations, we updated the discriminators using a history of generated images rather than the ones produced by the latest generators. Specifically, we maintained an image buffer that stored the 50 most recently created images [10]. We also initialized the weights from a Gaussian distribution N(0, 0.02). We replaced the negative log likelihood objective with a least-squares loss [12]. This loss function is more stable during training and produces higher quality results. Specifically, for a GAN loss $L_{\text{GAN}}(G, D, X, Y)$, we train the generator G to minimize $E_{x \sim p_{\text{data}}(x)}[(D(G(x)) - 1)^2]$ and the discriminator D to minimize $E_{y \sim p_{\text{data}}(y)}[(D(y) - 1)^2] + E_{x \sim p_{\text{data}}(x)}[D(G(x))^2]$. In the case of the model with the attention mechanism, the training was conducted in the same manner as before, but for a total of 100 epochs, due to the unavailability of sufficient computational resources.

2.3 Self Attention

Attention mechanisms have revolutionized natural language processing (NLP) and computer vision by allowing models to dynamically weight elements based on their relevance. In NLP, attention mechanisms significantly improve tasks like machine translation and text summarization by capturing global dependencies and addressing the limitations of fixed-length context vectors in traditional recurrent neural networks. The Transformer model by Vaswani et al. (2017) [2], employing self-attention, set new benchmarks in language translation and text generation. Similarly, attention mechanisms have transformed computer vision through Vision Transformers (ViTs), introduced by Dosovitskiy et al. (2020) [3]. ViTs apply self-attention to image patches, capturing long-range dependencies and achieving state-of-the-art performance in image classification. Furthermore, integrating self-attention into Generative Adversarial Networks (GANs) has enhanced image generation quality. Self-Attention GANs (SAGANs), demonstrated by Zhang et al. (2019) [4], model global dependencies more effectively, producing coherent and detailed images.

3 Dataset

To take advantage of CycleGAN's capability for unpaired image-to-image translation we created a dataset using the UTKFace dataset. Initially, we chose photos that were cropped to focus on the face region, minimizing irrelevant background information. Additionally, grayscale images were filtered out to ensure uniformity in color across the dataset. Then, all selected images were resized to 256x256 pixels, a common size used in many CycleGAN models [1] and suitable for most applications. For the training set, 4301 cropped images were selected for the 20s dataset and 2244 cropped images for the 50s dataset.

4 Implementation

4.1 Model Architecture

We used the implementation provided by the original CycleGan paper [1]. The Generator network begins with three convolutional layers that progressively downsample the input image to capture low-level features at multiple scales. The core of the generator consists of nine residual blocks, which allow the network to learn identity mappings and preserve important features across layers. Finally, to reconstruct the image to its original resolution, the generator employs two deconvolutional layers that upsample the feature maps. The discriminator network uses a series of convolutional layers to progressively downsample the input image and extract features. The final layer produces a single scalar to determine if the input image is real or fake.

4.2 Our approach

We also implemented the original CycleGAN model with the addition of self-attention in the generator. This idea was inspired by Self-Attention GANs (SAGANs) [4] as the authors of that paper combine generative adversarial networks with self-attention to help with modeling long-range, multi-level dependencies across image regions. Their model managed to set the best Inception score and Frechet Inception distance on the ImageNet dataset. Since then, numerous papers have implemented attention with GANs and some have achieved great results using self-attention with CycleGANs [5], [6]. The authors of those papers observed that the CycleGAN's output images' backgrounds were affected in unwanted ways, leading to unrealistic translations, so they incorporated self-attention into the image translation framework to encourage the generation of more

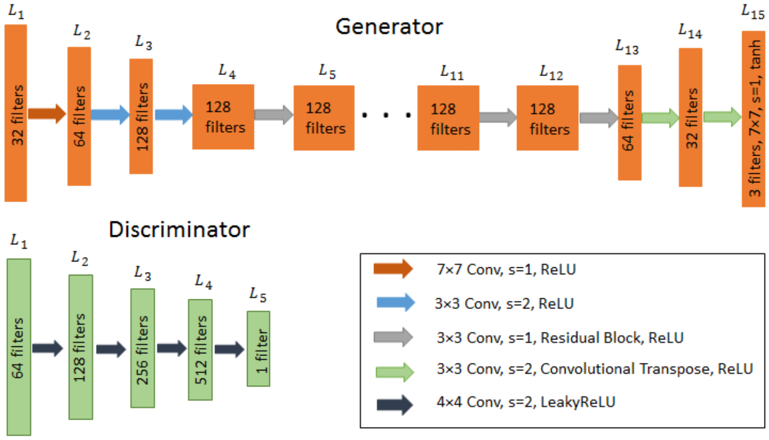


Figure 1: The original CycleGan model architecture.

realistic images compared to classic vanilla GANs by helping the model focus on specific objects without altering the background. Essentially, self-attention works as a mask to separate the foreground from the background of the image so the model can focus only on the foreground.

As mentioned earlier, for our experiments, we used a cropped dataset containing only faces as images. While there is no need to use self-attention to separate the foreground from the background of those images, self-attention should still help our model focus on facial features that alter as someone grows older. Thus, we integrated self-attention in our architecture to explore the possibility of generating more realistic results since the attention mechanism helps the model to add or remove those particular facial features.

For the generator, ideally, we wanted to implement two self-attention layers to allow the network to capture global dependencies and improve feature representation after both the downsampling and upsampling stages. However, due to low computational resources, we only applied one self-attention layer after the downsampling convolutional layers. For the discriminator network, we faced several challenges. GANs are known to be very hard to train and that was no exception in our case. The main challenge we faced is that the discriminator became too powerful too quickly. This can cause many issues such as the vanishing gradient problem for the generator, as it will receive very small gradients during backpropagation. To remedy this issue, we explored several changes to the vanilla CycleGAN model. Although we initially implemented a self-attention layer in the discriminator, we ultimately decided to remove it, as its presence enabled the discriminator to learn at an even more accelerated rate. Miyato et al. [7] proposed applying spectral normalization to the discriminator network to constrain the Lipschitz constant of the discriminator by restricting the spectral norm of each layer with a relatively small computational cost. Furthermore, the authors of SAGAN [4] proposed using spectral normalization in both the generator and discriminator networks. We also briefly experimented with different learning rates for the two networks. Heusel et al. [8] proposed using different learning rates for the generator and the discriminator to balance the training progress and help the two networks converge to a local Nash equilibrium.

5 Evaluation

Initially, we utilized DeepFace [11] to evaluate the age of our results in the test set. DeepFace is a facial recognition system developed by Facebook's AI Research group. It employs a convolutional neural network (CNN), to perform tasks such as face verification, recognition, and age estimation with high accuracy. As a baseline, we aimed for the generated images of older individuals (generated old) to be estimated between 50-60 years old and the generated images of younger individuals (generated young) to be estimated between 20-30 years old. It is important to acknowledge certain limitations associated with age detection using the DeepFace library. The accuracy of age prediction heavily depends on the quality of the input image. Additionally, the pre-trained models may not perform equally well across all demographics due to biases in the training data.

Another way we evaluate the accuracy of the generated aged faces is by using the Frechet Inception Distance [8]. We adopted the idea of using FID from SAGAN [4]. FID is a more principled and comprehensive metric, and has been shown to be more consistent with human evaluation in assessing the realism and variation of the generated sample. In a high level, FID calculates the Wasserstein-2 distance between the generated images and the real images in the feature space of an Inception-v3 network. However, the small number of photos in our test set led to numerical instability issues, specifically with the covariance matrices becoming near-singular. To mitigate this, we added a small value 0.000001 to the diagonal of the covariance matrices. This regularization significantly enhanced the numerical stability of the FID calculation, ensuring more reliable and consistent evaluation results for our face aging models.

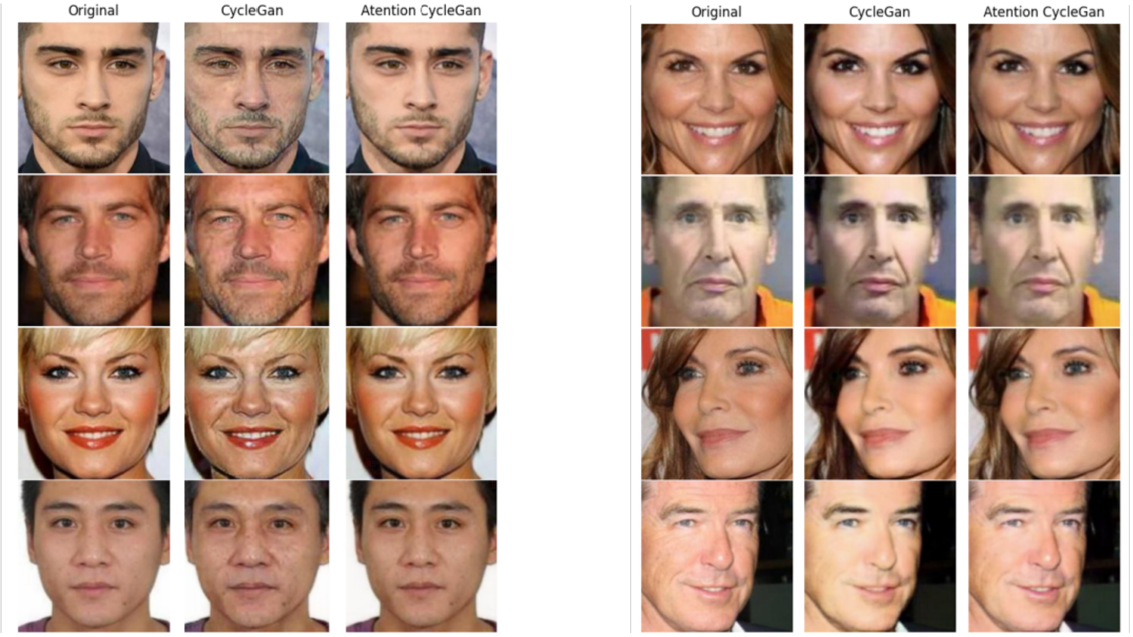
To assess the quality of our projects, we utilize the Peak Signal-to-Noise Ratio (PSNR) [14] and the Structural Similarity Index Measure (SSIM) [14] as primary metrics. PSNR is a widely adopted metric in digital image processing due to its simplicity and proven validity. It provides a quantitative measure of the peak error between the original and generated images, enabling straightforward comparisons. SSIM offers a more comprehensive assessment by taking into account the human visual system's sensitivity to contrast, luminance, and structure. This metric ranges from 0 to 1, where 1 indicates a perfect match between the compared images.

One commonly referenced metric for evaluating the generated images by CycleGANs is the Inception Score [4], which focuses on the classification confidence of generated images and their diversity across different classes. It does not directly measure how realistic the generated images are. This means that an image could be classified with high confidence as a certain class but still look unrealistic or contain artifacts. For instance, a generated image might be confidently classified as representing a 50-year-old, but could still contain unrealistic characteristics, such as distorted facial features, which IS would not penalize. Additionally, since the Inception v3 model, which is used to compute IS, is trained on ImageNet, it classifies images into one of the 1,000 ImageNet classes. These classes do not specifically include detailed human characteristics, such as different age groups or nuanced facial features. Despite its popularity, we chose not to use IS in our project for these reasons, as we need a more accurate assessment of the realism of the generated images.

6 Results

Its evident from the results the CycleGAN model successfully transforms the original images of individuals from their 20s to their 50s and vice versa. As we can see from the graphs, the generator loss fluctuates but converges over time, while the discriminator loss remains low and stable. Both cycle consistency and identity losses initially fluctuate but decrease and stabilize, indicating improved performance in generating realistic aged/younged images while maintaining consistency and original features.

In contrast, the Attention CycleGAN fails to produce significant transformations to either age group. Looking at the loss plots, we can see that the discriminator and generator losses do not seem to converge. The model appears to have learned to reconstruct the original image with a high degree of fidelity, as evidenced by the reduction in identity and cycle losses. However, it has not yet learned to perform the age transformation as intended, as the generator and discriminator do not appear to be learning effectively. This outcome suggests a need for further exploration and experimentation with the implementation of the self-attention mechanism in the CycleGAN model. The graphs are presented at the end of the report.



(a) Transforming images from 20s to 50s.

(b) Transforming images from 50s to 20s.

Figure 2: Examples of image transformation using CycleGAN and Attention CycleGan.

The results we observed in the generated images of CycleGAN and Attention CycleGAN are corroborated by their corresponding Fréchet Inception Distance (FID) metrics, with a baseline value of 0 indicating a perfect match to the original images. For CycleGAN, the FID scores are 99.418 for the O2Y transformation and 94.731 for the Y2O transformation, suggesting a deviation from the original images. In addition, Attention CycleGAN shows FID scores of 97.631 for the O2Y transformation and 104 for the Y2O transformation, indicating a slightly higher deviation.

In our evaluation of the CycleGAN and self-attention CycleGan models for face aging, both models demonstrated good performance in terms of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). These metrics, which are critical for assessing image quality and identity preservation, showed very high scores for both models. This outcome was anticipated, as PSNR and SSIM are not directly influenced by the aging process but rather by the model's ability to maintain the original identity and quality of the photo. The high PSNR and SSIM values indicate that both the CycleGAN and Attention CycleGAN have effectively learned to preserve the essential features and identity of the faces while they reconstruct the image, ensuring that the images remain recognizable and of high quality.

Table 1: Performance metrics for different image transformation models. O2Y describes the transformation from old to young, and Y2O from young to old.

Metric	CycleGAN (O2Y)	CycleGAN (Y2O)	AttnGAN (O2Y)	AttnGAN (Y2O)
FID	99.418	94.731	97.631	104
PSNR	28	29	31	32
SSIM	0.877	0.872	0.943	0.949

Regarding the performance of the estimator, the results for the CycleGAN-generated images revealed notable trends. For the "Generated young" category, the classifier estimated that 25% of the images fell within the 20-30 age range. In contrast, for the "Generated old" category, the classifier estimated only 1 out of 20 images to be within the 40-50 age range.

Table 2: Distribution of Predictions for CycleGAN.

	20-30	30-40	40-50
Generated young	5	12	3
Generated old	7	12	1

The evaluation of the Attention CycleGAN-generated images produced results that were approximately the same as those previously observed. For the "Generated young" category, the classifier estimated that 20% of the images fell within the 20-30 age range. Similarly, for the "Generated old" category, the classifier estimated that only 1 out of 20 images were within the 40-50 age range.

Table 3: Distribution of Predictions for Attention CycleGAN.

	20-30	30-40	40-50
Generated young	4	10	6
Generated old	10	9	1

7 Future work

We emphasize again the need for further experimentation with the implementation of the self-attention mechanism in the CycleGAN model. As mentioned earlier, the greatest challenge we faced was the discriminator learning too quickly in most of our training sessions, making it difficult for the generator to receive meaningful feedback. Freezing the discriminator for a few epochs during certain generator training epochs has been explored in GAN literature, but we unfortunately lacked both the necessary time and computational resources to implement this approach.

Moreover, finding optimal, separate learning rates for the two networks to slow down the discriminator's learning could potentially improve the model's stability and results. To avoid mode collapse, we could consider using the Wasserstein GAN (WGAN) loss [13], as it provides more stable gradients. Additionally, employing the standard GAN cross-entropy loss for the discriminator, rather than the Mean Squared Error (MSE) loss that we implemented, may yield better results.

Lastly, the self-attention CycleGAN that we implemented was trained for only 100 epochs, compared to the 200 epochs of the original model. We believe that incorporating a second attention layer in the generator, particularly after the upsampling stages, could help the model better capture global dependencies and thus improve the quality of the generated images.

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, arXiv: arXiv:1703.10593

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. arXiv:1706.03762, 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929, 2020.

[4] Han Zhang, Ian Goodfellow, Dimitris Metaxas and Augustus Odena, Self-Attention Generative Adversarial Networks, arXiv:1805.08318, 2019.

[5] Hao Tang, Hong Liu, Dan Xu, Philip H. S. Torr and Nicu Sebe, AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks, arXiv:1911.11897, 2021.

[6] Youssef A. Mejjati, Christian Richardt, James Tompkin, Darren Cosker and Kwang In Kim, Unsupervised Attention-guided Image to Image Translation, arXiv:1806.02311, 2018.

[7] Takeru Miyato, Toshiki Kataoka, Masanori Koyama and Yuichi Yoshida, Spectral Normalization for Generative Adversarial Networks, arXiv:1802.05957, 2018.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, arXiv:1706.08500, 2018.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, Generative Adversarial Networks, arXiv:1406.2661, 2014.

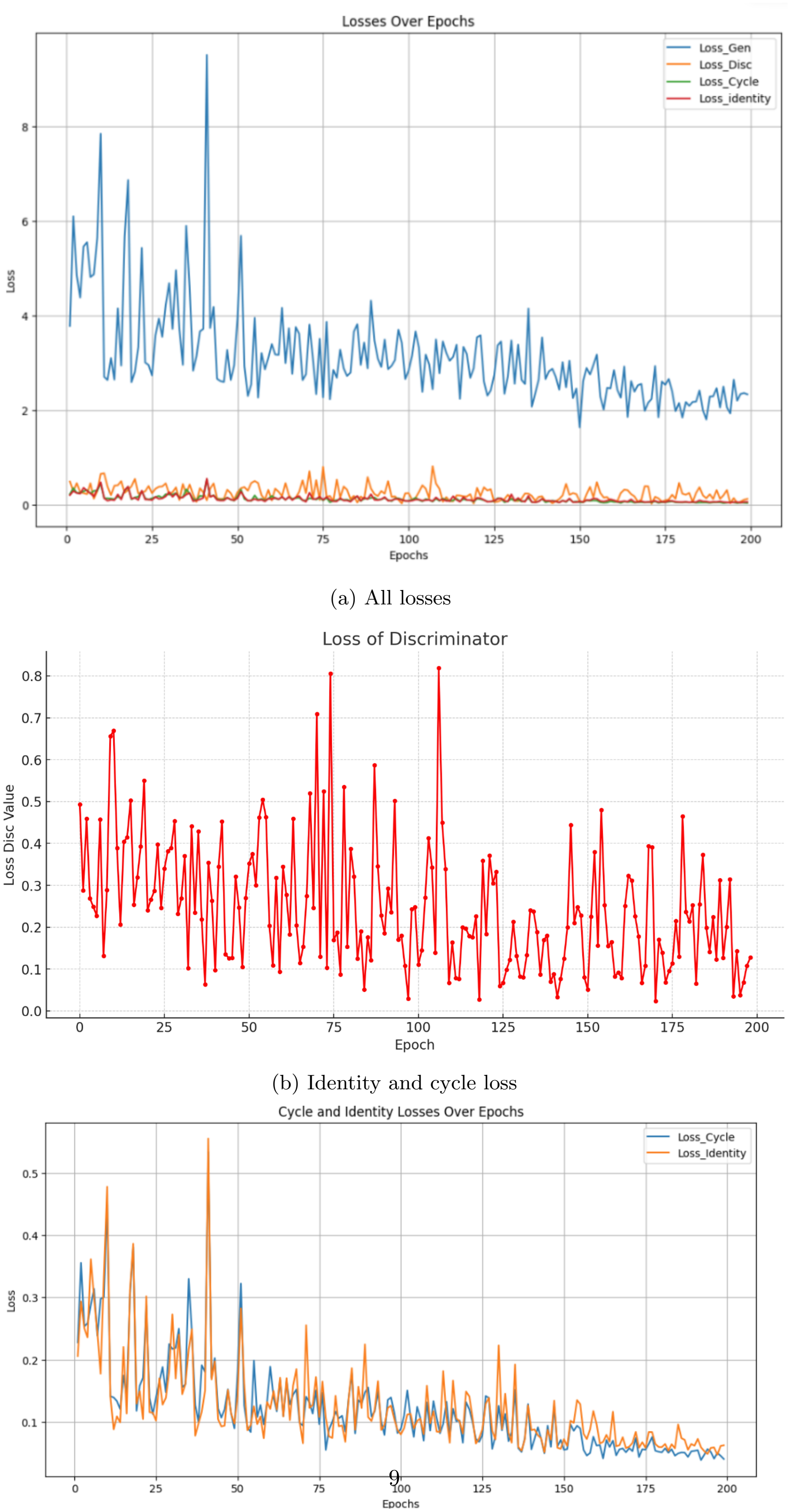
[10] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In CVPR, 2017

[11] S. I. Serengil and A. Ozpinar, HyperExtended LightFace: A Facial Attribute Analysis Framework, 2021

[12] M. Mathieu, C. Couprie, and Y. LeCun. Deep multiscale video prediction beyond mean square error. In ICLR, 2016. 2

[13] Martin Arjovsky, Soumith Chintala and Léon Bottou, Wasserstein GAN, arXiv:1701.07875, 2017.

[14] A. Horé and D. Ziou, Image Quality Metrics: PSNR vs. SSIM, 2010.

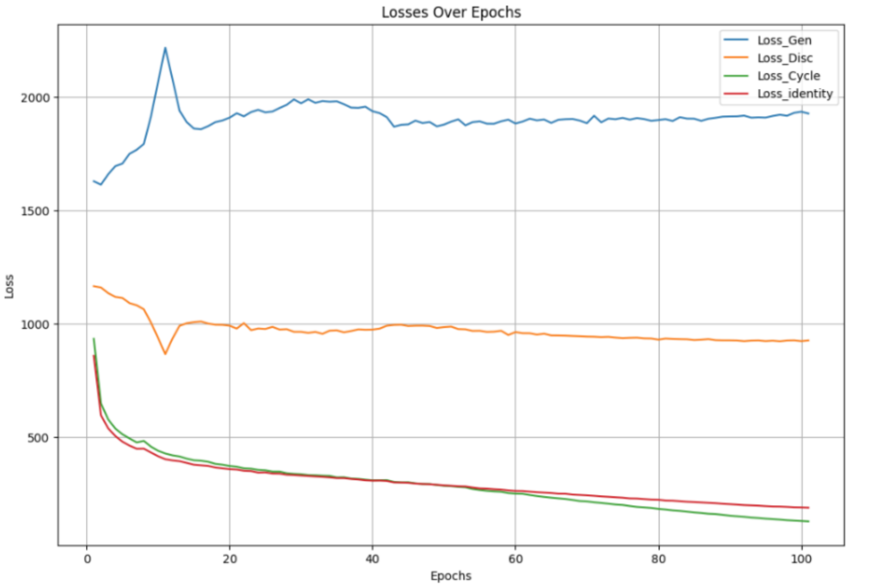


(a) All losses

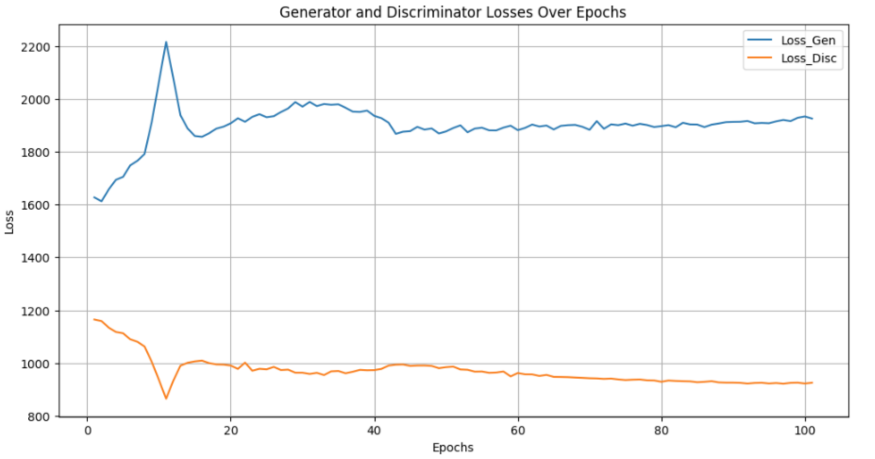
(b) Identity and cycle loss

(c) Identity and cycle loss

Figure 3: Losses per epoch for Cycle GAN



(a) All losses



(b) Identity and cycle loss

Figure 4: Identity and cycle losses per epoch for Cycle GAN