Final Project Report

CS424 – Generative AI For Vision
AY 2024/2025 Term 2

Koo Jing Xuan, Xavier	xavier.koo.2021@scis.smu.edu.sg
Aaron Kwah Boon Hou	aaron.kwah.2021@scis.smu.edu.sg
Ho Zi Kai, Maurice	maurice.ho.2021@scis.smu.edu.sg
Hsieh Pei Ling	plhsieh.2022@scis.smu.edu.sg

CycleGAN for Image-to-Image Translation: Implementation, Optimization & Novel Applications

Koo Jing Xuan, Xavier | Aaron Kwah Boon Hou | Ho Zi Kai, Maurice | Hsieh Pei Ling

Abstract. Image-to-image translation using Generative Adversarial Networks (GANs) has witnessed rapid advancements, with CycleGAN emerging as a key framework for unsupervised domain transfer. In this project, we explore and enhance CycleGAN through two distinct tasks: (1) photorealistic face-to-cartoon translation and (2) young-to-old facial age progression. To address limitations in vanilla CycleGAN, such as poor structural preservation, training instability, and lack of global context, we introduce architectural innovations, including dual self-attention modules, perceptual VGG-based loss functions, edge-aware enhancements, and spectral normalization in discriminators. Additionally, we adopt training strategies like warm-up cosine annealing, AMP for mixed-precision training, and gradient clipping for improved stability. Quantitative and qualitative evaluations demonstrate significant improvements across both tasks. For Task 1, our enhanced model achieved lower GMS scores—indicating higher visual realism—compared to the vanilla CycleGAN, particularly in the Real-to-Cartoon direction. For Task 2, our ablation studies show that the integration of VGG loss and self-attention yields the most visually convincing and age-accurate transformations, with an average age gap reduction from 20.73 to 7.54. Our experiments validate that these enhancements not only improve visual fidelity but also align closely with semantic expectations of aging and style translation, pushing the boundaries of what unsupervised translation frameworks can achieve in facial domains.

1. Introduction

Image-to-image translation has become a significant domain within generative adversarial networks (GANs). CycleGAN, an unsupervised framework within this domain, leverages cycle-consistency losses to ensure credible transformations by learning reversible mappings between domains. This project comprises two tasks aimed at exploring and enhancing CycleGAN's capabilities:

- **Task 1: Face-to-Cartoon Translation:** Translating photorealistic facial images into cartoon-style images and vice versa.

- **Task 2: Young-to-Old Age Progression:** Transformation of youthful faces into realistically aged appearances and vice versa. We utilize the AgeDB dataset as a novel application of CycleGAN. It is a comprehensive and diverse dataset encompassing broad age distributions and varied conditions.

2. Photorealistic Face-to-Cartoon Translation CycleGAN

2.1. Model Architecture and Improvements

Our enhanced CycleGAN builds upon the original CycleGAN framework with several architectural improvements. The vanilla CycleGAN often struggles with preserving detailed

structural information during translation, particularly for the facial features, when converting between photorealistic and cartoon domains. In addition, the standard implementations can often exhibit training instability, which leads to suboptimal convergence. Our model addresses these challenges through several targeted enhancements. With these new enhancements, the size of our entire model increased to approximately 25.01 million parameters.

2.1.1. Generator Design

The generator follows an encoder-decoder structure with a bottleneck of residual blocks but introduces two key innovations: 1. Self-attention modules 2. Edge-aware processing. Our improved generator design is illustrated in Fig 1. Each residual block contains two 3×3 convolutional layers with instance normalization and a skip connection, which is inspired by the design in He et al. (2016).

We incorporate self-attention modules at two strategic locations in the generator: after downsampling and after the residual blocks. This helps to address the limitations of convolutional networks, which are unable to model long-range dependencies due to their local nature. (Wang et al., 2018). Self-attention modules are memory-intensive, and as a result, we implemented a channel reduction factor of 16 and scaled dot-product attention for numerical stability (Vaswani et al., 2017). This allows our model to selectively focus on relevant facial features during the translation process, preserving structural integrity.

We subsequently added an improved edge detection module for the generator that translates real images to cartoon styles (G_AB) to enhance cartoon-like features. The edge enhancement is applied during the forward pass with a controlled contribution weight of 15%. This edge-aware processing specifically addresses the challenge of creating defined outlines characteristic of cartoon art styles, which vanilla CycleGAN often struggles to produce.

It is essential to highlight that edge awareness is exclusively implemented in the **G_AB generator (real-to-cartoon)** and **not in the G_BA generator (cartoon-to-real)**, as sharp edges are a defining characteristic of cartoon styles and are not representative of photorealistic images.

2.1.2. Discriminator Design

We utilized a PatchGAN discriminator (Isola et al., 2017) and enhanced it with spectral normalization to boost training stability, as shown in Fig 2.

Spectral normalisation was applied to all convolutional layers of the discriminator, ensuring the network's Lipschitz constant is constrained by normalizing the spectral norm of each weight matrix (Miyato et al., 2018). This approach stabilizes GAN training by mitigating issues such as exploding gradients and mode collapse, which are typical in conventional CycleGAN training.

The discriminator processes patches instead of the whole image, offering two key benefits: It requires fewer parameters (about 0.5 million) and delivers a more detailed supervision signal by determining whether each image patch is real or fake.

2.1.3. Loss Function

Our training objective consists of three main components, with specific improvements to the loss-weighting strategy.

- **Adversarial Loss (MSE)**. For improved stability, we employ Mean Squared Error (MSE) loss rather than the standard Binary Cross-Entropy (BCE) loss.
- **Cycle Consistency Loss (L1)**. We use it to ensure that translating an image to the target domain and back preserves the original content.
- **Identity Loss (L1).** This encourages the preservation of color composition when an image from the target domain is used as input.

A key innovation in our approach is the use of direction-specific loss weighting. Rather than applying uniform weights across all loss components, we employ a specialized weighting strategy for each translation direction.

2.1.4. Training Methodology

We implement several optimizations to improve training stability and convergence. We used **Adam Optimizer with Weight Decay (β1=0.5, β2=0.999, and weight decay of 1e-5**) for all three types of optimizers (i.e., Generators, Discriminator A, and Discriminator B). The addition of weight decay helps prevent overfitting and improves generalization by penalizing large weights.

We also implemented a **Warm-up Cosine Annealing Scheduler**. This scheduler gradually increases the learning rate during the first few epochs (warm-up phase) and then decreases it following a modified cosine curve. The warm-up phase helps stabilize early training, while the cosine annealing helps fine-tune the model in later epochs, addressing the common issue of learning rate sensitivity in GAN training.

During training, we realized that the loss started to become NaN as a result of exploding gradients. We then applied **Gradient Clipping** with a maximum norm of 1.0 to prevent exploding gradients. In addition, we implement **checks for NaN values** in loss calculations and skip backward passes when NaNs are detected so that we can maintain training stability even in challenging situations.

2.2. Experimental Results

2.2.1. Dataset and Implementation

We used a dataset consisting of photorealistic face images and corresponding cartoon-style images. The training set included 3,200 images from each domain, with an additional 800 images for validation. No data augmentation was done during the dataset's training because during our experimentations with minor data augmentation, we obtained a worse-off GMS of 4.05286. We can see the examples of our results, and sample images can be seen in Fig 3.

For Implementation details, we experimented with batch sizes ranging from 2 to 6, ultimately settling on 4. We set the learning rate to 0.0001 and trained for 150 epochs at an image resolution of 256×256.

2.2.2. Quantitative Evaluation

We evaluated our model using three metrics:

1. **Inception Score (IS):** Measures both quality and diversity of generated images
2. **Fréchet Inception Distance (FID):** Measures similarity between real and generated image distributions
3. **Geometric Mean Score (GMS):** Calculated as √(FID/IS), combining both metrics

Our best model achieved these results at Epoch 128, and we can compare them to the vanilla CycleGAN, which had a similar training duration and number of images.

Translation Direction	Enhanced Model			Base Vanilla CycleGAN		
	IS	FID	GMS	IS	FID	GMS
Real to Cartoon	2.15	34.22	3.99	4.04	163.32	6.35
Cartoon to Real	3.67	43.31	3.43	2.47	152.62	7.86
Average	-	-	**3.71**	**-**	**-**	**7.10**

Table 1. *Task 1 Enhanced and Vanilla CycleGAN Model Results*

2.2.3. Qualitative Results

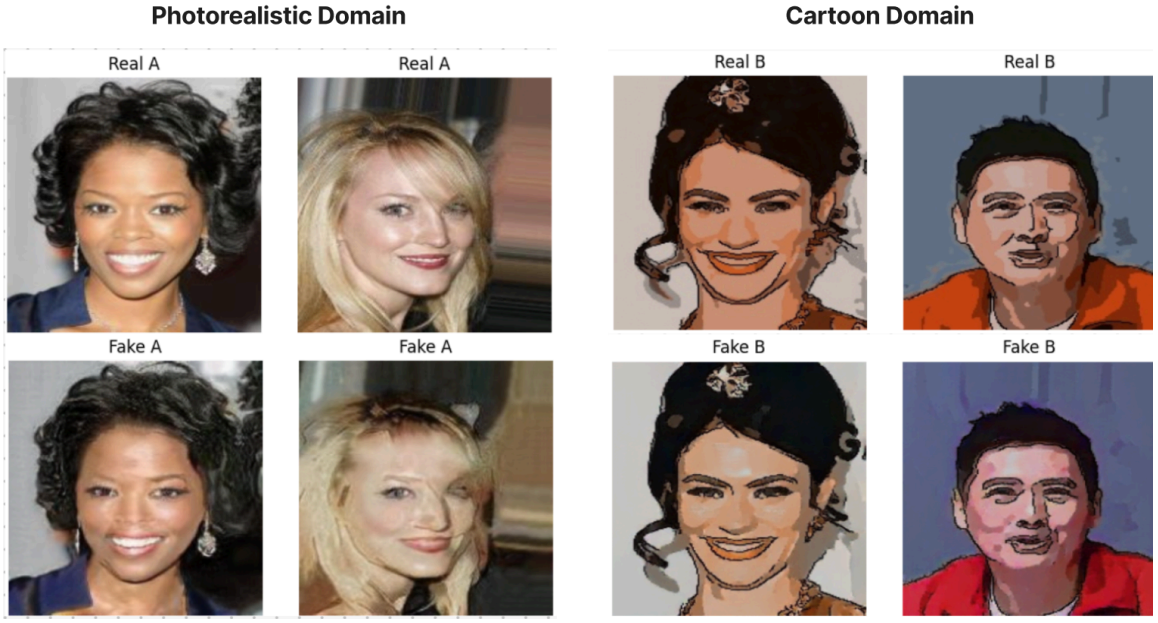


Figure 4+5. *Task 1 Photorealistic/ Cartoon Domain Comparison*

The original photorealistic images (Real A) and their cartoon-translated versions (Fake A) exhibit striking similarities. In both cases, the facial structures, expressions, and defining features remain largely unchanged. The woman with dark curly hair retains her smile and facial proportions, while the blonde's unique characteristics and smile are preserved. Nevertheless, a closer look uncovers subtle variations: the Fake A images display slightly smoother skin textures along with minor changes in lighting and contrast, as well as modifications in certain facial features, such as the nose of the dark-haired woman and the right eye of the blonde.

In the cartoon domain, the original illustrations (Real B) and their generated counterparts (Fake B) also show good correspondence. The cartoon woman with the smile maintains her distinctive outlined features and facial expression in the Fake B version. In contrast, the cartoon subject with the orange/red shirt preserves the basic structure but shows more notable color shifts with increased reddish tones. The Fake B images maintain the characteristic outlined, simplified style of cartoons, though they sometimes introduce slight variations in coloration and outline intensity.

2.2.4. Ablation Studies

We trained four variants of our model with identical hyperparameters, only varying the attention configuration:

1. **No Attention**: Baseline model without any attention mechanisms
2. **Single Attention (G_AB only)**: Attention at the bottleneck only in the Real→Comic generator
3. **Single Attention (Both)**: Single attention at the bottleneck in both generators
4. **Dual Attention (Both)**: Attention after downsampling and at the bottleneck in both generators

All models were trained for 150 epochs using 3200 training images and validated on 800 images. We used the same learning rate schedule (warmup cosine annealing), edge enhancement parameters, and loss weights across all variants.

Model Configuration	**Real→Comic**			**Comic→Real**			**Average**
	IS	**FID**	**GMS**	**IS**	**FID**	**GMS**	**GMS**
No attention	2.24	53.31	4.87	3.65	49.23	3.67	4.27
Single Attention (G_AB only)	2.98	88.31	5.44	3.72	51.85	3.73	4.59
Single Attention (Both)	2.17	39.48	4.17	3.48	43.11	3.52	3.85
Dual Attention (Both)	2.15	34.22	3.99	3.67	43.31	3.43	3.71

Table 3. *Task 1 Attention Mechanism Summary*

The Real→Comic transformation (G_AB) showed significant improvement from our dual attention mechanism implementation. The GMS improved from 4.87 (no attention) to 3.99 (dual attention)—an 18% improvement. This direction benefited most from the combination of dual attention and increased model capacity, particularly in the FID score, which improved dramatically from 53.31 to 34.22.

The Comic→Real transformation (G_BA) also demonstrated improvements, with GMS increasing from 3.67 to 3.43 (a 6.5% improvement). Interestingly, the single attention in

both generator configurations already provided significant benefits for this direction, suggesting that Comic→Real translation requires less complex attention mechanisms.

The transition from no attention to dual attention indicates a clear trend of improvement. However, it is important to note that the single attention in the "G_AB only" configuration actually performed worse than the baseline. This might suggest to us that asymmetric attention can create an imbalance in the CycleGAN's cycle consistency, underscoring the significance of balanced architectural changes in both generators. The Real→Comic direction benefits the most dramatically, likely due to the complexity of determining which facial features should be simplified versus emphasized in stylization.

2.3. Limitations & Possible Improvements

Despite achieving impressive results with our attention-enhanced CycleGAN model, several limitations remain. The model occasionally struggles with extreme facial poses and expressions, particularly when translating to comic style, where it can produce inconsistent stylization. Memory constraints limited our ability to implement more extensive attention mechanisms or deeper networks that might further improve performance. Also, even after using cycle-consistency and identity loss, the CycleGAN model may still struggle to retain person-specific features during transformation. This is due to the unpaired training setup, where there's no explicit supervision guiding the model to preserve identity.

Future improvements could include applying adaptive instance normalization for better style control and addressing the issue of inconsistent stylization when translating extreme facial expressions to comic style. For the challenge of retaining person-specific features during transformation, potential solutions include incorporating face landmark detection to guide attention mechanisms toward key facial features, using multi-scale discriminators to handle details at different resolutions, and adopting a progressive training approach that gradually increases resolution to improve fine details and overall coherence in both transformation directions.

3. Young-to-Old Age Progression CycleGAN

3.1. Model Architecture and Improvements

3.1.1. Generator Design

The generator adopts a ResNet-like structure and contains approximately 29.11 million parameters. It is designed to process RGB images (3 channels for both input and output), ensuring compatibility with standard image formats. The architecture begins with reflection padding to minimize edge artifacts, followed by an initial convolutional layer. Downsampling layers reduce the spatial dimensions of the image by a factor of two, resulting in a progressive feature resolution pathway of $64 \rightarrow 128 \rightarrow 256$.

1. **Feature Extraction and Transformation**: This stage applies convolutional operations to extract both local and global features from the input image. Through progressive downsampling, the image is encoded into high-dimensional feature representations.

2. **Residual Processing with Self-Attention**: The encoded features are passed through 9 residual blocks to promote effective feature reuse and preserve spatial structure. A self-attention mechanism is integrated within this stage to model long-range dependencies, allowing the network to better understand complex patterns and relationships across the image.

3. **Reconstruction**: The final stage resamples the feature maps (256 → 128 → 64) to restore the original image resolution. A concluding convolutional layer with Tanh activation maps the output to the range [-1, 1], which is standard for normalized image generation tasks.

3.1.2. Discriminator Design

The discriminator is built on the PatchGAN architecture, which evaluates the realism of overlapping image patches rather than the entire image at once. This localized assessment helps the model focus on fine-grained texture and detail. The architecture consists of five convolutional layers, each followed by LeakyReLU activations to maintain non-linearity while allowing minor gradient flow even when neurons are not active.

Incorporating **spectral normalization** on the weight matrices of each layer helps constrain the Lipschitz constant, leading to more stable training dynamics and better convergence. The PatchGAN structure further enhances texture fidelity by encouraging the generator to produce realistic local details rather than relying solely on global coherence.

3.1.3. Loss Function

The model employs a combination of multiple loss functions to guide the training process comprehensively:

1. **Adversarial Loss** (MSE): Encourages the generator to produce images indistinguishable from real ones by fooling the discriminator.
2. **Cycle-Consistency Loss** (L1): Ensures that translating an image to the target domain and back yields the original image, preserving structural consistency.
3. **Identity Loss** (L1): Encourages the generator to preserve color composition and identity when an image from the target domain is fed into the generator.
4. **Perceptual Loss**: Computed using a pre-trained VGG19 network, it extracts deep features from layers 5, 10, and 19 to compare high-level representations rather than raw pixels.
5. **Feature Matching Loss**: This loss encourages the generator to match the feature maps produced by the discriminator for real and generated images, improving the discriminator's ability to capture subtle patterns, reducing mode collapse, and leading to better image quality.

Using a combination of pixel-wise (L1) and feature-wise (perceptual) losses leads to visually sharper and semantically faithful image translations. The perceptual loss, in particular, helps maintain texture and content realism by capturing high-level abstractions, while adversarial and cycle-consistency losses ensure the model remains grounded in realism and content preservation. Identity loss further refines color and style consistency, especially in tasks like style transfer or domain translation.

3.1.4. Training Methodology

C**heckpoint loading and saving functionality** was introduced to allow for training continuity across sessions, ensuring efficient resource use and enabling the model to resume training without starting from scratch. To improve computational efficiency and reduce memory usage, **mixed precision training** was implemented using PyTorch's Automatic Mixed Precision (AMP). This method leverages lower precision computations where possible while maintaining numerical stability. A **gradient scalar** was also employed to manage the scaling of gradients, ensuring stable and efficient training.

The inclusion of advanced loss functions further enhanced the model's ability to **generate high-quality images**. A **perceptual loss** based on VGG feature extraction was added to improve the visual fidelity of generated images, making them more visually realistic. Additionally, a **feature matching loss** was introduced to align the discriminator's representations of real and generated images, helping to improve the model's discrimination capabilities.

I**mage pooling** was incorporated during discriminator training to enhance GAN stability even further. This technique allows the discriminator to use historical samples, improving the model's performance by preventing mode collapse, a common problem in GAN training. In addition, **learning rate scheduling** was implemented, allowing the optimizer's parameters to be adaptively adjusted throughout training to improve convergence and training stability.

Finally, **validation monitoring** was added to track potential overfitting, allowing for better performance tracking and adjustment. **Logging** was also implemented, providing detailed insights into various loss components during training, ensuring that the training process can be easily monitored and evaluated for optimization.

3.1.5. Evaluation Metrics

The evaluation framework has been implemented by importing InsightFace age estimation. This trained model detects faces in both real and generated images and estimates their ages. Then, we calculate the absolute differences between the average ages of real and generated images in both directions (young-to-old and old-to-young). The evaluation produces three key metrics: old age gap, young age gap, and total average gap, providing a more relevant assessment of the age transformation quality. Additionally, the image has to be identified as a face to be considered in the calculation matrix. This ensures the evaluation method is aligned with the task that we are achieving.

3.2. Experimental Results

3.2.1. Dataset and Implementation

We employed the AgeDB dataset, a novel application of CycleGAN not mentioned in the original paper, introducing a previously unexplored dataset for CycleGAN-based transformations. It is a comprehensive age estimation benchmark containing over 16,000 face images of 568 distinct subjects, with ages ranging from 1 to 101 years. The images include age annotations and identity labels, with variations in pose, expression, and lighting

conditions. The dataset's broad age distribution and balanced gender representation make it particularly valuable for developing robust, fair, and generalizable models.

For our CycleGAN age transformation task, we filtered and sorted the AgeDB dataset to enhance its relevance. It was divided into two groups based on age ranges: the Young group consists of individuals aged 18 to 32, while the Old group includes individuals aged 68 to 91. To maintain the quality and relevance of the data for the age transformation task, most of the grey images were removed. The final prepared dataset contained 2000 images in each age domain, with 1600 images allocated for training and 400 for testing.

The data loading pipeline has been enhanced to improve model training through the incorporation of more robust data augmentation techniques. Specifically, the transformations applied to the training and validation datasets have been separated to allow for more aggressive augmentations on the training data while maintaining the integrity of the validation set.

For the training dataset, the following augmentations were introduced:

- **RandomHorizontalFlip()**: This transformation generates mirror images of the training samples, effectively doubling the variety of sample orientations and providing the model with a broader range of spatial variations.

- **ColorJitter**: This transformation was applied with carefully calibrated parameters—brightness (0.2), contrast (0.2), saturation (0.2), and hue (0.1). It introduces controlled variations in lighting conditions and color properties.

These augmentations help the model to better generalize by exposing it to a broader range of image variations during the training phase. In contrast, the validation set has been kept simple, with only essential preprocessing steps applied, including resizing, tensor conversion, and normalization. This ensures fair and unbiased evaluation, allowing the model to be assessed on a consistent set of conditions.

Additionally, the batch size was reduced from 5 to 4 in order to maintain training stability, given the increased complexity resulting from the augmented images. This adjustment aims to improve model performance and prevent potential issues related to memory or convergence during training.

3.2.2. Quantitative Evaluation

We evaluated three models as we improve from the previous model starting from the Baseline then adding VGG Loss, lastly adding both VGG and Self-Attention—on training loss, identity preservation, and age estimation performance (See Fig. 6 and Fig.7).

Model	Avg Generator Loss	Avg Identity Loss
Baseline	2.59	0.095
VGG Loss	5.76	0.108
VGG Loss + Self-Attn	5.89	0.117

Table 4. *Task 2 Model Evaluation*

The Baseline model achieved the lowest average generator loss at 2.59 and the lowest identity loss at 0.095, indicating more stable training and better identity retention numerically. In comparison, adding the VGG Loss into the baseline exhibited a higher average generator loss of 5.76 and an identity loss of 0.108, while the VGG Loss + Self-Attention model recorded the highest losses at 5.89 and 0.117, respectively. Despite these higher quantitative losses, the VGG Loss + Self-Attention model produced significantly better qualitative results, suggesting that the inclusion of self-attention enhances the perceptual quality of the generated images even if it slightly compromises training loss metrics.

3.2.3. Qualitative Results

The Baseline model produced relatively blurry outputs with weak transformations, where aging or youth cues were barely noticeable. This suggests a limited capacity to capture the nuanced visual changes associated with age progression or regression. In contrast, the addition of VGG Loss into the model showed improvements in output clarity, with better skin texture and more faithful preservation of facial features, resulting in more coherent and visually appealing transformations. Lastly, adding self-attention on top of the VGG Loss model further enhanced the results by capturing more global context, such as the appearance of wrinkles and hair graying, leading to more realistic and convincing age transformations. These outputs more accurately represented visibly older or younger versions of the input faces, demonstrating the effectiveness of incorporating self-attention in refining age-related visual cues. See Fig.8 for comparison.

3.2.4. Ablation Studies

Comparing with Real Old Age 69.47 and 35.63 with model variant:

Model Variant	VGG Loss	Self-Attn	Visual Realism	Avg Fake Young Age	Avg Fake Old Age	Avg Age Gap
Baseline	X	X	Low	56.75	49.13	20.73
VGG Loss	✔️	X	Moderate	55.35	54.45	17.37
VGG Loss + Self-Attn	✔️	✔️	Best	45.16	63.92	7.54

Innovation	Baseline	After	Effect
+ VGG Loss	Flat textures, low realism	Sharper textures, better aging cues	Enforces perceptual realism
+ Self-Attention	Mostly local-only features	Global structure and symmetry	Captures long-range dependencies

Table 5. *Task 2 Ablation Comparison*

The final VGG + Self-Attention model generated the best balance between realism and transformation accuracy:

- The evaluation method confirms that the aging direction is correct and realistic
- It is the most convincing and detailed visually compared to the 3 models

3.2.5. Comparison Experiments & Error Analysis

To validate the effectiveness of our CycleGAN model for young-to-old and old-to-young face translation, we performed a series of controlled experiments that facilitate comprehensive comparisons. These experiments are designed to benchmark our model against alternative models, helping us better understand the reasons behind any observed performance gaps. We conducted two comparative experiments:

- **Basic CycleGAN on Same Dataset**

For this experiment, we trained a basic vanilla CycleGAN architecture directly on the same AgeDB dataset. This serves as a baseline to show and compare the limitations of a basic implementation. We employed a highly simplified model consisting of a generator with only two convolutional layers—an initial 7×7 convolution with ReLU activation, followed by another 7×7 convolution and a Tanh output. This minimal design is meant to demonstrate a baseline rather than a production-ready architecture. For the discriminator, we applied a single downsampling convolution layer with a LeakyReLU, followed by one more convolution layer. Again, this is an intentionally lightweight design. Because both the generator and discriminator have so few layers, the total parameter count is notably low (≈0.04 million parameters). This limited capacity directly impacted the quality of the age-translated outputs, as shown in Fig. 9.

The generated faces frequently have strong artifacts or "negative" colorations, as seen in the figure (top row: Real Young, second row: Fake Old, third row: Real Old, and fourth row: Fake Young). The "aged" or "rejuvenated" features do not accurately depict actual age progression or regression, and facial details are poorly preserved. These artifacts highlight the two-layer generator and discriminator's limited representational power. Because it has so few layers, the model produces low-fidelity transformations rather than learning high-level aging cues like wrinkles, changes in hair color, and variations in facial shape.

- **Pretrained CycleGAN on UTKface, Fine-Tuned on the Same Dataset**

In this experiment, we started by pre-training our best CycleGAN model on the UTKface dataset to leverage its larger and more varied data. Subsequently, we used the pre-trained model and fine-tuned it on the AgeDB dataset as the final domain for evaluation. The experiment goal was to determine if transfer learning from UTKFace could enhance the realism and accuracy of age translation on Agedb.

From the results shown in Fig. 10, after training on the UTKface dataset, facial structures are somewhat retained, and the generated images still exhibit mild color distortions and uneven texture details. For the final outputs from the AgeDB transfer learning training, the model captures some aging cues (e.g., wrinkles, skin tone changes) better than the vanilla approach. However, artifacts and blurry regions remain noticeable. Employing age estimation once again on the generated images yielded the following results:

	Average Age	Age Gap
Real Old (Agedb)	69.47	18.15
Fake Old (Generated)	51.32	
Real Young (Agedb)	35.63	18.59
Fake Young (Generated)	54.22	
Average		**18.37**

Table 6. *Task 2 Age Comparisons*

When these two gaps are combined, the average age gap is 18.37, which is better than a pure CycleGAN but still shows that the model has trouble creating features that are plausibly old or young and is not as good as our best-performing model. Interestingly, compared to the actual youthful domain, the Fake youthful photos are still significantly older.

3.3. Limitations & Possible Improvements

Limited Dataset Diversity

The AgeDB subset used (ages 18–32 and 68–91) covers only two narrow age bands and omits intermediate stages, reducing the model's ability to learn gradual, continuous aging patterns. We could incorporate additional age brackets (e.g., 33–50, 51–67) or use a continuous age-label conditioning scheme to better capture the full spectrum of facial aging. We could also augment with other age-annotated datasets (e.g., MORPH, CACD) to increase subject diversity and demographic balance.

Face Detection Failures

The evaluation pipeline relies on off-the-shelf face detectors, which occasionally fail on extreme poses, occlusions, or poor lighting. This leads to dropped samples (nocountold, nocountyoung), skewing age-gap metrics and limiting robustness. Further improvements to mitigate this could be integrating a more robust, multi-stage detection framework (e.g., MTCNN followed by RetinaFace) or training a custom detector fine-tuned on AgeDB to reduce false negatives.

4. Conclusion

Our enhanced CycleGAN significantly improves facial image translation by incorporating self-attention and perceptual loss. These changes lead to sharper, more realistic outputs in the age transformation task. Ablation studies confirm the value of each component, especially in reducing age gap errors and improving visual consistency. While challenges remain—such as limited dataset diversity and face detection issues—our approach offers a strong foundation for future improvements in unsupervised generative modeling.

Appendix

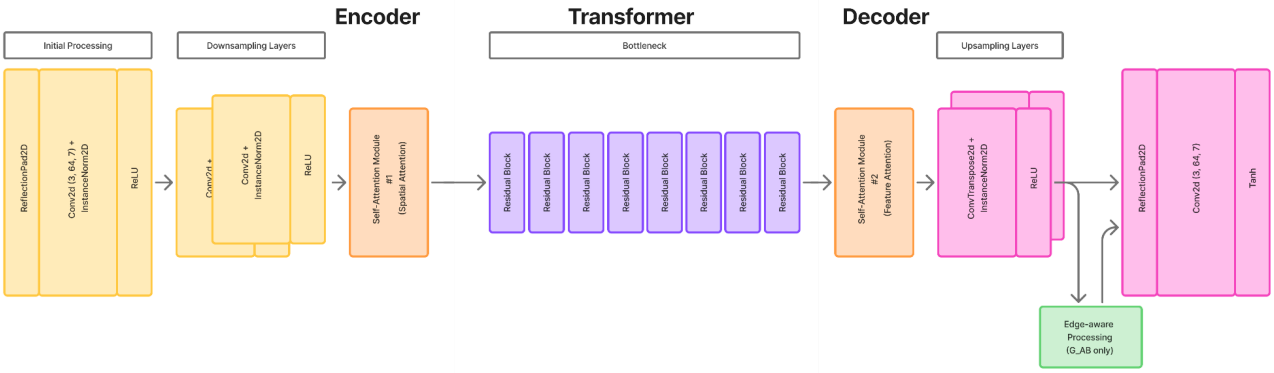


Figure 1. Task 1 Generator Architecture

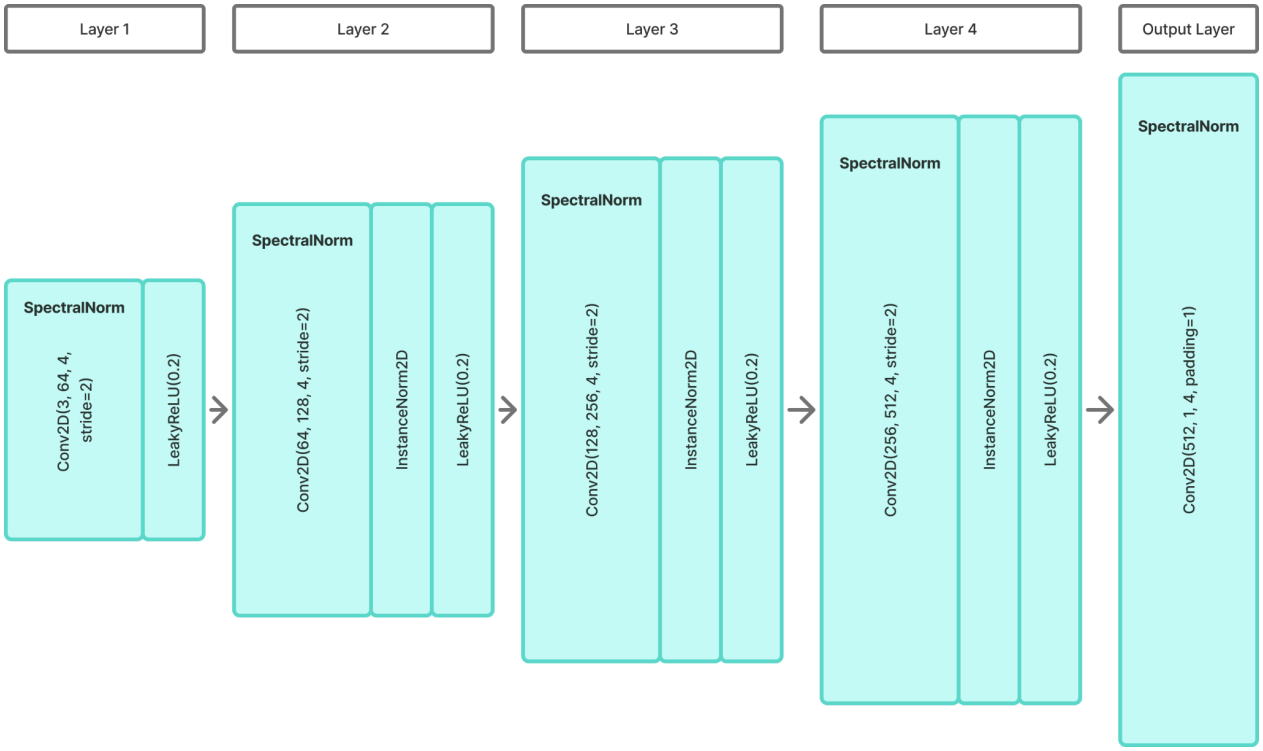


Figure 2. Task 1 Discriminator Architecture

Figure 3. Task 1 Data Augmentation Generated Image

Figure 4. Task 1 Real Realistic to Fake Realistic Comparison

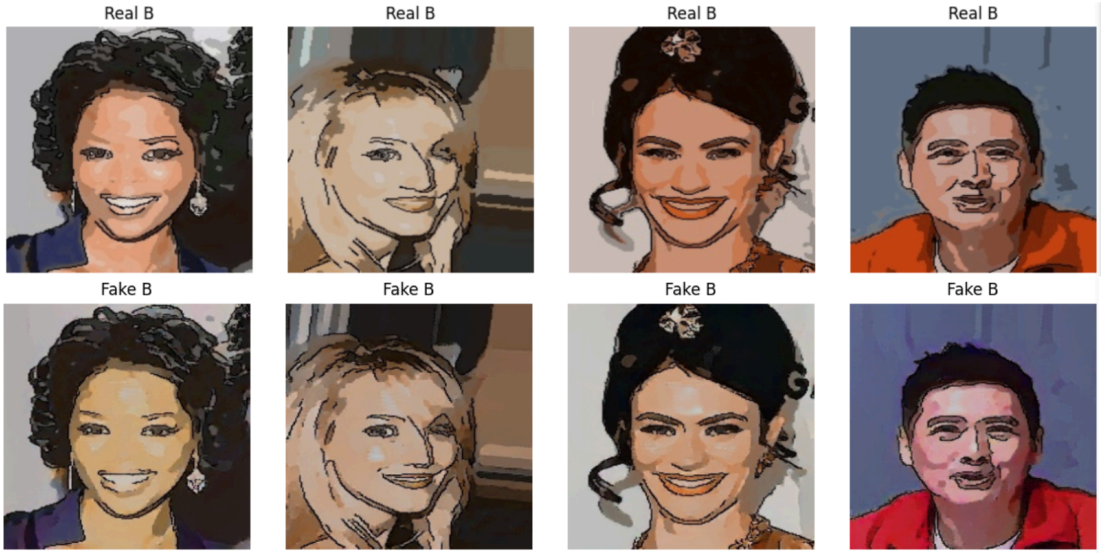


Figure 5. Task 1 Real Comic to Fake Comic Comparison

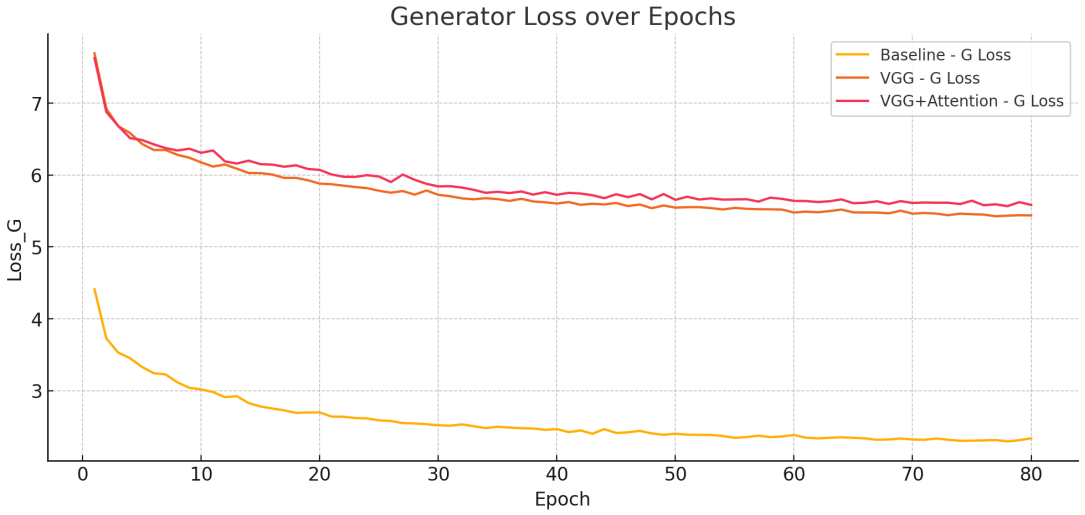


Figure 6. Generator Loss over Epochs

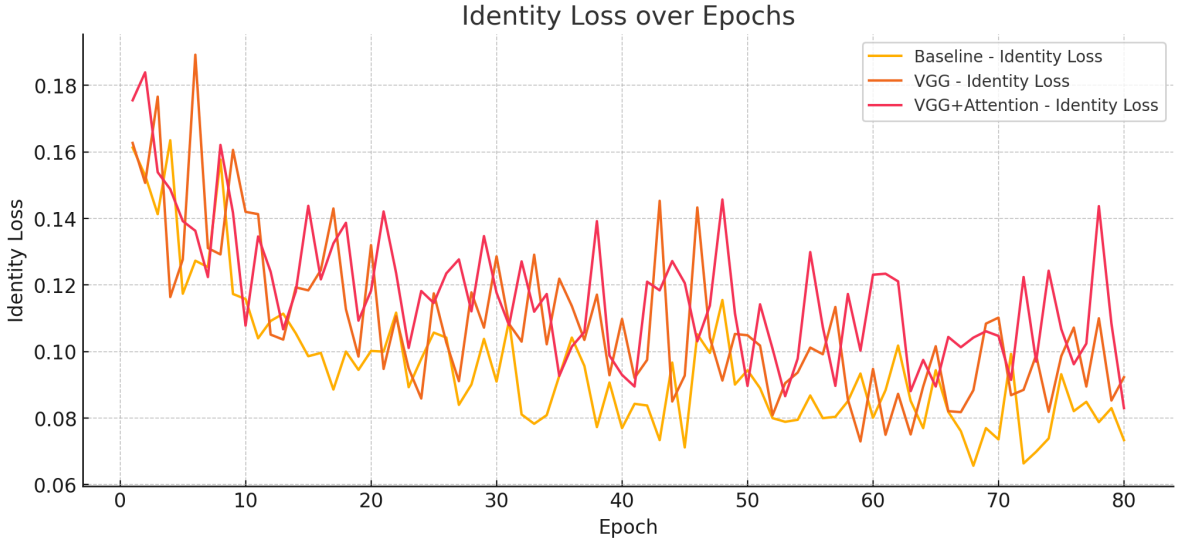


Figure 7. Identity Loss over Epochs

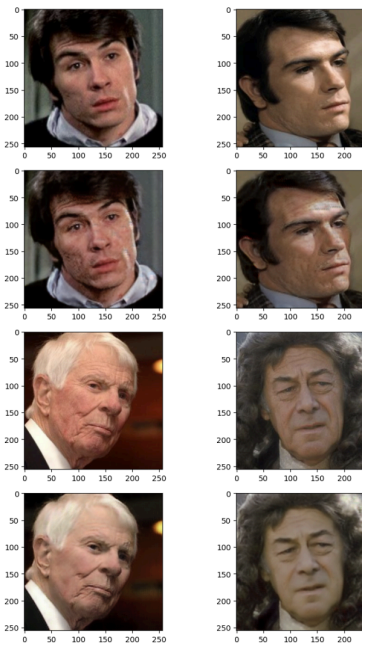
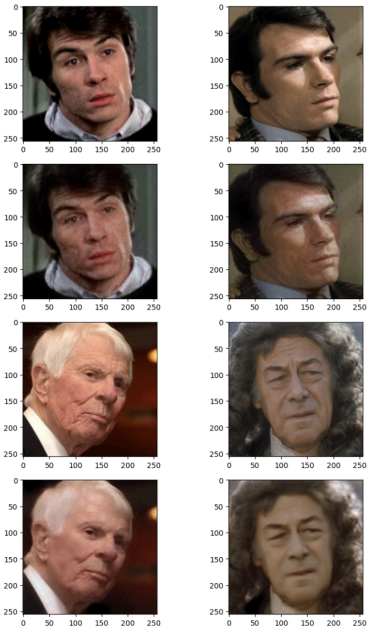
	Baseline	VGG Loss	VGG Loss + Self-Att
Real Young Fake Old Real Old Fake Young			
Description	Blurry outputs, weak transformation. Aging/youth cues are barely visible.	Improved skin texture, facial feature fidelity.	More global context captured (e.g., wrinkles + hair graying). More realistic transformations — visibly older/younger versions.

Figure 8. Comparison of Qualitative Results for Task 2

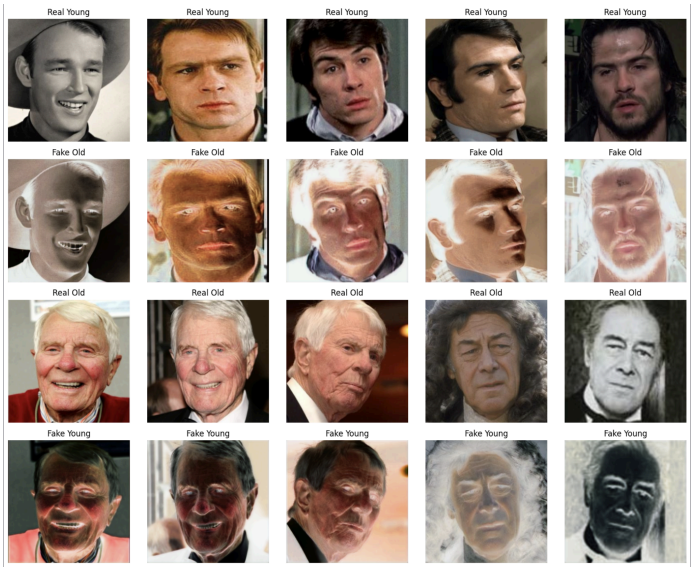


***Figure 9.* Basic CycleGAN on Same Dataset Output**

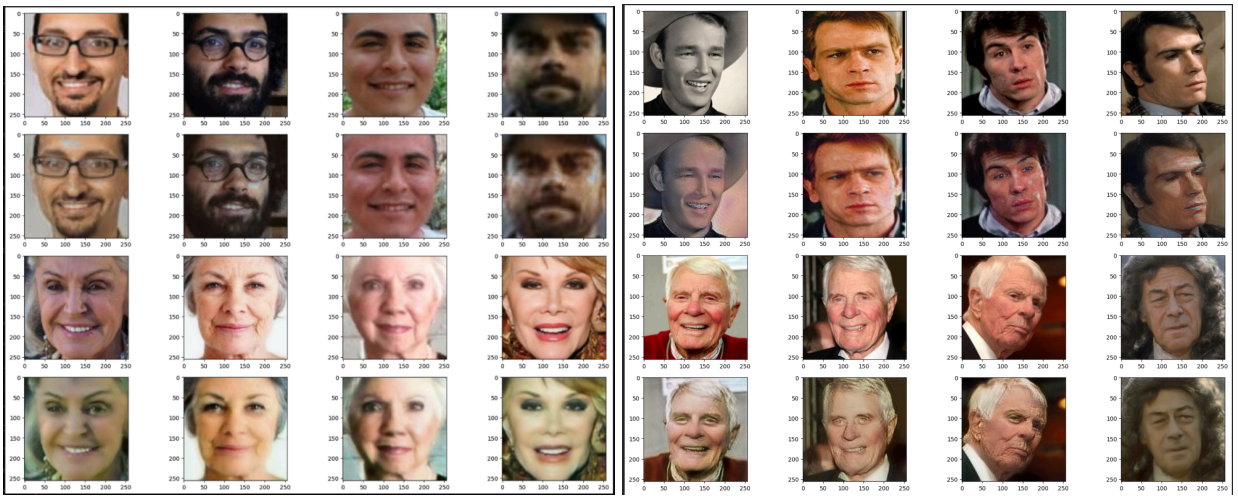


Figure 10. UTKface Training & AgeDB + Transfer Learning Training Output

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1125-1134).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2223-2232).

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In International Conference on Learning Representations.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794-7803).

Papers with Code - AgeDB Dataset (no date) Dataset | Papers With Code. Available at: https://paperswithcode.com/dataset/agedb (Accessed: 7 April 2025).