Generative AI-Assisted Face Synthesis for Forensic and Criminal Investigations

Shraddha Pandey*[1] and Ishaan Shankar Bansal[2] and Noel Jeygar Robert[3]
[1]*School of Computer Science and Engineering Vellore Institute of Technology,* Chennai, India
shraddha.pandey2025@gmail.com
[2]*School of Computer Science and Engineering Vellore Institute of Technology,* Chennai, India
ishaan.s.bansal@gmail.com
[3]*Assistant Professor Senior Grade 1, School of Computer Science and Engineering Vellore Institute of Technology* Chennai, India
noeljeygar.robert@vit.ac.in

Abstract—Forensic investigations frequently rely on facial sketches to identify individuals involved in criminal activities. However, since the human face is so diverse and variable, it is tough to predict the person from a forensic sketch expert's sketch as they often lack realism and can be influenced by eyewitness biases. Recent advancements in generative AI, particularly Gen- erative Adversarial Networks (GANs) have revolutionized image synthesis, enabling the generation of highly realistic facial images from sketches. Hence, we aim to leverage this technology and make this process more efficient and foolproof using GANs. In this paper, we propose advancements to existing technology by using a combination of PatchGAN discriminator and a U-Net generator.
Index Terms—Face Recognition, Face Synthesis, Generative AI, Generative Adversarial Networks, Image-to-Image Translation

INTRODUCTION

In investigative fields, facial reconstruction from forensic sketches is an essential tool that helps law enforcement agencies to identify potential suspects and missing persons. Traditionally, it is a manual and extensive process. First, forensic sketch artists manually convert witness descriptions into sketches, and then the sketches are compared against photographic databases. However, this method is highly sub- jective and prone to human errors, which may lead to un- reliable identifications. In recent times, with the emergence of Generative Artificial Intelligence (Gen-AI), we have an automated alternative which is to enable automatic synthesis of photorealistic faces for sketches. This not only is a more efficient method, but also more fool-proof, something that would for sure be appreciated by law enforcers.

Recent advancements in deep learning, particularly Gener- ative Adversarial Networks (GANs), have revolutionized the field of image synthesis [20]. By learning the underlying distribu- tion of facial features, GANs can generate high-quality images from low-detail sketches. In criminal and forensic applications, such AI-driven approaches can significantly improve identifi- cation rates by producing highly realistic reconstructions and aiding investigators in narrowing down potential matches more effectively. So, the primary goal of this study is to develop a Generative AI-based framework for forensic face synthesis that improves the accuracy and reliability of sketch-to-face translation [17].

LITERATURE SURVEY

Different types of computer-based sketch synthesis systems have been proposed in recent years. In [1] [2], they utilize edge detection as means to convey the lines and structure of the face. This is not based on learning and hence fails to convey the texture or color of the skin. In that paper, face shape was extracted from a photo and using rule-based exaggerations they were made to look more similar to sketches. Freeman et al. [3] introduced an example-based system capable of translating line drawings into various artistic styles, while Chen et al. [4] proposed a similar example-based approach for generating cartoon-like faces. These systems all depended on face shape extraction using alignment algorithms, an example of which is the Active Appearance Model (AAM) [5]. Before the introduction of Artificial Intelligence, specifically Generative Artificial Intelligence, the best performing model for this task was created using Eigen Transformations and Principal Component Analysis (PCA). Research in [6] demonstrated that a synthesized image created using Eigen Vectors and Transformations could closely match an artist-drawn sketch. However, the limitation in this case was that the synthesized face could only resemble the sketch under two specific con- ditions, which were: 1. the face photo had to be accurately reconstructed using PCA from training samples, and 2. the photo-to-sketch transformation had to be approximated as linear. The actual problem here lies in the fact that these two conditions are often difficult to meet simultaneously, particularly when hair is involved and which is commonly present in most real-world facial images and introduces additional complexity. X.Wang and X.Tang in [7] present a method for face photo-sketch synthesis and recognition using a multiscale Markov Random Fields (MRF) model. Their work highlights the inherent complexity of face synthesis, as traditional grammar and mathematical models are insufficient to capture

V. Agarwal et al. (eds.), *Proceedings of the Global Innovation and Technology Summit "AAROHAN 3.0"_Engineering Track (GITS-EAS 2025)*, Advances in Engineering Research 295,
https://doi.org/10.2991/978-94-6239-644-9_10

the intricate details of human faces, such as tone, texture, and shading. To combat this, the authors advocate for a patch-matching based approach, where the face is treated as overlapping patches and analyzed locally, instead of as a global structure, which is too complex. The underlying assumption here being that if two photo patches are similar, their corresponding sketch patches should also exhibit similarity.

In recent years, the studies regarding face synthesis and recognition have shifted from Edge Detection algorithms to Generative-AI based approaches [21]. The first ones to come were Generative Adversarial Networks (GANs) [14]. They gained promi- nence due to their ability to model complex data distributions. A 2023 study by Devakumar et al. [8] introduced Deep Convolutional GANs (DCGANs) for transforming forensic sketches into photorealistic images. The model was trained on preprocessed sketches and was able to generate high- resolution images of suspects. This approach was the most novel at the time of its release, obtaining an average Structural Similarity Index Measure (SSIM) of 0.587 as compared to the existing models that averaged an SSIM of 0.554. Similarly, Xiaodong Luo et al. [9] proposed a Dual-channel Generator- based GAN (DualG-GAN) for text-to-face synthesis. After advancements in GANs, newer models were developed for image generation tasks, which were transformer-based and diffusion-based models [18]. A recent study by Ivanovska et al. [10] introduced a Morphing Attack Detection (MAD) method using Denoising Diffusion Probabilistic Models (DDPMs). ”Morphing attacks” refer to a type of biometric security vulnerability where an attacker combines images of multiple individuals to create a synthetic image that can fool face recognition systems, allowing unauthorized access. This study showcased the potential of diffusion based-models in forensic applications [16].

METHODOLOGY

A. Preprocessing

Once we got our hands on the most used dataset for this use-case - the CUHK Face Sketch Database (CUFS) - we had to preprocess the data to make sure that the input data that we will provide to the model is well-structured and optimal. The dataset was created at the Chinese University of Hong Kong and consists of 606 photo-sketch pairs with the assumption that all the images are in the frontal pose without any occlusions. Since we were working with a limited size dataset, data augmentation techniques were applied which allowed us to increase the size and diversity of the dataset. Other than that, basic preprocessing methods were also applied to ensure consistent and accurate images.

B. GAN Models vs. Transformer-based Models vs. Diffusion- based Models

In today’s time, the most advanced technology for generative AI are Generative Adversarial Networks (GANs), Transformer-based Models, and Diffusion-based Models.

Transformer-based models, first described in a 2017 paper

[19] from Google, are neural networks that have revolutionised the field of image generation by leveraging self-attention mechanisms. Since transformers work in a sequential manner, they are able to learn the context of the input. If given an input image divided into patches, self-attention computes contextual relationships between patches and can be mathematically defined as:

$$\text{Attention}(Q,K,V) = \text{sofmax}\ QK^T / \sqrt{dxv} \quad (1)$$

Where Q,K,V are the Query, Key, and Value matrices derived from the input embeddings respectively and dk is the dimensionality of the key vectors.

Diffusion-based models on the other hand, work in a completely different way. They destroy training data through the successive addition of Gaussian noise and then learns to recover the data by reversing this noise process. In fact, diffusion-based models are considered to be the leading generative-AI model as of right now, being used in booming technology like Dall-E 2, OpenAI’s image generation model. We tested all three models on our dataset while keeping everything else intact (ie. number of epochs was kept as 100 in all three cases to ensure fairness) to test and find out which model works the best. We evaluated all of the models using the Structural Similarity Index Measure (SSIM) and L2 normalization, and the results are tabulated as shown below in

TABLE I: Comparison of different Generative AI models

Model	Average SSIM Score	Average L2 Norm Score
GAN	0.5879	53.6164
Transformer-Based	0.5850	63.4930
Diffusion-Based	0.2915	126.030

In terms of both the SSIM and L2 scores, the GAN model outperforms both the transformer-based and diffusion-based models, indicating a better perceptual similarity in synthesized images and a better ability to generate finer facial details with lower pixel-level distortion. GANs do require slightly more time and computational efficiency, but in sensitive matters such as forensic investigations, we cannot compromise on the effectiveness and so we finalise to utilize GANs as our model.

C. Choice of Model Architecture and Justification

Once we realised that out of all the state-of-art technologies, GANs was the best performing one, we decided to enhance the basic structure of a GAN model as shown in Fig. 1. to use in our work.

A GAN consists of two competing neural networks: a Gen- erator G and a discriminator D, which can be mathematically represented as shown below:

$$\min \max V(D, G) = \mathrm{E}_{x \sim p\,data(x)} [\log D(x)] + \mathrm{E}_{z \sim pz}(z) [\log(1 - D(G(z)))]$$

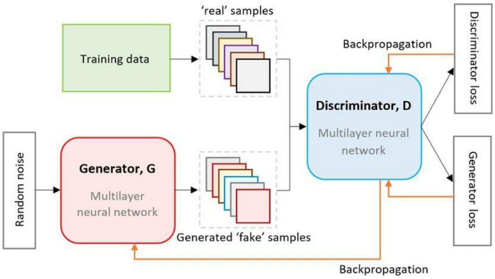

Fig. 1: Architecture of GAN [11]

The generator aims to create images indistinguishable from real data, while the discriminator learns to distinguish between real and generated samples. This adversarial training process enables GANs to capture fine-grained facial details. However, a normal vanilla GAN works on a global level which has been proved to not work well in face synthesis tasks [22]. Instead, we must implement some sort of patch-matching based approach, where the face is treated as overlapping patches and analyzed locally [6]. So, we explore the PatchGAN discriminator instead of a traditional binary classifier.

The PatchGAN discriminator works on a local level, which helps to ensure that the images are processed locally. It eval- uates N x N patches of an image rather than the entire image, which helps in texture consistency, something very important when dealing with faces and skin. In the architecture diagram of PatchGAN given below in Fig. 2., we can visualize how it divides and then predicts as compared to the conventional discriminator.

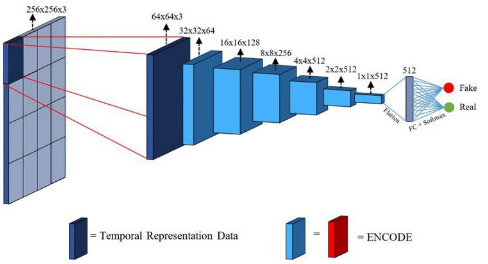

Fig. 2: Architecture of PatchGAN discriminator [12]

Mathematically, given an image X of size H x W, PatchGAN partitions it into H/N x W/N overlapping patches.
To create a more complex GAN model which would easily be able to maintain spatial features and structure and syn- thesized faces, we also explored different types of generator. Theoretically, the one that seemed the best for our task was a U-Net generator which would help to preserve features. A U-Net generator follows an encoder-decoder based archi tecture with skip connections. The encoder helps to extract hierarchical features using convolutional layers, while the decoder reconstructs the face while integrating features from the encoder using skip connections. This also helps us to work on a local level as the U-Net GAN uses a segmentation network as the discriminator. So, the segmentation network predicts two classes: real and fake, and in doing so, the

discriminator gives the generator region-specific feedback. Fig. 3. shows the architecture of a U-Net GAN:

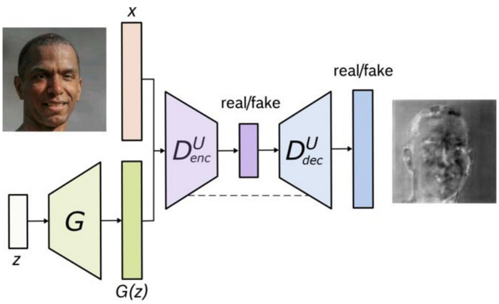

Fig. 3: Architecture of U-Net GAN

Having the best generator and discriminators for our task individually, we also had to ensure that the combination would work well, hence, with some permutation we tested the following four combinators of networks:

- Vanilla Generator + Vanilla Discriminator
- Vanilla Generator + PatchGAN Discriminator
- U-Net Generator + Vanilla Discriminator
- U-Net Generator + PatchGAN Discriminator

D. Model Performance Evaluation

Each model combination was evaluated using Structural Similarity Index Measure (SSIM) score and L2 normalization. SSIM assesses perceptual similarity, accounting for luminance, contrast, and structural consistency. So, a higher SSIM value indicates better preservation of perceptual features, while a lower SSIM value indicates worse preservation of features. On the other hand, a lower L2 normalization score signifies closer similarity between generated and real images, while a higher L2 normalization score signifies more difference between generated and real images. This is because L2 normalization measures pixel-wise difference. So, for the four combinations, the average SSIM and L2 scores were as follows:

TABLE II: Model Performance Evaluation

Model Combination	Average SSIM Score	Average L2 Norm Score
Vanilla G + Vanilla D	0.5800	52.2161
Vanilla G + PatchGAN D	0.5820	55.9411
U-Net G + Vanilla D	0.5632	60.7534
U-Net G + PatchGAN D	0.5879	53.6164

Among the tested architectures, the U-Net generator with the PatchGAN discriminator achieved the highest SSIM score, indicating superior perceptual quality, while maintaining a reasonable L2 score. Therefore, we decide to choose this combination for our final model.

E. Evaluation of number of training epochs

The number of training epochs plays a crucial role in deter- mining the performance of a deep learning model. While an insufficient number of epochs leads to underfitting, excessive training may cause overfitting, where the model memorizes the training data but generalizes poorly on unseen samples. To evaluate the optimal number of training epochs for our face synthesis model, we trained the best-performing GAN architecture (U-Net generator + PatchGAN discriminator) for four different numbers of epochs - 100, 300, 500, and 1000. We compared the results using SSIM and L2 normalization values, which are presented in Table III.

TABLE III: Comparison of model performance at different training epochs.

Epochs	Average SSIM Score	Average L2 Norm Score
100	0.5862	62.0919
300	0.6124	54.3063
500	0.6309	50.8770
1000	0.6138	53.5260

From the table, we can observe the following trends:

- At 100 epochs, the model starts generating recognizable facial structures but lacks sharp details.
- At 300 epochs, the structural similarity improves, and the generated images appear more refined.

- At 500 epochs, performance peaks with the highest SSIM scores and lowest L2 errors, indicating that the model has learned a strong mapping from sketches to realistic faces.
- At 1000 epochs, SSIM slightly drops while L2 errors increase, suggesting that the model starts overfitting.

So, the inference here is that the best results are achieved at 500 epochs, where the balance between feature learning and generalization is optimized.

I. Discussion

The latest work we were able to find in this area was published in 2023 by Devakumar et al. [8]. At the time of its publication, it surpassed the average SSIM as compared to the existing models, and with our approach, we have beat the average SSIM of all of the existing work including the most recent one.

TABLE IV: Comparison of existing work with our proposed idea

Proposed Idea	Average Score
PatchGAN discriminator + U-net generator at 500 epochs (our proposed idea)	0.6309
Deep Convolutional GANs (DCGANs)	0.587
Previous Work	0.554

I. Experimental Results

Fig. 4. below shows the experimental results for the four different combinations of generator and discriminator that we used. Figure IV. (a) (b) (c) (d) respectively represent the faces synthesized using the combinations of Vanilla Gener- ator + Vanilla Discriminator, U-Net Generator + PatchGAN Discriminator, U-Net Generator + Vanilla Discriminator, and Vanilla Generator + PatchGAN discriminator. These were compared by keeping all other hyperparameters constant to ensure fairness, and in this case the number of epochs was fixed to 100. As discussed in the results earlier, and as can be seen from the figure above as well the combination of U- Net generator and PatchGAN discriminator provided the best results in this case. In the figures, we have also displayed the SSIM and L2 Norm scores for each of the faces, and the average used for our work was calculated by summing up those values and dividing them by five.

Fig. 5. below shows the experimental results given by the U- Net generator + PatchGAN discriminator at different numbers of epochs. Figure V. (a) (b) (c) (d) respectively represent the output at 100, 300, 500, and 1000 epochs. In the figures, we have also displayed the SSIM and L2 Norm scores for each of the faces, and the average used for our work was calculated by summing up those values and dividing them by five. From the output of the model as well as the scores, we can see that : at 100 epochs, the model starts generating recognizable facial structures but lacks sharp details. At 300 epochs, the structural similarity improves, and the generated images appear more refined. At 500 epochs, performance peaks with the highest SSIM scores and lowest L2 errors, indicating that the model has learned a strong mapping from sketches to realistic faces. At 1000 epochs, SSIM slightly drops while L2 errors increase, suggesting that the model starts overfitting.

From Fig. 6., we were able to infer some things that helped us decide that 500 epochs was the optimal number of epochs for our model. In the initial phase of 1 to 50 epochs, the generator starts with extremely high loss at around 46 while the discriminator starts strong at just around 0.35. In this phase, the discriminator remains relatively stable, whilst the generator rapidly improves its quality as the loss drops to around 20 by epoch number 30. Between epochs 50 to 300, the discriminator still keeps strong and the generator continues to show gradual improvement, by decreasing its average loss from about 20 to around 15, with its lowest going till 13. In the final phase of 300 to 500 epochs, both losses stabilize in a relatively narrow band with final losses being 0.48 for the discriminator and 15.43 for the generator. This showed that there was a healthy competition as the discriminator never collapsed to 0 and the generator showed continuous improvement. At 500 epochs, we saw the best loss curve as can be seen from the figure.

Fig. 7. which shows the training loss curve over the other number of epochs. Figure VII. (a), (b), and (c) respectively show the training loss curve at 100, 300, and 1000 epochs. From the curves, we can infer that 100 and 300 epochs both

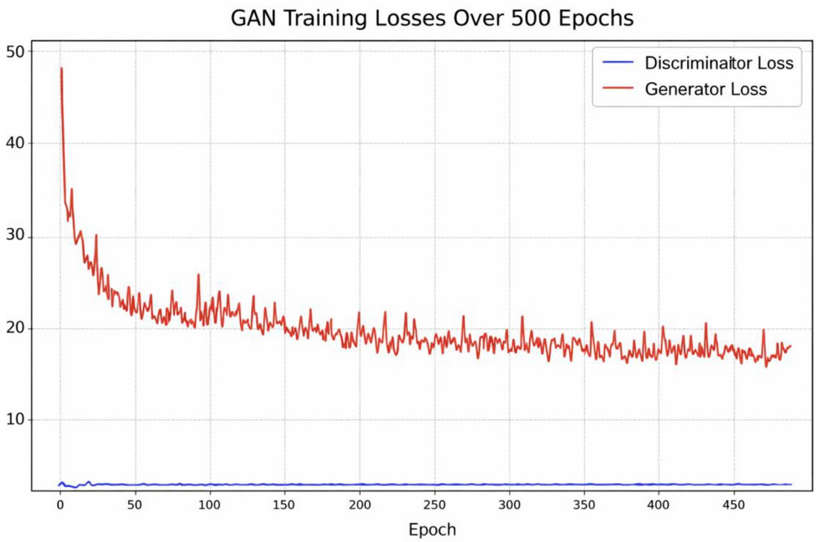

Fig. 4: Experimental results of different generator–discriminator combinations for sketch-to-face synthesis.

turned out to be insufficient for convergence as the losses never really got the time to stabilize, while 1000 became redundant as we saw diminishing returns.

I. CONCLUSION

This paper aimed to develop a Generative AI-Assisted Face Synthesis system for forensic and crime investigation, utilizing GAN-based models to reconstruct realistic human faces from sketches. The study explored various architectures, comparing traditional GANs with newer generative AI models like Transformers and Diffusion Models.

The key findings and achievements of this paper are as follows:

- Best Performing Model: The U-Net Generator + Patch- GAN Discriminator outperformed all other configura- tions, achieving the highest SSIM (0.5879) and lowest L2 error (53.6164).
- Optimal Training Duration: Results indicated that 500 epochs provided the best trade-off between underfitting and overfitting, improving feature preservation without introducing artifacts.
- GAN vs. Transformer vs. Diffusion: The GAN-based model demonstrated superior performance over both Transformer-based and Diffusion-based models, particu- larly in reconstructing fine facial details.

II. FUTURE SCOPE

While this paper has successfully demonstrated the potential of generative AI for forensic applications, and created a new benchmark in its field, we believe that there are still improve- ments to be made. First of all, the dataset could be expanded by integrated additional large-scale forensic sketch datasets or real-world police sketches. Secondly, this model and its benchmark can also be used to develop a real-time forensic face reconstruction tool that can be deployed in investigative agencies. This tool could also contain a user-friendly interface where forensic experts are able to input sketches and fine-tune the generated images interactively as they wish or as described by the eye-witness in real time.

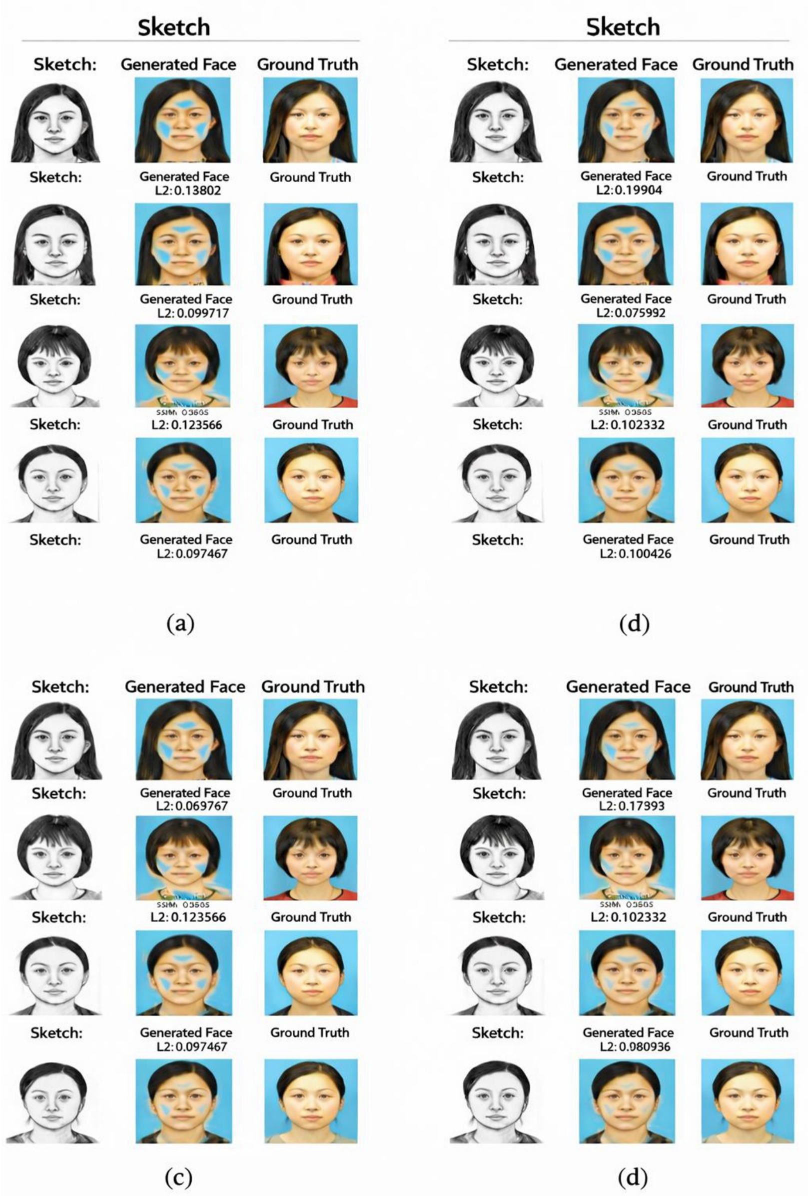

Fig. 5: Generated faces from first 5 sketches in the test set for (a) Vanilla Generator + Vanilla Discriminator (b) U-Net Generator + PatchGan Discriminator (c) U-Net Generator + Vanilla Discriminator (d) Vanilla Generator + PatchGAN Discriminator at 100 epochs

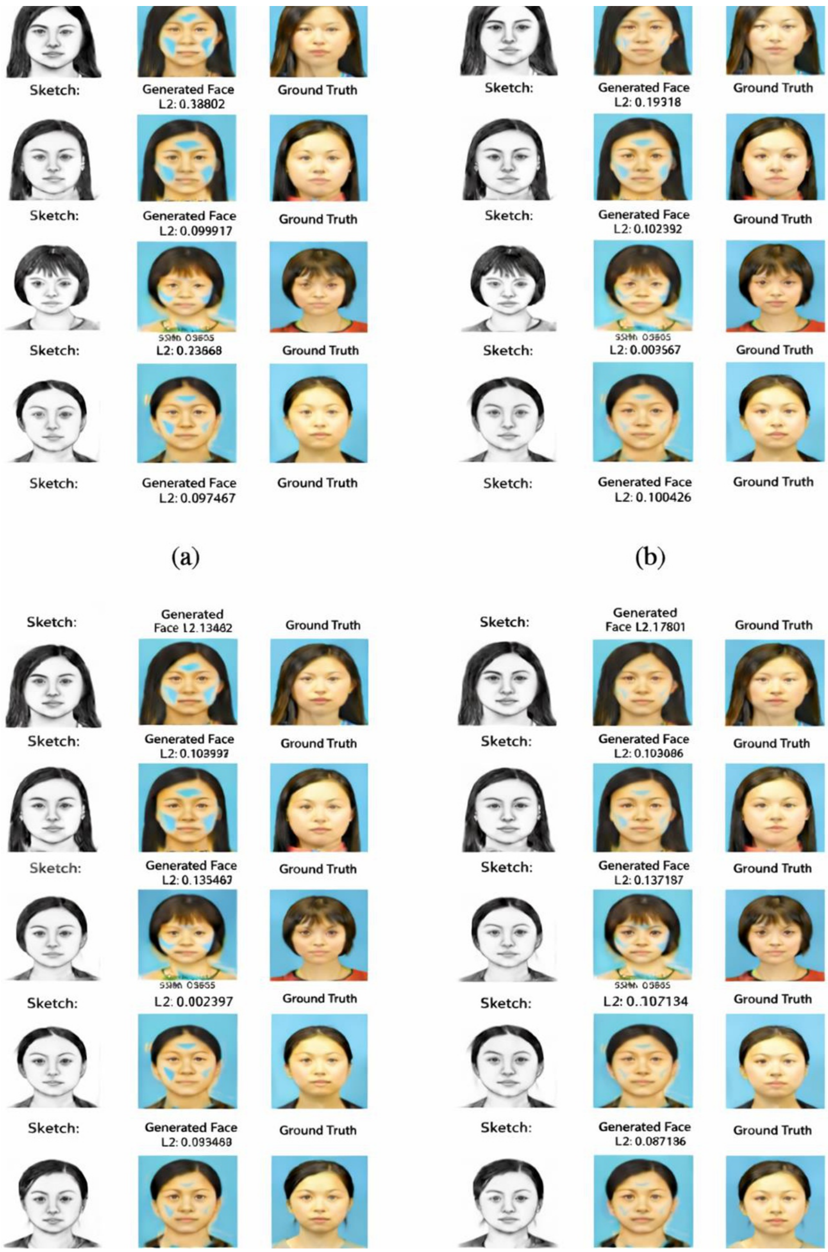

Fig. 6: Visualization of generated face images at different training epochs — (a) 100 epochs, (b) 300 epochs, (c) 500 epochs, and (d) 1000 epochs — compared with the original sketch and ground truth image.

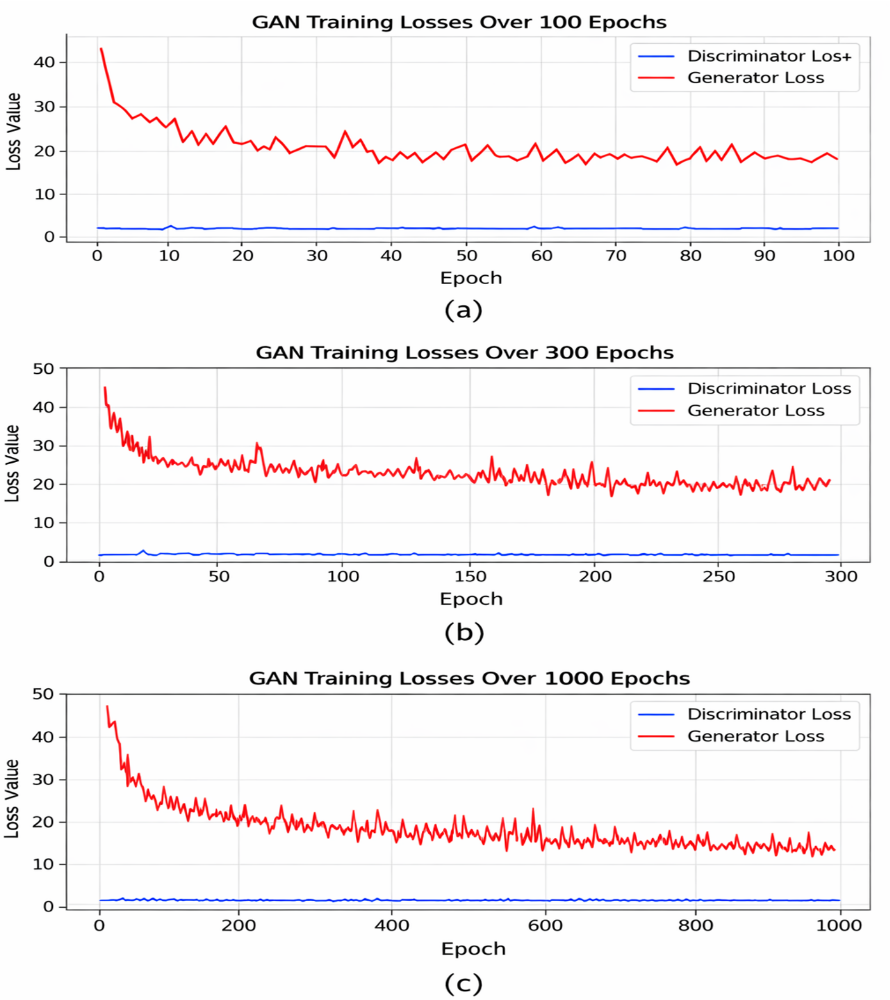

Fig. 7: GAN training loss curve over (a) 100 epochs (b) 300 epochs (c) 1000 epochs

References

[1] Koshimizu, H., Tominaga, M., Fujiwara, T., Murakami, K. (1999). On Kansei facial processing for computerized facial caricaturing system Picasso. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1999, 286–290.

[2] Iwashita, S., Takeda, Y., Onisawa, T. (1999). Expressive facial caricature drawing. Proceedings of the IEEE International Conference on Fuzzy Systems, 1999, 1375–1380.

[3] Freeman, W. T., Tenenbaum, J. B., Pasztor, E. (1999). An example- based approach to style translation for line drawings. Technical Report, MERL.

[4] Chen, H., Xu, Y., Shum, H., Zhu, S., Zheng, N. (2001). Example-based facial sketch generation with non-parametric sampling. Proceedings of the IEEE International Conference on Computer Vision, 2001, 433–438. 60

[5] Cootes, T. F., Edwards, G. J., Taylor, C. J. (1998). Active appearance models. Proceedings of the European Conference on Computer Vision, 1998, 484–498.

[6] Tang, X., Wang, X. (2003). Face sketch synthesis and recognition. Proceedings of the IEEE International Conference on Computer Vision, 2003, 687–694.

[7] Wang, X.,Tang, X. (2009). Face photo-sketch synthesis and recogni- tion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(11), 1955–1967. https://doi.org/10.1109/TPAMI.2008.181

[8] Sreedev Devakumar and Greeshma Sarath. 2023. Forensic Sketch to Real Image Using DCGAN. Procedia Comput. Sci. 218, C (2023), 1612–1620. https://doi.org/10.1016/j.procs.2023.01.139

[9] Luo X, He X, Chen X, Qing L, Zhang J. DualG-GAN, a Dual-channel Generator based Generative Adversarial Network for text-to-face syn- thesis. Neural Netw. 2022 Nov;155:155-167. doi: 10.1016/j.neU- Net.2022.08.016. Epub 2022 Aug 19. PMID: 36058021.

[10] Ivanovska, M., S ˇtruc, V. (2023, April). Face morphing attack detection with denoising diffusion probabilistic models. In 2023 11th International Workshop on Biometrics and Forensics (IWBF) (pp. 1-6). IEEE.

[11] Little, C., Elliot, M., Allmendinger, R., Samani, S. (2021). Generative adversarial networks for synthetic data generation: A comparative study. arXiv. https://doi.org/10.48550/arXiv.2112.01925

[12] Ganokratanaa, T., Aramvith, S., Sebe, N. (2020). Unsu- pervised anomaly detection and localization based on deep spatiotemporal translation network. IEEE Access, PP, 1–1. https://doi.org/10.1109/ACCESS.2020.2979869

[13] W. Zhang, X. Wang and X. Tang. (2011). Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[14] Fard, A. P., Mahoor, M. H., Lamer, S. A., Sweeny, T. (2023). GANalyzer: Analysis and manipulation of GANs latent space for controllable face synthesis. arXiv preprint arXiv:2302.00908. https://doi.org/10.48550/arXiv.2302.00908

[15] Arif, D., Mehmood, Z., Ullah, A. Winberg, S. (2024). High-definition image formation using multi-stage cycle generative adversarial network with applications in image forensic. Arabian Journal for Science and En- gineering, 49(3), 3887–3896. https://doi.org/10.1007/s13369-023-08193- x

[16] Gu¨era, D., Meng, L., Delp, E. J. (2021). GANs and artificial facial expressions in synthetic portraits. Applied System Innovation, 5(4), 63. https://doi.org/10.3390/asi5040063

[17] Sun, J., Yu, H., Zhang, J. J., Dong, J., Yu, H. Zhong, G. (2022). Face image-sketch synthesis via generative adversarial fusion. Neural Networks, 154, 179–189. https://doi.org/10.1016/j.neU-Net.2022.07.013

[18] Shen, Y., Zhou, B., Luo, P., Tang, X. (2018). FaceFeat-GAN: a two- stage approach for identity-preserving face synthesis. arXiv preprint arXiv:1812.01288. https://arxiv.org/abs/1812.01288

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

[20] Chakraborty, C., Nagarajan, S. M., R. R., K. U., Devarajan, G. G. (2025). Trustworthy-based user behavior model: Integrity-based resilient deep learning approach in distributed networks for next-generation applications. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2025.3530772

[21] R. R., H. S., Nagarajan, S. M., Devarajan, G. G., Omar, M., Bashir, A. K. (2024). Threat detection and mitigation for tactile internet driven consumer IoT-healthcare system. IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/TCE.2024.3370193

[22] Banerjee, S., Das, R., Rathinam, R., Dhanalakshmi, R. (2023). Real-time fire detection in unmanned ground vehicles integrating YoloV5 and AWS IoT. 2023 International Conference on Sys- tem, Computation, Automation and Networking (ICSCAN), 1-6. https://doi.org/10.1109/ICSCAN58655.2023.10394971

Additional recent studies have explored improvements in sketch-to-face synthesis and recognition systems. Information-theoretic encoding approaches have improved face photo-sketch recognition accuracy [13]. GAN-based methods have been widely studied for controllable face synthesis and manipulation of latent space representations [14]. Advanced GAN architectures have also been applied for high-definition image generation and forensic reconstruction tasks [15]. Studies highlight the role of synthetic facial expressions and AI-generated portraits in improving facial modeling techniques [16]. Recent research demonstrates improved sketch-to-image translation using adversarial fusion frameworks for enhanced identity preservation [17]. Two-stage generative frameworks such as FaceFeat-GAN have also shown promising results for identity-preserving face synthesis [18]. Transformer-based architectures and modern generative frameworks continue to influence image generation and deep learning research [19–22].