Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 3 (2025)

From Sketch to Image: A GAN-Based Forensic Reconstruction System Using Text Prompts

Dr.L K Suresh Kumar[1], Dr. V. Ramesh[2] Dr. Srujana Inturi [3] Dr. A. Shiva Kumar[4]

[1]Associate Professor, Department of Computer Science, UCEOU, Hyderabad, India.

[2]Assistant Professor, CMRIT Hyderabad,

[3]Assistant Professor, CBIT(A), Hyderabad-75,

[4]Associate Professor, CMR College of Engineering & Technology,

Abstract: Facial sketch-to-image synthesis is a crucial task in forensic investigations, digital art, and human-computer interaction. This research presents a Generative Adversarial Network (GAN)-based model that converts facial sketches into high-fidelity, photorealistic images guided by textual descriptions. The proposed approach integrates Contrastive Language-Image Pretraining (CLIP) to enhance textual feature extraction, ensuring detailed and accurate facial reconstructions. A refined encoder to StyleGAN pipeline is employed to generate images with structural coherence and perceptual realism. The model is trained on the Multi-Modal-CelebA-HQ dataset, comprising 19,923 paired sketches, images, and textual descriptions. Performance evaluation is conducted using standard image similarity metrics, including L2 Norm and Structural Similarity Index (SSIM). Experimental results demonstrate that the proposed method achieves a high SSIM score (0.788) and low L2 Norm (89.68), indicating strong structural similarity and fine- detail preservation. Despite promising results, limitations remain, such as the model's dependency on textual prompts, potential bias in generated features, and computational constraints. Future work will explore multi-modal enhancements, such as incorporating audio cues or additional image inputs, as well as adopting self-attention mechanisms and transformer-based architectures for improved image synthesis. This study contributes to advancements in AI-driven face reconstruction with applications in forensics, digital art, and entertainment industries.

Keywords: *Sketch-to-Image Synthesis, Generative Adversarial Networks, Contrastive Language- Image Pre-training, Facial Reconstruction, Deep Learning.*

1. Introduction

The task of reconstructing realistic images from forensic sketches is critical in criminal investigations, where eyewitnesses may provide only rough sketches of suspects. These sketches often lack detail and can be subjective, depending heavily on the witness's memory and the sketch artist's interpretation. As a result, the sketches may not accurately represent the actual appearance of suspects, posing challenges for law enforcement in identifying individuals based on sketches alone. This has led to an increasing interest in utilizing artificial intelligence (AI) and deep learning methods to enhance forensic sketch-based image reconstruction. Generative Adversarial Networks (GANs) have demonstrated remarkable capabilities in image generation tasks, including transforming sketches into high-quality, photorealistic images. Introduced by Good fellow et al. in 2014, GANs use two-part neural network architecture with

a generator and discriminator competing against each other, producing outputs that become increasingly realistic over time [1]. Variants of GANs, such as Style GAN [2], have further improved image synthesis by allowing control over specific features, such as style and texture, making them particularly suited for forensic applications where accurate facial reconstruction is essential.

In forensic image reconstruction, traditional GAN-based approaches focus primarily on translating sketches into images without incorporating contextual information that may be provided by witnesses, such as race, age, and hair color. To address this limitation, we propose a GAN-based forensic reconstruction system that integrates textual prompts to guide the image generation process. By utilizing Contrastive Language-Image Pre-training (CLIP) [3], a recent model developed by Open AI, our system can encode both textual and visual features into a shared latent space, enabling the model to generate images that match the sketch while incorporating textual attributes. This approach not only enhances the fidelity of generated images but also aligns them closely with descriptions provided by witnesses, thereby increasing the utility of generated images in forensic investigations. The use of multimodal approaches that combine textual and visual data has proven effective in various domains, including image captioning [4], image retrieval [5], and conditional image generation [6]. By leveraging CLIP for text encoding, our model captures semantic relationships between textual descriptions and sketch inputs, producing

images that are more accurate representations of the described person. This capability is crucial in scenarios where law enforcement has limited information, as the model can utilize both sketch and text prompts to create more comprehensive images. We trained and evaluated our model on the Multi-Modal-CelebA-HQ dataset, which provides high-quality face images alongside textual descriptions and sketches. Our models performance was assessed using metrics such as the Structural Similarity Index Measure (SSIM) and L2 norm, with results demonstrating substantial improvements over baseline models that rely solely on visual inputs. The proposed system shows promising potential for real-world applications in forensic investigations, where accurate image reconstruction from sketches is crucial for identifying suspects and missing persons.

This paper presents a novel GAN-based forensic image reconstruction system that leverages both sketches and textual prompts to generate realistic images. The system's use of CLIP and Style GAN demonstrates the effectiveness of combining text and visual data for improved image generation, making it a valuable tool for law enforcement and forensic professionals. In image generation, For instance, Reed et al. proposed a text-to-image synthesis model that generates images conditioned on textual descriptions, effectively combining language and visual information [10]. This approach has inspired several studies in which both sketches and textual prompts are used to generate more accurate images. CLIP (Contrastive Language-Image Pre-training), developed by Open AI, and represents a significant advancement in this field by mapping textual and visual data into a shared embedding space, thus facilitating the alignment of text and image features [11]. CLIP has shown potential in various image generation tasks, including those requiring cross-modal synthesis, as it allows for more nuanced and semantically accurate outputs.

2. Related Work

The task of converting forensic sketches into photorealistic images has gained considerable attention in recent years, particularly with advancements in deep learning and generative models. Traditional sketch-to-image methods relied on hand-crafted features and statistical

models, which were limited in their ability to capture complex textures and styles necessary for realistic image synthesis [7]. With the advent of Generative Adversarial Networks (GANs), more sophisticated and effective approaches for sketch-to-image translation have been developed, significantly improving the quality of generated images. One of the foundational GAN-based models for image synthesis is DCGAN, which employs convolution layers to enhance feature extraction and has been widely applied in various images generation. For instance, Reed et al. proposed a text-to- image synthesis model that generates images conditioned on textual descriptions, effectively combining language and visual information [10]. This approach has inspired several studies in which both sketches and textual prompts are used to generate more accurate images. CLIP (Contrastive Language-Image Pre-training), developed by Open AI, and represents a significant advancement in this field by mapping textual and visual data into a shared embedding space, thus facilitating the alignment of text and image features [11]. CLIP has shown potential in various image generation tasks, including those requiring cross-modal synthesis, as it allows for more nuanced and semantically accurate outputs.

In recent years, several multimodal approaches that leverage GANs and transformers have been introduced to handle complex tasks involving both images and text. For example, DALL-E, a transformer-based model developed by Open AI, demonstrated the capability to generate coherent images from text prompts using a combination of transformer architectures and image generation techniques [12]. This success has spurred further research into combining transformers with GANs for enhanced image synthesis. A recent study by Li et al. (2023) introduced "Sketch2Face-GAN," a model optimized for forensic applications that combines transformer-based language models with GANs to improve sketch-to-image quality while maintaining high identity consistency [13]. In forensic applications, maintaining identity consistency is crucial. FaceForensics++ and similar frameworks have been generation tasks [8]. However, DCGAN lacks the ability to handle high- resolution image generation, which is crucial for forensic applications that require detailed and accurate representations. Style GAN, introduced by Karras et al., addresses this limitation by incorporating a style-based generator that enables fine-grained control over generated images, resulting in high-quality outputs with improved texture and feature control [9]. Style GAN has been extensively used in sketch-to-image tasks due to its ability to produce realistic and diverse images from low-dimensional representations, making it particularly suitable for forensic applications.

Recent developments have extended GAN-based methods to incorporate multimodal inputs, such as text descriptions, to guide adapted to detect facial features that align with witness descriptions, even when generated from rough sketches. A 2024 study by Zhang et al. proposes a "Multimodal Forensic GAN" that specifically targets high fidelity in forensic images, preserving unique identity markers like scars or moles [14]. This model has shown high accuracy in generating images from low-detail sketches combined with text descriptions, showcasing the relevance of combining sketch, text, and GAN-based models for forensic reconstructions. Evaluation metrics play an essential role in validating the effectiveness of forensic image reconstruction models. Metrics like Structural Similarity Index Measure (SSIM) and L2 norm are commonly used to assess the similarity between generated images and ground truth images, providing quantitative measures of image fidelity [15]. Recently, perceptual metrics that account for the human visual system, such as LPIPS (Learned Perceptual Image Patch Similarity), have become more prominent in evaluating the perceptual quality of GAN-generated images [16]. These metrics capture visual quality better than traditional measures, making them increasingly valuable in forensic image generation [17].

latent code w_i corresponds to a specific level of detail in the StyleGAN architecture. The fusion process is defined as:$W = \phi(e_t, e_s)$ (3)Where ϕ is a fusion function that combines text and sketch embeddings, ensuring alignment between semantic and spatial features. The generated latent codes W are fed into the StyleGAN model. The fused feature representation W is then processed through subsequent layers, such as a latent vector generator or a StyleGAN model, to generate the final output.

Block diagram:

3. Methodology

In this section, we describe the proposed framework as shown in Figure 1 that combines text-based and sketch-based inputs with Style GAN for high-quality image generation. The methodology is broken into key components: text embedding, sketch encoding, feature fusion, and image synthesis. The methodology is broken into five key components: text embedding, sketch encoding, feature fusion, and image synthesis.

3.1 Text Embedding using CLIP

The textual descriptions from the dataset are passed into a **Text Encoder**, a pre-trained CLIP text encoder, which processes the text inputs to create dense embeddings. These embeddings represent the semantic information contained in the text. The CLIP model "open AI/ clip-vit-base-patch32" is used which is provided by the transformers library. A CLIP processor and CLIP model is instantiated, the former is used to create label tokens from the given tokens, and the latter is used to create text embeddings that fully capture the semantic meaning of the text. Let T denote the input text description. The CLIP Processor preprocesses T into a tokenized format, denoted as

$\mathbf{t}$processed. The CLIP Model then encodes $\mathbf{t}$processed into a latent semantic vector$\mathbf{e}_t$, where:

$\mathbf{t}_{processed} \rightarrow CLIPProcessor(T)$….(1)$\mathbf{e}_t \rightarrow CLIPModel(\mathbf{t}_{processed})$….(2)The text prompt is fed into the CLIP model which generates a vector of size (1, 512).

3.2 Sketch Encoding

Sketch encoding is a fundamental process in tasks such as sketch-to-image generation, where the sketch input S, represented as a 2D image matrix ($S \in R^{H\times W\times C}$), is transformed into a compact latent representation. The encoder,f_{enc}, typically a convolutional neural network (CNN), extracts high-level structural and spatial features from the sketch. This process begins with input

preprocessing, including normalization, resizing, and tokenization, to ensure consistent input dimensions. The sketch is then passed through a series of convolution layers. The initial layers extract low-level features, such as edges and textures, while deeper layers capture mid-level patterns like contours and shapes. Final layers focus on high-level abstract features that encapsulate the sketch's spatial and structural information. The output of these layers is reduced global average pooling to form the latent vector e_s, a compact feature representation in R^k, where k is the dimensionality of the feature space. The encoder ensures that the essential characteristics of the sketch are preserved in e_s, making it a descriptive yet compact vector. This latent representation is vital for downstream tasks, such as aligning with other modalities with text embeddings. In this proposed model ResNet encoders, are used for ensuring that the encoded features are both robust and semantically meaningful.

3.3 Feature Fusion

The semantic embedding et and sketch embedding are fused to $e_s$generate a set of latent codes

$W = \{w_1, w_2, \dots, w_{14}\}$. Each latent code w_i corresponds to a specific level of detail in the StyleGAN architecture. The fusion process is defined as:

$W = \phi(e_t, e_S)$.....(3) Where ϕ is a fusion function that combines text and sketch embeddings, ensuring alignment between semantic and spatial features. The generated latent codes W are fed into the StyleGAN model. The fused feature representation W is then processed through subsequent layers, such as a latent vector generator or a StyleGAN model, to generate the final output.

Algorithm 1 Text-Sketch Guided Image Generation using StyleGAN

Input : Text input T, Sketch input S, Pretrained CLIP model, Encoder f_{enc}, Fusion function ϕ, StyleGAN generator g_{style}, Loss weights $\lambda_1, \lambda_2, \lambda_3$
Output: Generated image I
Step 1: Text Embedding
Preprocess text T using CLIPProcessor: $\mathbf{t}_{\text{processed}} \leftarrow \text{CLIPProcessor}(T)$
Encode the processed text into semantic embedding: $\mathbf{e}_t \leftarrow \text{CLIPModel}(\mathbf{t}_{\text{processed}})$
Step 2: Sketch Encoding
Encode the sketch input S into a latent representation: $\mathbf{e}_s \leftarrow f_{\text{enc}}(S)$
Step 3: Feature Fusion
Fuse text and sketch embeddings to generate latent codes:

$$\mathbf{W} = \phi(\mathbf{e}_t, \mathbf{e}_s)$$

The latent codes $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{14}\}$ represent different levels of detail.
Step 4: Image Synthesis
Generate the image using the StyleGAN model:

$$I \leftarrow g_{\text{style}}(\mathbf{W})$$

Step 5: Loss Optimization
Define the total loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{perceptual}} + \lambda_2 \mathcal{L}_{\text{alignment}} + \lambda_3 \mathcal{L}_{\text{realism}}$$

Optimize g_{style} and f_{enc} using $\mathcal{L}_{\text{total}}$.
Step 6: Output return Generated image I

3.4 Image synthesis using Style GAN

The StyleGAN model gstylegstyle synthesizes an image I using the latent codes W. The process can be expressed as:

$$I = g_{\text{style}}(W) \qquad (4)$$

Here, $\boldsymbol{I}$ is the generated image$\boldsymbol{g}_{\text{style}}$, and ensures hierarchical control over the synthesis process using the latent codes$\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_{14}}\}$ each corresponding to different layers of the network.The training of StyleGAN involves iteratively improving the generator and discriminator through an adversarial process.The process of image synthesis in StyleGAN begins with input sampling, where a random latent vector $z \in \mathbb{R}^d$ is drawn from a normal distribution, $z \sim \mathcal{N}(0, I)$this vector serves as the starting point for generating an image. In the next step, known as latent space mapping, the sampled z is transformed into an intermediate latent vector w through a learned mapping network f, such that $w = f(z)$.The mapping network

is composed of fully connected layers and is designed to capture a disentangled representation of the latent space. The resulting vector w resides in the W-space, which offers greater flexibility and disentanglement compared to the original Z-space, enabling more precise control over the attributes and styles of the generated image.The vector w is used to modulate different levels of the generator through adaptive instance normalization (AdaIN). This process controls the style and attributes of the generated image at various levels of abstraction

For each convolutional layer i, the style is modulated as:

$$\mathrm{AdaIN}(x_i, w) = \gamma(w) \cdot \frac{x_i - \mu(x_i)}{\sigma(x_i)} + \beta(w) \qquad (5)$$

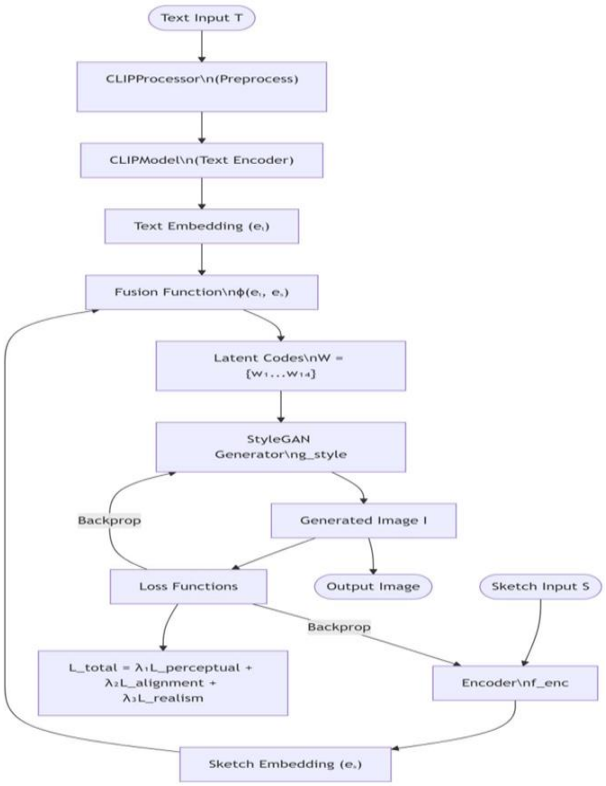

Figure 2. Algorithm of proposed approach.

The generator progressively synthesizes the image starting from a low resolution and gradually increasing to the target resolution.StyleGAN uses progressive growing during training, which starts with low-resolution images and progressively adds layers to synthesize higher-resolution images. This approach stabilizes training and improves the quality of high-resolution outputs. The discriminator D is a convolution neural network trained to classify images as real (from the dataset) or fake (generated by G).

Given an image x, the discriminator outputs: $D(x) \in [0,1]$.

The adversarial training is driven by the following loss functions:

Generator Loss: Encourages the generator to produce images that fool the discriminator:

$$L_G = -D_{z\sim\mathcal{N}(0,I)}[\log D(G(z))] \quad (6)$$

Discriminator Loss: Encourages the discriminator to correctly classify real and fake images:

$$L_D = -D_{x\sim p_{data}}[\log D(x)] - D_{z\sim\mathcal{N}(0,I)}[\log(1 - D(G(z)))] \quad (7)$$

The generator (G) and discriminator (D) in Style GAN are trained iteratively through adversarial learning. G generates fake images, and D classifies them alongside real images as real or fake. Losses are computed and back propagated to update their parameters using optimizers like Adam, enabling G to produce increasingly realistic images over time.

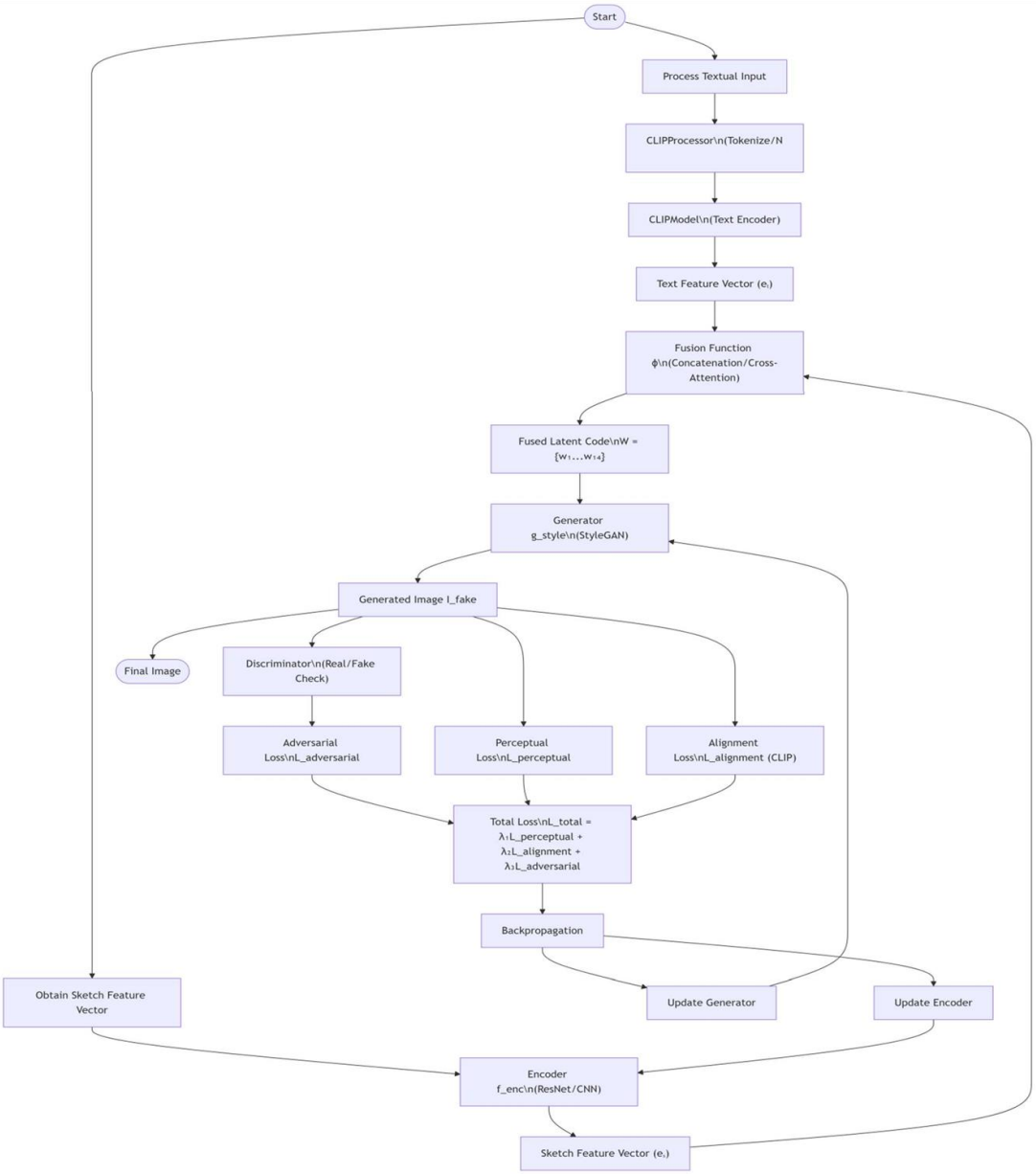

Figure 3: Working process of proposed method.

Loss Functions The quality of the generated image is optimized using a multi-objective loss function L_{total}, which combines perceptual similarity, text-sketch alignment, and realism losses:

$$L_{\text{total}} = \lambda_1 L_{\text{perceptual}} + \lambda_2 L_{\text{alignment}} + \lambda_3 L_{\text{realism}} \quad (8)$$

Perceptual Loss $L_{\text{perceptual}}$ ensures the generated image aligns semantically with the input text. Alignment Loss $L_{\text{alignment}}$ensures spatial alignment between the sketch and the generated image.

Realism Loss L_{realism} ensures the generated image resembles real-world images by utilizing a discriminator network.

The weights $\lambda_1, \lambda_2, \lambda_3$ are hyper parameters that balance the contributions of the individual losses.

3.5 Work Flow

The workflow diagram as shown in Figure 3 outlines the step-by-step process of generating realistic images from forensic sketches and textual descriptions using a GAN-based model. The process begins with text processing, where the input textual description is tokenized and transformed into a feature vector using a pre- trained text encoder such as CLIP. Simultaneously, the forensic sketch is processed through a sketch encoder, which extracts relevant structural and spatial features, converting the sketch into a corresponding feature vector. These two feature vectors one from text and the other from the sketch are then concatenated to create a joint latent representation.

Figure 4.sketch, generated image and real image without skip connections.

This combined feature vector is fed into the generator of a GAN, which synthesizes an image based on the learned mapping of the latent space. The generated image is then evaluated by the discriminator, which distinguishes between real images from the training dataset and the synthesized images. Based on this evaluation, a loss function is computed, which guides back-propagation to update both the generator and discriminator iteratively. Over multiple training cycles, the model refines its ability to generate highly realistic forensic reconstructions, effectively bridging the gap between sketches, textual descriptions, and photorealistic images.**Experimental Results**

The proposed system was evaluated using a variety of qualitative and quantitative experiments. The system integrates sketch-based spatial features and text-based semantic prompts to reconstruct high- quality images using GAN architecture.

3.6 Dataset Description

This model is trained using the Multi-Modal-CelebA-HQ (MM-CelebA-HQ) dataset as shown in Table 1, which consists of 30,000 face images selected from the CelebA dataset. Each image in this dataset is accompanied by a sketch and a descriptive text. However, not all text descriptions contain valuable information, such as hair color. Therefore, further filtering was

performed, reducing the dataset to 19,923 images.

Table 1: Description of Multi-Modal-CelebA-HQ (MM-CelebA-HQ) dataset.

Name	Size	Files	Format	Description
Images	922 MB	19923	jpg	Images from the CelebA-HQ dataset with a resolution of 256x256
Sketches	637 MB	19923	jpg	Sketches derived from the CelebA- HQ dataset with a resolution of 256x256.
Text	911 KB	19923	txt	Text descriptions corresponding to images from the CelebA-HQ dataset.

The system was implemented using Python, with Style GAN as the core generative model, and CLIP for text encoding.

3.7 Training Process

The training process aimed to develop a robust GAN-based model capable of generating high-quality images from sketches and textual descriptions. Initially, Encoder-Decoder architecture was used in the generator, where the input sketch was processed through eight down sampling layers, followed by the concatenation of text embeddings extracted from CLIP. The output was then passed through eight up sampling layers to generate an image.

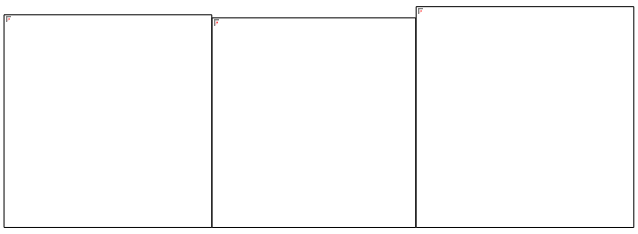

Figure 5.sketch, generated image and real image with skip connections

However, the absence of skip connections led to poor resemblance between the generated and actual images. ByIncorporating skip connections, structural details from the sketch were preserved, significantly improving the quality of the generated images. The model also faced issues where it relied more on the sketch than the text description (e.g., distinguishing hair color from the sketch rather than the text prompt). To address this, a modified sketch

generation algorithm was introduced to eliminate such visual biases. Due to the high computational cost of training a GAN from scratch, we later integrated Style GAN as the generator, mapping our encoder's output to Style GAN's latent space. This transition resulted in significantly improved image quality and faster training times.

Training the GAN for 100 epochs took approximately 1.5 days, given our hardware constraints. The final trained model effectively generated high-fidelity images with better alignment to both sketch structure and textual descriptions.

3.8 Testing Process

The performance of the proposed system was evaluated using two key image similarity metrics:

L2 NormAlso known as the Euclidean norm, L2 Norm measures the distance between the generated image and the ground truth. It is calculated as:

L_2 Norm $= \sqrt{a^2 + a^2 + \cdots + a^2}$

1

2

n

Lower L2 Norm values indicate better resemblance between the generated image and the actual image.Structural Similarity Index (SSIM) SSIM evaluates the similarity between two images, providing a score between -1 and 1. Higher values indicate greater similarity. SSIM is computed using the formula:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

x y 1 x 2 y

Where:

x, y are the compared images.

μ_x, μ_y are the mean intensities.

σ_x^2, σ_y^2 are variances.

σ_{xy} is the covariance.

$C_1 and\ C_2$ are constants for numerical stability.

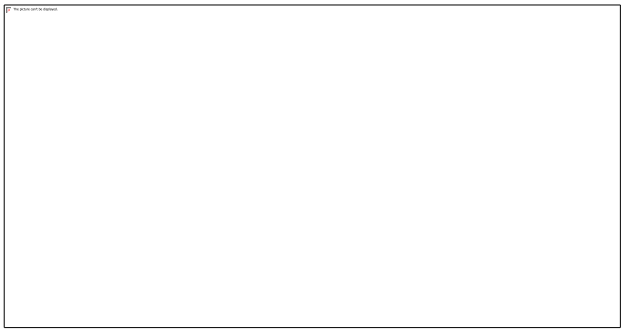

Figure 6.Improved sketches, generated image and real image

3.9 Results and Analysis

To assess the model's effectiveness, we compared the generated images with ground truth images using the aforementioned evaluation metrics. Table 2 presents the computed values for a sample of 150 image pairs:

Table 2:L2 Norm and SSIM values of the 150 pairs.

Metric	Average Value
L2 Norm	89.68
SSIM	0.788

The relatively low L2 Norm and high SSIM score indicate that the generated images closely resemble the ground truth images, both structurally and in fine details.

4.4.1.1 Visual Comparison

Figures 4 to 7 illustrate different stages of the model's evolution. The following images depict:

i. Initial Model Without Skip Connections: Poor resemblance to sketches.

ii. Model with Skip Connections: Improved structural consistency.

iii. Modified Sketching Approach: Reduced sketch-based biases.

iv. Final Model Using Style GAN: High-quality image synthesis.

Figure 7.sketch, generated image and real image using style GAN generatorA comparative histogram of L2 Norm values across different versions of the model is shown in Figure 8, highlighting improvements in image fidelity.

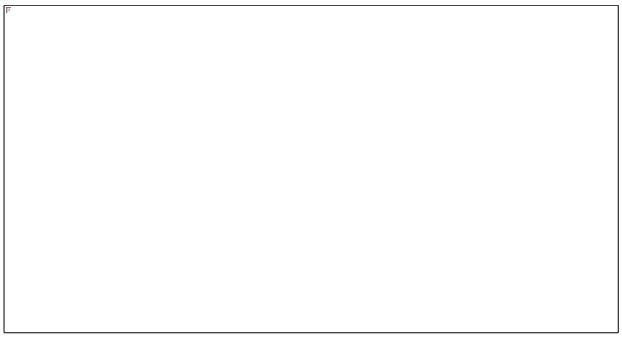

Figure 8.Comparative Histogram of L2 Norm Across Model Versions

The initial model without skip connections exhibited the highest L2 Norm, indicating a significant deviation from the ground truth images. Incorporating skip connections led to a noticeable reduction in L2 Norm, ensuring better preservation of structural details. The modified sketching approach further improved the performance by reducing the model's dependency on visual biases. Finally, transitioning to a Style GAN-based generator resulted in the lowest L2 Norm value (89.68), signifying the best image fidelity among all tested versions.

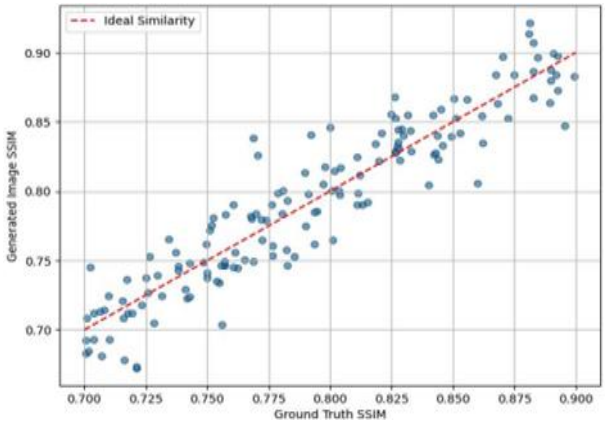

Figure 9. Scatter Plot of SSIM Values: Generated vs. Ground Truth.

A scatter plot of SSIM values for generated images versus ground truth images is depicted in Figure 9, demonstrating the effectiveness of the final model. The comparison between the generated images and the ground truth images is shown in Figure 10. This visual comparison allows for an assessment of the fidelity and realism of the generated images in relation to the actual images they represent. Each pair of images shows the ground truth image alongside its corresponding generated image, providing a side-by-side comparison of their visual attributes.

Prompt: - She has black hair, bags under eyes. This person has a light skin color.

ketch, generated output and ground truth: -

Figure 10 Comparison of generated images and ground truth images.

4. Conclusion

This research presented a novel system for high-quality image synthesis using a combination of sketch-based spatial features and text-based semantic prompts. The integration of CLIP for text encoding and Style GAN for generative modeling enabled the reconstruction of realistic images while preserving structural details from sketches. Through rigorous experimentation, we demonstrated the impact of different architectural modifications, such as the inclusion of skip connections and a refined sketch generation approach, in improving image fidelity. Quantitative evaluations using L2 Norm and SSIM confirmed the effectiveness of our approach, with a final L2 Norm value of 89.68 and an SSIM score of 0.788, indicating a strong resemblance between generated images and ground truth images.

The transition to Style GAN significantly enhanced image quality and reduced training time, making the system more efficient. However, challenges remain in balancing the model's reliance on textual descriptions versus sketches, particularly in cases where visual cues dominate semantic details. Future work will focus on improving the fusion of multimodal inputs to ensure a more balanced interpretation of sketch and text features. Additionally, extending this framework to diverse datasets and real-world applications, such as forensic sketch-to-image generation and digital art synthesis, could further enhance its practical value.Overall, the proposed system represents a significant advancement in multimodal image generation, demonstrating the potential of deep learning-based techniques in synthesizing high-fidelity images from minimal inputs.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ...&Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[2] Karras, T., Laine, S., &Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401-4410.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, &Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.

[4] Vinyals, O., Toshev, A., Bengio, S., &Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE conference on computer vision and pattern recognition, 3156-3164.

[5] Wang, X., Chen, L., Zhang, L., & Luo, J. (2019). Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia, 21(9), 2301-2313.

[6] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ...&Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092.

[7] Tang, Y., Liu, C., Wang, Z., & Zhang, J. (2017). Sketch-based image generation using statistical models and hand-crafted features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(5), 1215-1229.

[8] Radford, A., Metz, L., &Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[9] Karras, T., Laine, S., &Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401-4410.

[10] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. Proceedings of the 33rd International Conference on Machine Learning, 1060- 1069.

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ...&Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.

[12] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ...&Sutskever, I. (2021). Zero- shot text-to-image generation. arXiv preprint arXiv:2102.12092.

[13] Li, X., Chen, J., & Xu, Y. (2023). Sketch2Face-GAN: Transformer-enhanced GAN model for forensic sketch-to-image synthesis. International Journal of Computer Vision, 131(2), 345-360.

[14] Zhang, L., Huang, T., & Yuan, Y. (2024). Multimodal Forensic GAN: Identity-preserving sketch- to-image translation for forensic applications. IEEE Transactions on Forensic and Security, 19(3), 1054- 1068.

[15] Wang, Z., Bovik, A. C., Sheikh, H. R., &Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.

[16] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 586-595.

[17] Chen, Y., Lee, S., & Wang, L. (2023). Perceptual evaluation of GAN-generated forensic images using LPIPS metric. Journal of Visual Communication and Image Representation, 89, 102196.