Pertanika J. Sci. & Technol. 34 (1): 117 - 142 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Building upon Adaptive GAN Training: Dual-stage GANs for Enhanced Forensic Face Sketch Synthesis

Muhamad Faris Che Aminudin and Shahrel Azmin Suandi*

Intelligent Biometric Group, School of Electrical and Electronic Engineering, USM Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Pulau Pinang, Malaysia

ABSTRACT

The synthesis of forensic face sketches is a crucial component of law enforcement, assisting in suspect identification based on eyewitness accounts. Conventional approaches, such as relying on forensic artists or composite sketch software, often suffer from subjectivity and inefficiency, leading to inconsistencies in quality and accuracy. This research introduces a novel method leveraging a dual-stage Generative Adversarial Network (GAN) architecture, conditioned on textual descriptions, to automate the forensic sketch generation process. The first stage produces a preliminary sketch that establishes the foundational facial structure, while the second stage enhances the sketch with intricate details like facial hair and accessories. Additionally, an adaptive stop training mechanism is implemented to terminate training when the generator and discriminator exhibit stagnation, thereby optimising computational efficiency. By incorporating GloVe and LSTM embeddings for encoding textual descriptions, our model effectively interprets complex linguistic inputs. The proposed framework is assessed on forensic sketch datasets, demonstrating superior performance over traditional techniques both qualitatively and quantitatively. This approach not only streamlines forensic sketch creation but also improves accuracy and realism, positioning it as a valuable asset in criminal investigations.

Keywords: Artificial Intelligence, forensic face sketch, generative adversarial network, image synthesis, sketch images

ARTICLE INFO

Article history:
Received: 18 February 2025
Accepted: 06 August 2025
Published: 06 February 2026

DOI: https://doi.org/10.47836/pjst.34.1.06

E-mail addresses:
muhdfaris@student.usm.my (Muhamad Faris Che Aminudin)
shahrel@usm.my (Shahrel Azmin Suandi)
* Corresponding author

INTRODUCTION

Forensic face sketching is an essential tool in criminal investigations, aiding law enforcement in suspect identification based on eyewitness accounts (Mohana Kumar et al., 2023). Traditionally, forensic artists manually create sketches, but this process is

e-ISSN: 2231-8526

subjective, time-consuming, and heavily dependent on an artist's skill level. Witnesses may struggle with memory recall, and different artists may interpret descriptions inconsistently, leading to variations in the sketches (Natarajan et al., 2024). Furthermore, the scarcity of skilled forensic artists presents challenges, particularly for law enforcement agencies with limited resources.

In recent years, deep learning has emerged as a promising alternative to automate forensic sketch generation (Khatoon & Umar, 2022; Martis et al., 2024). Generative Adversarial Networks (GANs) have demonstrated remarkable capabilities in synthesising high-quality images by training a generator to create realistic images while a discriminator evaluates their authenticity (Reed et al., 2016). GANs have been extensively applied to tasks such as image synthesis (Colleoni et al., 2024; Liu et al., 2024; Yildiz et al., 2024), style transfer (Gao et al., 2024; Khowaja et al., 2024), and data augmentation. However, forensic sketch generation presents unique challenges, including the need to accurately capture fine facial details and map descriptive text to visual features.

Text-to-image synthesis has been explored in multiple domains, with models such as StackGAN (H. Zhang et al., 2017), AttnGAN (Xu et al., 2018), and DF-GAN (Tao et al., 2022) improving image quality and resolution through hierarchical generation and attention mechanisms. However, these models primarily focus on generating natural scenes rather than forensic sketches. Additionally, text-to-face synthesis methods like Text2FaceGAN (Nasir et al., 2019) and DLT-GAN (Chen et al., 2024) manipulate facial attributes in existing images rather than generating sketches from scratch. Existing methods for forensic sketch generation lack the precision required to capture subtle facial features such as scars, wrinkles, and specific facial structures, which are crucial for accurate suspect identification.

To address these limitations, we propose a dual-stage GAN architecture for forensic sketch synthesis. Our method generates sketches directly from textual descriptions and refines them through a two-stage process:

- **Stage I: Coarse Sketch Generation**

  In this stage, the model generates low-resolution sketches that capture the fundamental facial structure based on high-level textual descriptions.

- **Stage II: Fine Detail Refinement**

  Building upon the coarse sketches, this stage refines the images by adding fine-grained details such as scars, facial hair, and other distinguishing features, guided by detailed textual descriptions.

A key innovation in our approach is the adaptive stop training mechanism, which monitors the convergence of the generator and discriminator losses, halting training when performance stabilises. This mechanism optimises computational efficiency while

preventing overfitting. Additionally, our approach integrates Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) and Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) embeddings to transform textual descriptions into meaningful feature vectors, improving the model's ability to understand and generate realistic facial features.

This research introduces several significant contributions:

- **Dual-stage GAN Architecture**

  A two-stage GAN framework that effectively separates coarse and fine feature generation, enhancing the quality and realism of forensic sketches.

- **Advanced Text Embeddings**

  The use of GloVe and LSTM embeddings to transform textual descriptions into informative vectors that guide the GAN, improving alignment between descriptions and generated images.

- **Adaptive Stop Training Mechanism**

  A dynamic training termination process based on real-time monitoring of losses and FID evaluation, optimising training efficiency, and preventing overfitting.

- **Comprehensive Evaluation**

  Demonstrating significant improvements over existing methods through extensive qualitative and quantitative evaluations on the CUFS and CUFSF datasets.

The remainder of this paper details the proposed methodology, including data preparation, GAN architecture, text embedding strategies, and training mechanisms. Next part presents experimental results, evaluating the model's performance through qualitative and quantitative comparisons. Finally the conclusions and potential future research directions is discussed.

METHODS

In this section, a novel approach is presented for generating forensic sketches from textual descriptions using a dual-stage Generative Adversarial Network (GAN) architecture. The method is designed to address the challenges inherent in translating complex textual descriptions into detailed facial images, capturing both global structures and fine-grained details essential for accurate identification. The key components of our proposed method include comprehensive data preparation, the dual-stage GAN architecture, advanced text embedding strategies, and an adaptive stop training mechanism to optimise the training process.

Overview

Our approach consists of the following main components:

- **Data Preparation**: Creating a high-quality dataset with annotated facial attributes and corresponding textual descriptions.
- **Text Embedding**: Utilising GloVe embeddings and an LSTM network to encode textual descriptions into meaningful vectors.
- **Dual-stage GAN Architecture**: Designing a two-stage GAN where Stage-I generates a coarse sketch capturing basic facial structures, and Stage-II refines the sketch by adding fine-grained details.
- **Adaptive Stop Training Mechanism**: Implementing a dynamic training termination process based on monitoring of losses and the FID score to prevent overfitting and optimise computational resources.

An overview of our proposed method is illustrated in Figure 1.

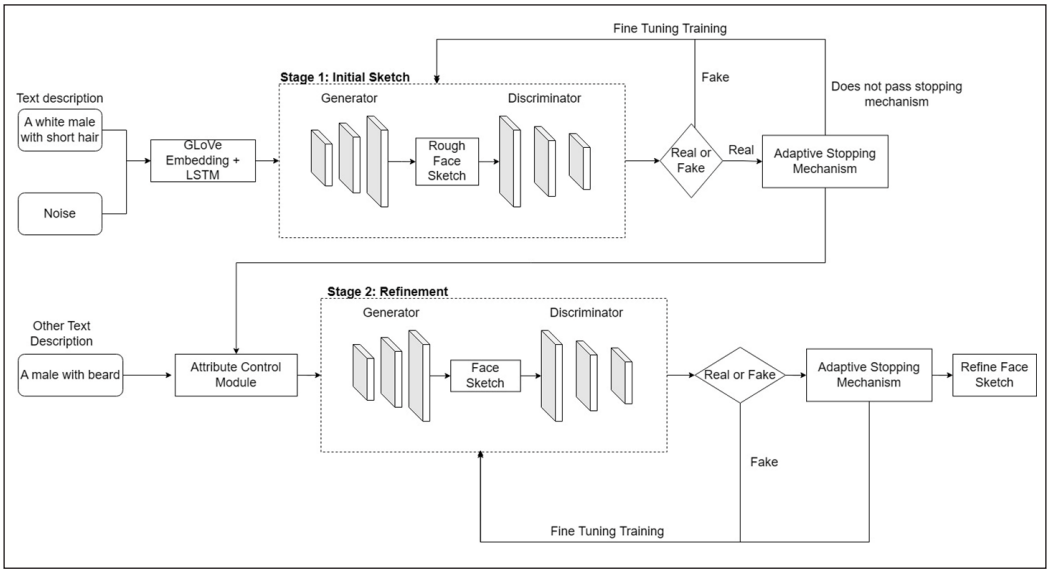

Figure 1. Overview of the proposed dual-stage GAN framework for forensic sketch synthesis conditioned based on textual description

Data Preparation

Effective data preparation is fundamental to the success of training robust GANs, particularly for specialised tasks like forensic sketch synthesis. Our data preparation process involves several critical steps.

Dataset Selection

The CUHK Face Sketch Database (CUFS) (X. Wang & Tang, 2009) and the CUHK Face Sketch FERET Database (CUFSF) (W. Zhang et al., 2011) were utilised due to their comprehensive collection of face sketches paired with corresponding photographs. CUFS contains 606 face images from the CUHK student database, the AR database, and the XM2VTS database, while CUFSF includes 1,194 face images from the FERET database. These datasets offer diversity in terms of ethnicity, age, gender, and facial features, which is crucial for training a model that generalises well across different demographics.

The combined dataset consists of 1,800 face sketches. The demographic distribution is as follows:

- Gender: Approximately 55% male and 45% female.
- Ethnicity: Includes Caucasian, Asian, African, and other ethnicities.
- Age Range: Predominantly between 18 to 60 years old.

Image Preprocessing

To ensure consistency across the training data, we performed several preprocessing steps. We used the Dlib library (King, 2009) to detect 68 facial landmarks for each sketch. These landmarks include key points around the eyes, nose, mouth, and facial outline. We aligned the faces based on the positions of the eyes and the centre of the mouth using an affine transformation. This process standardises the facial orientation and positioning, reducing variability due to head tilt or rotation.

After alignment as shown in Figure 2 we cropped the images to a standardised size centred on the face. The cropping window was determined

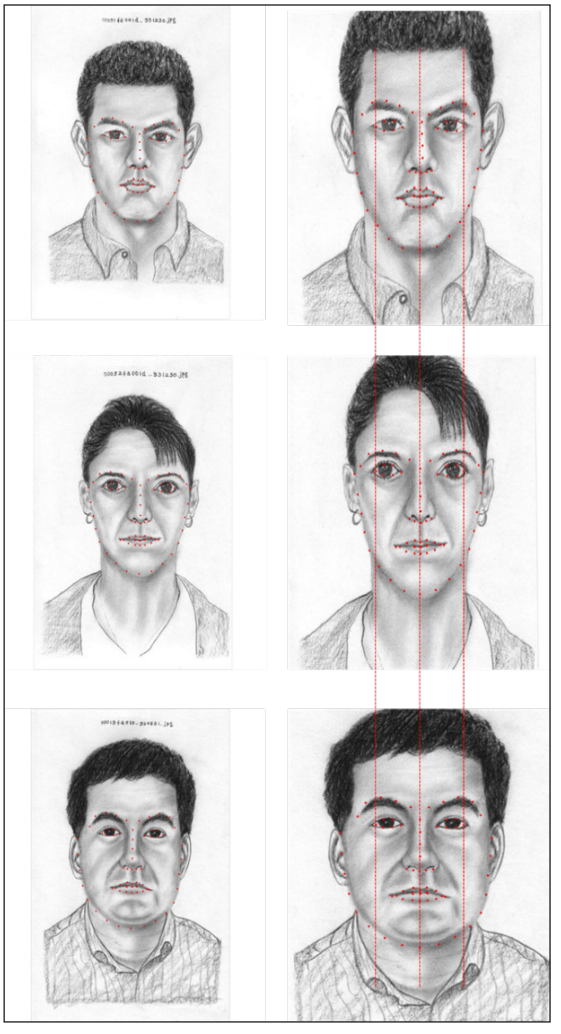

Figure 2. Image processing cropping and aligning all face sketches

based on the facial landmarks to include the entire face and some surrounding context (e.g., hair and ears). The cropped images were then resized to 128 × 128 pixels for Stage-II training and 64 × 64 pixels for Stage-I training using bicubic interpolation.

Pixel values were normalised to the range [-1, 1] to match the output range of the generator's activation functions (tanh). Normalisation is essential for stabilising the training process and ensuring that different input features contribute equally to the model's learning.

Facial Attribute Annotation

Accurate facial attribute annotation is crucial for generating meaningful textual descriptions. Six annotators independently labelled each face sketch with 47 predefined facial attributes as shown in Table 1. The attributes were selected based on their relevance to forensic identification and included both categorical and continuous variables.

Table 1
List of facial attributes for annotation

Gender	Eyeglasses	Nose Width
Ethnicity	Moustache	Nose Length
Age Group	Beard	Nose Shape
Face Shape	Sideburns	Mouth Width
Skin Tone	Earrings	Lip Thickness
Hair Colour	Necklace	Smile Presence
Hair Length	Hat	Facial Expression
Hair Style	Eye Colour	Eye Shape
Forehead Size	Eyebrow Thickness	Eyebrow Shape
Cheekbone Structure	Eyebrow Position	Facial Marks (e.g., scars)
Jawline	Eye Distance	Eye Orientation
Chin Size	Eye Bags	Eye Asymmetry
Ear Size	Eye Size	Facial Hair Density
Neck Thickness	Nose Bridge	Facial Accessories

Textual Description Generation

Using the annotated attributes, we generated detailed and coherent textual descriptions for each face sketch. The process involved:

A set of templates was developed to convert attribute combinations into natural language descriptions. The templates were designed to ensure variability and naturalness in the descriptions. For example:

- "The person is a young adult Asian female with a round face shape and fair skin tone. She has long, straight black hair parted in the middle. Her eyes are almond-shaped and

brown, with thin eyebrows positioned slightly above the eyes. She is smiling softly, revealing thin lips. She is wearing small earrings and no glasses."

- "A Caucasian male with a prominent jawline and wrinkled skin. He has short grey hair and a full beard. His blue eyes are deep-set under thick eyebrows. He wears round eyeglasses and has a noticeable scar on his left cheek."

To enhance the variability and richness of the descriptions, we employed natural language processing techniques:

- **Sentence Structure Variation**: Rearranging sentence components to create different sentence structures.
- **Attribute Emphasis**: Varying the emphasis on certain attributes, e.g., "He has a *very* prominent jawline."

We generated an average of 5 unique descriptions per face sketch, resulting in a total of approximately 9,000 textual descriptions.

Embedding Representation

To bridge the gap between textual descriptions and image generation, we converted the text into numerical vectors using advanced embedding techniques. Each word in the textual description was converted into a 300-dimensional GloVe vector (Pennington et al., 2014), which captures semantic relationships between words. GloVe embeddings were chosen due to their balance between performance and computational efficiency.

The sequence of word embeddings was processed by a single-layer LSTM network (Hochreiter & Schmidhuber, 1997; Voditel et al., 2023) with a hidden size of 256. The LSTM captures the sequential and contextual information of the text, allowing the model to understand the relationships between words in the description.

Mathematically, for a sequence of word embeddings $\{w_1, w_2, ..., w_t\}$, the Long Short-Term Memory (LSTM) network updates its hidden state (h_t) and cell state (c_t) at each time step t as follows:

$$i_t = \sigma(W_i w_t + U_i h_{t-1} + b_i) \quad [1]$$

$$f_t = \sigma\left(W_f w_t + U_f h_{t-1} + b_f\right) \quad [2]$$

$$o_t = \sigma(W_o w_t + U_o h_{t-1} + b_o) \quad [3]$$

$$\tilde{c_t} = \tanh(W_c w_t + U_c h_{t-1} + b_c) \quad [4]$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \quad [5]$$

$$h_t = o_t \odot \tanh(c_t) \quad [6]$$

where σ represents the sigmoid activation function, tanh is the hyperbolic tangent function, and ⊙ denotes element-wise multiplication.

The final hidden state (h_T) is used as the sentence embedding $t \in R^{256}$. To prevent the embedding magnitudes from becoming too large and destabilising training, we normalised the sentence embeddings to have unit length:

$$t \leftarrow \frac{t}{|t|} \quad [7]$$

Each face sketch was paired with its corresponding textual descriptions, forming the training pairs (t, I). The dataset was split into training, validation, and testing sets with a ratio of 70%, 15%, and 15%, respectively, ensuring that individuals in the test set were not present in the training set to evaluate the model's generalisation ability.

In practical forensic scenarios, eyewitness descriptions may be incomplete, ambiguous, or even contain conflicting information. Our approach is specifically designed to be robust to such real-world challenges. When presented with vague or partial descriptions where certain attributes are omitted or described in nonspecific terms, the model processes the available attributes and treats unspecified elements as unknown leveraging the sequential modeling capabilities of the LSTM-based text encoder to flexibly interpret partial input.

For contradictory descriptions, in which mutually exclusive attributes such as having in description saying have a beard and no beard, the system prioritises explicit, non-conflicting cues, and defaults to the most neutral or common representation for attributes marked as ambiguous or in conflict. This decision is handled during preprocessing, where attribute conflicts are resolved by setting them as undetermined, ensuring the model does not generate anatomically implausible or unrealistic sketches. Additionally, Gaussian noise vectors injected at both stages introduce further variability, which, when combined with multiple inference passes, allows the model to generate diverse but plausible outputs that reflect the inherent uncertainty in ambiguous witness accounts.

During inference, to capture this uncertainty, we sample the noise vector multiple times for each description and generate several candidate sketches. The output with the highest text-image semantic alignment score is selected as the final result. This strategy ensures that the model can accommodate the ambiguity often present in natural language descriptions, producing sketches that remain faithful to the input while minimising the impact of vagueness or contradictions.

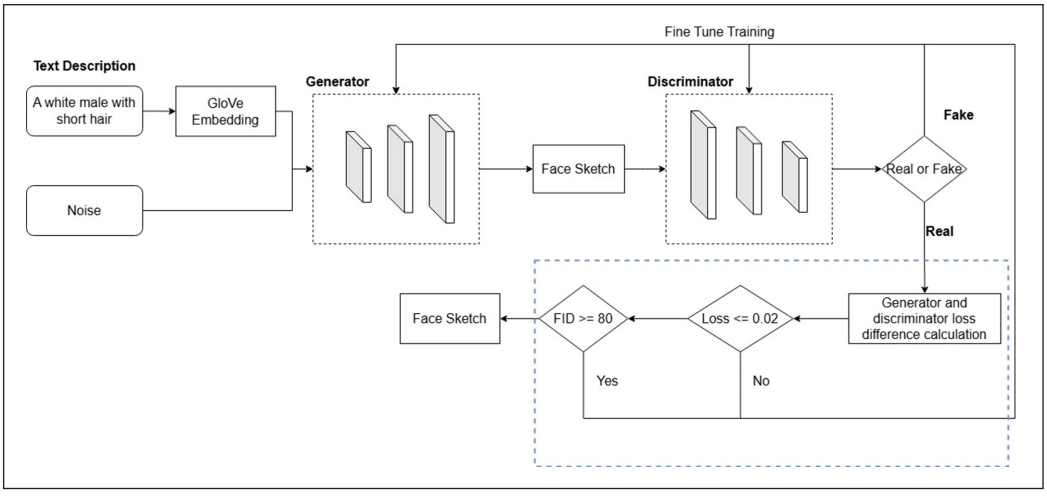

Figure 3. Stage-I Generator (G_1) and Discriminator (D_1) for coarse sketch generation

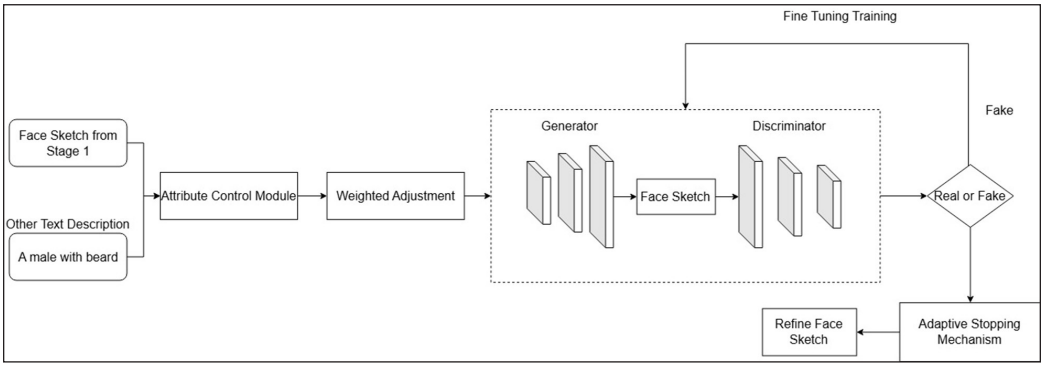

Figure 4. Architecture of Stage-II Generator (G_2) and Discriminator (D_2). The generator refines coarse sketches by adding fine details, while the discriminator evaluates the high-resolution images conditioned on the textual descriptions

Dual-stage GAN Architecture

Our framework improves upon existing single-stage approaches by employing a progressive two-stage generation process. The first stage constructs a coarse facial structure capturing the global features, while the second stage refines this output by adding fine-grained details. This separation enables the model to better capture both broad facial layouts and intricate attributes. Additionally, the adaptive stop training mechanism enhances training efficiency and prevents overfitting, addressing challenges that prior work often overlook.

The overall architecture is designed to generate high-resolution forensic sketches by progressively refining images through these two stages: coarse sketch generation followed by fine detail refinement. This architecture is illustrated in Figures 3 and 4.

Stage-I: Coarse Sketch Generation

In Stage-I, the generator (G_1) aims to produce a low-resolution coarse sketch that captures the basic facial structure based on the textual description. The generator (G_1) receives as input a concatenated vector of a noise vector $z \in R^{100}$ sampled from a normal distribution ($N(0,1)$) and the text embedding $t \in R^{256}$. This combined vector is projected and reshaped to form an initial feature map of size ($4 \times 4 \times 512$) . The generator then processes this feature map through a series of transposed convolutional layers (also known as deconvolutional layers) to progressively upscale the spatial dimensions to (64×64).

Each transposed convolutional layer is followed by batch normalisation (Ioffe & Szegedy, 2015) and ReLU activation, except for the final layer, which uses a tanh activation function to produce output pixel values in the range ([–1,1]). The architectural details of (G_1) are summarised in Table 2.

Table 2
Architecture of Stage-I Generator (G_1)

Layer	Kernel Size	Output Size	Activation
Fully Connected	-	4 × 4 × 512	ReLU
Transposed Conv	4 × 4, stride 2, pad 1	8 × 8 × 256	ReLU
Transposed Conv	4 × 4, stride 2, pad 1	16 × 16 × 128	ReLU
Transposed Conv	4 × 4, stride 2, pad 1	32 × 32 × 64	ReLU
Transposed Conv	4 × 4, stride 2, pad 1	64 × 64 × 3	Tanh

The discriminator (D_1) evaluates the authenticity of the generated images. It receives as input an image (I) either real or generated and the text embedding (t). The image is processed through a series of convolutional layers with spectral normalisation (Miyato et al., 2018) to stabilise training. LeakyReLU activation with a negative slope of 0.2 is used after each convolutional layer.

The text embedding (t) is projected through a fully connected layer and reshaped to match the spatial dimensions of the image features. It is then concatenated with the image features at an intermediate layer to condition the discriminator on the textual description. The architectural details of (D_1) are summarised in Table 3.

Table 3
Architecture of Stage-I Discriminator

Layer	Kernel Size	Output Size	Activation
Conv	4 × 4, stride 2, pad 1	32 × 32 × 64	LeakyReLU (0.2)
Conv	4 × 4, stride 2, pad 1	16 × 16 × 128	LeakyReLU (0.2)
Conv	4 × 4, stride 2, pad 1	8 × 8 × 256	LeakyReLU (0.2)
Concatenate text embedding	-	8 × 8 × 256	-
Conv	4 × 4, stride 2, pad 1	4 × 4 × 512	LeakyReLU (0.2)
Conv	4 × 4, stride 1, pad 0	1 × 1 × 1	Sigmoid

Stage-II: Fine Detail Refinement

In Stage-II as shown in Figure 4, the generator (G_2) refines the coarse sketch from Stage-I to produce a high-resolution image with fine details. The generator (G_2) takes as input the Stage-I output image (I_{S1}) a new noise vector ($z' \in R^{100}$) and the text embedding (t). The Stage-I image is processed through convolutional layers to extract image features. The text embedding and noise vector are projected and concatenated with the image features.

(G_2) employs residual blocks to allow for better gradient flow and to capture higher-level features. We use four residual blocks, each consisting of two convolutional layers with batch normalisation and ReLU activation. To focus on specific regions described in the text, we incorporate a self-attention mechanism. The self-attention layer allows the model to capture long-range dependencies and refine details by attending to relevant parts of the image.

Finally, transposed convolutional layers upscale the image to (128 × 128). The architectural details of (G_2) are summarised in Table 4.

Table 4
Architecture of Stage-II Generator (G_2)

Layer	Kernel Size	Output Size	Activation
Input Image Conv	3 × 3, stride 1, pad 1	64 × 64 × 64	ReLU
Concatenate text embedding and noise	-	64 × 64 × (64 + 256 + 100)	-
Residual Blocks (4 blocks)	-	64 × 64 × 512	ReLU
Self-Attention Layer	-	64 × 64 × 512	-
Transposed Conv	4 × 4, stride 2, pad 1	128 × 128 × 256	ReLU
Transposed Conv	4 × 4, stride 2, pad 1	256 × 256 × 128	ReLU
Conv	3 × 3, stride 1, pad 1	256 × 256 × 3	Tanh

The discriminator (D_2) is similar to (D_1) but adapted for higher-resolution images. It processes the input image through convolutional layers with spectral normalisation and incorporates the text embedding to condition its predictions. The architectural details of (D_2) are summarised in Table 5.

Table 5
Architecture of Stage-II Discriminator (D_2)

Layer	Kernel Size	Output Size	Activation
Conv	4 × 4, stride 2, pad 1	64 × 64 × 64	LeakyReLU (0.2)
Conv	4 × 4, stride 2, pad 1	32 × 32 × 128	LeakyReLU (0.2)
Conv	4 × 4, stride 2, pad 1	16 × 16 × 256	LeakyReLU (0.2)
Conv	4 × 4, stride 2, pad 1	8 × 8 × 512	LeakyReLU (0.2)
Concatenate text embedding	-	8 × 8 × (512 + 256)	-
Conv	4 × 4, stride 2, pad 1	4 × 4 × 1024	LeakyReLU (0.2)
Conv	4 × 4, stride 1, pad 0	1 × 1 × 1	Sigmoid

Loss Functions

We use the conditional adversarial for both stages to encourage the generator to produce realistic images that align with the textual descriptions. For generator (G_i) and discriminator (D_i) ((i = 1,2)), the adversarial loss functions are defined as:

- Generator Loss

$$\mathcal{L}_{G_i}^{adv} = -E_{z,t}[\log D_i\,(G_i(z,t),t)] \quad [8]$$

- Discriminator Loss

$$\mathcal{L}_{\mathcal{D}_i} = -E_{I_{real},t}[\log D_i\,(I_{real},t)] - E_{z,t}\left[\log\left(1 - D_i(G_i(z,t),t)\right)\right] \quad [9]$$

To enhance the semantic alignment between the generated images and the textual descriptions, we incorporate a matching-aware discriminator. The discriminator not only distinguishes between real and fake images but also evaluates whether the image matches the given text.

In Stage-II, we include a feature matching loss to stabilise training and encourage the generator to produce images with similar statistics to the real images.

$$\mathcal{L}_{G_2}^{\mathcal{FM}} = E_{I_{real},t,z}\left[\sum_{l=1}^{L}\frac{1}{N_l}|D_2^{(l)}(I_{real},t) - D_2^{(l)}(G_2(I_{S1},z,t),t)\,|_1\right] \quad [10]$$

where ($D_2^{(l)}$) represents the activation of the $(l) - th$ layer of discriminator (D_2), and (N_l) is the number of units in that layer. The total generator loss for Stage-II combines the adversarial loss and the feature matching loss:

$$\mathcal{L}_{G_2}^{total} = \mathcal{L}_{G_2}^{adv} + \lambda_{FM}\mathcal{L}_{G_2}^{\mathcal{FM}} \quad [11]$$

where (λ_{FM}) is a weighting factor set to 10 in our experiments.

Text-image Discriminator

To further ensure semantic consistency, we incorporate a text-image matching loss inspired by DAMSM (Deep Attentional Multimodal Similarity Model). The model learns to align word-level and sentence-level features between text and image. We compute the word-level matching loss using the cosine similarity between word embeddings and image region features, guided by an attention mechanism.

The sentence-level matching loss is computed by measuring the cosine similarity between the global image feature vector and the sentence embedding. The total matching loss combines both word-level and sentence-level losses:

$$\mathcal{L}_{\mathcal{DAMSM}} = \mathcal{L}_{word} + \mathcal{L}_{sent} \quad [12]$$

This loss encourages the generated images to be semantically aligned with the textual descriptions at both the word and sentence levels.

Adaptive Stop Training Mechanism

Training GANs requires careful balance to prevent overfitting or underfitting. We introduced an adaptive stop training mechanism that monitors the performance during training and halts the process when improvements become negligible.

Loss Monitoring

We track the generator and discriminator losses for both stages. If the absolute change in loss over a specified number of epochs (patience parameter) falls below a predefined threshold (ϵ_{loss}), training is stopped. We set (ϵ_{loss} = 0.002) and a patience of 5 epochs.

The Fréchet Inception Distance (FID) score is computed every 10 epochs. If the FID score does not improve by at least ($\epsilon_{FID} = 1.0$) over 3 evaluations, training is halted. This ensures that the quality of generated images is improving and prevents unnecessary computation when the model has converged.

Table 6
Training hyperparameters

Hyperparameter	Value
Learning Rate (Generator)	0.0002
Learning Rate (Discriminator)	0.0004
Optimiser	Adam
Adam β_1	0.5
Adam β_2	0.999
Batch Size	64
Weighting Factor (λ_{FM})	10
Weighting Factor (λ_{DAMSM})	5
Number of Epochs	Adaptive (up to 600)
Patience (Loss)	5 epochs
Patience (FID)	3 evaluations

The adaptive stop training mechanism is implemented as an early stopping callback in the training loop. It monitors the validation loss and FID score, and adjusts the learning rate if the improvement stagnates before halting training. The hyperparameters used during training are summarised in Table 6.

Training Procedure

Training is conducted separately for each stage. In Stage-I, (G_1) and (D_1) are trained jointly using the paired text embeddings and face sketches. After Stage-I converges (as determined

by the adaptive stop mechanism), we freeze the parameters of the text encoder and (G_1)to retain the learned representations.

In Stage-II, (G_2) and (D_2) are trained using the outputs from (G_1) and the same text embeddings. The feature matching loss are incorporated to enhance the quality of the refined images and ensure semantic alignment. The adaptive stop training mechanism is also applied in Stage-II. The overall training procedure is summarised in the Algorithm below:

Algorithm 1 Training Procedure for Dual-Stage GAN

Stage-I Training
while not converged **do**
  f**or** each minibatch do
    Sample $(\boldsymbol{t}, \boldsymbol{I_{real}})$
    Sample noise vector $\boldsymbol{z}$
    Generate fake image $\boldsymbol{I_{fake}} = G_1(\boldsymbol{z}, \boldsymbol{t})$
    Update (D_1) by minimizing $(\mathcal{L}_{\mathcal{D}_1})$
    Update (G_1) by minimizing $(\mathcal{L}_{\mathcal{G}_1}^{adv})$
  end for
  Apply adaptive stop training mechanism
end while
Freeze text encoder and (G_1)
Stage-II Training
while not converged **do**
  for each minibatch **do**
    Sample $(\boldsymbol{t}, \boldsymbol{I_{real}})$
    Sample noise vector $\boldsymbol{z}'$
    Generate Stage-I image $(\boldsymbol{I_{S1}} = G_1(\boldsymbol{z}, \boldsymbol{t}))$
    Generate fake image $(\boldsymbol{I_{fake}} = G_2(\boldsymbol{I_{S1}}, \boldsymbol{z}', \boldsymbol{t}))$
    Update (D_2) by minimizing $(\mathcal{L}_{\mathcal{D}_2})$
    Update (G_2) by minimizing $(\mathcal{L}_{\mathcal{G}_2}^{total})$
    Update (G_2) and text encoder by minimizing $(\lambda_{DAMSM}\mathcal{L}_{\mathcal{DAMSM}})$
  end for
  Apply adaptive stop training mechanism
end while

RESULTS AND DISCUSSION

In this section, we present a comprehensive evaluation of our proposed dual-stage GAN model for forensic sketch synthesis. We assess the model's performance using both quantitative metrics and qualitative analyses, comparing it with state-of-the-art approaches. The experiments are designed to demonstrate the effectiveness of our method in generating high-quality sketches that accurately reflect complex textual descriptions. Additionally, we conduct ablation studies to understand the contribution of each component of our model and perform a user study to evaluate the practical utility of the generated sketches.

Evaluation Metrics

To evaluate the performance of our model, we employ several widely used metrics in image generation tasks: the Inception Score (IS), the Fréchet Inception Distance (FID), the Structural Similarity Index Measure (SSIM), and the Learned Perceptual Image Patch Similarity (LPIPS). These metrics provide quantitative measures of the quality, diversity, and perceptual similarity of the generated images. In this subsection, we provide detailed explanations of each metric and justify their relevance to our task.

Inception Score (IS)

The Inception Score assesses the quality and diversity of generated images by utilising a pre-trained Inception v3 network (Szegedy et al., 2016). It evaluates how well the generated images can be classified into distinct categories, reflecting both image quality and diversity. A higher IS indicates that the generated images are clear the conditional label distribution $p(y|x)$ has low entropy and diverse the marginal distribution $p(y)$ has high entropy.

Mathematically, the Inception Score is calculated as:

$$\mathrm{IS} = \exp\left(E_{x\sim p_g}\left[D_{\mathrm{KL}}\left(p(y|x)|p(y)\right)\right]\right) \quad [13]$$

where D_{KL} denotes the Kullback-Leibler divergence, $p(y|x)$ is the conditional label distribution given the generated image x, and $p(y)$ is the marginal distribution over all classes. In our context, since the forensic sketches do not correspond to specific predefined classes, we adapt the Inception Score by grouping images into clusters based on facial attributes extracted by the Inception network, as suggested by Chong and Forsyth (2020).

Fréchet Inception Distance (FID)

The Fréchet Inception Distance (Chan & Sithungu, 2025; Heusel et al., 2017) measures the similarity between the distribution of generated images and real images by comparing their feature representations extracted from a pre-trained Inception network. FID considers both the mean and covariance of the feature representations, capturing differences in both content and style. Lower FID scores indicate that the generated images are closer to the real data distribution, implying higher fidelity and diversity.

The FID is calculated as:

$$\mathrm{FID} = |\mu_r - \mu_g|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \quad [14]$$

where μ_r and Σ_r are the mean and covariance of the real images' feature representations, and μ_g and Σ_g are those of the generated images. The square root of the covariance product is computed using the matrix square root.

Structural Similarity Index Measure (SSIM)

SSIM measures the perceptual similarity between two images, considering luminance, contrast, and structural information. It is designed to model the human visual system's perception of image quality (Z. Wang et al., 2004). SSIM values range from –1 to 1, with higher values indicating greater similarity.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad [15]$$

For two image patches x and y, SSIM is defined as:

where μ_x, μ_y are the mean pixel values, σ_x^2, σ_y^2 are the variances, σ_{xy} is the covariance between x and y, and C_1, C_2 are constant to stabilise the division with weak denominators.

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS evaluates the perceptual similarity between images using deep network features extracted from pre-trained models such as AlexNet, VGG, or SqueezeNet (R. Zhang et al., 2018). It computes the weighted L_2 distance between feature maps of the two images. LPIPS correlates well with human judgments of image similarity, especially in terms of perceptual quality. Lower LPIPS values indicate higher perceptual similarity.

$$\mathrm{LPIPS}(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left| w_l \odot \left(F_l^h(x) - F_l^h(y) \right) \right|_2^2 \quad [16]$$

Mathematically, LPIPS is computed as:

where $F_l^h(\cdot)$ represents the activation at layer l, spatial location (h,w), w_l is the learned weight vector for layer l, and H_l, W_l are the spatial dimensions of the feature maps.

Implementation Details

In this subsection, we provide a detailed description of the hardware and software environment, training hyperparameters, and the training procedure employed in our experiments.

Training Procedure

We trained the model for up to 600 epochs for both Stage-I and Stage-II, although the adaptive stop training mechanism often halted training earlier based on convergence criteria. On average, the adaptive mechanism stopped Stage-I training after 400 epochs and Stage-II after 480 epochs.

Each epoch involved processing all training samples once. For Stage-I, each epoch took approximately 30 minutes, while for Stage-II, each epoch took around 45 minutes due to the higher resolution and more complex architecture. The total training time when not stopped early was approximately 50 hours.

To ensure reproducibility, we fixed random seeds for NumPy and PyTorch using np.random.seed(42), torch.manual_seed(42). We also enabled deterministic behaviour in PyTorch by setting torch.backends.cudnn.deterministic = True, torch.backends.cudnn.benchmark = False. All experiments were repeated three times to account for random variations in training, and we report the mean and standard deviation of the metrics.

Quantitative Results

In this subsection, we present the quantitative results of our experiments. We compare our proposed method with several state-of-the-art models in text-to-image synthesis and forensic sketch generation, including StackGAN (H. Zhang et al., 2017), AttnGAN (Xu et al., 2018), and DF-GAN (Tao et al., 2022). We evaluate the models on the combined CUFS and CUFSF datasets, which provide a diverse set of facial sketches and textual descriptions.

Performance Comparison

The quantitative results are presented in Table 7. We report the Inception Score (IS), Fréchet Inception Distance (FID), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) for each model.

Our proposed method achieves the highest Inception Score of 16.32, indicating improved image quality and diversity compared to the baseline models. The FID score of 28.41 is significantly lower than that of StackGAN, AttnGAN, and DF-GAN, suggesting that our generated images are closer to the real data distribution. The SSIM and LPIPS

Table 7
Quantitative comparison of different models on the CUFS and CUFSF datasets

Model	IS ↑	FID ↓	SSIM ↑	LPIPS ↓
StackGAN	13.12 ± 0.05	49.5 ± 1.2	0.45 ± 0.02	0.35 ± 0.01
AttnGAN	13.85 ± 0.07	32.01 ± 1.0	0.52 ± 0.01	0.28 ± 0.01
DF-GAN	14.02 ± 0.06	29.01 ± 0.9	0.56 ± 0.02	0.24 ± 0.01
Proposed Method	16.32 ± 0.04	28.41 ± 0.8	0.61 ± 0.01	0.19 ± 0.01

scores also show that our method produces images with higher perceptual similarity to the ground truth sketches.

The improvements in IS and FID can be attributed to our dual-stage architecture, which progressively refines the images, and the advanced text embedding strategy that captures contextual information from the descriptions. The higher SSIM indicates that the structural content of the generated images closely matches that of the real images, which is crucial for forensic applications where specific facial features are important. The lower LPIPS score suggests that the perceptual quality of our images is higher, aligning better with human judgments of similarity.

Qualitative Results

In this subsection, we present qualitative comparisons of the images generated from different models. Visual examples are crucial for assessing the perceptual quality and fidelity of the generated sketches, especially in forensic contexts where subtle facial features are important.

Visual Comparisons

Figure 5 shows examples of generated images from our proposed method, along with the corresponding textual descriptions. Each example highlights different facial attributes and complexities in the descriptions.

Our model generates images with sharper features, better facial structure, and finer details such as facial hair, and accessories. For instance, when the description mentions "a short face Asian female with trapezoid face shape, they are not wearing glasses, with moustache and beard" our model accurately captures these attributes, whereas the baseline models may miss some details or produce blurry features.

In cases where the description includes accessories like glasses or earrings, our model effectively

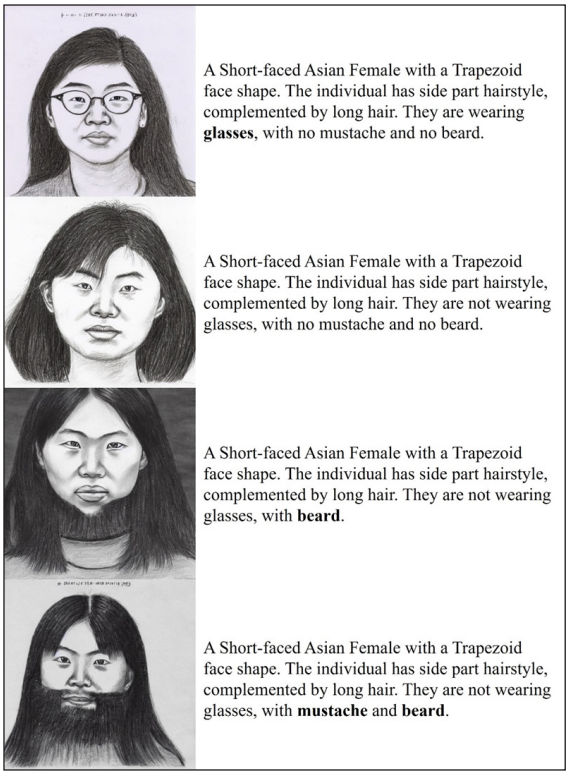

Figure 5. Qualitative comparison of generated images from different models. Our proposed method produces images with more accurate facial features and finer details that closely align with the textual descriptions

incorporates them into the generated images. The self-attention mechanism in Stage-II allows the model to focus on specific regions, enhancing the rendering of fine-grained details.

Importance in Forensic Context

The ability to generate sketches that closely match textual descriptions is critical in forensic applications. Accurate depiction of facial features can significantly aid in the identification of suspects. Our qualitative results demonstrate that our model is more effective in capturing essential features, which could enhance the utility of forensic sketches in investigations.

Ablation Studies

To understand the contribution of each component of our model, we conducted ablation studies by systematically removing or modifying parts of the model and observing the impact on performance.

Effect of Text Embedding Strategies

The performance of different text embedding methods was compared to evaluate the impact of our proposed GloVe + LSTM embedding. The results are presented in Table 8.

Table 8
Impact of different text embedding strategies

Embedding Strategy	IS ↑	FID ↓
GloVe (Average Pooling)	13.98 ± 0.05	102.3 ± 0.9
FastText	14.01 ± 0.06	93.4 ± 0.8
GloVe + LSTM (Proposed)	14.25 ± 0.04	68.7 ± 0.8

Our proposed GloVe + LSTM embedding outperforms the other strategies, highlighting the importance of capturing sequential and contextual information in textual descriptions. The LSTM processes the word embeddings sequentially, allowing the model to understand the relationships between words and phrases, which is essential for accurately mapping descriptions to facial features.

To further support these findings, we analysed the discriminator and generator loss curves for all three models, as shown in Figure 6. The LSTM-based model demonstrates smoother and faster convergence compared to both GloVe (with average pooling) and FastText, with the generator loss stabilising more effectively after epoch 500. This indicates that the GloVe + LSTM-based model is more efficient in learning to generate realistic images.

One of the key factors contributing to the superior performance of the LSTM-based GAN is its ability to capture sequential dependencies within the textual descriptions. LSTMs process each word in a sequence, retaining information from previous words,

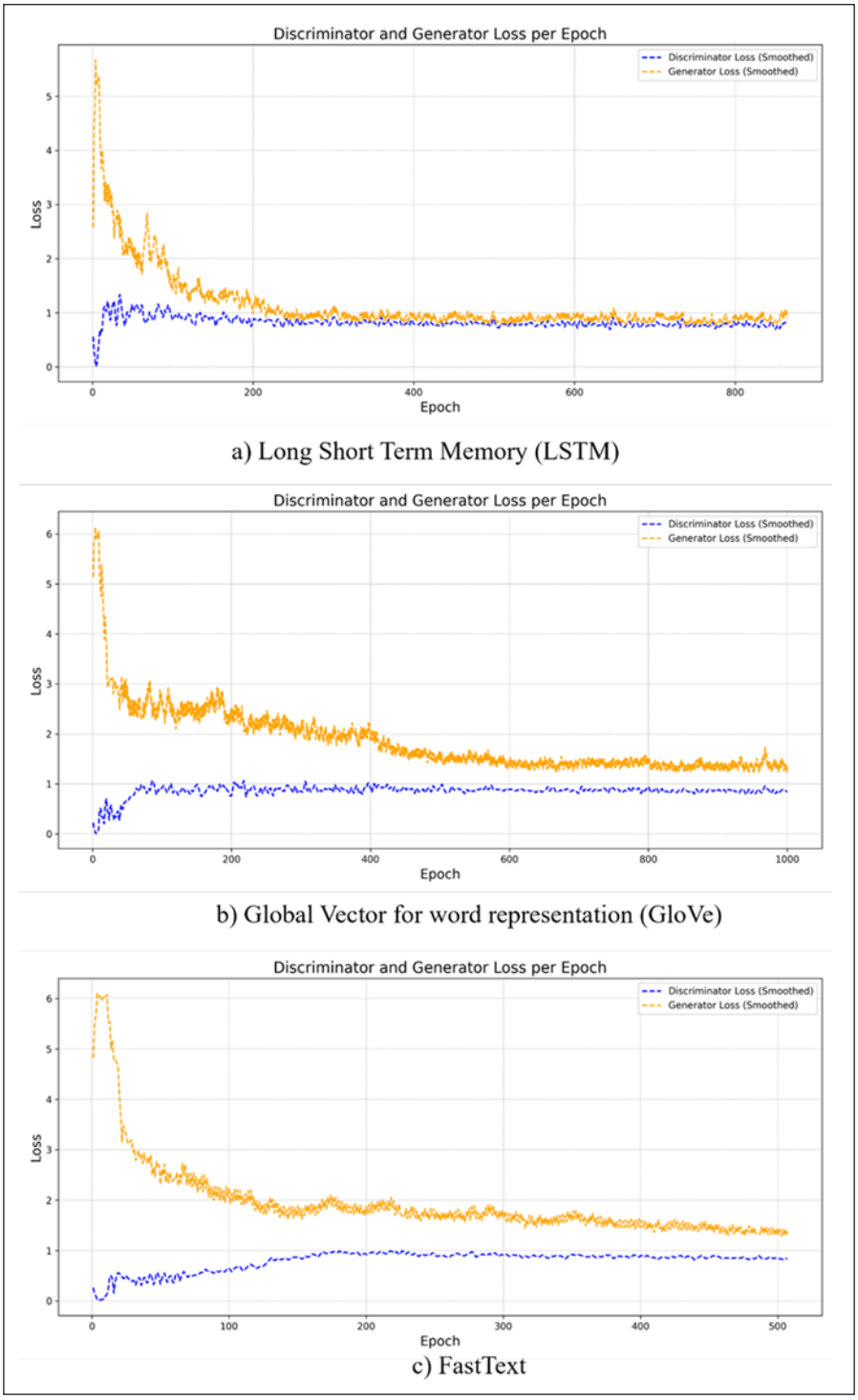

Figure 6. Discriminator and generator loss curves for LSTM-based GAN (top), GloVe-based GAN (middle), and FastText-based GAN (bottom)

which is crucial for understanding the full context of the description. This allows the GAN to generate more detailed and accurate sketches that reflect the nuances present in the text.

In contrast, GloVe embeddings with average pooling compress the entire description into a single vector by averaging the embeddings of individual words. This approach may lose important contextual information, especially in complex sentences where word order and dependencies matter. Similarly, while FastText embeddings capture subword information and can handle out-of-vocabulary words better, they still lack the ability to model sequential dependencies explicitly.

Figure 6 shows that the LSTM-based GAN (top) achieves lower discriminator and generator losses compared to the GloVe-based GAN (middle) and the FastText-based GAN (bottom). The loss curves of the LSTM-based GAN decrease more steadily and stabilise at lower values, indicating a more stable and effective training process. The lower Fréchet Inception Distance (FID) score of 68.7 achieved by the GloVe + LSTM embedding indicates that the images generated by this model are closer to the real data distribution compared to the other methods. The higher Inception Score (IS) of 14.25 also reflects the improved quality and diversity of the generated images.

In summary, the GloVe + LSTM-based GAN provides the best performance for generating realistic forensic sketches from textual descriptions. Its ability to capture sequential dependencies offers a significant advantage over both GloVe with average pooling and FastText embeddings, as reflected by its superior IS and FID scores, as well as the improved convergence observed in the loss curves. These findings highlight the importance of selecting an appropriate text encoder that can effectively model the complexity and context of textual descriptions in text-to-image synthesis tasks, particularly when working with detailed and context-dependent descriptions like those used in forensic sketch generation.

Effect of Adaptive Stop Training Mechanism

The impact of the adaptive stop training mechanism was evaluated by training the model with and without it. The results are shown in Table 9.

Table 9
Impact of adaptive stop training mechanism

Training Method	Epochs	IS ↑	FID ↓
Without Adaptive Stop	600	14.10 ± 0.05	49.3 ± 0.9
With Adaptive Stop (Proposed)	480 (average)	16.32 ± 0.04	28.4 ± 0.8

The adaptive stop mechanism not only reduced the training time by approximately 20% but also improved the model's performance by preventing overfitting. Without adaptive

stopping, the model continued to train beyond the point of convergence, leading to slight degradation in performance due to overfitting to the training data.

Effect of Dual-stage Architecture

The performance of our dual-stage GAN was compared with a single-stage GAN using the same total number of parameters. The single-stage GAN was designed to generate images directly at the final resolution. The results are presented in Table 10.

Table 10
Impact of dual-stage architecture

Architecture	IS ↑	FID ↓
Single-Stage GAN	9.783 ± 0.06	32.0 ± 1.0
Dual-Stage GAN (Proposed)	16.32 ± 0.04	28.4 ± 0.8

Table 11
Stage-wise quantitative evaluation on the CUFS and CUFSF datasets

Architecture	IS ↑	FID ↓
First-Stage GAN Output	9.51 ± 0.02	52.3 ± 1.1
Second-Stage GAN Output	16.32 ± 0.04	28.4 ± 0.8

The dual-stage architecture significantly outperforms the single-stage GAN, demonstrating the effectiveness of progressively refining the images. The first stage focusses on capturing the global facial structure, while the second stage refines the image by adding fine-grained details. This hierarchical approach allows the model to handle complex mappings from text to image more effectively.

Stage-wise Evaluation and Visualisation

To provide a clearer understanding of the dual-stage generation process, we independently evaluated and visualised the outputs after each stage of the model. Figure 7 illustrates representative examples of the progressive synthesis process. The Stage-I generator produces a structurally consistent, low-resolution sketch that establishes the basic facial layout, while Stage-II enhances this preliminary output by adding fine-grained details such as facial hair, glasses, scar, and subtle texture.

Quantitative evaluation was also performed at each stage. Table 11 reports the Inception Score (IS) and Fréchet Inception Distance (FID) for both Stage-I and Stage-II outputs. As expected, Stage-II demonstrates substantial improvements across all metrics, confirming that fine detail refinement contributes significantly to overall sketch realism and alignment with ground truth.

These results highlight the critical role of hierarchical generation: while Stage-I provides a strong structural foundation, Stage-II is essential for achieving high-fidelity, perceptually convincing sketches suitable for forensic applications.

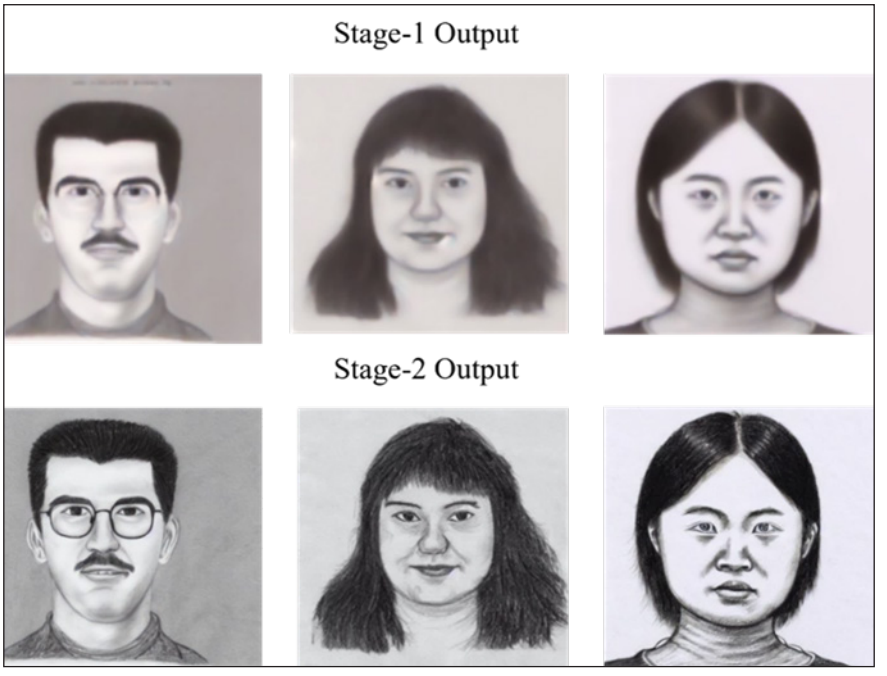

Figure 7. Representative outputs showing Stage-I (coarse) sketch and Stage-II (refined) sketch

Effect of Self-attention Mechanism

We evaluated the impact of the self-attention mechanism in Stage-II by training the model without it. The results are shown in Table 12.

Table 12
Impact of self-attention mechanism

Configuration	IS ↑	FID ↓
Without Self-attention	8.12 ± 0.05	41.5 ± 0.9
With Self-attention (Proposed)	16.32 ± 0.04	28.4 ± 0.8

Including the self-attention mechanism improves the model's ability to capture long-range dependencies and focus on relevant regions specified in the textual descriptions. This leads to better incorporation of fine details and enhances the overall image quality.

CONCLUSION

A novel dual-stage GAN architecture with an adaptive training mechanism was presented for forensic sketch synthesis guided by textual descriptions. Our approach effectively captures both global facial structures and fine-grained details, outperforming state-of-the-art methods in generating realistic and accurate forensic sketches.

The dual-stage architecture allows the model to first generate a coarse sketch capturing the basic facial structure and then refine it by adding detailed features, resulting in higher-

quality images. The adaptive training mechanism optimises computational efficiency by halting training when improvements plateau, preventing overfitting.

Our experiments demonstrated the effectiveness of GloVe + LSTM embeddings in capturing the contextual and sequential nature of textual descriptions, leading to more accurate image generation. The ablation study confirmed the significance of each component in our model.

While our model shows significant improvements, it relies on the quality and detail of textual descriptions. Future work will focus on handling ambiguous or less detailed descriptions by incorporating advanced natural language processing techniques. Optimising the model for real-time sketch generation in practical investigative scenarios is another area for exploration. Additionally, addressing potential biases in data and ensuring responsible deployment in forensic contexts are critical for ethical considerations.

Our proposed dual-stage GAN with adaptive training represents a significant advancement in forensic sketch generation. It offers law enforcement agencies a reliable and efficient tool for producing detailed and accurate forensic sketches based on witness descriptions.

ACKNOWLEDGEMENT

The authors acknowledge Universiti Sains Malaysia for providing the facilities and resources that supported this research.

REFERENCES

Chan, D. A., & Sithungu, S. P. (2025). Evaluating the suitability of inception score and Fréchet inception distance as metrics for quality and diversity in image generation. In *Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems* (pp. 79-85). Association for Computing Machinery. https://doi.org/10.1145/3708778.3708790

Chen, Z., Pang, Y., Jin, S., Qin, J., Li, S., & Yang, H. (2024). DLT-GAN: Dual-layer transfer generative adversarial network-based time series data augmentation method. *Electronics*, *13*(22), 4514. https://doi.org/10.3390/electronics13224514

Chong, M. J., & Forsyth, D. (2020). Effectively unbiased FID and inception score and where to find them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6069-6078). IEEE. https://doi.org/10.1109/CVPR42600.2020.00611

Colleoni, E., Sanchez Matilla, R., Luengo, I., & Stoyanov, D. (2024). Guided image generation for improved surgical image segmentation. *Medical Image Analysis*, *97*, 103263. https://doi.org/10.1016/j.media.2024.103263

Gao, H., Yang, X., Hu, Y., Wang, B., Xu, H., Liang, Z., Mu, H., Wang, Y., & Chen, Y. (2024). GANs-generated synthetic datasets for face alignment algorithms in complex environments. *Applied Soft Computing*, *167*, 112260. https://doi.org/10.1016/j.asoc.2024.112260

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6629-6640). Curran Associates Inc.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ioffe, S., & Szegedy, C. (2015). Batch normalisation: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 448-456). PMLR.

Khatoon, S., & Umar, M. S. (2022). Forensic sketch-to-photo transformation with improved Generative Adversarial Network (GAN). In *5th International Conference on Multimedia, Signal Processing, and Communication Technologies* (pp. 1-5). IEEE. https://doi.org/10.1109/IMPACT55510.2022.10029068

Khowaja, S. A., Nkenyereye, L., Mujtaba, G., Lee, I. H., Fortino, G., & Dev, K. (2024). FISTNet: FusIon of STyle-path generative Networks for facial style transfer. *Information Fusion*, *112*, 102572. https://doi.org/10.1016/j.inffus.2024.102572

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10*, 1755-1758.

Liu, Y., Li, Q., & Sun, Z. (2024). One-shot face reenactment with dense correspondence estimation. *Machine Intelligence Research*, *21*, 941-953. https://doi.org/10.1007/s11633-023-1433-9

Martis, J. E., Sannidhan, M. S., Pratheeksha Hegde, N., & Sadananda, L. (2024). Precision sketching with de-aging networks in forensics. *Frontiers in Signal Processing*, *4*, 1355573. https://doi.org/10.3389/frsip.2024.1355573

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalisation for generative adversarial networks. In *ICLR 2018 6th International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=B1QRgziT-

Mohana Kumar, S., Sowmya, B. J., Kavitha, H., Dayananda, P., Manjunath, R., Supreeth, S., & Shruthi, G. (2023). A deep learning-based approach for identification and recognition of criminals. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(3), 975-987.

Nasir, O. R., Jha, S. K., Grover, M. S., Yu, Y., Kumar, A., & Shah, R. R. (2019). Text2FaceGAN: Face generation from fine grained textual descriptions. In *IEEE Fifth International Conference on Multimedia Big Data* (pp. 58-67). IEEE. https://doi.org/10.1109/BigMM.2019.00-42

Natarajan, R., Mahadev, N., Gupta, S. K., & Alfurhood, B. S. (2024). An investigation of crime detection using artificial intelligence and face sketch synthesis. *Journal of Applied Security Research*, *19*(4), 542-559. https://doi.org/10.1080/19361610.2024.2302237

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1060-1069). PMLR.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818-2826). IEEE. https://doi.org/10.1109/CVPR.2016.308

Tao, M., Hao, Z., Zhang, Y., Tan, J., Wu, X., Zhao, D., & Yan, R. (2022). *DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis*. arXiv. https://doi.org/10.48550/arXiv.2008.05865

Voditel, P., Gurjar, A., Pandey, A., Jain, A., Sharma, N., & Dubey, N. (2023). Image captioning - A deep learning approach using CNN and LSTM network. In *3rd International Conference on Pervasive Computing and Social Networking* (pp. 343-348). IEEE. https://doi.org/10.1109/ICPCSN58827.2023.00062

Wang, X., & Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(11), 1955-1967. https://doi.org/10.1109/TPAMI.2008.222

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. https://doi.org/10.1109/TIP.2003.819861

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1316-1324). IEEE. https://doi.org/10.1109/CVPR.2018.00143

Yildiz, E., Yuksel, M. E., & Sevgen, S. (2024). A single-image GAN model using self-attention mechanism and DenseNets. *Neurocomputing, 596*, 127873. https://doi.org/10.1016/j.neucom.2024.127873

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision* (pp. 5908-5916). IEEE. https://doi.org/10.1109/ICCV.2017.629

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 586-595). IEEE. https://doi.org/10.1109/CVPR.2018.00068

Zhang, W., Wang, X., & Tang, X. (2011). Coupled information-theoretic encoding for face photo-sketch recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 513-520). IEEE. https://doi.org/10.1109/CVPR.2011.5995324