
RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

*Project Phase II Report On*

# Face Generation and Recognition in Forensic Science

*Submitted in partial fulfillment of the requirements for the award of the degree of*

## Bachelor of Technology

*in*

***Computer Science and Engineering***

**By**

**Jeffin Jitto (U2003100 )**

**Under the guidance of**

**Ms. Jisha Mary Jose**

**Department of Computer Science and Engineering**
**Rajagiri School of Engineering & Technology (Autonomous)**
**(Parent University: APJ Abdul Kalam Technological University)**
**Rajagiri Valley, Kakkanad, Kochi, 682039**
**May 2024**

# CERTIFICATE

*This is to certify that the project report entitled* ***"Face Generation and Recognition in Forensic Science"*** *is a bonafide record of the work done by* ***Jeffin Jitto (U2003100)*** *submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

**Ms. Jisha Mary Jose**
Asst. Professor
Dept. of CSE
RSET

**Dr. Tripti C.**
Associate Professor
Dept. of CSE
RSET

**Dr. Preetha K. G.**
Head of the Department
Professor
Dept. of CSE
RSET

# ACKNOWLEDGMENT

I wish to express my sincere gratitude towards **Dr. P. S. Sreejith**, Principal of RSET, and **Dr. Preetha K. G.**, Head of the Department of "Computer Science and Engineering" for providing me with the opportunity to undertake this project, "Face Generation and Recognition in Forensic Science".

I am highly indebted to my project coordinator, **Dr. Tripti C.** , Associate Professor, Department of CSE, for her valuable support.

It is indeed my pleasure and a moment of satisfaction for me to express my sincere gratitude to my project guide **Ms. Jisha Mary Jose**, Asst. Professor, Department of CSE, for her patience and all the priceless advice and wisdom she has shared with me.

Last but not the least, I would like to express my sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

**Jeffin Jitto**

# Abstract

Suspect identification can be challenging for forensic investigations since standard procedures are time-consuming and prone to mistakes. This calls for the creation of novel approaches utilizing developments in machine learning (ML) and artificial intelligence (AI). In order to overcome these obstacles, the proposed Forensic Face Creation and Recognition project will make use of sophisticated recognition algorithms and AI-based face generation models. The goal of the research is to create high-quality face images from textual descriptions by applying a fully trained Generative Adversarial Network (GAN) to text-to-image synthesis.

Image Generation, Text Guided Image Manipulation using Denoising Diffusion Probabilistic Models (DDPMs), and Dataset Matching are the three primary components of the process. Using a stable diffusion model, Image Generation quickly creates high-resolution images from word prompts by combining an autoencoder (VAE), U-Net, and text encoder. With the introduction of an alternate noise space for DDPMs, Text Guided picture Manipulation makes it possible to do meaningful picture altering tasks in response to text prompts. Convolutional neural networks (CNNs) are used in dataset matching to extract features and calculate similarity, which makes dataset alignment and comparison easier.

The suggested methodology gives law enforcement authorities effective tools for identifying suspects, which represents a substantial development in forensic investigations. The project intends to increase the efficiency of criminal investigations, accelerate the matching process with large datasets, and enhance the accuracy of facial sketches by utilizing AI and ML approaches. The approach's ability to produce coherent and contextually relevant face images is validated by experimental results, which also show the approach's potential for speeding up the conclusion of criminal cases, particularly unsolved cold cases. All things considered, the Forensic Face Creation and Recognition project is a promising first step in strengthening forensic science's technological innovation capabilities.

# Contents

**Acknowledgment** **i**

**Abstract** **ii**

**List of Figures** **vi**

**1 Introduction** **1**

1.1 Background . . . . . 1

1.2 Problem Definition . . . . . 1

1.3 Scope and Motivation . . . . . 1

1.4 Objectives . . . . . 2

1.5 Challenges . . . . . 2

1.6 Assumptions . . . . . 3

1.7 Societal / Industrial Relevance . . . . . 3

1.8 Organization of the Report . . . . . 3

1.9 Summary of the Chapter . . . . . 4

**2 Literature Survey** **5**

2.1 High-Resolution Image Synthesis with Latent Diffusion Models[1] . . . . . 5

2.2 A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks[2] . . . . . 5

2.3 E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs [7] . . . . . 8

2.4 DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN. [11] . . . . . 9

2.5 Attention-Modulated Triplet Network for Face Sketch Recognition. [16] . . 10

2.6 Comparison . . . . . 12

2.7 Summary of the Chapter . . . . . 13

**3 Requirements** **14**

3.1 Hardware Requirements . . . . . 14

3.2 Software Requirements . . . . . 14

3.3 Functional Requirements . . . . . 16

3.4 Summary of the Chapter . . . . . 17

**4 System Architecture** **18**

4.1 System Overview . . . . . 18

4.2 Architecture Diagram of the System . . . . . 19

4.3 Sequence Diagram . . . . . 20

4.4 Module Division . . . . . 20

4.4.1 Stable Diffusion Model: . . . . . 20

4.4.2 Denoising Diffusion Probabilistic Module: . . . . . 22

4.4.3 Database Matching using VGG 16 model: . . . . . 23

4.5 Work Schedule - Gantt Chart . . . . . 25

4.6 Work Breakdown and Responsibilities . . . . . 25

4.7 Summary of the Chapter . . . . . 26

**5 System Implementation** **27**

5.1 Datasets Identified . . . . . 27

5.2 Proposed Methodology/Algorithms . . . . . 28

5.2.1 Stable Diffusion model . . . . . 28

5.2.2 Denoising Diffusion Probabilistic model(DDPM) . . . . . 28

5.2.3 VGG16 convolutional neural network(CNN) . . . . . 29

5.3 User Interface Design . . . . . 29

5.4 Database Design . . . . . 30

5.5 Conclusion . . . . . 32

5.6 Summary of the Chapter . . . . . 32

**6 Results and Discussions** **33**

6.1 Overview . . . . . 33

6.2 Testing . . . . . 33

6.3 Graphical Analysis . . . . . 36

6.4 Discussion . . . . . 36
6.5 Summary of the Chapter . . . . . 37

**7 Conclusions & Future Scope** **38**
7.1 Future Scope . . . . . 38
7.2 Conclusion . . . . . 38
7.3 Summary of the chapter . . . . . 39

**Appendix A: Presentation** **43**

**Appendix B: Research Paper** **61**

**Appendix C: Vision, Mission, Programme Outcomes and Course Outcomes** **67**

**Appendix D: CO-PO-PSO Mapping** **72**

# List of Figures


2.1 Design of a Fully Trained GAN . . . . . 6
2.2 Encoder . . . . . 7
2.3 Decoder . . . . . 7
2.4 Discriminator . . . . . 8
2.5 Architecture diagram of E2F-GAN . . . . . 9
2.6 Network Architecture of SC-StyleGAN . . . . . 10
2.7 Architecture diagram of Attention Modulated Triplet Network . . . . . 11
2.8 Comparison . . . . . 12

4.1 Architecture Diagram . . . . . 19
4.2 Sequence Diagram . . . . . 20
4.3 Architecture diagram of Stable Diffusion Model . . . . . 22
4.4 DDPM . . . . . 23
4.5 Database Matching using Siamese networks . . . . . 23
4.6 Gantt Chart . . . . . 25
4.7 Work Breakdown and Responsibilities . . . . . 25

5.1 Login Page . . . . . 30
5.2 UI for entering the textual description . . . . . 30
5.3 Image Manipulation . . . . . 31
5.4 Image Manipulation . . . . . 31
5.5 Dataset Matching . . . . . 32

6.1 Image Generation: A blonde women with blue eyes wearing a scarf . . . . 33
6.2 Image manipulation: A Bald women . . . . . 34
6.3 Image manipulation: An Indian women . . . . . 34
6.4 Image Manipulation: A black haired women with brown eyes . . . . . 35
6.5 Database Matching . . . . . 35

# Chapter 1

# Introduction

## 1.1 Background

Traditional hand-drawn facial sketches in forensic science are time-consuming and subject to potential inaccuracies, highlighting a need for more efficient and accurate methods of suspect identification. High-profile criminal cases have underscored the limitations of traditional methods, emphasizing the need for innovative solutions that align with contemporary technological expectations and improve the overall efficiency of forensic investigations. Recent advancements in artificial intelligence (AI) and machine learning (ML) offer an unprecedented opportunity to transform facial generation and recognition technologies, leveraging vast datasets and sophisticated algorithms. The project aims to expedite the resolution of criminal cases, including cold cases that have remained unsolved, by providing law enforcement with more reliable tools for suspect identification.

## 1.2 Problem Definition

In the field of forensic science, the method of producing hand-drawn facial sketches remains time-intensive and is often inaccurate. Challenges in both sketching and recognition techniques can impede the identification of suspects and hinder the effectiveness of criminal investigations, highlighting the need for innovative solutions and improved methodologies for facial sketching and Recognition.

## 1.3 Scope and Motivation

The project aims to enhance and implement state-of-the-art facial recognition algorithms to ensure precise matching of generated sketches with existing databases, thereby improving the efficiency and accuracy of suspect identification. Moreover it encourage col-

laboration between forensic science agencies, law enforcement, and technology experts to foster knowledge exchange, share resources, and establish best practices in the development and implementation of improved facial generation and recognition methodologies. In the field of forensic science, the method of producing hand-drawn facial sketches remains time-intensive and is often inaccurate. Challenges in both sketching and recognition techniques can impede the identification of suspects and hinder the effectiveness of criminal investigations, highlighting the need for innovative solutions and improved methodologies for facial sketching and Recognition.

## 1.4 Objectives

- Develop and implement modern technologies that improves the precision of hand-drawn facial sketches, reducing inaccuracies and ensuring a more realistic representation of facial features.
- Automate or semi-automate the facial sketching process to expedite generation, reduce human error, and enhance efficiency based on eyewitness accounts.
- Integrate advanced artificial intelligence (AI) and machine learning (ML) algorithms for generating realistic and diverse facial images, enabling the creation of accurate composite sketches resembling actual suspects.
- Improve facial recognition algorithms to ensure accurate matching with existing databases, facilitating quicker and more reliable suspect identification.
- Establish standardized protocols for facial sketching and recognition processes to ensure consistency, reliability, and compatibility across forensic agencies and jurisdictions.
- Develop techniques accounting for variations in age progression and different facial expressions to maintain effectiveness over time and under diverse circumstances.

## 1.5 Challenges

- Overcoming the challenge of accurately portraying intricate facial features is crucial for the success of the forensic face generation and recognition project.

- Adapting the technology to account for the natural variations in age progression and diverse facial expressions presents a significant hurdle in achieving reliable and adaptable facial recognition.
- Ensuring effective collaboration between forensic science experts, law enforcement, and technology specialists is critical for the success of the project.

## 1.6 Assumptions

- The eye witness has a clear image of the culprit.
- The witness is able to distinguish facial features of the culprit.
- The witness is mentally sound.
- The witness has basic communication skills.

## 1.7 Societal / Industrial Relevance

- The project stands to enhance criminal justice outcomes by equipping law enforcement with more accurate tools for suspect identification. This improvement could expedite case resolutions and ensure the apprehension of perpetrators.
- Contributing to public safety, the project's advancements in suspect identification can bolster overall security. Efficient case resolutions aid in removing potentially harmful individuals from the community, enhancing public safety.
- Accurate facial generation and recognition technologies can contribute to the reduction of wrongful arrests and convictions, addressing a critical concern and safeguarding individuals from unjust legal consequences.

## 1.8 Organization of the Report

The report unfolds with an introduction spotlighting the necessity for an Automated Facial Synthesis and Recognition System in forensic science, driven by the shortcomings of manual sketching and the integration of cutting-edge technologies like NLP and deep

learning. A thorough literature review follows, examining existing methods and identifying challenges in traditional forensic practices. The system architecture is meticulously explained in Chapter 3, encompassing three acts that leverage advanced tools such as Stable diffusion, DDPM, and VGG16. Each detailed component is expounded upon in Chapter 4, ranging from the Stable diffusion for image generation to the Database Matching using VGG16 for efficient criminal identification. The chapter also includes a work schedule, Gantt chart, and work breakdown, providing a comprehensive project management overview. The report concludes in Chapter 5, delving into the potential impact of successful implementation, ethical considerations, and future scope, ultimately positioning the proposed model as a pioneering solution with the potential to revolutionize forensic science and criminal investigations.

## 1.9 Summary of the Chapter

The chapter introduces a project focused on advancing suspect identification in forensic science through modern technologies, particularly artificial intelligence (AI) and machine learning (ML). Traditional hand-drawn facial sketches face limitations, and recent advancements in AI and ML offer opportunities for more efficient and accurate methods. The project's objective is to enhance the precision of facial sketches, automate the sketching process, integrate advanced AI/ML for generating realistic facial images, and improve facial recognition algorithms. Standardizing protocols and addressing challenges in portraying intricate facial features and adapting to age progression and expressions are key considerations. The project assumes clear eyewitness images and emphasizes societal relevance by enhancing criminal justice outcomes, contributing to public safety, and reducing wrongful arrests. Effective collaboration between forensic science experts, law enforcement, and technology specialists is crucial for project success.

# Chapter 2

# Literature Survey

## 2.1 High-Resolution Image Synthesis with Latent Diffusion Models[1]

The paper provides an update on the results of text-to-image synthesis and class-conditional synthesis on ImageNet. It also includes a user study for inpainting and super resolution models. Additionally, it discusses denoising diffusion models and their generative process. The paper presents qualitative results on object removal and image inpainting, showcasing the capabilities of the generative approach.The document mentions the use of several models, including LDM-4 and LDM-8, as well as Pixel Baseline 2. These models are used for conditional image synthesis tasks and are compared to state-of-the-art methods across a wide range of tasks without task-specific architectures. Additionally, the paper discusses the use of diffusion models, such as Diffusion Probabilistic Models (DM), for density estimation and sample quality. These models are designed to work on a compressed latent space of lower dimensionality, making training computationally cheaper and speeding up inference with almost no reduction in synthesis quality. This model learns a space that is perceptually equivalent to the image space but offers significantly reduced computational complexity. The study addresses limitations and societal impact, acknowledging the potential ethical considerations of generative models, such as the democratization of technology, the potential for misuse in creating manipulated or misleading content, and the reproduction of biases present in the training data.

## 2.2 A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks[2]

The proposed methodology revolves around generating realistic facial images from text descriptions, driven by a new approach that fully trains the Generative Adversarial Networks (GANs) to convert text into visual content. The method involves two critical sections:

text encoding, where the textual data is transformed into semantic vectors, and image decoding, where realistic images are generated based on the encoded text embeddings. The overall architecture follows a two-stream design, with text encoding in the first part and image decoding in the second, both working in harmony to produce visually accurate outputs.

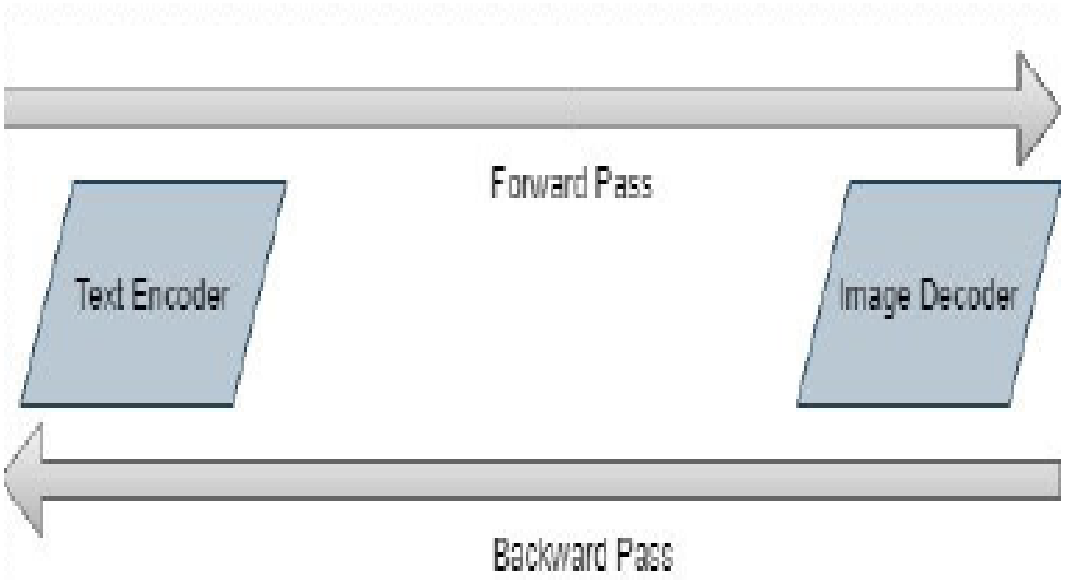



Figure 2.1: Design of a Fully Trained GAN

In the text encoding phase, the bidirectional Long Short-Term Memory (LSTM) model is employed to extract semantic features from input sentences. This allows for a more comprehensive understanding of the text, as it considers both forward and backward contexts. The outputs from the bidirectional LSTM are concatenated, forming a semantic vector to which noise is added to produce the final input for image generation.

The image decoding process takes the generated semantic vectors and translates them into realistic images. This is achieved through a convolutional neural network (CNN) that incorporates 6 deconvolution layers. These layers progressively upsample feature maps to increase image quality and size. Fine-tuning is performed between the blocks to enhance training parameters and improve image generation accuracy.

The network architecture employs a two-stream discriminator, where one discriminator is equipped with an attention mechanism to focus on specific facial features, while the other lacks this mechanism. Both discriminators assess the originality of facial region and feature synthesis. Training involves minimizing cross-entropy losses for both

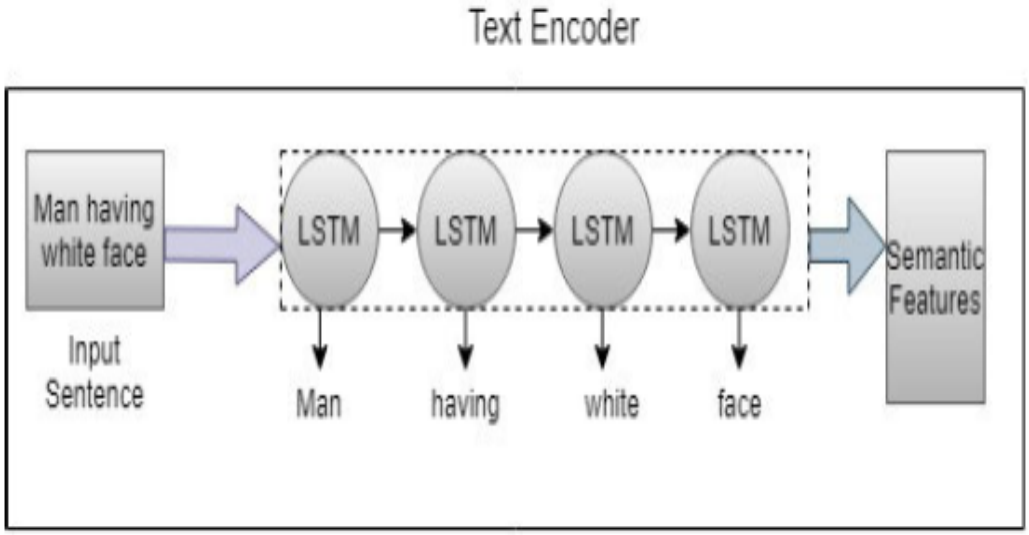



Figure 2.2: Encoder

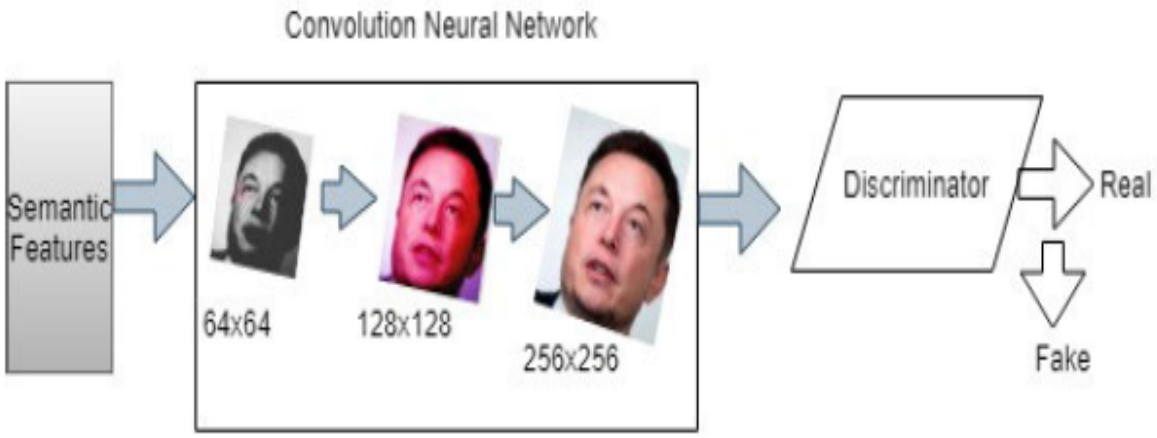



Figure 2.3: Decoder

discriminators in an adversarial manner.

In summary, this methodology utilizes fully trained GANs, bidirectional LSTMs for text encoding, and a well-structured CNN for image decoding. The use of two-stream discriminators enhances the originality evaluation of generated images. This approach produces high-quality, contextually relevant images from textual descriptions, with applications ranging from creative fields to law enforcement.

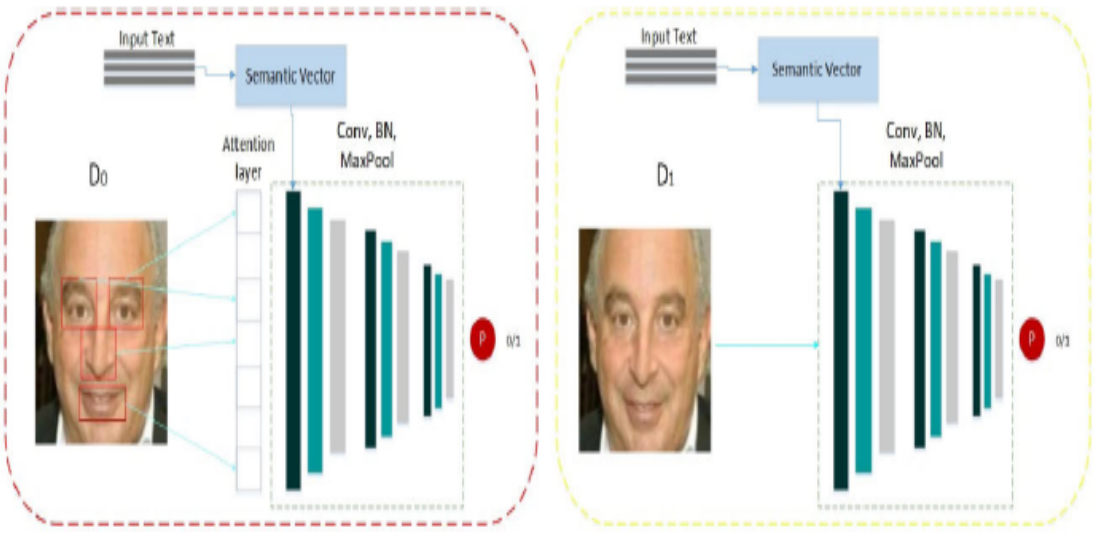



Figure 2.4: Discriminator

## 2.3 E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs [7]

The Eyes-to-Face GAN (E2F-GAN) is a novel face inpainting model introduced in the paper, which usees a two-module architecture: a coarse module and a refinement module. In the coarse module features are extracted from the periocular region to generate an initial coarse output with help of an edge predictor module. This coarse output undergoes refinement in the subsequent module to enhance the overall inpainting quality.

Key features of the E2F-GAN model include a decoder with seven layers, which includes an attention layer and six upsampling convolution layers. Additionally, a Channel

and Spatial Attention Block (CSAB) is used to emphasize important features and miti- gate the impact of redundant ones.

The model's performance is evaluated on the E2Fdb dataset, specifically generated for this purpose. The evaluation encompasses both qualitative and quantitative analyses, employing statistical metrics and measuring the model's ability to preserve demographic and biometric features. Overall, the E2F-GAN method demonstrates a coarse-to-fine architecture, integrates an edge predictor module, and utilizes attention mechanisms for effective face inpainting.

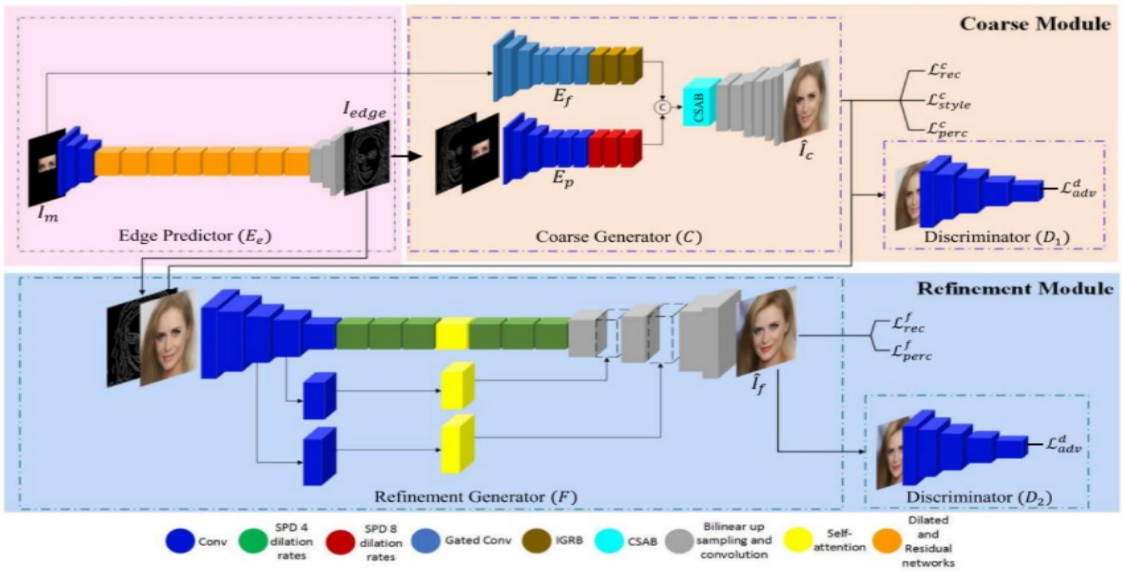



Figure 2.5: Architecture diagram of E2F-GAN

## 2.4 DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN. [11]

The system's core is the Spatially Conditioned StyleGAN (SC-StyleGAN) framework, an extension of the original StyleGAN architecture, tailored for portrait image generation and editing. SC-StyleGAN is structured with two pivotal sub-networks: the spatial encoding network and the synthesis network. The former maps input conditions, comprising sketches and semantic maps, to intermediates representing the results of coarse and middle style-controlled layers. The synthesis network utilizes pre-trained layers from the original StyleGAN synthesis network, taking the spatially encoded intermediates as input to generate the final synthesized image. Users are afforded greater precision and ease in expressing desired results through the incorporation of two input modalities: sketches and semantic maps. Sketches contribute structural features, while semantic maps define region boundaries, collectively simplifying the drawing and editing process. The system boasts additional features, including a data-driven suggestive drawing interface, global selection, and local guidance and editing control. These features aim to enhance the user experience and provide heightened flexibility in the generation and editing of portrait images. The paper validates the system's efficacy through comprehensive evaluations, both qualitative and quantitative, as well as a user study. These assessments confirm the system's superiority in terms of generation ability, usability, and expressiveness when

compared to existing solutions. The synergy of SC-StyleGAN and the user-friendly interface equips non-professional users with a powerful tool for nuanced and creative portrait image generation and editing

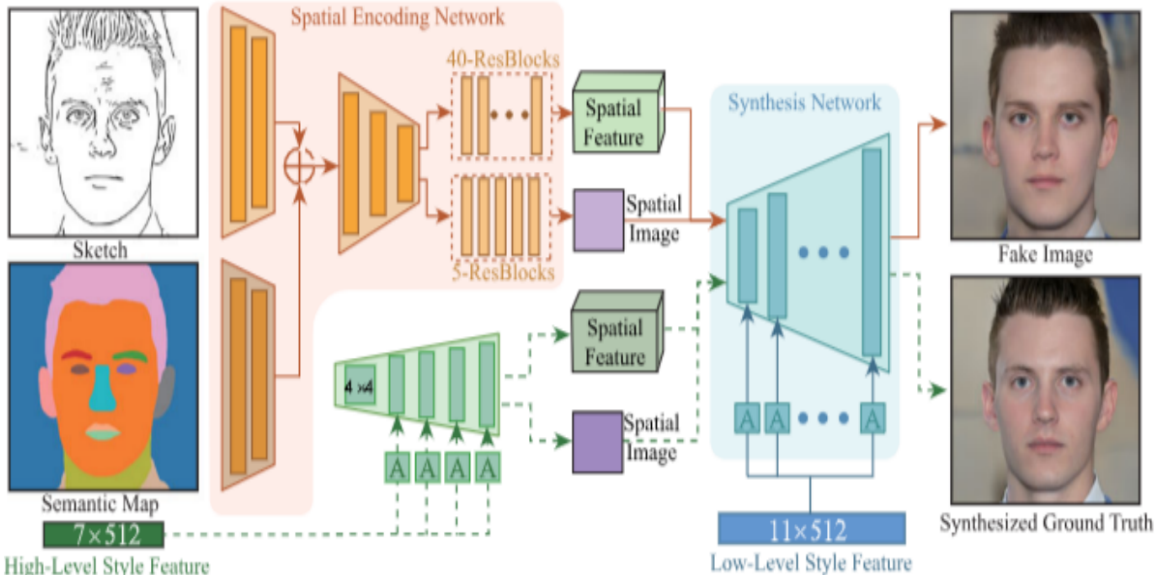



Figure 2.6: Network Architecture of SC-StyleGAN

## 2.5 Attention-Modulated Triplet Network for Face Sketch Recognition. [16]

This study focuses on face sketch recognition, a kind of cross-modality recognition where a given face sketch is used to discover a related photo from a face photo dataset. When dealing with composite face sketches, existing face photo-sketch identification techniques—such as synthesizing photographs from sketches and utilizing feature descriptors and mapping methods—have drawbacks. Siamese networks and triplet networks are two examples of deep learning algorithms that have been presented to extract similar features in a shared area for face photo-sketch recognition. In order to increase recognition efficiency and accuracy, the study suggests a unique method that combines a triplet network with an attention module and a spatial pyramid pooling layer. The attention module is designed to extract shape features that are comparable across various picture modalities, such as photos and sketches. The purpose of the spatial pyramid pooling layer is to accommodate varying image sizes and lessen the impact of picture noise. With an accuracy higher than 81% for Set B in the UoM-SGFS dataset, the suggested approach outperforms the stateof-the-art findings on composite face photo-sketch datasets. By assisting in the search for

similar regions in images, the network's attention mechanism lowers cross-modality differences and increases identification accuracy. The UoM-SGFS dataset, a sizable face photo-sketch dataset comprised of 300 color face pictures and 300 color face sketches produced with the aid of EFIT-V software, is used in this research. The dataset includes two sets, Set A and Set B, with modified face attributes in Set B to increase the similarity between the face photo and face sketch. The paper also mentions testing the model on the e-PRIP dataset. The study also mentions testing the model on hand-drawn sketches; however, data augmentation is employed to increase the number of training samples, and the dataset only consists of 188 image pairs. The E-PRIP and CUFS dataset experimen8 tal results are also included in the study, demonstrating the superiority of the suggested strategy over alternative approaches. The experimental results show that the proposed method outperforms other methods in terms of recognition accuracy on UoM-SGFSA, UoM-SGFSB, and e-PRIP datasets. The technique also shows how well the attention module captures the similar features between hand-drawn sketches and pictures of faces.

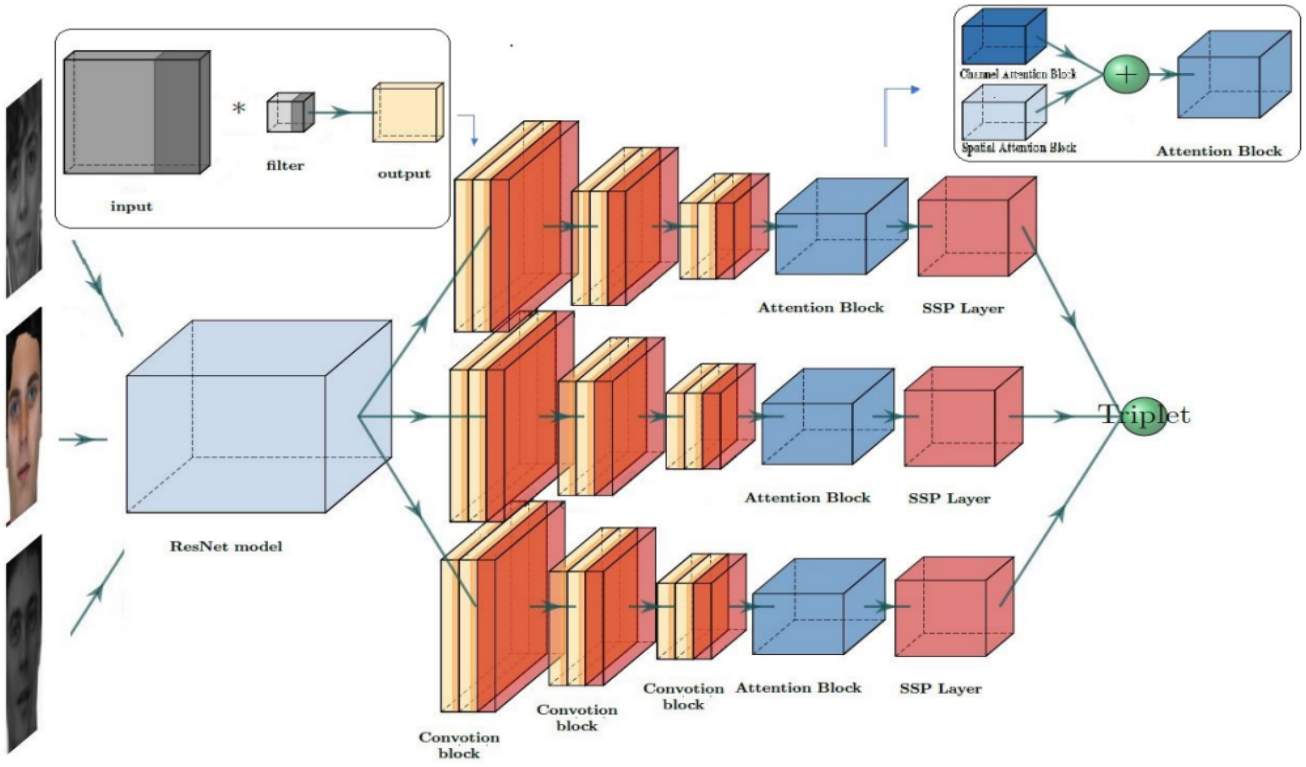



Figure 2.7: Architecture diagram of Attention Modulated Triplet Network

## 2.6 Comparison

| Reference Papers | Insights |
|---|---|
| High-Resolution Image Synthesis with Latent Diffusion Models [1]. | Latent Diffusion Model (LDM) with various parameters and conditioning mechanisms for tasks such as text-to-image synthesis, semantic image synthesis, and image inpainting. |
| Realistic Image Generation of Face from Text description using the fully trained Generative Adversarial Network [2]. | Focus on generating realistic face images from text descriptions. Utilizes a fully trained GAN for image synthesis. |
| E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs [3]. | The paper introduces E2F-GAN, a deep learning model for periocular-based face inpainting, employing a coarse module, a refinement module, facial landmarks, and edges to address challenges while outperforming existing methods in quantitative and identity metrics. |
| DrawingInStyles: Portrait Image Generation and Editing with Spatially Conditioned StyleGAN [4]. | SC-StyleGAN & DrawingInStyles is used for the generation of images from sketches. |
| Attention modulated triplet network for face sketch recognition [5]. | Triplet network+ Attention module+ SPP layer for face sketch recognition. |

Figure 2.8: Comparison

## 2.7 Summary of the Chapter

The five papers cover diverse aspects of facial image generation and manipulation. The first paper provides an update on the results of text-to-image synthesis and class-conditional synthesis on ImageNet. The second paper focuses on producing realistic facial images from textual descriptions using fully trained Generative Adversarial Networks (GANs). E2F-GAN concentrates on inpainting facial features, particularly eyes, utilizing edge-aware coarse-to-fine GANs for accurate reconstruction. DrawingInStyles explores portrait image generation and editing through StyleGAN, incorporating spatial conditioning to control specific style elements. The fifth paper introduces an attention-modulated triplet network for face sketch recognition, enhancing accuracy by integrating attention mechanisms within the recognition process. Together, these papers contribute to the advancement of techniques in facial image synthesis, inpainting, style control, and sketch recognition, showcasing the diversity and innovation within the field.

# Chapter 3

# Requirements

## 3.1 Hardware Requirements

- **Random Access Memory (RAM):**

  A minimum of 12 GB RAM is recommended for smooth execution of generation training. Sufficient RAM is essential to handle the computational load during the generation of facial images.

- **Storage:**

  A minimum of 256 GB of available storage space on your system is needed for training and running the program successfully. This storage capacity is required for storing the software, model files, and any generated datasets.

- **GPU:**

  A minimum of 12 GB and above GPU is required to train and run the program successfully.GPU is essential to accelerate the processing of complex computations required to generate the image and manipulate it.

## 3.2 Software Requirements

- **Operating System:**

  The project is compatible with Windows operating systems, including Windows 7 and higher versions.

- **Integrated Development Environment (IDE):**

  The project is designed to be developed and executed using Visual Studio Code or such applications like pycharm to facilitate code development and management.

- **Programming Language:**

  The project is implemented in Python, and thus, the recommended version for compatibility is Python 3.10.

- **Pytorch:**

  Pytorch version 2.3 with cuda 11.8 is used to run and train the program smoothly. It enables to use the hardware of system efficiently.

- **CUDA:**

  Cuda toolkit version 12 or above can be used to run and train the program smoothly. It enables us to access the GPU of system and accelerate the computations to train and run the program smoothly.

### 3.3 Functional Requirements

- **Description:**

  The primary use case is in the domain of forensic science, specifically focusing on the generation of facial images for investigative purposes. This use case involves the application of the project to generate realistic facial images from textual descriptions, aiding law enforcement agencies in suspect identification investigations.
  **Scenario:**

- **Input Data:**

  Law enforcement obtains textual description from witnesses with great facial details.

- **Application:**

  The application is employed to generate high-resolution and realistic facial images based on the provided low-resolution inputs. The generator network of DDPM utilizes advanced deep learning techniques to enhance facial features.

- **Enhancement Output:**

  The output consists of high-quality facial images that provide a clearer depiction of the individuals in question. This enhancement aids investigators in identifying key facial features such as scars, tattoos, or distinctive marks.

- **Investigative Analysis:**

  Enhanced facial images are analyzed by forensic experts and investigators, contributing to the development of leads or the identification of potential suspects. The improved images serve as valuable assets in criminal investigations.
  **Benefits:**

- **Time and Resource Efficiency:** Investigators save time and resources that would otherwise be spent manually enhancing images. The automated enhancement process expedites the analysis phase.

- **Forensic Applications:** It enhances the ability to extract meaningful information from visual evidence and contributing to the resolution of criminal investigations.

## 3.4 Summary of the Chapter

The project requires Windows OS (7 and above), Visual Studio Code for development, and Python 3.10. With a minimum of 8 GB RAM and 128 GB storage, it focuses on forensic science, generating high-quality facial images with the help of textual descriptions to aid law enforcement in suspect identification. Law enforcement provides detailed textual inputs, and the application, driven by diffusion models, enhances facial features to produce clearer images. These enhanced images are valuable for investigative analysis, contributing to leads and potential suspect identification. The project's benefits include time and resource efficiency, automating image enhancement, and enhancing the extraction of meaningful information from visual evidence in criminal investigations.

# Chapter 4

# System Architecture

## 4.1 System Overview

Forensic science continually evolves to incorporate cutting-edge technologies. This proposal outlines the development of an Automated Facial Synthesis and Recognition System aimed at transforming witness descriptions into detailed facial images and subsequently matching them against a criminal database. This system aims to enhance the efficiency and accuracy of suspect identification in criminal investigations and also save time in doing so.

The traditional process of manually sketching faces based on witness descriptions is time-consuming and subjective. In response to this challenge, our proposed system leverages state-of-the-art technologies, such as Natural Language Processing (NLP) and deep learning-based image synthesis, to automate the generation of detailed facial images from textual descriptions. Furthermore, the system integrates seamlessly with a comprehensive criminal database, enabling rapid and accurate matching against a diverse set of known individuals.

In our proposed system the first act includes software known as Stable Diffusion Model which crafts lifelike portraits, conjuring diverse faces from a whisper of randomness.

Act two focuses on generating an edited image using text as a guide with DDPM model. It begins with the initialization of a random noise vector drawn from a standard normal distribution, laying the groundwork for subsequent image synthesis. Throughout multiple diffusion steps, this noise vector undergoes iterative denoising, guided by predictions from neural networks. These predictions inform adjustments made to the noise

vector at each timestep, complemented by directional cues provided by the networks. Simultaneously, stochastic elements from noise vectors sampled from a standard normal distribution contribute variability to the update process, ensuring diverse image outcomes. Upon completion of the diffusion steps, the final noise vector represents the generated image, crafted through the coordinated interplay of denoising, prediction, and stochastic updating mechanisms. In scenarios where text guides the image generation process, such as method used here, textual input is incorporated into the DDPM framework, shaping predictions and steering the synthesis of images that reflect the semantic content specified in the text prompt.

After obtaining the final edited image we then use VGG16 model for database matching to check and generate a suspected culprit list by matching the obtained image with those of database.

But this framework isn't just a technological marvel. It whispers promises of revolutionizing investigations, aiding in the pursuit of justice in cases from robberies to missing persons. Its iterative nature allows it to learn from its mistakes, constantly sharpening its gaze.

## 4.2 Architecture Diagram of the System

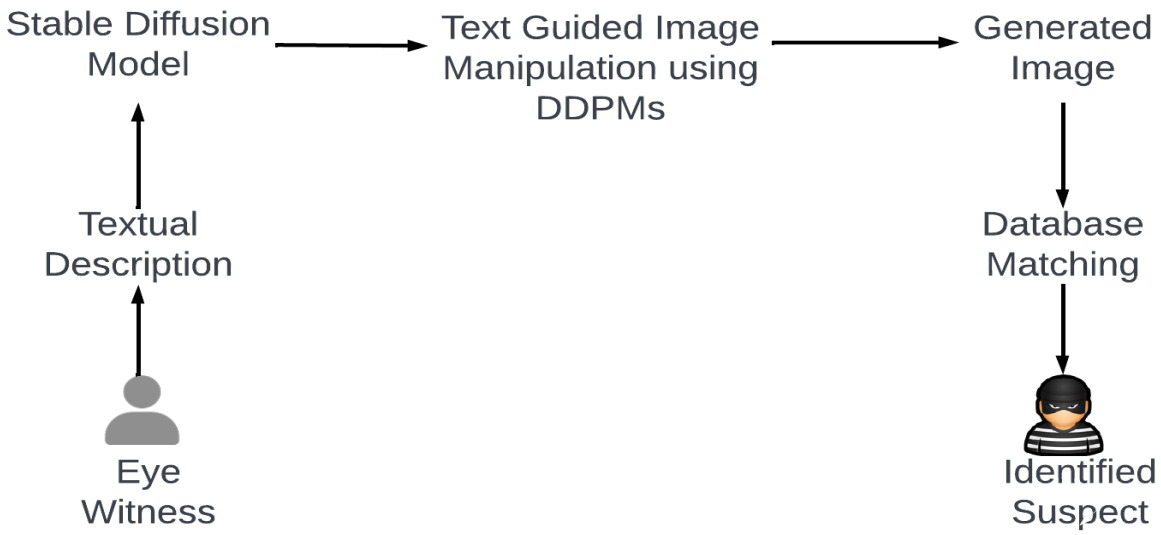



Figure 4.1: Architecture Diagram

## 4.3 Sequence Diagram

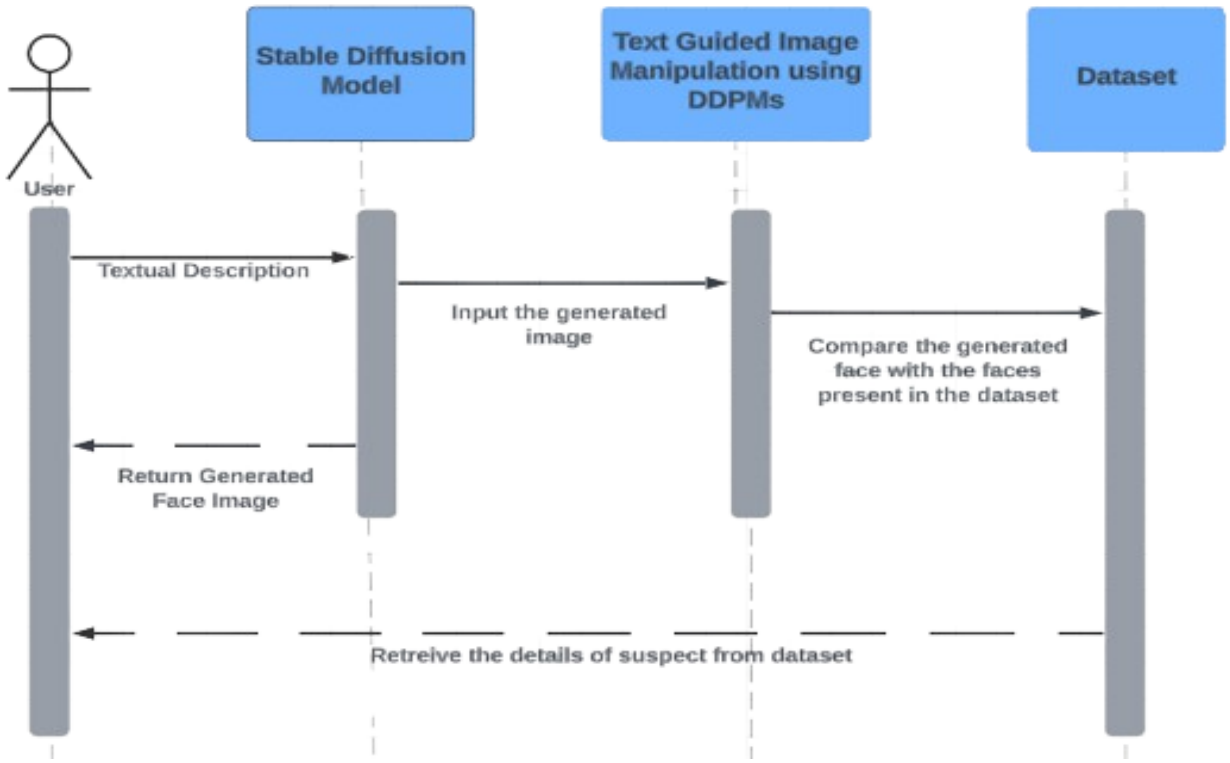



Figure 4.2: Sequence Diagram

## 4.4 Module Division

### 4.4.1 Stable Diffusion Model:

Image Generation includes the use of stable diffusion model to generate images using text as input. It takes text or voice, describing the image of suspected culprit, as input prompt which is processed and used by the model.The stable diffusion model represents a significant advancement in image synthesis and manipulation techniques. By leveraging latent diffusion, this model enables rapid generation of high-resolution images while consuming minimal computational resources. There are three main components of the stable diffusion model :

Autoencoder (VAE):

The autoencoder, based on the Variational Autoencoder (VAE) architecture, serves as

the initial stage in the latent diffusion process. Comprising an encoder and a decoder, the VAE transforms high-dimensional image data into a lower-dimensional latent space representation. During training, the encoder converts input images into compact latent representations, while the decoder reconstructs denoised images from these latent representations. The VAE's ability to efficiently compress image data facilitates subsequent processing by the U-Net.

U-Net:

The U-Net architecture plays a pivotal role in the latent diffusion process by predicting denoised representations of noisy latents generated during training. By employing a conditional model that incorporates information from the text encoder, the U-Net generates noise predictions for input latents, effectively enhancing the quality of latent representations. The encoder-decoder structure of the U-Net design has a skip connection which facilitates the transformation of noisy latents into refined representations suitable for image generation.

Text Encoder:

The text encoder, exemplified by CLIP's Text Encoder, contributes to the latent diffusion process by transforming input prompts into embeddings that guide the denoising process of the U-Net. By mapping textual descriptions to latent space embeddings, the text encoder provides contextual information that aids in generating coherent and contextually relevant images. Leveraging pre-trained models such as CLIP's Text Encoder ensures robust and effective guidance for the latent diffusion process.

During the inference process, the stable diffusion model employs the trained autoencoder and U-Net components to generate high-resolution images from input prompts. The autoencoder decodes denoised latents into image space, while the U-Net refines noisy latents using guidance from the text encoder. This iterative process of latent diffusion enables the rapid generation of high-quality images with reduced memory and compute requirements, making it suitable for various creative applications.

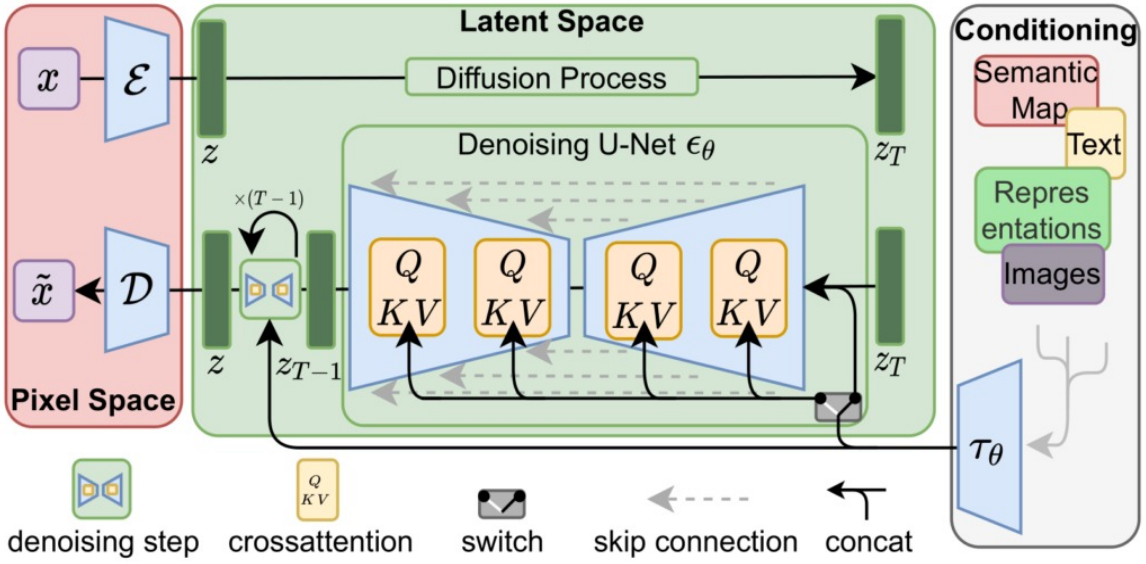



Figure 4.3: Architecture diagram of Stable Diffusion Model

### 4.4.2 Denoising Diffusion Probabilistic Module:

Firstly, the generated image from stable diffusion model is treated as the "target sample" you want to achieve. Then, you start with a random image or noise as your starting point, just like starting with a blank canvas. This noise is gradually transformed through a series of steps to resemble the target photograph. Each step makes the noise more similar to the target image, much like applying layers of paint to a canvas to create a picture.

The inversion algorithm iteratively adjusts the noise image to minimize the difference between it and the target photograph. This adjustment is guided by a loss function, which quantifies how far off the noise image is from the target. The algorithm uses gradient-based optimization techniques to tweak the noise image in the direction that reduces this difference, much like refining a sketch to make it look more like the final painting.

As the optimization process progresses, the noise image gradually transforms into an edited version of the original photograph, reflecting the desired changes in style, appearance, or content. In essence, the DDPM inversion algorithm allows you to reverse the editing process, starting with the final result and working backward to recreate it from scratch.

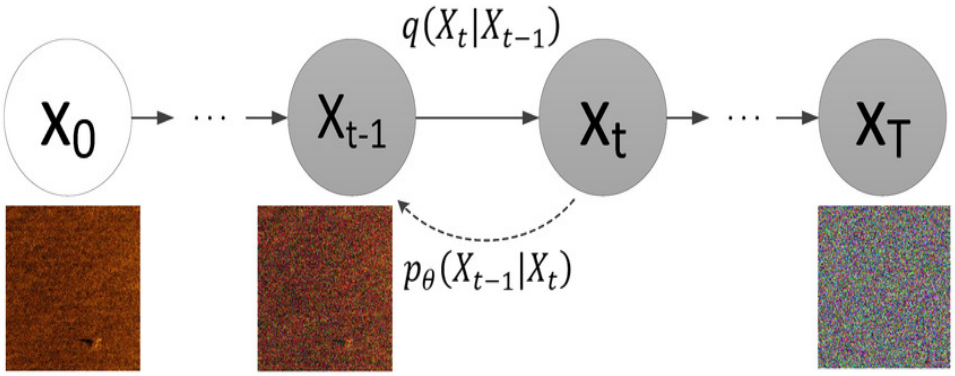



Figure 4.4: DDPM

### 4.4.3 Database Matching using VGG 16 model:

This module is used to match the generated image with an existing criminal or any useful database present to find the culprint faster and in a more efficient manner.

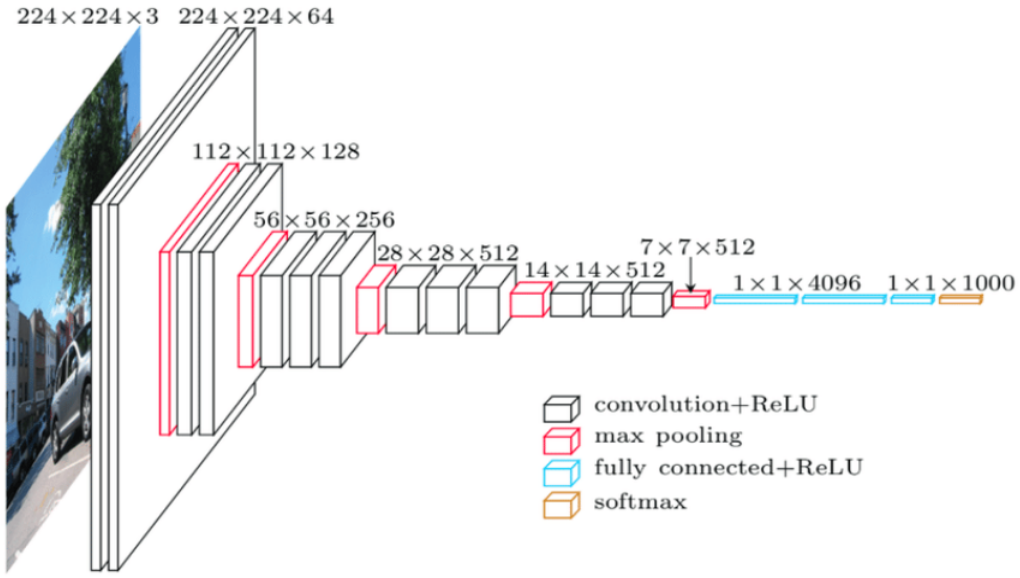



Figure 4.5: Database Matching using Siamese networks

VGGNet, renowned for its simplicity and effectiveness, adopts a standardized 224×224 image input, maintaining consistency by center-cropping each image in the ImageNet competition. Employing 3×3 convolutional filters, the model maximizes receptive field coverage while minimizing computational complexity. It incorporates 1×1 convolution filters for linear transformations at the input stage. The Rectified Linear Unit (ReLU) activation function, a pivotal innovation from AlexNet, expedites training by delivering zero outputs for negative inputs and retaining linearity for positive inputs. With a fixed

convolution stride of 1 pixel, VGG preserves spatial resolution post-convolution. Unlike AlexNet's Local Response Normalization, VGG opts for ReLU across its hidden layers, eschewing the former's increased training time and memory consumption. Pooling layers follow convolutional stages, diminishing feature map dimensionality and parameter count amid escalating filter numbers. VGGNet culminates in three fully connected layers, featuring two 4096-channel layers and a final layer with 1000 channels, corresponding to ImageNet's classes.

## 4.5 Work Schedule - Gantt Chart

| | October | November | December | January | February | March | April |
|---|---|---|---|---|---|---|---|
| Research and Analysis | | | | | | | |
| Dataset Matching | | | | | | | |
| DDPM | | | | | | | |
| Stable Diffusion | | | | | | | |
| Integration of Modules | | | | | | | |
| Modifications based on suggestion | | | | | | | |

Figure 4.6: Gantt Chart

## 4.6 Work Breakdown and Responsibilities


Figure 4.7: Work Breakdown and Responsibilities

## 4.7 Summary of the Chapter

Using state-of-the-art technologies, the Automated Facial Synthesis and Recognition system aims at accelerating suspect identification. The method quickly produces high-resolution facial images from textual descriptions by utilizing a steady diffusion model. The autoencoder (VAE), U-Net, and text encoder are a few of the many parts that go into this process and combine to convert input prompts into finely detailed visuals. Moreover, the incorporation of Denoising Diffusion Probabilistic Models (DDPMs) enables significant image modifications in response to text cues, augmenting the variety and caliber of forensic image manipulation assignments. The technology aligns and analyzes datasets from many sources using Convolutional Neural Networks (CNNs) like VGG16, offering insights into patterns and trends in the data for better decision-making in criminal investigations.

Overall, the system's novel methodology promises to transform the identification of suspects by providing quick and precise facial image generation from textual descriptions. It improves the effectiveness and precision of forensic investigations by automating crucial procedures and utilizing cutting-edge technologies, thereby aiding in the pursuit of justice in criminal cases.

# Chapter 5

# System Implementation

The three main components of the suggested methodology are Dataset Matching, Text-guided Image Manipulation using DDPM Inversion, and Image Generation. Using a stable diffusion model with an Autoencoder (VAE), U-Net, and Text Encoder, Image Generation quickly produces high-resolution images from text or voice input while using the least amount of processing power. Text-guided Image Manipulation offers a wide range of creative possibilities by utilizing Denoising Diffusion Probabilistic Models (DDPMs) to facilitate various editing procedures while maintaining image structure and semantics. With the use of pre-trained CNNs like VGG16, datasets are aligned through feature extraction and similarity computation, enabling tasks like image retrieval and object recognition and enabling thorough data analysis and interpretation across various study fields. This process is known as dataset matching.

## 5.1 Datasets Identified

With 5.85 billion CLIP-filtered image-text pairs, the LAION-5B dataset represents a substantial progress in the field of publicly available image-text datasets. It is 14 times bigger than the LAION-400M dataset. In addition to samples in English and more than 100 other languages, a significant percentage of the dataset contains texts (such as names) that cannot be assigned to a particular language. It also offers detection scores for watermark and NSFW content, nearest neighbor indices, and an online interface for exploration and subset building. Distributed image downloading, CLIP inference for similarity scoring, and the petabyte-scale Common Crawl dataset are the next steps in the acquisition pipeline. The final dataset is produced by filtering rules that guarantee inappropriate image-text pairs are eliminated.

## 5.2 Proposed Methodology/Algorithms

This paper involves mainly use of 3 algorithms , First being the Stable Diffusion model also known as Latent Diffusion model which consists of 3 components, then it uses DDPM (Denoising Diffusion Probabilistic model) for image manipulation and finally VGG16 convolutional neural network(CNN) for dataset matching and obtain a suspected culprint list.

### 5.2.1 Stable Diffusion model

The stable diffusion model incorporates three integral components. Firstly, the autoencoder, utilizing the Variational Autoencoder (VAE) architecture, serves as the initial stage in the latent diffusion process. Comprising an encoder and decoder, it transforms high-dimensional image data into a lower-dimensional latent space representation. During training, the encoder compresses input images into compact latent representations, while the decoder reconstructs denoised images from these latent representations. This compression of image data by the VAE facilitates subsequent processing by the U-Net.

Secondly, the U-Net architecture plays a crucial role in the latent diffusion process by predicting denoised representations of noisy latents generated during training. Incorporating information from the text encoder, it employs a conditional model to generate noise predictions for input latents, thereby enhancing the quality of latent representations. The U-Net transforms noisy latents into refined representations suitable for image generation which is characterized by an encoder-decoder structure with skip connections,

Lastly, the text encoder, exemplified by CLIP's Text Encoder, contributes to the latent diffusion process by transforming input prompts into embeddings that guide the denoising process of the U-Net. By mapping textual descriptions to latent space embeddings, the text encoder provides contextual information crucial for generating coherent and contextually relevant images. Leveraging pre-trained models such as CLIP's Text Encoder ensures robust and effective guidance for the latent diffusion process.

### 5.2.2 Denoising Diffusion Probabilistic model(DDPM)

In Denoising Diffusion Probabilistic Models (DDPMs), images are generated using a sequence of white Gaussian noise samples. These noise samples can be seen as the latent code correlated with the generated image, similar to how latent codes are used in Gener-

ative Adversarial Networks (GANs). However, the native noise space in DDPMs lacks a convenient structure, making it challenging to use for editing tasks.

To address this limitation, For Denoising Diffusion Probabilistic models (DDPMs), we provide a different latent noise space that allows for a variety of editing operations with straightforward techniques. Additionally, we present a technique for obtaining these editable noise maps from any image, be it artificially or authentically created. The editable noise maps deviate from a typical normal distribution and are not statistically independent across time-steps, in contrast to the native noise space in DDPMs. Nevertheless, simple modifications on these maps result in significant adjustments of the output image, including shifting or color adjustments, and they enable the flawless reconstruction of any desired image.

Moreover, modifying these noise maps in conjunction with a new text prompt in text-conditional models modifies the semantics while maintaining the structure. Unlike the more constrained DDIM inversion method, this attribute allows text-based alteration of actual images utilizing the varied sampling methodology of DDPMs. Furthermore, we show how the quality and diversity of current diffusion-based editing techniques can be improved by including this alternate noise space into them.

### 5.2.3 VGG16 convolutional neural network(CNN)

It begins by loading the VGG16 model and extracting the output from its 'fc1' layer, representing a fully connected layer. Image features are then extracted using this model through a resizing operation to (224, 224) pixels, conversion to RGB format, and preprocessing to align with VGG16 requirements. These features, comprising a 4096-dimensional vector, are utilized for subsequent similarity calculations. The algorithm proceeds by computing the similarity between query images and the entire dataset, employing Euclidean distance metrics on their feature vectors. The top two similar images are identified, and their similarity percentages are calculated and displayed alongside the images. This process iterates over all query images, ensuring comprehensive retrieval results.

## 5.3 User Interface Design

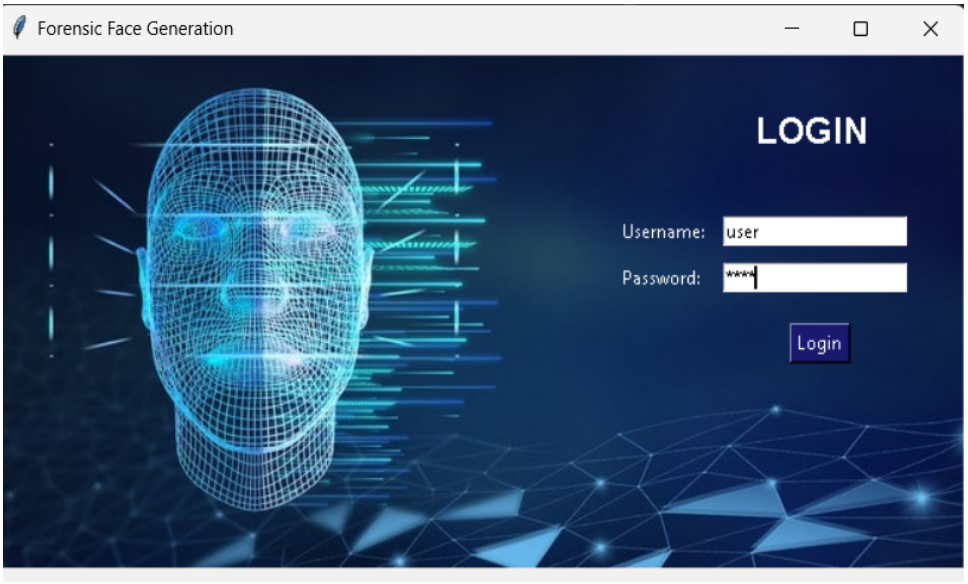



Figure 5.1: Login Page

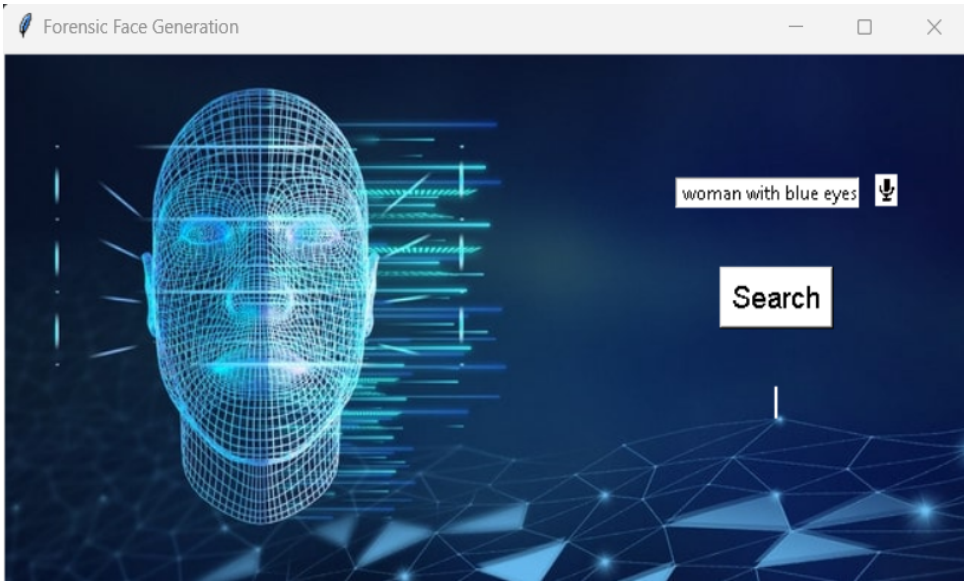



Figure 5.2: UI for entering the textual description

## 5.4 Database Design

The LAION-5B dataset's database design adheres to a stable and expandable architecture for effective management and querying of the enormous archive of image-text pairs. Important columns like URL, TEXT, WIDTH, HEIGHT, LANGUAGE, similarity, pwatermark, and punsafe are included in the database schema, making it simple to get pertinent data for study. Indexing strategies are used to maximize performance, especially for frequently searched columns such as language and URL. Furthermore, the database

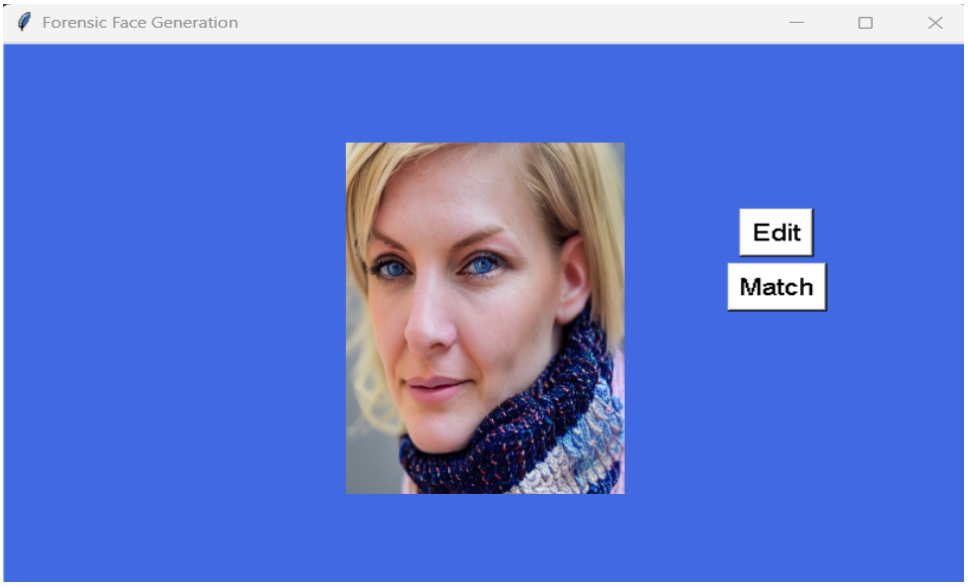



Figure 5.3: Image Manipulation

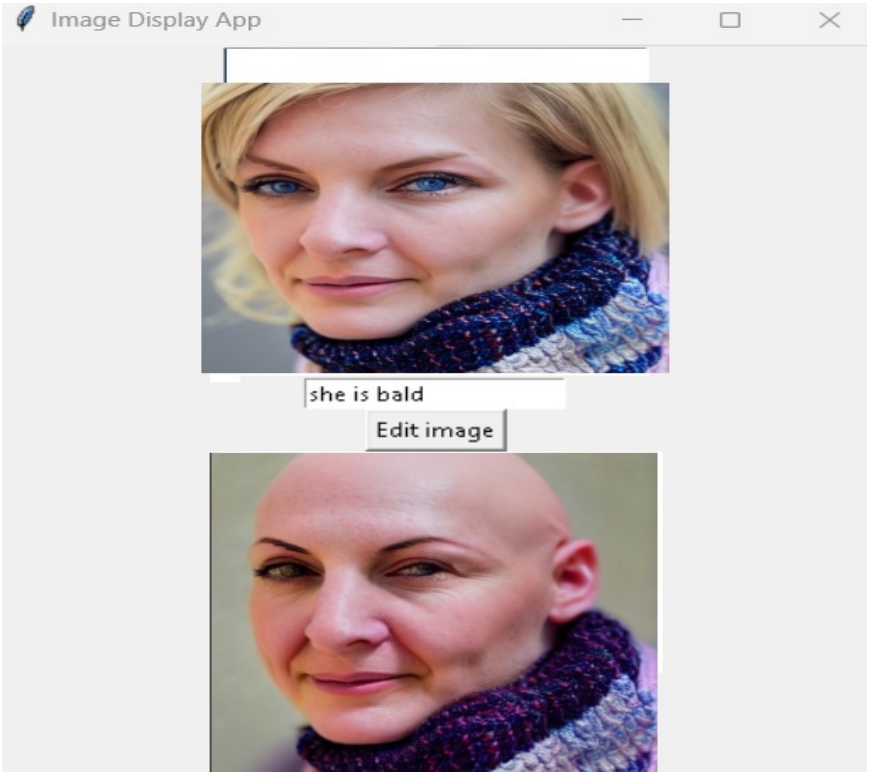



Figure 5.4: Image Manipulation

architecture includes safety features that guarantee users can filter out potentially harmful information, like watermarks and NSFW detection scores. Scalability and efficiency are ensured by the distributed architecture of the database infrastructure, which permits simultaneous data processing during the acquisition and post-processing phases. Over-

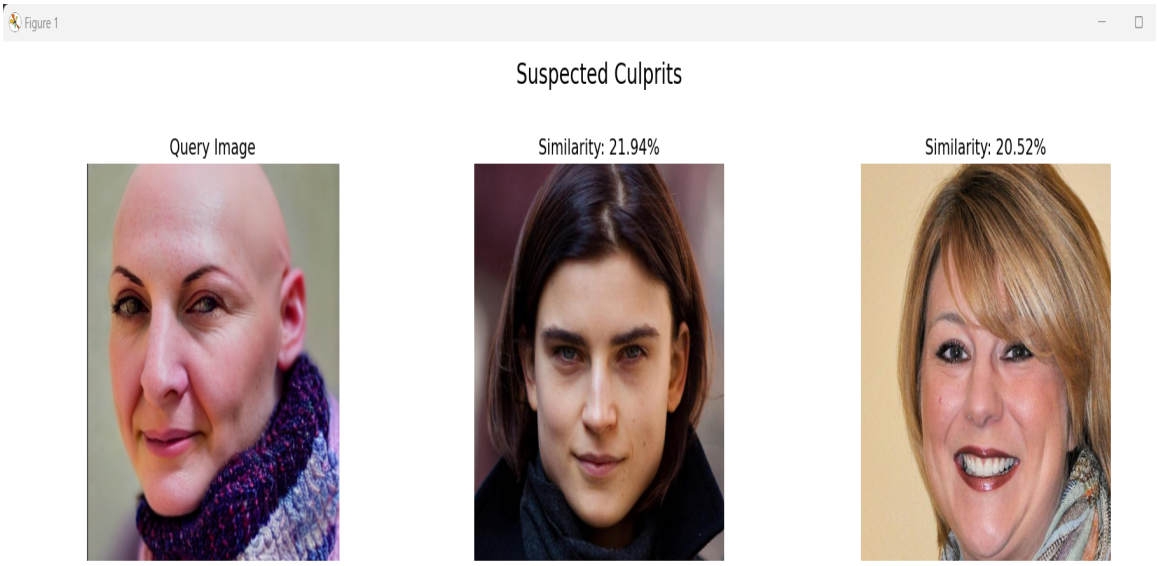



Figure 5.5: Dataset Matching

all, the LAION-5B dataset is readily available due to the database design's emphasis on speed, scalability, and security.

## 5.5 Conclusion

The proposed algorithms aid in the creation and editing of images while maintaining their logical structure and significance. The design of the user interface facilitates easy interaction, enabling users to explore and alter photographs with ease. Additionally, the database design places a high priority on efficiency, safety, and scalability, guaranteeing reliable image-text pair maintenance . The structure of the database ensures that large amounts of data may be handled securely and effectively.

## 5.6 Summary of the Chapter

A thorough approach for image production, alteration, and dataset matching is described in this chapter. Three main methods are introduced: the VGG16 CNN for dataset matching, the DDPM for picture editing, and the Stable Diffusion model. Because of the user-friendly interface design, manipulating images and exploring datasets is made simple. The LAION-5B dataset's database design places a high priority on security, scalability, and efficiency to provide dependable administration of large numbers of image-text pairs.

# Chapter 6

# Results and Discussions

## 6.1 Overview

To evaluate the text-to-face synthesis system's performance, it was carefully built and tested on a variety of datasets. This chapter presents a detailed analysis of the obtained results. We measured the system's effectiveness, found areas of success, faced difficulties, and investigated possible improvements by using particular metrics. Analyzing the system's advantages and disadvantages helps develop methods for improving its performance in real-world scenarios and offers insightful information on its usefulness.

## 6.2 Testing

Highlights image generated using the stable diffusion model and also shows the image manipulated using DDPM inversion an database matching with help of VGG16 model below.

Figure 6.1: Image Generation: A blonde women with blue eyes wearing a scarf

Image generated successfully using stable diffusion model above with the help of user prompts.

Figure 6.2: Image manipulation: A Bald women

Generated image edited successfully using DDPM inversion above using text prompts from user.

Figure 6.3: Image manipulation: An Indian women

Generated image edited successfully using DDPM inversion above using text prompts from user.

Figure 6.4: Image Manipulation: A black haired women with brown eyes

Generated image edited successfully using DDPM inversion above using text prompts from user.

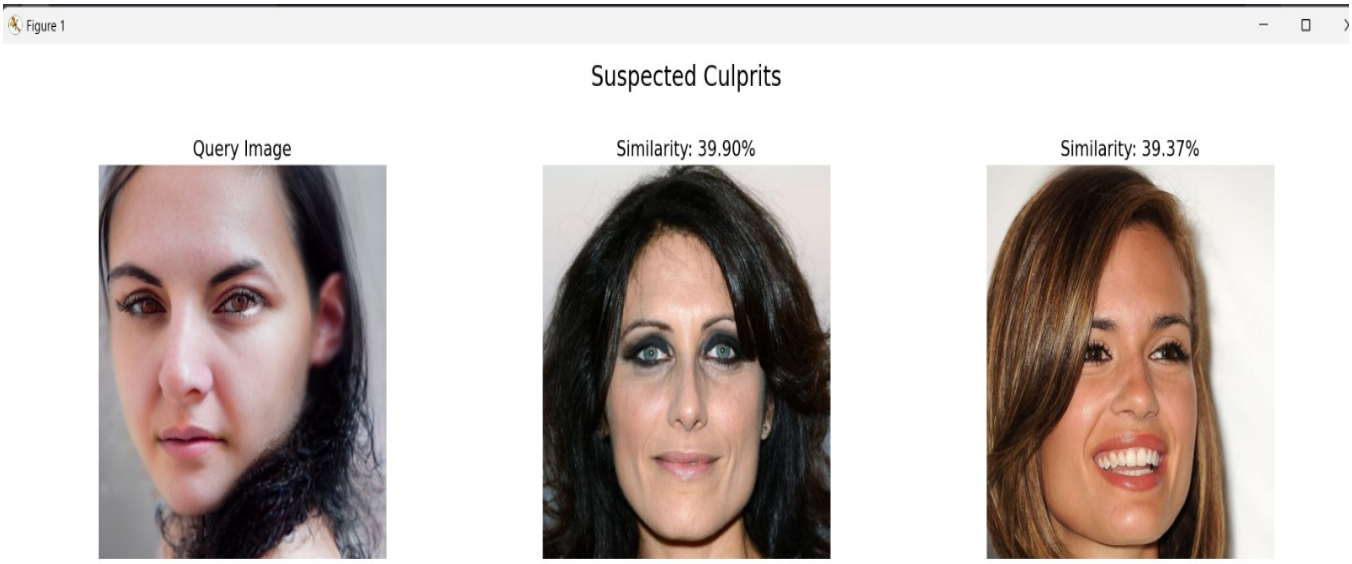
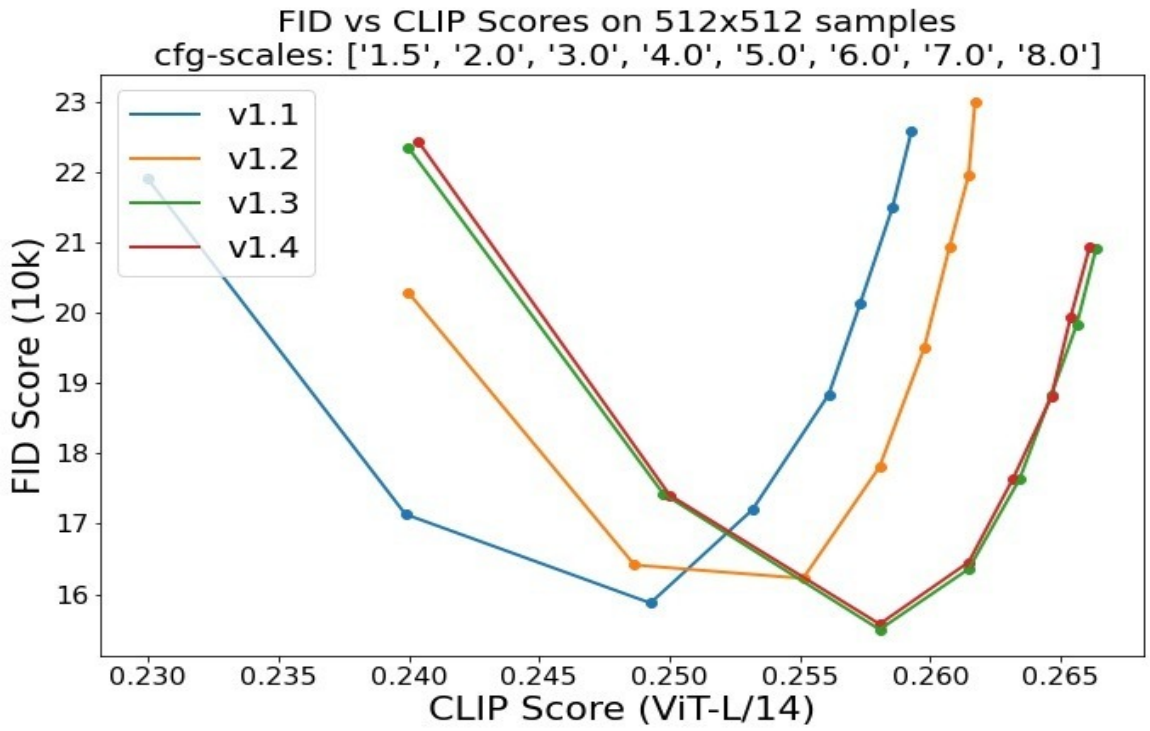



Figure 6.5: Database Matching

Edited image successfully matched with database to obtain a list of suspected culprit list.

### 6.3 Graphical Analysis




The above graph represents the FID(Frechet Inception Distance) vs CLIP(Contrastive Language-Image Pre-Training) scores at different checkpoints obtained while fine tuning the stable diffusion model.

### 6.4 Discussion

The input from user is taken as text or voice and processed to be used by text encoder to be put into latent space to iteratively refine a noisy image towards a target image through a series of steps. Each step involves applying diffusion processes to the image, gradually reducing the noise and improving the image quality. At each step of the diffusion process, stochastic sampling is used to add controlled amounts of noise to the image. This helps to explore the space of possible images and generate diverse outputs.The model's parameters are optimized to minimize the difference between the generated image and the target image, conditioned on the provided text prompt.Once the optimization process is complete, the final generated image is produced based on the text prompt.

Further the image is edited by first starting from random noise and transformed

through a series of steps to resemble the target photograph. Each step makes the noise more similar to the target image which is adjusted by inversion algorithm which is guided by loss function which quantifies how far noise image is from target.It then uses gradient-based optimization techniques to tweak the noise image in the direction that reduces this difference which gradually transforms into an edited version of the original photograph, reflecting the desired changes in style or appearance.

The manipulated or edited image is then passed through VGG16 model to match it with a database where on basis of similarity percentage calculated, we obtain a list of suspected culprit list having the images with most similalrity percentage.

## 6.5 Summary of the Chapter

The performance of the text-to-face synthesis system is thoroughly evaluated in this chapter. Testing shows that the system can effectively provide a variety of outputs by creating and manipulating images in response to user requests. Visual analysis, especially with regard to metrics such as FID and CLIP scores, shows how the system evolves and improves over several checkpoints. The complexities of the system's operations—from latent space refinement to inversion algorithm-guided image editing—are discussed in detail. The discussion also includes the integration of the VGG16 model for database matching, emphasizing the system's capacity to recognize comparable photos. All things considered, the chapter provides incisive analysis and talks that open the door for additional advancements and practical uses of the system.

# Chapter 7

# Conclusions & Future Scope

## 7.1 Future Scope

The successful implementation of this model holds the potential to revolutionize forensic practices, particularly in cases where traditional methods face limitations. Enabling law enforcement to generate facial images based on eyewitness testimonies not only expedites the investigative process but also contributes to more accurate suspect identifications. As the progress towards achieving this goal, it is essential to emphasize the ethical considerations surrounding the use of such technology, ensuring responsible and lawful applications in adherence to privacy standards and legal regulations. Ultimately, the envisioned model stands as a promising advancement in the field of forensic science, offering a novel solution to enhance the resolution and efficiency of criminal investigations.

This versatile multi-modal image generation and manipulation framework driven by textual descriptions, opens avenues for future advancements. Research directions include enhancing diversity and controllability, employing novel techniques to elevate image quality, and refining visual-linguistic similarity learning for precise manipulation.Its applicability can extend beyond faces, encompassing objects, scenes, and artistic creations. Further exploration may address scalability, optimizing the framework for real-time synthesis on large datasets. These pursuits promise to propel the field of image synthesis, offering innovative solutions for diverse applications.

## 7.2 Conclusion

In conclusion, the objective is to develop a robust and effective model capable of generating facial images of suspects based on eyewitness descriptions. This innovative approach seeks to bridge the gap between verbal eyewitness accounts and visual representation, offering law enforcement agencies a powerful tool for criminal investigations. By harness-

ing the capabilities of advanced deep learning techniques, my model aspires to translate textual descriptions provided by eyewitnesses into realistic facial images. This endeavor aims to enhance the resolution of criminal cases by providing investigators with visual representations that can significantly aid in suspect identification.

## 7.3 Summary of the chapter

This chapter presents a model intended to transform forensic practices by generating facial images of suspects based on eyewitness descriptions. Through advanced deep learning techniques, the model aims to expedite investigations and enhance the accuracy of suspect identifications, addressing limitations in traditional methods. Ethical considerations are underscored to ensure responsible and lawful applications, in compliance with privacy standards and legal regulations. Additionally, the chapter discusses future research avenues, highlighting the model's adaptability in generating diverse image types beyond faces. Ultimately, the project aims to bridge the gap between verbal descriptions and visual representations, offering law enforcement agencies an invaluable tool to bolster criminal investigations.

# References


[1] Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022 pp. 10674-10685.

[2] M. Z. Khan et al., "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[3] A. Hassanpour et al., "E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs," in IEEE Access, vol. 10, pp. 32406-32417, 2022, doi: 10.1109/ACCESS.2022.3160174.

[4] W. Su, H. Ye, S. -Y. Chen, L. Gao and H. Fu, "DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN," in IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 10, pp. 4074-4088, 1 Oct. 2023, doi: 10.1109/TVCG.2022.3178734.

[5] L. Fan, X. Sun and P. L. Rosin, "Attention-Modulated Triplet Network for Face Sketch Recognition," in *IEEE Access* , vol. 9, pp. 12914-12921, 2021, doi: 10.1109/ACCESS.2021.3049639.

[6] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1316-1324, doi: 10.1109/CVPR.2018.00143.

[7] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629.

[8] X. Hou, X. Zhang, Y. Li and L. Shen, "TextFace: Text-to-Style Mapping Based Face Generation and Manipulation," in IEEE Transactions on Multimedia, vol. 25, pp. 3409-3419, 2023, doi: 10.1109/TMM.2022.3160360.

[9] U. Osahor and N. M. Nasrabadi, "Text-Guided Sketch-to-Photo Image Synthesis," in IEEE Access, vol. 10, pp. 98278-98289, 2022, doi: 10.1109/ACCESS.2022.3206771.

[10] X. Chen, L. Qing, X. He, J. Su and Y. Peng, "From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance," in IEEE Access, vol. 6, pp. 14567-14575, 2018, doi: 10.1109/ACCESS.2018.2803787.

[11] X. Luo, X. He, X. Chen, L. Qing and H. Chen, "Dynamically Optimized Human Eyes-to-Face Generation via Attribute Vocabulary," in IEEE Signal Processing Letters, vol. 30, pp. 453-457, 2023, doi: 10.1109/LSP.2023.3268792.

[12] I. Chanpornpakdi and T. Tanaka, "The Role of the Eyes: Investigating Face Cognition Mechanisms Using Machine Learning and Partial Face Stimuli," in IEEE Access, vol. 11, pp. 86122-86131, 2023, doi: 10.1109/ACCESS.2023.3295118.

[13] Y. Li, X. Chen, F. Wu, and Z.-J. Zha, "Linestofacephoto: Face photo generation from lines with conditionalself-attention generative adversarial networks," in Proc. 27th *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.

[14] Chen, S.-Y., "DeepFaceEditing: Deep Face Generation and Editing with Disentangled Geometry and Appearance Control", ¡i¿arXiv e-prints¡/i¿, 2021. doi:10.48550/arXiv.2105.08935.

[15] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B.Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8798–8807.

[16] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. 2020. DeepFaceDrawing: deep generation of face images from sketches. ACM Trans. Graph. 39, 4, Article 72 (August 2020).

[17] C. Galea and R. A. Farrugia, "Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning," in *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.

[18] X. Li, X. Yang, H. Su, Q. Zhou and S. Zheng, "Recognizing Facial Sketches by Generating Photorealistic Faces Guided by Descriptive Attributes," in *IEEE Access*, vol. 6, pp. 77568-77580, 2018, doi: 10.1109/ACCESS.2018.2883463.

[19] M. Luo, H. Wu, H. Huang, W. He and R. He, "Memory-Modulated Transformer Network for Heterogeneous Face Recognition," in *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2095-2109, 2022, doi: 10.1109/TIFS.2022.3177960.

[20] Y. Guo, L. Cao, C. Chen, K. Du and C. Fu, "Domain Alignment Embedding Network for Sketch Face Recognition," in *IEEE Access*, vol. 9, pp. 872-882, 2021, doi: 10.1109/ACCESS.2020.3047108.

# Appendix A: Presentation

# Face Generation and Recognition in Forensic Science

Gayathri Ravi (U2003083)
Heynes Joy (U2003095)
Jeffin Jitto (U2003100)
Jocelyn Joshy (U2003103)

RSET

May 3, 2024

Guide: Ms. Jisha Mary Jose





## Contents I


- Problem Definition
- Project Objective
- Novelty of Idea and Scope of Implementation
- Literature Survey
- Proposed Method
- Architecture Diagram
- Sequence Diagram
- Modules
- Stable Diffusion Module
- DDPM Module
- Database Matching using VGG-16
- Results

# Contents II


- Work Division
- Conclusion
- Future Scope
- Paper Publication
- References






# Problem Definition

- In the field of forensic science, the method of producing hand-drawn facial sketches remains time-intensive and is often inaccurate. Challenges in both sketching and recognition techniques can impede the identification of suspects and hinder the effectiveness of criminal investigations, highlighting the need for innovative solutions and improved methodologies for facial sketching and Recognition.

## Project Objective

- Develop an automated face generation and recognition system for forensic science that accurately translates witness descriptions into detailed facial images and matches them against a comprehensive criminal database, with the aim of improving the efficiency and accuracy of suspect identification in criminal investigations.





## Novelty of idea and scope of implementation

- Enhancing traditional forensic methods with advanced algorithms. This can aid in the identification of suspects, victims, or missing persons by comparing facial features with existing databases.
- Applying face generation and recognition to revisit unsolved cases, potentially leading to the identification of previously unknown suspects or victims.

# Literature Survey I

| Reference Papers | Insights |
|---|---|
| High-Resolution Image Synthesis with Latent Diffusion Models [1]. | Latent Diffusion Model (LDM) with various parameters and conditioning mechanisms for tasks such as text-to-image synthesis, semantic image synthesis, and image inpainting. |
| Realistic Image Generation of Face from Text description using the fully trained Generative Adversarial Network [2]. | Focus on generating realistic face images from text descriptions. Utilizes a fully trained GAN for image synthesis. |
| AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks [3]. | Generation of high quality images from textual description using AttnGAN |
| StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks [4]. | Utilizes Stacked Generative Adversarial Networks (StackGAN) for high-resolution text-to-image synthesis which improves image quality in each stage of stack. |
| TextFace: Text-to-Style Mapping Based Face Generation and Manipulation [5]. | Uses a framework named TextFace for text-to-face generation and manipulation, incorporating various techniques. |
| Text-Guided Sketch-to-Photo Image Synthesis [6]. | Implements a text-guided sketch-to-photo image synthesis model using a combination of technologies,including a GAN |





# Literature Survey II

| Reference Papers | Methods |
|---|---|
| E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs [7]. | The paper introduces E2F-GAN, a deep learning model for periocular-based face inpainting, employing a coarse module, a refinement module, facial landmarks, and edges to address challenges while outperforming existing methods in quantitative and identity metrics. |
| From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance [8]. | The paper introduces an eyes-to-face synthesis approach using a conditional generative adversarial network (GAN) to address face occlusion in surveillance systems, demonstrating its effectiveness in preserving identity and generating realistic faces, with potential applications in criminal identification and tracking. |
| Dynamically Optimized Human Eyes-to-Face Generation via Attribute Vocabulary [9]. | The paper presents EA2F-GAN, a two-stage solution incorporating attribute vocabulary for dynamically optimizing the generation of realistic faces from human eyes, addressing the limitations of existing methods and outperforming state-of-the-art approaches. |
| The Role of the Eyes: Investigating Face Cognition Mechanisms Using Machine Learning and Partial Face Stimuli [10]. | The paper, utilizing machine learning and partial face stimuli, demonstrated that covering the eyes significantly impairs face cognition, highlighting their crucial role, while also exploring task substitutions to reduce workload. |

# Literature Survey III

| Reference Papers | Insights |
|---|---|
| DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN [11]. | SC-StyleGAN & DrawingInStyles is used for the generation of images from sketches. |
| Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks [12]. | Generation of face photos from line drawings or sketches using conditional self-attention generative adversarial networks (cGANs) |
| DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control [13]. | The method disentangles the geometry and appearance of a face image, enabling flexible face editing tasks such as changing appearance, replacing geometry with a sketch, and editing both geometry and appearance. |
| High-resolution image synthesis and semantic manipulation with conditional GAN [14]. | The method for generating high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks (conditional GANs). |
| Deepfacedrawing: Deep generation of face images from sketches [15]. | This is a deep learning system trained on a novel dataset, generating realistic face images from abstract sketches. It specializes in front-facing portraits without accessories, offering applications like face morphing and copy-paste. |





# Literature Survey IV

| Reference Papers | Insights |
|---|---|
| Attention modulated triplet network for face sketch recognition [16]. | Triplet network+ Attention module+ SPP layer for face sketch recognition. |
| DCNN, Morphed faces, transfer learning [17]. | Deep CNN followed by triplet embedding and triplet loss function.<br>Extended UoM-SGFS is used.<br>3D face morphable model to enable the generation of synthetic face photos and sketches. |
| Recognizing Facial Sketches by Generating Photorealistic Faces Guided by Descriptive Attributes [18]. | A Multi-modal Conditional GAN (MMC-GAN) incorporates visual sketch and semantic facial attributes for image generation. A two-path generator to enhance details by learning global and local facial features. An identity-preserving constraint is introduced to ensure consistency between sketches and facial images. |
| Memory-Modulated Transformer Network for Heterogeneous Face Recognition [19]. | Memory-Modulated Transformer Network (MMTN) for heterogeneous face recognition. A memory module explores prototypical style patterns in the reference domain, while a style transformer module blends the content of the input image with the style of the reference image. |
| Domain Alignment Embedding Network for Sketch Face Recognition [20]. | Domain alignment embedding network (DAEN) for sketch face recognition. DAEN method incorporates domain-related query sets and support sets to incorporate domain information and compute the domain alignment embedding loss in each training episode. CNN is used for feature extraction. |

## Methodology

- The description from eye-witness will be used as input, which will be processed further using Stable Diffusion model to generate the face image.
- The generated image can then be manipulated using Denoising Diffusion Probabilistic model.
- The manipulated image will then be matched with database using VGG-16 model to identify the suspect.





## Modules

- Stable Diffusion Module
- Denoising Diffusion Probabilistic Module
- Database Matching using VGG16

# Stable Diffusion Module

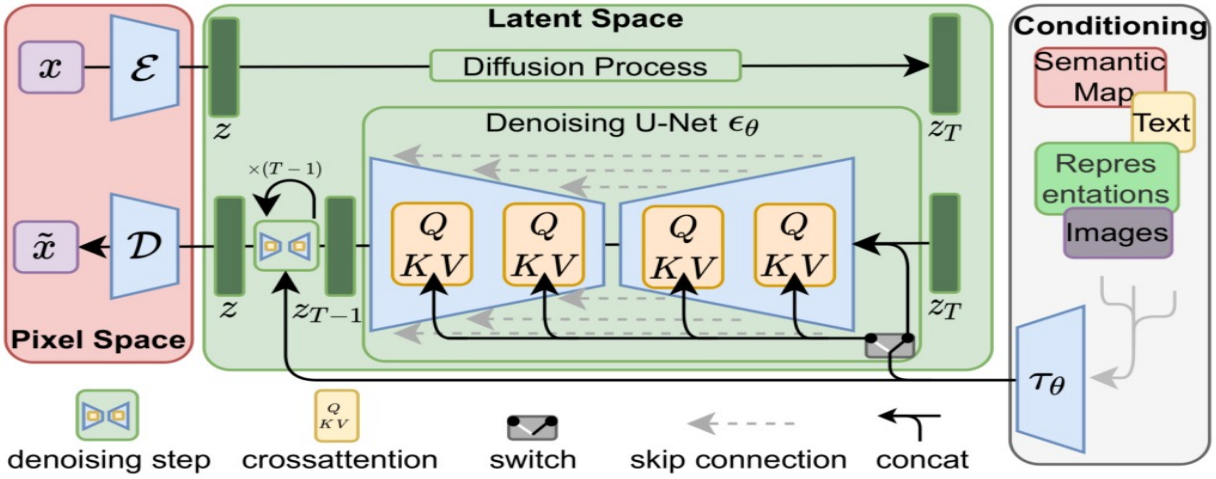



Figure 1: Stable Diffusion Module





# Denoising Diffusion Probabilistic Module

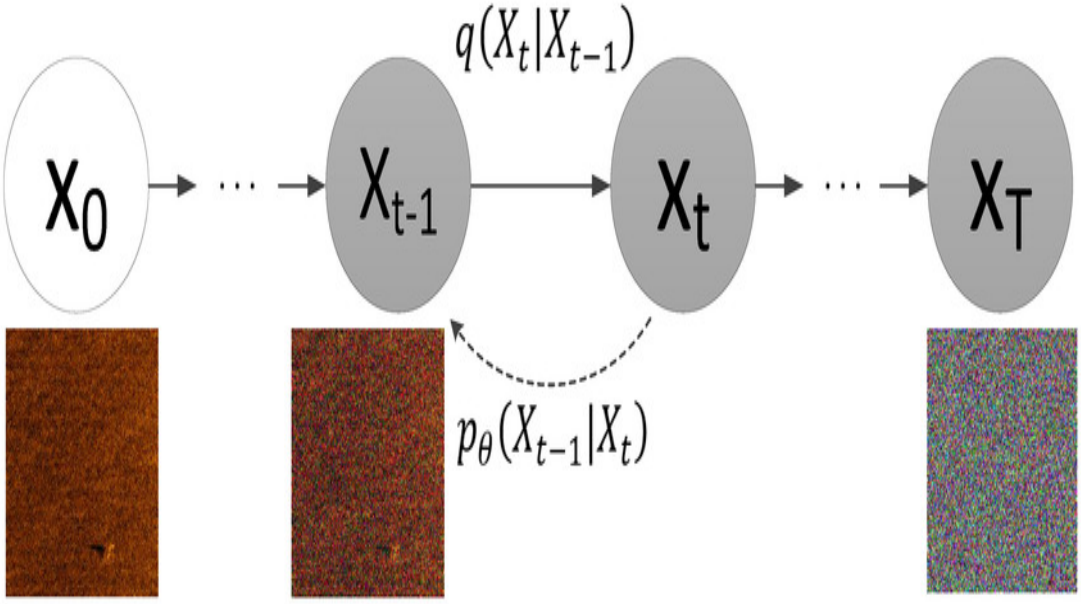



Figure 2: DDPM Module

# Database Matching using VGG16

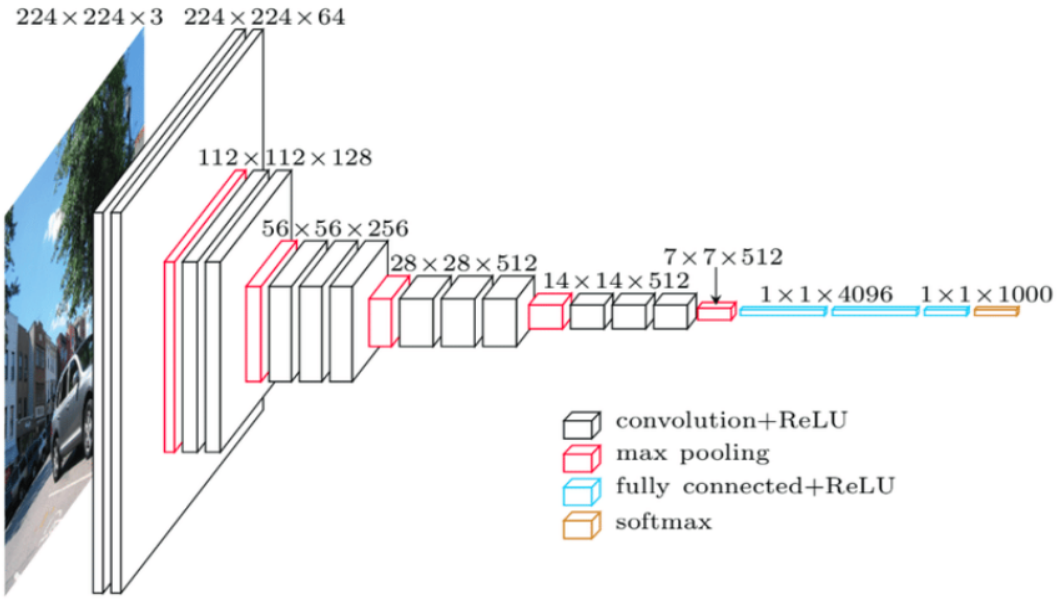



Figure 3: Database Matching using VGG16





# Architecture Diagram

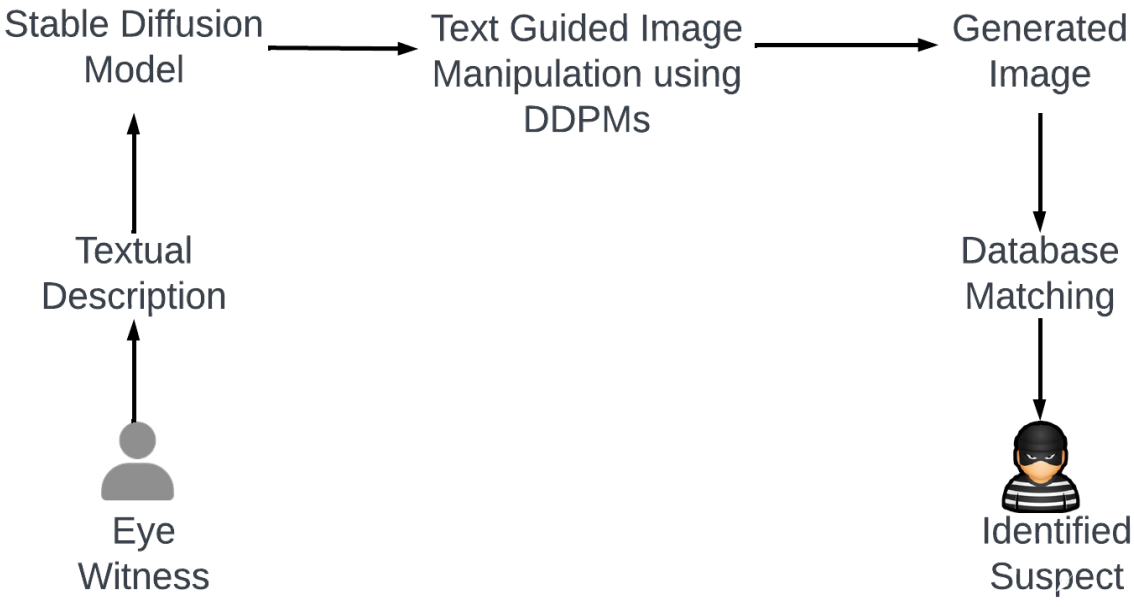



Figure 4: Architecture Diagram

# Sequence Diagram

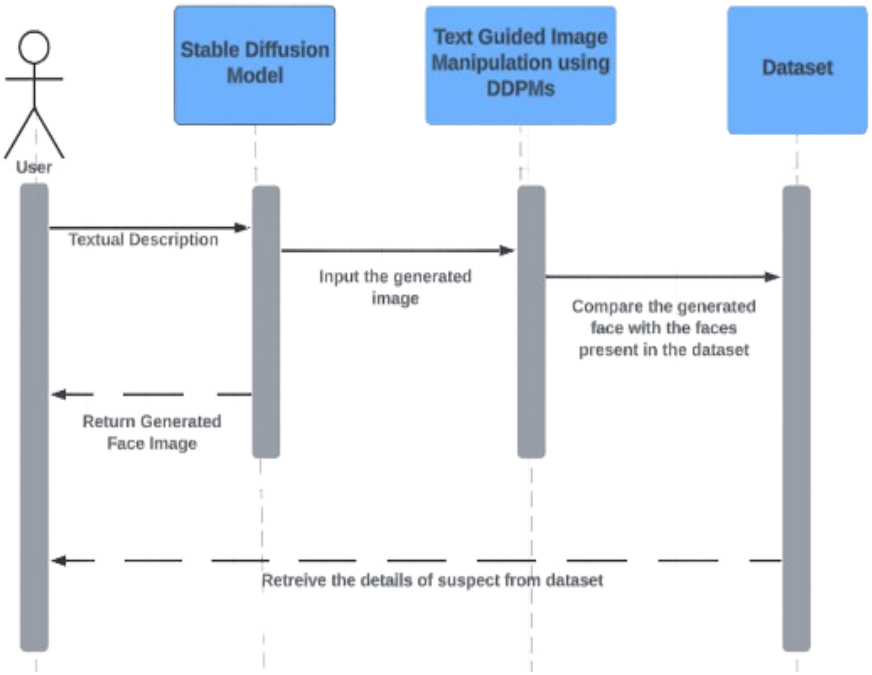



Figure 5: Sequence Diagram





# Work Breakdown


Figure 6: Work Breakdown

## Results

- Image generated using the stable diffusion model

Figure 7: Image Generation: A blonde women with blue eyes wearing a scarf





## Results

- Generated image edited successfully using DDPM inversion using text prompts from user

Figure 8: Image manipulation: A Bald women

# Results

Figure 9: Image manipulation: An Indian women





# Results

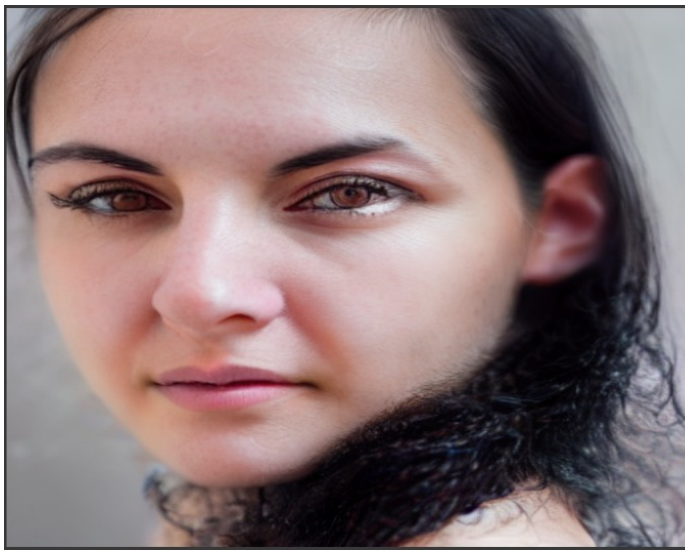


Figure 10: Image Manipulation: A black haired women with brown eyes

# Results

- Edited image successfully matched with database to obtain a list of suspected culprit list.

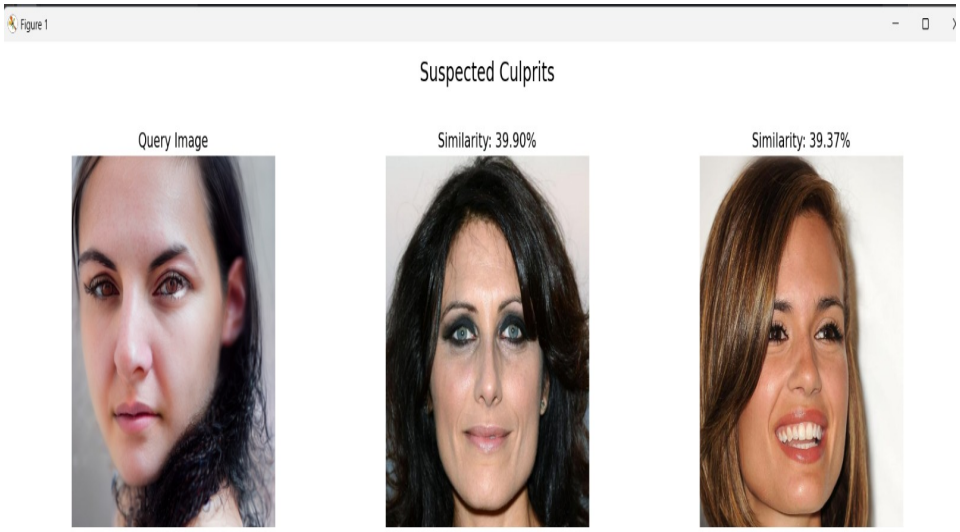



Figure 11: Database Matching





# Conclusion

- We successfully created a model that will generate an image of the suspect's face based on the description given by the eye-witness which would help in resolution of criminal cases.

## Future Scope

- The successful implementation of this model holds the potential to revolutionize forensic practices, particularly in cases where traditional methods face limitations.
- Extend the applicability beyond faces, encompassing objects, scenes, and artistic creations.





## Reference I


1. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022 pp. 10674-10685.
2. M. Z. Khan et al., "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.
3. T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1316-1324, doi: 10.1109/CVPR.2018.00143.

## Reference II


4. H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629.
5. X. Hou, X. Zhang, Y. Li and L. Shen, "TextFace: Text-to-Style Mapping Based Face Generation and Manipulation," in IEEE Transactions on Multimedia, vol. 25, pp. 3409-3419, 2023, doi: 10.1109/TMM.2022.3160360.
6. U. Osahor and N. M. Nasrabadi, "Text-Guided Sketch-to-Photo Image Synthesis," in IEEE Access, vol. 10, pp. 98278-98289, 2022, doi: 10.1109/ACCESS.2022.3206771.






## Reference III


7. A. Hassanpour et al., "E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs," in IEEE Access, vol. 10, pp. 32406-32417, 2022, doi: 10.1109/ACCESS.2022.3160174.
8. X. Chen, L. Qing, X. He, J. Su and Y. Peng, "From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance," in IEEE Access, vol. 6, pp. 14567-14575, 2018, doi: 10.1109/ACCESS.2018.2803787.
9. X. Luo, X. He, X. Chen, L. Qing and H. Chen, "Dynamically Optimized Human Eyes-to-Face Generation via Attribute Vocabulary," in IEEE Signal Processing Letters, vol. 30, pp. 453-457, 2023, doi: 10.1109/LSP.2023.3268792.

## Reference IV


10. I. Chanpornpakdi and T. Tanaka, "The Role of the Eyes: Investigating Face Cognition Mechanisms Using Machine Learning and Partial Face Stimuli," in IEEE Access, vol. 11, pp. 86122-86131, 2023, doi: 10.1109/ACCESS.2023.3295118.
11. W. Su, H. Ye, S. -Y. Chen, L. Gao and H. Fu, "DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN," in IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 10, pp. 4074-4088, 1 Oct. 2023, doi: 10.1109/TVCG.2022.3178734.






## Reference V


12. Y. Li, X. Chen, F. Wu, and Z.-J. Zha, "Linestofacephoto: Face photo generation from lines with conditionalself-attention generative adversarial networks," in Proc. 27th *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.
13. Chen, S.-Y., "DeepFaceEditing: Deep Face Generation and Editing with Disentangled Geometry and Appearance Control", ¡i¿arXiv e-prints¡/i¿, 2021. doi:10.48550/arXiv.2105.08935.
14. T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B.Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8798–8807.

## Reference VI


15. Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. 2020. DeepFaceDrawing: deep generation of face images from sketches. ACM Trans. Graph. 39, 4, Article 72 (August 2020).
16. L. Fan, X. Sun and P. L. Rosin, "Attention-Modulated Triplet Network for Face Sketch Recognition," in *IEEE Access* , vol. 9, pp. 12914-12921, 2021, doi: 10.1109/ACCESS.2021.3049639.
17. C. Galea and R. A. Farrugia, "Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning," in *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.






## Reference VII


18. X. Li, X. Yang, H. Su, Q. Zhou and S. Zheng, "Recognizing Facial Sketches by Generating Photorealistic Faces Guided by Descriptive Attributes," in *IEEE Access*, vol. 6, pp. 77568-77580, 2018, doi: 10.1109/ACCESS.2018.2883463.
19. M. Luo, H. Wu, H. Huang, W. He and R. He, "Memory-Modulated Transformer Network for Heterogeneous Face Recognition," in *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2095-2109, 2022, doi: 10.1109/TIFS.2022.3177960.
20. Y. Guo, L. Cao, C. Chen, K. Du and C. Fu, "Domain Alignment Embedding Network for Sketch Face Recognition," in *IEEE Access*, vol. 9, pp. 872-882, 2021, doi: 10.1109/ACCESS.2020.3047108.

## Paper Publication

- Yet to be published





# THANK YOU

# Appendix B: Research Paper

# Face Generation and Recognition in Forensic Science

Gayathri Ravi
Dept. of Computer Science and Engineering
Rajagiri School of Engineering & Technology
Rajagiri Valley P O, Kochi, Kerala, India
gayathricravi@gmail.com

Heynes Joy
Dept. of Computer Science and Engineering
Rajagiri School of Engineering & Technology
Rajagiri Valley P O, Kochi, Kerala, India
heynzjoy2002@gmail.com

Jeffin Jitto
Dept. of Computer Science and Engineering
Rajagiri School of Engineering & Technology
Rajagiri Valley P O, Kochi, Kerala, India
jeffinjitto2002@gmail.com

Jocelyn Joshy
Dept. of Computer Science and Engineering
Rajagiri School of Engineering & Technology
Rajagiri Valley P O, Kochi, Kerala, India
jocjos1203@gmail.com

Ms. Jisha Mary Jose
Dept. of Computer Science and Engineering
Rajagiri School of Engineering & Technology
Rajagiri Valley P O, Kochi, Kerala, India
jisham@rajagiritech.edu.in

***Abstract*—Suspect identification can be challenging for forensic investigations since standard procedures are time-consuming and prone to mistakes. This calls for the creation of novel approaches utilizing developments in machine learning (ML) and artificial intelligence (AI). In order to overcome these obstacles, the proposed Forensic Face Creation and Recognition project will make use of sophisticated recognition algorithms and AI-based face generation models. The goal of the research is to create high-quality face images from textual descriptions by applying a fully trained Generative Adversarial Network (GAN) to text-to-image synthesis. Image Generation, Text Guided Image Manipulation using Denoising Diffusion Probabilistic Models (DDPMs), and Dataset Matching are the three primary components of the process. Using a stable diffusion model, Image Generation quickly creates high-resolution images from word prompts by combining an autoencoder (VAE), U-Net, and text encoder. With the introduction of an alternate noise space for DDPMs, Text Guided picture Manipulation makes it possible to do meaningful picture altering tasks in response to text prompts. Convolutional neural networks (CNNs) are used in dataset matching to extract features and calculate similarity, which makes dataset alignment and comparison easier. The suggested methodology gives law enforcement authorities effective tools for identifying suspects, which represents a substantial development in forensic investigations. The project intends to increase the efficiency of criminal investigations, accelerate the matching process with large datasets, and enhance the accuracy of facial sketches by utilizing AI and ML approaches. The approach's ability to produce coherent and contextually relevant face images is validated by experimental results, which also show the approach's potential for speeding up the conclusion of criminal cases, particularly unsolved cold cases. All things considered, the Forensic Face Creation and Recognition project is a promising first step in strengthening forensic science's technological innovation capabilities.**

***Index Terms*—Text to Face Image Generation, Image Manipulation, Dataset Matching, DDPM , Stable Diffusion Model**

## I. INTRODUCTION

Forensic Science is time-consuming and prone to potential errors, which highlights the need for more efficient and accurate methods to identify suspects. High-profile criminal cases have highlighted the limitations of traditional methods and the need for innovative solutions that meet today's technological expectations and improve the overall effectiveness of forensic investigations. Recent advances in artificial intelligence (AI) and machine learning (ML) offer an unprecedented opportunity to transform face generation and recognition technologies using large datasets and complex algorithms. The goal of the project is to accelerate the resolution of criminal cases, including unsolved cold cases, by providing law enforcement agencies with a more reliable means of identifying suspects. A subset of text-to-image synthesis, text generation has potential in various research fields and has wide applications, especially in public safety. However, due to the limited availability of datasets, research into the nature of text-to-face has been limited. Most of the existing work in this area is based on semi-trained generative adversarial networks (GANs), where a pre-trained text encoder extracts semantic features from input sentences. These features are then used to train the image decoder. A fully trained GAN that trains both the text encoder and the image decoder simultaneously helps produce more accurate and efficient results instead of training both separately. By combining data from LFW, CelebA and local

or collected sources, you can create a dataset that helps create and recognize images generated from text using GANs. This dataset is tagged based on predefined categories. The visual results further confirm the effectiveness of this approach in creating face images that match the provided text description. The main goal of the Forensic Face Creation and Recognition project is to solve the problems of suspect identification using new technologies. This requires the development of sophisticated tools, including artificial intelligence-based face generation models and advanced recognition algorithms. The goal is to improve the accuracy of facial sketches, streamline the matching process with extensive databases, and ultimately optimize the efficiency of criminal investigations.

## II. RELATED WORK

Various methods have been developed in the fields of image generation, manipulation and recognition to push the boundaries of AI's understanding and synthesis of visual content. One notable contribution is TediGAN[2], which uses StyleGAN to generate and process a text-based face image. TediGAN introduces modules such as StyleGAN Inversion and Visual-Linguistic Similarity Learning to achieve accurate attribute transformation while maintaining semantic coherence. Using these components, TediGAN provides a comprehensive solution for versatile image synthesis that transforms the ability of generative models to understand text input and translate it into visually accurate results.

Fully trained Generative Adversarial Networks (GAN) have also been investigated. create realistic face images from text descriptions. These networks use two-way long-short-term memory (LSTM) models to encode text and convolutional neural networks (CNNs) to decode images. By encoding text input into semantic vectors and decoding them into realistic images, these methods demonstrate the potential of AI-based systems to bridge the gap between text and visual content and open up new opportunities for creative expression and content generation.

In addition, techniques such as E2F-GAN[7] offer innovative approaches to face painting - the process of filling in missing or damaged areas in facial images. E2F-GAN uses coarse-grained architecture and attention mechanisms to improve the quality of painted images and effectively preserve demographic and biometric features. This method represents a significant advance in image restoration techniques and offers practical solutions for situations where facial images may be incomplete or damaged.

DrawingInStyles[11] uses a spatially conditioned style GAN to create and edit portraits, giving users precise control over synthesized. pictures images through input methods such as sketches and semantic maps. By incorporating these ways, DrawingInStyles allows users to express their creative vision more flexibly and accurately, and shows the potential of AI-based systems in creative fields such as digital art and design.

For example, face sketch research. recognition An attention-modulated triplet network[16]that focuses on improving detection performance and accuracy by combining triplet networks with attention mechanisms. By combining attention mechanisms, these methods reduce mode differences and improve recognition accuracy, contributing to the development of robust systems for face photo sketch recognition in various fields, including law enforcement and biometric authentication.In general, related work on image generation, manipulation and recognition shows the continued benefits of AI-based approaches. Advances with implications ranging from creative expression to practical applications in various fields. Together, these methods help expand the capabilities of artificial intelligence to understand and synthesize visual content, paving the way for future innovations in this field.

## III. PROPOSED METHODOLOGY

### *A. Image Generation*

Image Generation includes the use of stable diffusion model to generate images using text as input. It takes text or voice, describing the image of suspected culprit, as input prompt which is processed and used by the model. The stable diffusion model represents a significant advancement in image synthesis and manipulation techniques. By leveraging latent diffusion, this model enables rapid generation of high-resolution images while consuming minimal computational resources. There are three main components of the stable diffusion model :

**1. Autoencoder (VAE):**
The autoencoder, based on the Variational Autoencoder (VAE) architecture, serves as the initial stage in the latent diffusion process. Comprising an encoder and a decoder, the VAE transforms high-dimensional image data into a lower-dimensional latent space representation. During training, the encoder converts input images into compact latent representations, while the decoder reconstructs denoised images from these latent representations. The VAE's ability to efficiently compress image data facilitates subsequent processing by the U-Net.

**2. U-Net:**
The U-Net architecture plays a pivotal role in the latent diffusion process by predicting denoised representations of noisy latents generated during training. By employing a conditional model that incorporates information from the text encoder, the U-Net generates noise predictions for input latents, effectively enhancing the quality of latent representations. The U-Net architecture, characterized by an encoder-decoder structure with skip connections, facilitates the transformation of noisy latents into refined representations suitable for image generation.

**3. Text Encoder:**
The text encoder, exemplified by CLIP's Text Encoder, contributes to the latent diffusion process by transforming input prompts into embeddings that guide the denoising process of the U-Net. By mapping textual descriptions to latent space embeddings, the text encoder provides contextual

information that aids in generating coherent and contextually relevant images. Leveraging pre-trained models such as CLIP's Text Encoder ensures robust and effective guidance for the latent diffusion process.

During the inference process, the stable diffusion model employs the trained autoencoder and U-Net components to generate high-resolution images from input prompts. The autoencoder decodes denoised latents into image space, while the U-Net refines noisy latents using guidance from the text encoder. This iterative process of latent diffusion enables the rapid generation of high-quality images with reduced memory and compute requirements, making it suitable for various creative applications.

### *B. Text guided Image manipulation using DDPM inversion*

In Denoising Diffusion Probabilistic Models (DDPMs), images are generated using a sequence of white Gaussian noise samples. These noise samples can be seen as the latent code associated with the generated image, similar to how latent codes are used in Generative Adversarial Networks (GANs). However, the native noise space in DDPMs lacks a convenient structure, making it challenging to use for editing tasks.

To address this limitation, we propose an alternative latent noise space for DDPMs that enables a wide range of editing operations using simple methods. A new method was introduced to extract editable noise maps for a real or synthetically generated image. The editable noise maps are not independent across the timesteps and it also does not follow a standard normal distribution unlike the native noise maps in DDPMs. The straightforward transformations on these maps lead to meaningful manipulations, like color edits and shifting, of output image. They also allow for perfect reconstruction of any image.

Furthermore, in text-conditional models, fixing these noise maps while changing the text prompt alters the semantics while preserving the structure. This property enables text-based editing of real images using the diverse sampling scheme of DDPMs, contrasting with the more limited DDIM inversion approach. Additionally, we demonstrate how integrating this alternative noise space into existing diffusion-based editing methods can enhance their quality and diversity.

### *C. Dataset Matching*

Dataset matching plays a crucial role in numerous data-driven applications, enabling the harmonization and comparison of datasets sourced from diverse origins. It involves aligning datasets based on their attributes, allowing for meaningful comparisons and analyses. One common approach to dataset matching involves feature extraction and similarity calculation using Convolutional Neural Networks (CNNs) like VGG16.

The process begins by extracting features from images using a pre-trained CNN model VGG16 that transforms the raw image. data into high-dimensional feature vectors. These feature vectors store multiple representations of the images, allowing them to be quantitatively compared.

The similarity is then calculated based on the Euclidean distance or other similarity metrics applied to the extracted feature vectors. By measuring the distance between the feature vectors, the images are classified according to whether they are similar to the query image. It enables the identification of images that closely match the query image in terms of visual content.

In the context of research, dataset matching is a key technique for tasks such as image retrieval, object recognition, and content-based recommendation systems. By aligning data and determining their similarity, researchers gain insight into patterns, trends, and relationships in the data. In addition, dataset matching enables versatile analysis, where data from different categories (such as text and images) can be compared and integrated to obtain a complete picture.

In general, the presented data set matching method shows the usefulness of CNNs when extracting features and in similarity computing, which lays the foundation for advanced data analysis and interpretation in various fields of research. By systematically comparing datasets, researchers can unlock the full potential of their data, facilitating informed decision-making and information discovery.

## IV. RESULTS AND DISCUSSIONS

A software for generating and manipulating images using text prompts was successfully built and tested. The result and analysis are shown below.

Fig. 1. Image Generation: "A blonde women with blue eyes wearing a scarf"

Fig. 2. Image manipulation: A Bald women

## V. CONCLUSION

In conclusion, the objective is to develop a robust and effective model capable of generating facial images of suspects

Fig. 3. Image manipulation: An Indian women

Fig. 4. Image Manipulation: "A black haired women with brown eyes"

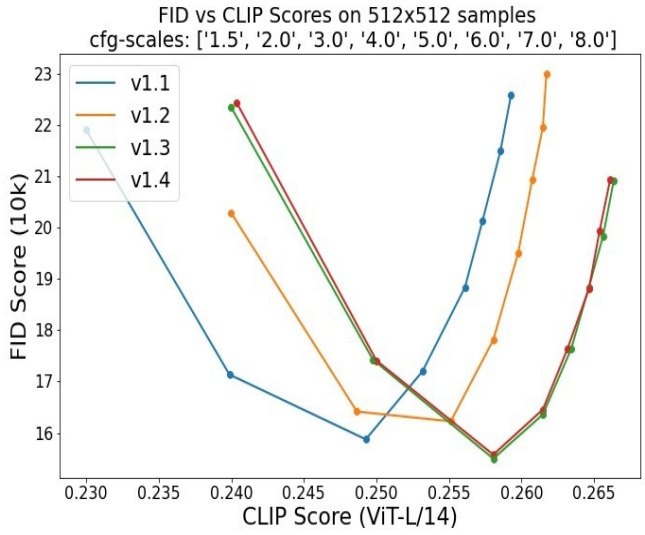



Fig. 5. The above graph represents the FID(Frechet Inception Distance) vs CLIP(Contrastive Language-Image Pre-Training) scores at different checkpoints obtained while fine tuning the stable diffusion model.

based on eyewitness descriptions. This innovative approach seeks to bridge the gap between verbal eyewitness accounts and visual representation, offering law enforcement agencies a powerful tool for criminal investigations. By harnessing the capabilities of advanced deep learning techniques, our model aspires to translate textual descriptions provided by eyewitnesses into realistic facial images. This endeavor aims to enhance the resolution of criminal cases by providing investigators with visual representations that can significantly aid in suspect identification.

## REFERENCES


[1] Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022 pp. 10674-10685.

[2] W. Xia, Y. Yang, J. -H. Xue and B. Wu, "TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 2256-2265, doi: 10.1109/CVPR46437.2021.00229.

[3] M. Z. Khan et al., "A Realistic Image Generation of Face From Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[4] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1316-1324, doi: 10.1109/CVPR.2018.00143.

[5] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629.

[6] X. Hou, X. Zhang, Y. Li and L. Shen, "TextFace: Text-to-Style Mapping Based Face Generation and Manipulation," in IEEE Transactions on Multimedia, vol. 25, pp. 3409-3419, 2023, doi: 10.1109/TMM.2022.3160360.

[7] U. Osahor and N. M. Nasrabadi, "Text-Guided Sketch-to-Photo Image Synthesis," in IEEE Access, vol. 10, pp. 98278-98289, 2022, doi: 10.1109/ACCESS.2022.3206771.

[8] A. Hassanpour et al., "E2F-GAN: Eyes-to-Face Inpainting via Edge-Aware Coarse-to-Fine GANs," in IEEE Access, vol. 10, pp. 32406-32417, 2022, doi: 10.1109/ACCESS.2022.3160174.

[9] X. Chen, L. Qing, X. He, J. Su and Y. Peng, "From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance," in IEEE Access, vol. 6, pp. 14567-14575, 2018, doi: 10.1109/ACCESS.2018.2803787.

[10] X. Luo, X. He, X. Chen, L. Qing and H. Chen, "Dynamically Optimized Human Eyes-to-Face Generation via Attribute Vocabulary," in IEEE Signal Processing Letters, vol. 30, pp. 453-457, 2023, doi: 10.1109/LSP.2023.3268792.

[11] I. Chanpornpakdi and T. Tanaka, "The Role of the Eyes: Investigating Face Cognition Mechanisms Using Machine Learning and Partial Face Stimuli," in IEEE Access, vol. 11, pp. 86122-86131, 2023, doi: 10.1109/ACCESS.2023.3295118.

[12] W. Su, H. Ye, S. -Y. Chen, L. Gao and H. Fu, "DrawingInStyles: Portrait Image Generation and Editing With Spatially Conditioned StyleGAN," in IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 10, pp. 4074-4088, 1 Oct. 2023, doi: 10.1109/TVCG.2022.3178734.

[13] Y. Li, X. Chen, F. Wu, and Z.-J. Zha, "Linestofacephoto: Face photo generation from lines with conditionalself-attention generative adversarial networks," in Proc. 27th *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.

[14] Chen, S.-Y., "DeepFaceEditing: Deep Face Generation and Editing with Disentangled Geometry and Appearance Control", ¡i¿arXiv e-prints¡/i¿, 2021. doi:10.48550/arXiv.2105.08935.

[15] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B.Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8798–8807.

[16] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. 2020. DeepFaceDrawing: deep generation of face images from sketches. ACM Trans. Graph. 39, 4, Article 72 (August 2020).

[17] L. Fan, X. Sun and P. L. Rosin, "Attention-Modulated Triplet Network for Face Sketch Recognition," in *IEEE Access* , vol. 9, pp. 12914-12921, 2021, doi: 10.1109/ACCESS.2021.3049639.

[18] C. Galea and R. A. Farrugia, "Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning," in *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1421-1431, June 2018, doi: 10.1109/TIFS.2017.2788002.

[19] X. Li, X. Yang, H. Su, Q. Zhou and S. Zheng, "Recognizing Facial Sketches by Generating Photorealistic Faces Guided by Descriptive Attributes," in *IEEE Access*, vol. 6, pp. 77568-77580, 2018, doi: 10.1109/ACCESS.2018.2883463.

[20] M. Luo, H. Wu, H. Huang, W. He and R. He, "Memory-Modulated Transformer Network for Heterogeneous Face Recognition," in *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2095-2109, 2022, doi: 10.1109/TIFS.2022.3177960.

[21] Y. Guo, L. Cao, C. Chen, K. Du and C. Fu, "Domain Alignment Embedding Network for Sketch Face Recognition," in *IEEE Access*, vol. 9, pp. 872-882, 2021, doi: 10.1109/ACCESS.2020.3047108.

# Appendix C: Vision, Mission, Programme Outcomes and Course Outcomes

# Vision, Mission, Programme Outcomes and Course Outcomes

**Institute Vision**

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

**Institute Mission**

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

**Department Vision**

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

**Department Mission**

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

**Programme Outcomes (PO)**

Engineering Graduates will be able to:

**1. Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

**2. Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

**3. Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

**4. Conduct investigations of complex problems**: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**5. Modern Tool Usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

**6. The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**7. Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

**8. Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**9. Individual and Team work**: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

**10. Communication**: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

**11. Project management and finance**: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

**12. Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

**Programme Specific Outcomes (PSO)**

A graduate of the Computer Science and Engineering Program will demonstrate:

**PSO1: Computer Science Specific Skills**
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

**PSO2: Programming and Software Development Skills**
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

**PSO3: Professional Skills**
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

**Course Outcomes (CO)**

**Course Outcome 1:** Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

**Course Outcome 2:** Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

**Course Outcome 3:** Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

**Course Outcome 4:** Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

**Course Outcome 5:** Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

**Course Outcome 6:** Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

# Appendix D: CO-PO-PSO Mapping

## CO-PO AND CO-PSO MAPPING

| | PO 1 | PO 2 | PO 3 | PO 4 | PO 5 | PO 6 | PO 7 | PO 8 | PO 9 | PO 10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | | |
| CO 2 | 2 | 2 | 2 | | 1 | 3 | 3 | 1 | 1 | | 1 | 1 | | 2 | |
| CO 3 | | | | | | | | | 3 | 2 | 2 | 1 | | | 3 |
| CO 4 | | | | | 2 | | | 3 | 2 | 2 | 3 | 2 | | | 3 |
| CO 5 | 2 | 3 | 3 | 1 | 2 | | | | | | | 1 | 3 | | |
| CO 6 | | | | | 2 | | | 2 | 2 | 3 | 1 | 1 | | | 3 |

3/2/1: high/medium/low

## JUSTIFICATIONS FOR CO-PO AND CO-PSO MAPPING

| MAPPING | LOW/MEDIUM/HIGH | JUSTIFICATION |
|---|---|---|
| 100003/ CS722U.1-PO1 | M | Knowledge in the area of technology for project development using various tools results in better modeling. |
| 100003/ CS722U.1-PO2 | M | Knowledge acquired in the selected area of project development can be used to identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions. |
| 100003/ CS722U.1-PO3 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO4 | M | Can use the acquired knowledge in designing solutions to complex problems. |
| 100003/ CS722U.1-PO5 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |

| | | |
|---|---|---|
| 100003/ CS722U.1-PO6 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.1-PO7 | M | Project development based on societal and environmental context solution identification is the need for sustainable development. |
| 100003/ CS722U.1-PO8 | L | Project development should be based on professional ethics and responsibilities. |
| 100003/ CS722U.1-PO9 | L | Project development using a systematic approach based on well defined principles will result in teamwork. |
| 100003/ CS722U.1-PO10 | M | Project brings technological changes in society. |
| 100003/ CS722U.1-PO11 | H | Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms. |
| 100003/ CS722U.1-PO12 | H | Knowledge for project development contributes engineering skills in computing & information gatherings. |
| 100003/ CS722U.2-PO1 | H | Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains. |
| 100003/ CS722U.2-PO2 | H | Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals. |
| 100003/ CS722U.2-PO3 | H | Identifying, formulating and analyzing the project results in a systematic approach. |

| | | |
|---|---|---|
| 100003/ CS722U.2-PO5 | H | Systematic approach is the tip for solving complex problems in various domains. |
| 100003/ CS722U.2-PO6 | H | Systematic approach in the technical and design aspects provide valid conclusions. |
| 100003/ CS722U.2-PO7 | H | Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development. |
| 100003/ CS722U.2-PO8 | M | Identification and justification of technical aspects of project development demonstrates the need for sustainable development. |
| 100003/ CS722U.2-PO9 | H | Apply professional ethics and responsibilities in engineering practice of development. |
| 100003/ CS722U.2-PO11 | H | Systematic approach also includes effective reporting and documentation which gives clear instructions. |
| 100003/ CS722U.2-PO12 | M | Project development using a systematic approach based on well defined principles will result in better teamwork. |
| 100003/ CS722U.3-PO9 | H | Project development as a team brings the ability to engage in independent and lifelong learning. |
| 100003/ CS722U.3-PO10 | H | Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms. |
| 100003/ CS722U.3-PO11 | H | Identification, formulation and justification in technical aspects provides the betterment of life in various domains. |

| | | |
|---|---|---|
| 100003/ CS722U.3-PO12 | H | Students are able to interpret, improve and redefine technical aspects with mathematics, science and engineering fundamentals for the solutions of complex problems. |
| 100003/ CS722U.4-PO5 | H | Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems. |
| 100003/ CS722U.4-PO8 | H | Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations. |
| 100003/ CS722U.4-PO9 | H | Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| 100003/ CS722U.4-PO10 | H | Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products. |
| 100003/ CS722U.4-PO11 | M | Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices. |
| 100003/ CS722U.4-PO12 | H | Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development. |
| 100003/ CS722U.5-PO1 | H | Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.5-PO2 | M | Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to |

| | | |
|---|---|---|
| | | comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. |
| 100003/ CS722U.5-PO3 | H | Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments. |
| 100003/ CS722U.5-PO4 | H | Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. |
| 100003/ CS722U.5-PO5 | M | Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages. |
| 100003/ CS722U.5-PO12 | M | Students are able to interpret, improve and redefine technical aspects and contribute their engineering skills in computing and information engineering domains like network design and administration, database design and knowledge engineering. |
| 100003/ CS722U.6-PO5 | M | Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life. |
| 100003/ CS722U.6-PO8 | H | Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. |
| 100003/ CS722U.6-PO9 | H | Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems |

| | | |
|---|---|---|
| 100003/ CS722U.6-PO10 | M | Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components. |
| 100003/ CS722U.6-PO11 | M | Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data. |
| 100003/ CS722U.6-PO12 | H | Students will be able to associate with a team as an effective team player, applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| 100003/ CS722U.1-PSO1 | H | Students are able to develop Computer Science Specific Skills by modeling and solving problems. |
| 100003/ CS722U.2-PSO2 | M | Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills. |
| 100003/ CS722U.3-PSO3 | H | Working in a team can result in the effective development of Professional Skills. |
| 100003/ CS722U.4-PSO3 | H | Planning and scheduling can result in the effective development of Professional Skills. |
| 100003/ CS722U.5-PSO1 | H | Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems. |
| 100003/ CS722U.6-PSO3 | H | Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills. |