Face Recognition under Occlusion using Robust Autoencoder and Siamese Network

A

Project report submitted for B.Tech. Project

Integrated Post Graduate in Masters Of Technology

in

Information Technology

By

Vaidya Pranay Sunil : 2020IMT-108

Under the Supervision of

Dr.Santosh Singh Rathore

Department of Computer Science

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR

GWALIOR, INDIA

DECLARATION

I hereby certify that the work, which is being presented in the report/thesis, Face Recognition under Occlusion using Robust Autoencoder and Siamese Network, in fulfillment of the requirement for the award of the degree of Bachelor of Technology and submitted to the institution is an authentic record of my/our own work carried out during the period May-2023 to Aug-2023 under the supervision of Dr. Santosh Singh Rathore. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Dated: **Signature of the candidate**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dated: **Signature of supervisor**

Acknowledgements

I am highly indebted to Dr. Santosh Singh Rathore, for his esteemed mentorship, and for allowing me to freely explore and experiment with various ideas in the course of making this project a reality. The leeway I was given went a long way towards helping cultivate a genuine hunger for knowledge and keeping up the motivation to achieve the best possible outcome. I can genuinely say that this Bachelor's Thesis Project (BTP) made me explore many areas of machine learning that are new to me, and kindled an interest to further follow up on some of those areas. Moreover, the semi-successful completion of this project has brought with it great satisfaction and more importantly, confidence in my ability to produce more high-quality non-trivial artificial intelligence systems that can make a difference in the real-world. I would like to sincerely express my gratitude to this prestigious institution for providing me and my colleagues with the opportunity to pursue this BTP. It is an honor to be able to work on such an important academic project under the guidance and support I am provided with. I am grateful for the resources and facilities provided by this institution, which have been instrumental in enabling me to conduct my research and complete this project. Moreover, I deeply appreciate the efforts of my professors in mentoring and fairly evaluating our works.

Vaidya Pranay Sunil

Abstract

Addressing the issue of recognizing partially occluded faces is complex in real-world scenarios. While In recent years many techniques focus on restoring occluded image details, they often overlook the overall facial expressions and structural information. In for project we present a novel approach. Leveraging the power of a robust convolutional autoencoder based on the UNet architecture, we address occlusions by effectively deoccluding the images. Our method involves passing the occluded image through the autoencoder, which removes the occlusions and yields a deoccluded version. This deoccluded image then serves as a foundation for recognition through a Siamese network. In this recognition step, we employ advanced feature embedding techniques using the Xception and MobileNetV3 architectures. Together, our approach creates a comprehensive solution for occluded face recognition. Our dataset deals with 2 type of occlusions, including items like masks and sunglasses.

Contents

List of Figures

List of Tables

1

Introduction

This chapter offers an overview of the subject matter by presenting background information on face recognition technique with and without occlusion. Moreover, this chapter entails the motivating factors that instigated the investigation of this topic.

1.1 Background

Face recognition has received a lot of attention, which is primarily due to the rapid development of deep learning techniques and the expanding size of public face data sets. Face recognition has become a key technology with a variety of uses, including user authentication, security, and human-computer interaction.

The traditional methods are Eigenfaces [6], Fisherfaces [7], and Local Binary Patterns (LBP) [8] were the pillars of earlier face recognition methods. These techniques centred on dimensionality reduction methods for recognition and facial feature extraction. Face recognition has changed drastically with the deep learning techniques, particularly CNNs for example VGG-Face [9] that based on the VGG architecture and can recognize a large number of individuals with high accuracy and FaceNet [10] which maps face images into a high-dimensional space where face similarity can be measured by the Euclidean distance. they have attained state-of-the-art performance in a variety of face recognition tasks and can automatically train hierarchical feature representations from raw pixel data. Siamese networks are a type of neural network architecture designed for one-shot learning. They learn to distinguish between pairs of images, making them suitable for scenarios with limited training data. Combination of Siamese network and convolutution methods have drastically increase the performance of face recogintion

Face recognition still confronts a number of difficulties, nevertheless, in spite of recent developments. The existence of occlusions on the face is one of the biggest obstacles. Occlusions happen when items like sunglasses, masks, scarves, or even hair partially occluded face features. These occlusions can have a significant negative effect on how well facial recognition algorithms work, resulting in false matches and misidentifications. Even the best methods like VGG-Face [9] and FaceNet [10] struggle when faces wear masks or accessories, like scarves and sunglasses. It's a tough challenge that's hard to crack.

1.2 Occluded Face Recognition

A face that is partially hidden or covered by items like glasses, masks etc. is known as an occluded face. It makes it challenging for face recognition systems to accurately identify and match faces, which poses challenges. Face recognition technology needs to be handled carefully if it is to perform well in real-world scenarios

Deep learning-based face recognition techniques have been popular and robust algorithms performance of face recognition can be significantly impacted by occluded objects, which causes a steady decline in accuracy when handling occluded faces.

Occluded Face recognition is a rapidly developing field, and there is ongoing research aimed at improving the accuracy and efficiency of face Recognition algorithms. Instead of relying solely on the full face, focusing on specific facial features that are less likely to be occluded, such as the eyes or the nose, can lead to better recognition rates. this Process can be achieve by providing a region of interest (ROI) containing the face, Facial landmark detection to identify key points on the face, such as the eyes, nose, mouth, and other facial landmarks Using MTCNN [11] and Attention mechanism [12] . These points serve as anchors to locate specific facial features accurately and feature extraction. deep learning-based methods have shown great promise in feature extraction and recognition, offering significant advancements in face recognition systems.

Occlusions can be handled effectively using reconstruction-based techniques [13] [14], especially when the occluded regions exhibit certain patterns or when there is sufficient data in the visible portions to generate precise estimations. The complexity of the occlusions and the calibre of the information at hand, though, determine how well these strategies work. An example of a Reconstruction-based technique is an Autoencoder [15]. They are great at extracting patterns and representations from data, which makes them suitable for occluded face recognition applications. The network can learn to reconstruct the full face even when some parts are obscured by training an Autoencoder on Unoccluded faces. They leverage available information to estimate and reconstruct occluded regions,

contributing to improved recognition accuracy in challenging scenarios.
Autoencoders and Siamese networks [16] are two examples of novel techniques. De-occlusion is made easier by Autoencoders, which restore occluded regions for better recognition. Siamese networks acquire strong feature embeddings, improving recognition when objects are obscured. These techniques are essential for improving.

1.3 Motivation

Occluded face recognition is highly motivated by its usefulness and potential for use in practical contexts. Faces are the most common way that people may be recognised, and they are crucial to social interactions, security, and authentication systems. Facial recognition is difficult in real-world situations, though, when faces are partially hidden by clothing, masks, or other occlusions.

(i) Real-World Relevance: The practical use of obstructed face recognition in several real-world contexts serves as its primary source of inspiration. Faces are essential for identifying people, yet in daily life, they are sometimes partially hidden by clothing, masks, or other objects. For applications in security, surveillance, access control, and other fields, it is crucial to create strong face recognition systems that can manage occlusion

(ii) Enhanced Security: By properly recognising people even when their faces are partially veiled, occluded face recognition may considerably enhance security measures. This has substantial consequences for surveillance and law enforcement, since it is critical for ensuring public safety to identify possible threats and suspects in crowded or concealed circumstances.

(iii) User-Friendly identification: Occluded face recognition is essential for maintaining a seamless and safe user experience as facial identification becomes more widely used in devices like smartphones and laptops. It lessens user annoyance and improves device security by identifying people even with partial blockages.

(iv) Challenging Scenarios: Researchers and developers encounter interesting hurdles when dealing with occluded face recognition. Innovative methods and algorithms are needed to address these problems, advancing the fields of computer vision and deep learning research

(v) Enhanced User Privacy: In applications like access control and authentication, partial face recognition lessens the need to acquire entire facial data, improving user privacy

1.4 Report Organization

(i) **Chapter 1**: Introduce you to general concept of face recognition, why it is challenging and what is our motive behind this project.

(ii) **Chapter 2:** Describes the previous works that have been carried out in the field of Face recognition under occlusion which he have studied in order to gain an all round knowledge of the subject we are working upon. It also highlighted the limitations of work done in this field and formulated the research objective.

(iii) **Chapter 3:** Describe the problem statement accourding to research gap and objective of our thesis.

(iv) **Chapter 4:** Describes the approach we have taken. It also contains the architecture of our modified network, newer activation functionsand loss which we have used and also provide an explanation for why they have been chosen.

(v) **Chapter 5:** Describe the experiments and result for our developed model. it also shows the comparison with existing model.

(vi) **Chapter 6:** Describe Conclusion and future work regarding our project

.

2

Literature Review

This chapter responds to the significant amount of research assigned to understanding the various de-occlusion and face recognition methods with and without occlusion. We examine some of the literature and briefly review the development of former proposed methods and the research gaps.

2.1 Review on Existing Occluded face Recognition Methods

Over the years, researchers have introduced numerous methods aimed at achieving improved accuracy in recognizing occluded faces. The field of occluded face recognition remains an active area of research due to its novelty and critical importance. The following studies represent significant contributions by researchers in the field of occluded face recognition.

The paper **Face Recognition with Occlusion** [17] introduces a regression-based method to improve face recognition accuracy in the presence of occlusions like masks and sunglasses. The study proposes a novel occlusion detection approach that combines information from the original and residual images, which captures occluded details. The research also demonstrates that utilizing non-occluded facial regions for recognition yields better results than using reconstructed images. This dual approach of detection and feature selection offers insights for enhancing face recognition in real-world scenarios with occlusions.

The paper **FROM: Face Recognition with Occlusion Masks** [18] presents a novel approach to enhance face recognition performance in the presence of occlusions.which is designed to detect and correct corrupted features caused by occlusions using dynamically learned masks.using mask decoder and occlusion pattern predictor. The robust architecture of FROM incorporates a Feature Pyramid Network (FPN) that extracts a comprehensive feature map. This map is subsequently utilized for convolutional decoding, aided by landmark techniques, to generate masks. These masks are pivotal in extracting occluded areas from images, playing a crucial role in enhancing face recognition.

The paper **Reconstruction of Partially Occluded Face PCA (principal component analysis)** [19] introduces a quick recursive PCA method for addressing face occlusions. It starts with a training phase involving face normalization and PCA to obtain eigenfaces. During testing, occluded faces are restored through iterative analysis and

synthesis steps. The damaged face is normalized, PCA coefficients obtained, and reconstruction performed. The algorithm's effectiveness is demonstrated through faster convergence compared to classical PCA compensation, resulting in natural-looking reconstructed faces. The paper **Occlusion Robust Face Recognition Based on Mask Learning With PSDN** [20] introduces a novel approach for occlusion robust face recognition using a Mask Learning strategy with a Pairwise Differential Siamese Network (PDSN). Inspired by human visual perception, which disregards occlusions, the method establishes a mask dictionary through a PDSN that captures the differences between occluded and non-occluded facial features. This dictionary yields Feature Discarding Masks (FDMs) to eliminate corrupted feature elements during recognition. Results from both synthesized and real-world occluded face datasets highlight the method's effectiveness, showcasing its strong generalization capabilities for broader face recognition tasks.

The paper introduces a **robust LSTM-Autoencoders (RLA)** [21] model for restoring partially occluded faces in real-world scenarios. Consisting of two LSTM components, the RLA model focuses on occlusion-robust face encoding and recurrent occlusion removal. The multi-scale spatial LSTM encoder captures latent representations by reading facial patches of various scales, enhancing occlusion-robustness.Supervised and adversarial CNNs are integrated to refine the robust LSTM autoencoder, boosting identity-related details in recovered faces.

The paper **Robust Deep Auto-encoder for Occluded Face Recognition** [22] addresses the challenge of occlusions in face recognition by introducing a deep learning-based approach, specifically the Double Channel SSDA (DC-SSDA) model. Unlike previous methods, DC-SSDA doesn't require prior knowledge of occlusion types or locations.the proposed method demonstrates to various occlusion types and locations, leading to substantial performance enhancement in recognizing occluded faces.

The paper **Fully convolutional Deep Stacked Denoising Sparse Auto encoder network for partial face reconstruction** [23] introduces a novel approach called Self-Motivated Feature Mapping (SMFM) for partial face reconstruction using a combination

of Fully Convolutional Networks (FCN) and Deep Stacked Denoising Sparse Autoencoders (DS-DSA). The proposed method focuses on generating feature maps from FCN and utilizes DS-DSA for partial face reconstruction.

The paper **Self-restrained triplet loss for accurate masked face recognition** [24] addresses the challenge of recognizing masked faces. They introduce the Embedding Unmasking Model a simple yet powerful cnn model which enhances existing face recognition systems. They also propose a novel loss function called Self-restrained Triplet (SRT) to make EUM generate embeddings similar to unmasked faces of the same individuals.

The paper **Occluded Face Recognition by Identity-Diversity Inpainting** [13] presents an innovative approach to addressing occluded face recognition challenges using identity-diversity inpainting. By combining generative adversarial networks (GAN) with a pre-trained CNN recognizer, the proposed method effectively reconstructs occluded faces while maintaining accurate identity cues. The integration leverages GAN for face inpainting and CNN for representation, enhancing both tasks simultaneously. The approach outperforms four existing methods in experimental comparisons, demonstrating its efficacy in occluded face recognition.

The survey paper [25] comprehensively explores occluded face recognition methods. It addresses challenges posed by factors like occlusion. The study categorizes approaches based on manual design features and deep learning For occluded face recognition, it analyzes 2D and 3D methods. Under 2D, the survey delves into robust feature extraction and subspace regression. 3D methods are considered the future mainstream due to inherent robustness; the survey discusses surface, local surface, and deep learning-based approaches

2.2 Research Analysis of Existing Methods

The field of occluded face recognition has seen significant advancements through a series of research papers focused on addressing challenges posed by occlusions. Research in occluded face recognition employs diverse techniques to boost accuracy when faced with

obstructions. These approaches acknowledge the vital role occlusions play in real-world challenges. Researches have introduced robust strategies to tackle occlusions and improve recognition outcomes.

Several papers propose the utilization of deep learning models, such as convolutional neural networks (CNNs) and autoencoders, to restore occluded facial regions. By leveraging the power of these models, the studies introduce techniques like recursive PCA, identity-diversity inpainting, and mask learning to effectively recover facial features hidden by occlusions. Some of them use combination of tow or more approach

Matched-filter sensing needs a smaller number of samples to provide effective performance for detection, however, it also requires foreknowledge of the signal- which isn't always feasible. Moreover, in the case of primary user emulation attacks where an attacker node imitates the primary user's signal, this method has been undependable. This makes matched filter sensing impractical in real-world circumstances.

Additionally, methods like landmark techniques, adversarial training (e.g., Generative Adversarial Networks or GANs), and novel loss functions are utilized to enhance the accuracy and robustness of the models. For instance, models like "FaceNet" [10] and "DeepID" [26]incorporate these techniques to achieve accurate face recognition while effectively countering the impact of occlusions.

A Famous strategy is to combining feature extraction, restoration, and recognition in one approach. Methods like Pairwise Differential Siamese Networks (PDSN) [20], robust LSTM-Autoencoders (RLA) [21], and the Embedding Unmasking Model (EUM) [24] use complex structures to tackle occlusion detection, restore features, and achieve accurate recognition all at once. This comprehensive approach recognizes the intricate nature of occluded face recognition tasks.

In summary, the research papers show progress in recognizing faces even when they're partly covered. They emphasize the importance of dealing with such challenges and suggest new methods involving features, fixing, and recognizing. All of this work aims to make better face recognition systems for handling real situations of occlusion.

2.3 Research Gaps

Many study focus on various types of occlusion but real world scenario which involve various combination of occlusion, with varying lighting condition this situation require more complex approach to ensure robust recognitioin even in real world scenario

For applications like surveillance and security, real-time face recognition is essential. Many existing models are not optimized for real-time processing due to there complexity behaviour. Approach should we lightweight enough without compromising accuracy and performance

Every model and approach which we have discuss above work only on occluder on which they have train in real life there might be some unseen occlusion then that approach may get fail. In this paper by Haibo Qiu in 2022 [18] have use numerous type of occluder for training and testing purpose have succeed in tackle this problem.

Many of the model were too complex to handle there weights on larger and different datasets.

3

Problem Statement based on Identified Research Gaps

3.1 Problem Statement

In the field of face recognition, the presence of occlusions such as masks, sunglasses, or facial hair poses a significant challenge to accurate identification, especially in real-world scenarios. While several methods have been proposed to address occluded face recognition, there remain critical research gaps that need to be tackled for more effectivly.

The main problem revolves around the fact that existing approaches often excel at specific occlusion types or controlled settings but struggle to adapt to the complex and dynamic nature of real-world occlusions. Furthermore, the challenge of handling unseen occlusions and ensuring real-time performance adds to the complexity of the problem.

This system should be able to adapt to new, unseen occlusions and perform efficiently in real-time applications without compromising accuracy. Moreover, privacy concerns related to masked face recognition need to be addressed to ensure ethical and secure deployment.

The goal is to make this system work well in real-life situations, from security scenarios to everyday settings. Plus, it's important to think about privacy concerns when recognizing faces with masks.

3.2 Thesis Objective

- To Devlop robust autoencoder model that remove occlusion and reconstruct image with higher performance, similarity and faster rate.
- To Devlop Siamese network for better face recognition process for a image which is de-occluded using autoencoder.
- To Evaluate their performance using the qualitative comparison ,quantitative comparison and performance matrix and graph.

4

Proposed Methodology

This chapter provides a comprehensive discussion of the methodology employed in the project, including the mathematical operations and architecture of our models that were implemented.

4.1 Autoencoder

Autoencoders are a type of neural network-based unsupervised learning approach designed for data representation and reconstruction. [15]They are structured to compress input data into a lower-dimensional representation and then decode it back to the original input space. In the context of neural networks, the role of autoencoders is to enforce network compression. an autoencoder consists of three main components: an encoder, a bottleneck, and a decoder.

- The encoder is the initial part of the autoencoder. It compresses input data into a lower-dimensional space known as the "encoding" or "latent space." [26] This space captures crucial features of the input through linear and non-linear operations. The encoder's job is to distill the most important aspects of the data into a compact form, containing enough details for the decoder to recreate the original data precisely.

$$h = f(x) = \sigma(Wx + b) \tag{4.1}$$

- The bottleneck, or "encoding layer," is pivotal in the Autoencoder design. It holds the compact input data representation. By constraining dimensionality, it captures vital features while discarding less crucial data. This layer's dimensionality is usually much smaller than the input's. It emphasizes feature selection and dimension reduction, prompting a concise and meaningful data representation.

$$h = Bottleneck(f(x)) \tag{4.2}$$

- The decoder is the next part of the autoencoder. It reconstructs the original data from the compressed version obtained from the bottleneck. It strives to produce output that looks like the input. The decoder has layers that transform the compressed data to match the original. Its job is to undo the compression by the encoder, making the compressed data look like the original input.

$$r = g(h) = \sigma(W'h + b') \tag{4.3}$$

4.2 System Modeling

4.2.1 Architecture

Our model consists of two key stages: the **autoencoder** and the **Siamese network**. In the upcoming sections, we will explored each of these components individually. Now, let's discuss how these two stages collaborate when it comes to recognizing occluded images

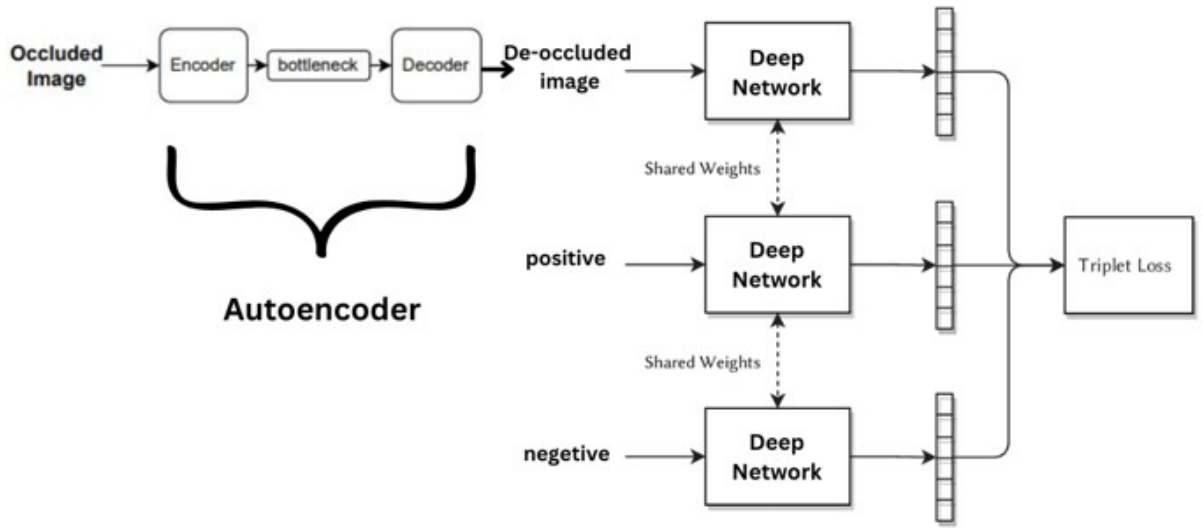

Figure 4.1: Final Model

In Figure 4.1 The process begins with the occluded image being fed into our autoencoder, which effectively eliminates the occlusions from the image. This de-occluded image then serves as the anchor image, playing a crucial role in the subsequent steps. Our Siamese network comes into play here, determining the minimum Euclidean distance between the anchor image and both positive and negative images. Through this comparison, we arrive at a recognition decision for the occluded image.

4.2.2 Autoencoder Architecture

The autoencoder we've employed is a type called a robust convolutional autoencoder, designed with three main components: an encoder, a code conversion module, and a decoder.

4.2.2.1 Encoder

The encoding phase of the U-Net architecture [27] involves capturing the crucial features of an input image and representing them in a compressed form. This is done through a series of convolutional layers that progressively extract higher-level features from the input image.

In figure 4.2 Module 1 incorporates conv2D operation with a stride of (1, 1) and a ReLU activation function along with Batch Normalization which helps gradients to flow better during backpropagation, leading to faster convergence and allowing the network to train more effectively. module 1 is followed by MaxPooling layer and droupout layers to reduce the spatial dimensions of the feature maps and focus on important information and to prevent overfitting respectively. process in carry till we reduce our latent space to minimal.

4.2.2.2 code Conversion

As the bottleneck, we have introduce "code conversion" This step is particularly designed to handle occluded face images, where visual information loss can be significant. To tackle this challenge, we introduce a specialized code "module 2" after full convolutional encoding. This module bridges the gap between encoding and decoding, ensuring effective restoration of occlusion-induced losses. By transforming extracted codes into a uniform length format through full connection transforms, we utilize all image pixels to reconstruct lost details. This strategic enhancement not only empowers the decoder for better restoration but also improves the encoding process itself. Our approach, inspired by the U-Net architecture [27], results in superior image quality during decoding.

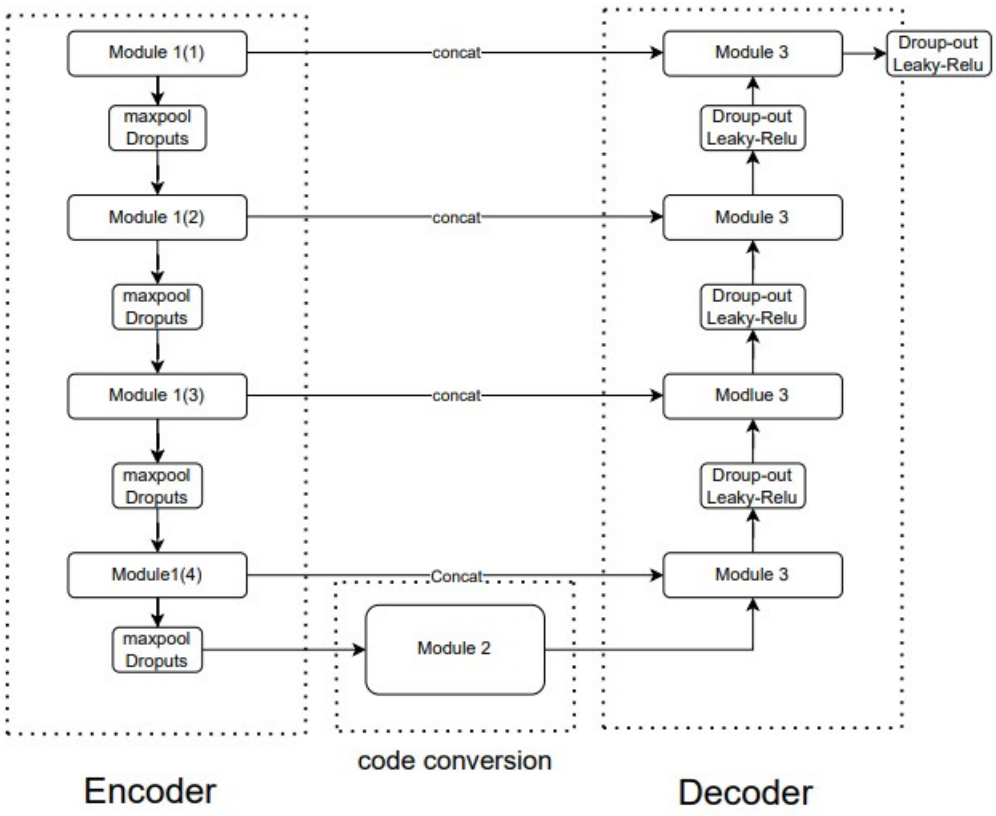

Figure 4.2: End to End De-Occlusion model using Autoencoder

4.2.2.3 Decoder

The decoding phase of the U-Net focuses on reconstructing the input image from the compressed representation stored in the bottleneck. This phase aims to create a restored image that closely resembles the original while remove the impact of occlusions.

In Figure 4.2 Module 3 incorporate of Conv2DTranspose layers with a stride of (2,2) and activation function leaky ReLU which will prevent problem of dying ReLU. Module 3 will increasing the spatial dimensions. The feature maps from the encoding phase and decoding phase are concatenated to combine global and local information necessary for accurate reconstruction. or we can simplly say that they are sharing weight to reduce the pixel loss. so that we get accurate restoration of image. Convolutional layers gradually expand the feature maps, refining the details to match the original image dimensions. Dropout

layers continue to prevent overfitting, ensuring the reconstruction process doesn't rely too heavily on specific features.

4.2.2.4 Reconstruction

The ultimate objective of our Autoencoder is to generate a reconstructed image that effectively removes occlusions and retains the key features.
During training, the autoencoder aims to minimize the mean squared error (MSE) which also called Pixel Loss between the reconstructed image and the original input image.

4.2.2.5 Activation functions

Rectified Linear Unit

ReLU function outputs the input value x if it's positive or zero, and it outputs zero for any negative input value. In other words, ReLU introduces non-linearity by allowing only positive values to pass through, effectively "activating" the neuron if the input is positive.

$$f(x) = \ max(0, x) \tag{4.4}$$

Leaky Rectified Linear Unit

Leaky ReLU is a variation of the Rectified Linear Unit (ReLU) activation function used in neural networks. It introduces a small, non-zero slope for negative inputs, which helps prevent the "dying ReLU" problem where neurons become inactive during training.The function is defined as:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{cases} \tag{4.5}$$

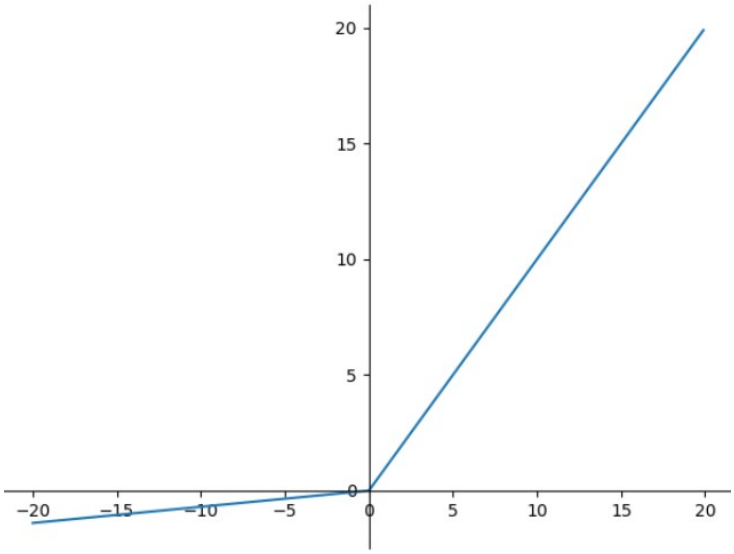

Figure 4.3: graph of Leaky-ReLU [1]

4.2.2.6 Pixel Loss

To ensure that the essential facial identity information is maintained, we introduce the concept of pixel loss. This loss metric serves the purpose of quantifying the preservation of key facial features in the generated image. The pixel loss operates by computing the normalized Euclidean distance between the generated output image, denoted as , and the target image, denoted as y. Both the generated output image and the target image y possess a structured arrangement with dimensions C x H x W. Mathematically, the pixel loss can be expressed as follows:

$$\text{Pixel Loss} = \frac{1}{C \times H \times W} ||(\hat{y} - y)||^2 \tag{4.6}$$

The pixel loss helps the network create images that look more like the target face by guiding the optimization process.

4.2.3 Siamese Network

A Siamese network for face recognition [16] is a type of neural network architecture designed to learn and compare similarities between pairs of images. It's commonly used for tasks like facial verification and recognition. [28]The network consists of two identical subnetworks that share weights. Each subnetwork processes one image from a pair, and the outputs are compared to determine if the images belong to the same person or not. This architecture is effective in learning features and handling variations in lighting, pose, and expression. It's particularly useful for tasks where labeled training data is limited, as it doesn't require a large amount of labeled pairs for training.

4.2.3.1 Architecture

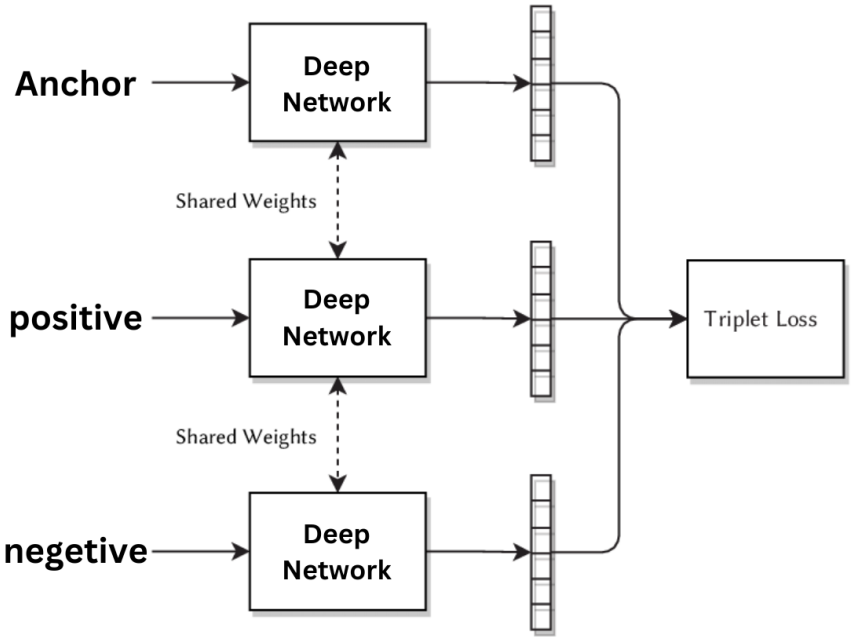

Figure 4.4: Siamese Network Architecture [2]

We're designing a Siamese Network that works with 3 input images: anchor, positive, and negative. The images are encoded into feature vectors using the provided Deep Network. A distance layer then calculates the differences between anchor and positive, as well as anchor and negative pairs. This helps in distinguishing between similar and dissimilar images, which is essential for tasks like face recognition or similarity comparisons.

Deep Network

The Encoder turns images into feature vector. We use a pre-trained model either Xception (extension of inception v3) or Mobilenetv3. By using transfer learning we can speed up training and needs less data. The Model connects to Fully Connected (Dense) layers and Batch normalization. The last layer uses L2 Normalisation to adjust data. L2 keeps sums of squares up to 1 in each row.

$$L2Norm = \|x\|_2 = \sqrt{\sum_{i=0}^{n} |x_i|^2} \tag{4.7}$$

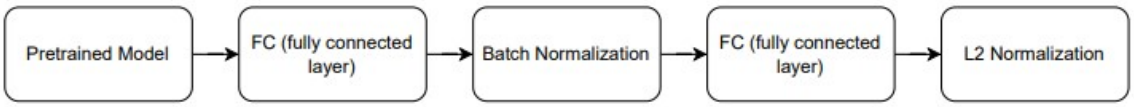

Figure 4.5: Deep Network

4.2.3.2 Models

Xception

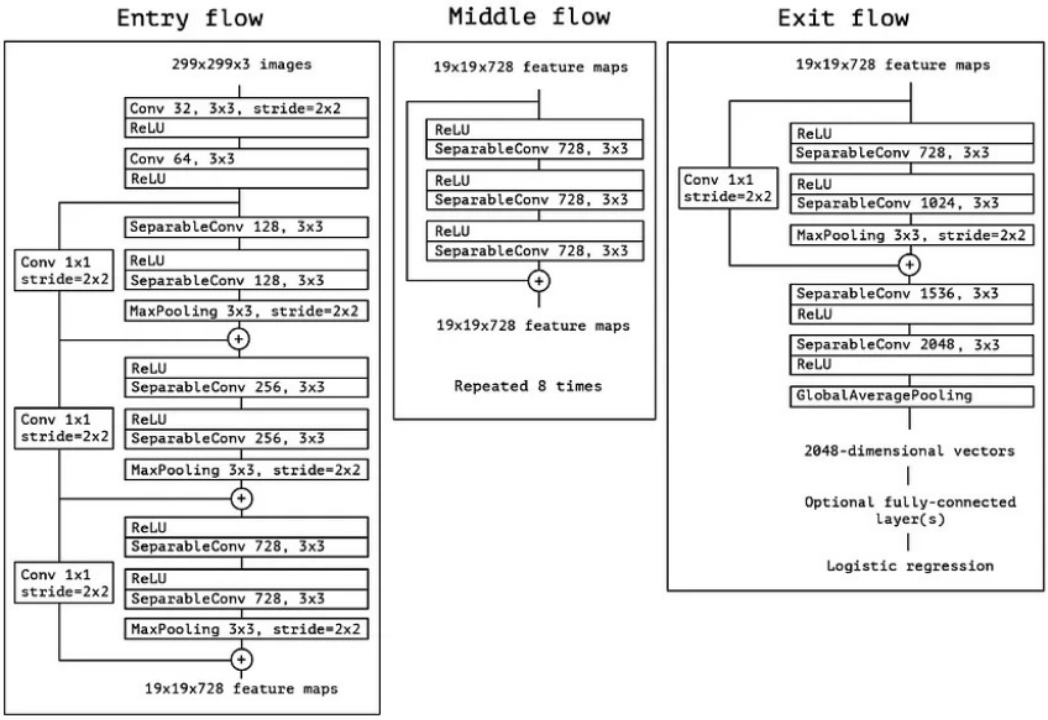

Figure 4.6: Xception Architecture [3]

Xception, [29] short for "Extreme Inception," is a deep neural network architecture that builds upon the principles of the Inception. Xception employs depthwise separable convolutions, where it first performs separate convolutions on each input channel followed by combining them using pointwise convolutions. Xception reduces both parameters and computation compared to traditional convolutions, making it efficient. It learns hierarchical features through multiple layers, capturing complex patterns for image recognition. The architecture excels in accuracy and efficiency, making it suitable for various computer vision tasks. we set the trainable parameter false till middle flow where we get our feature map.

MobilenetV3

MobileNetV3 [30] designed for lightweight and efficient neural networks. It reduces computations while maintaining accuracy. MobileNetV3 employs depthwise separable convolutions. Instead of standard convolutions, it uses two separate layers: depthwise convolution and pointwise convolution. MobileNetV3 introduces new activation functions like "h-swish" and "hard-swish." These optimize computation by combining activation and multiplication in one step. MobileNetV3 starts with initial layers that perform basic processing on the input image, such as convolution and normalization.The architecture includes bottleneck blocks. Blocks combine 1x1 pointwise convolutions, depthwise separable convolutions, and "h-swish" activations. Squeeze-and-excitation (SE) blocks are a subset of some blocks. These improve essential features by adjusting channels according to their significance. h-swish function is define as:-

$$\text{h-swish}\,(x) = x\frac{\text{ReLU6}\,(x+3)}{6} \tag{4.8}$$

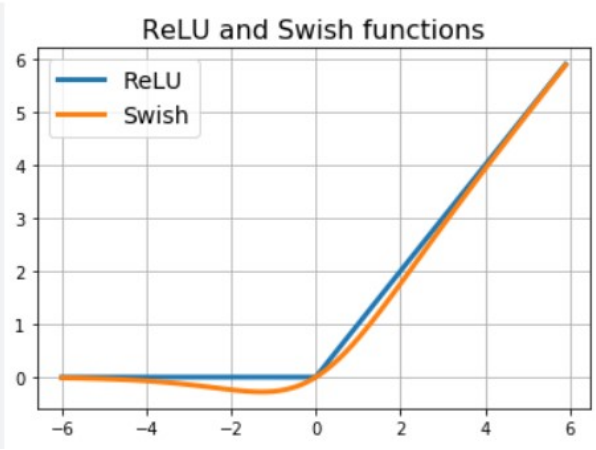

Figure 4.7: swish function [4]

4.2.3.3 Triple Loss

Triplet loss trains a machine-learning model to differentiate between items. It employs groups of three, or triplets, comprising an anchor, a similar (positive), and a dissimilar (negative) item. [31] This helps the model learn to understand similarities and distinctions

between items. In this scenario, The triple loss function aims to reduce the gap between embeddings of anchor and positive face images, while simultaneously increasing the gap between embeddings of anchor and negative face images. This encourages the model to better discriminate between similar and dissimilar faces.

Triplet loss calculates the distances among anchor, positive, and negative samples. The distance matrix is computed using the Euclidean distance metric, helping to measure the dissimilarity and similarity between these samples effectively.

euclidean distance formula:-

$$\mathrm{p} = \|f_i^a - f_i^p\|_2^2 \tag{4.9}$$

$$\mathrm{n} = \|f_i^a - f_i^n\|_2^2 \tag{4.10}$$

Total Triple Loss Function is give my:-

$$\text{Loss} = \sum_{i=0}^{N} \left[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha\right]_+ \tag{4.11}$$

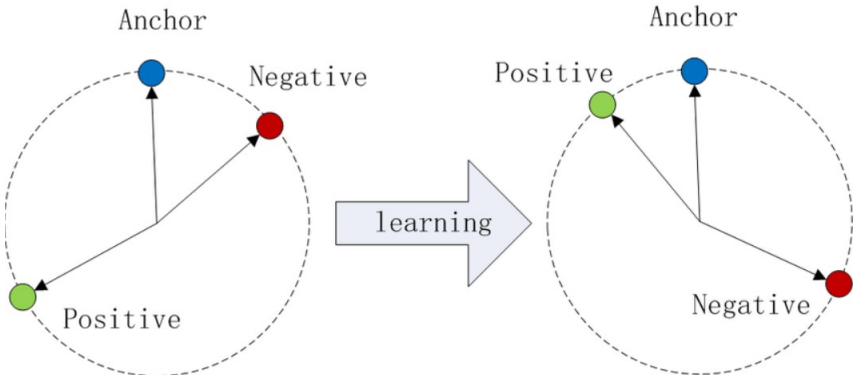

Figure 4.8: Triple Loss [5]

5

Experiment and Results

The contents of this chapter encompass the description of the experiment set-up that was employed in the project, as well as a detailed account of the outcomes and results that were obtained.

5.1 Experiment setup

5.1.1 Dataset

Dateset which is use is our Projects are Images that are Generated Using Style GAN-2, 10,000 images With occlusion and without occlusion to train our autoencoder. then LFW (labeled faces in wild), it consist of 1680 of identities with 6107 of total images.

5.1.2 Data Preparation

To make face De-occlusion work better, we need to add synthetic occlusion like sunglasses and masks. This will help for better training of autoencoder.This helps the program learn how to recognize faces even when they're covered. By including various types of occlusions, the autoencoder becomes adept at handling challenging conditions, resulting in improved face recognition accuracy and reliability.

To tackle this issue we proposed mtcnn. MTCNN will detect the Landmarks such as eyes, nose, mouth,etc. and the According to landmarks we will add occlusion like sunglasses on eyes and mask on mouth and nose etc.

Algorithm : Adding Occlusion Using MTCNN

---

1) Load MTCNN face detection model

2) Load face image and occluder image with transparency

3) Detect face and Landmarks using MTCNN

4) **For** each detected face:

5) Calculate occluder size and position based on landmark positions

6) Resize occluder image to fit landmark

7) Overlay occluder on face image

8) save modified face image

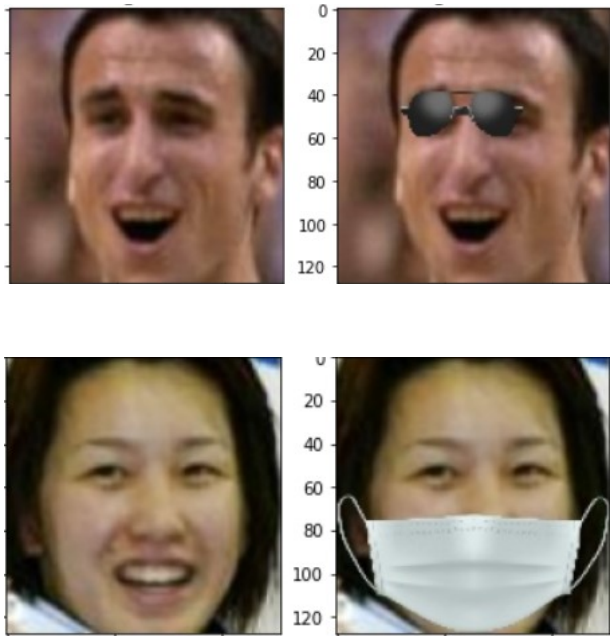

Figure 5.1: occlusion Added

5.1.3 Training

According to our Proposed Architecture our model is divided into 2 stage. Autoencoder and siamese network. firstly we need to train autoencoder. then our siamese network using De-occluded image as anchor image we will then train our proposed Siamese network.

Training Autoencoder

Before training the Autoencoder, we prepare our images through preprocessing. The input images should be resized to dimensions of (128, 128). While loading the images, they come in the BGR format, so we convert them to RGB. Finally, we ensure that the image is normalized and converting its data type to a 32-bit floating-point format. These steps help ready the images for effective training.

Algorithm : Training Autoencoder

---

1) Input: Occluded images {X} ; Clear images {Y}

2) repeat:

3) $\{X\} \rightarrow Encoder \rightarrow C1$

4) $C1 \rightarrow CodeConversion \rightarrow C2$

5) $C2 \rightarrow Decoder \rightarrow \{\hat{Y}\}$

6) $Updateautoencodersusing\{Y;\hat{Y}\}byminimizingLossFunction$

7) **until**$AutoencoderTrained$

8) **return**

Table 5.1: Hyper-parameters for Autoencoder

Hyper-parameters	Value
Learning Rate	0.001
epoch	200
Optimizer	Adam

Training Siamese Network

Images undergo identical preprocessing as in the autoencoder. Subsequently, we divide them into train and test sets for training. By using these sets, we construct triplets comprising anchor, positive, and negative face data. We then generate batches of face data, preprocess them for pretrained model, and pass them through a specific network model. This batch data serves as the foundation for training the model.

Table 5.2: Hyper-parameters for Siamese Network

Hyper-parameters	Value
Learning Rate	0.001
epsilon	0.1
Batch Size	256
epoch	50
Optimizer	Adam

5.2 Results and Discussion

5.2.1 Qualitative Analysis

Our baseline model employs a standard Autoencoder for de-olusion reconstruction. In contrast, our proposed model introduces an innovative approach called the Robust Convolutional Autoencoder using the U-Net architecture.

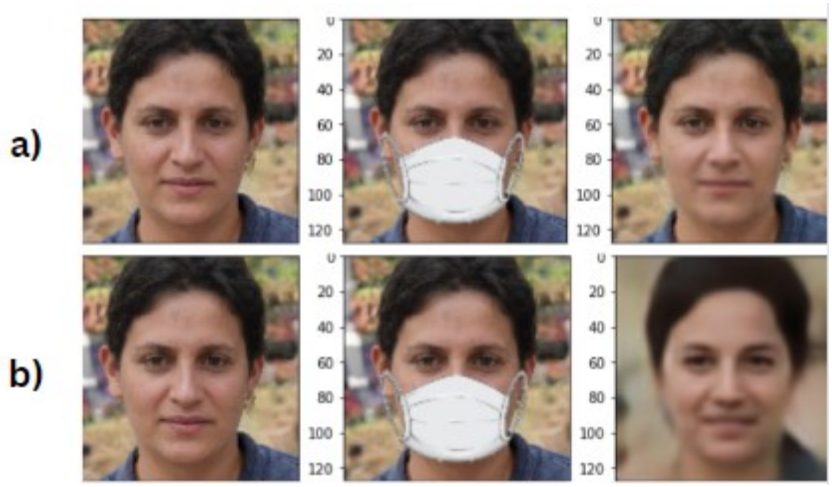

Figure 5.2: a - our proposed model
b- AE

Figure 5.3: a - our proposed model
b - AE

5.2.2 Quantitative Analysis

we need to measure the effectiveness of the proposed approach. we introduce 2 quantifier to measure the effectiveness of the proposed model in face restoration the effectiveness of the proposed approach in facial restoration

5.2.2.1 PSNR

PSNR [32] also Known as Pick signal to noise ratio. PSNR states how well the reconstructed image matches the original. If we have a clean image (I) and a occluded image (K) both with a size of m x n, we calculate the mean square error (MSE) between them.and MAX is maximum pixel value This helps us understand the quality of the restored image compared to the initial image. PSNR is a common way to measure image quality by comparing pixel differences. It's a popular method for evaluating images objectively. Range from 10 to 30 percent consider to be moderate reconstruction. 30 above consider as a good reconstruction.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i,j) - K(i,j))^2 \tag{5.1}$$

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \tag{5.2}$$

Table 5.3: Comparsion Table for PSNR

Occluders	AE	SSDA	DCSSDA	our(proposed)
sunglasses	24.21	26.98	28	**31.06**
mask	22.54	24.04	25.67	**29**
both	21.20	22.63	25.01	**28.93**

5.2.2.2 SSIM

SSIM [32] also known as structural similarity. this matric is use to measure the similarity between two images. It views image composition's structural information as separate from brightness and contrast. This captures object structures in the scene. It models image

distortion as a mix of brightness, contrast, and structure. we estimate brightness using the mean value, contrast using the standard deviation, and structural similarity using covariance. The range of SSIM vary from 0 to 1. when 2 image are indentical means SSIM tends to 1.

In the following function μ is mean, σ is variance, σ_{xy} is covariance

$$\mu = \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x(i,j) \tag{5.3}$$

$$\sigma_x = \left(\frac{1}{HW-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [x(i,j) - \mu_x]^2 \right)^{\frac{1}{2}} \tag{5.4}$$

$$\sigma_{xy} = \frac{1}{HW-1} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [x(i,j) - \mu_x] [y(i,j) - \mu_y] \tag{5.5}$$

Brightness, contrast and structural similarity are define as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{(\mu_x)^2 + (\mu_y)^2 + C_1} \tag{5.6}$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{(\sigma_x)^2 + (\sigma_y)^2 + C_2} \tag{5.7}$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{(\sigma_x)^2 + (\sigma_y)^2 + C_2} \tag{5.8}$$

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{5.9}$$

where . C1; C2, and C3 are constants used to maintain stability. The SSIM of two images can be define as:

$$\text{SSIM}(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \tag{5.10}$$

In practice we usually set $\alpha = \beta = \gamma = 1$ and C3 = C2/2. by taking these parameter values our final SSIM of 2 images can be define as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5.11}$$

Occluders	AE	our(proposed)
sunglasses	0.80	**0.96**
mask	0.74	**0.91**
both	0.71	**0.93**

Table 5.4: Comparsion Table for SSIM

5.2.3 Results for Recognition

we have build our Siamese network architecture is 2 different ways we leveraged the power of pretrained models Xception and MobileNetV3. Transfer learning involves taking advantage of the knowledge gained by these models during their initial training and by fine tuning we enable them to extract relevant features from input images. these model will generate embedded feature vectors to perform face recognition. We'll evaluate their accuracy, compare their performance materics such as Precision score, Recall F1 score and there limitations.

5.2.3.1 Metrics

Accuracy

represents the ratio of correctly predicted instances to the total number of instances in a dataset, providing a measure of overall model correctness.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision

It measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It highlights the model's ability to avoid false positive predictions.

$$\text{Precision} = \frac{TP}{TP+TN}$$

Recall

It is a metric that measures the proportion of correctly predicted positive instances out of all actual positive instances in the dataset.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score

It provides a balance between the two metrics and is particularly useful when dealing with imbalanced datasets. F1 score is the harmonic mean of precision and recall.

$$F1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2.3.2 Comparison

Table 5.5 presents a comparative analysis of various models and their corresponding experimental results. The purpose is to identify the most effective model that is well-suited for optimizing the performance of the face recognition system. All models were trained for a duration of 50 epochs.

Table 5.5: Comparsion Table between models

Model	Accuracy	Precision	Recall	F1 score
Xception	81.43 %	77.43%	84.77%	81.23%
MobilenetV3	79.12%	68.41%	91.70%	78.57%

5.2.3.3 Graph and Confusion Matrix

Graph of Autoencoder During Training

graphical representation of the accuracy and loss curves during the model training. the progressive reduction in the loss values with each epoch. This implies the the reduction in the dissimilarity between the restores image and the original images.

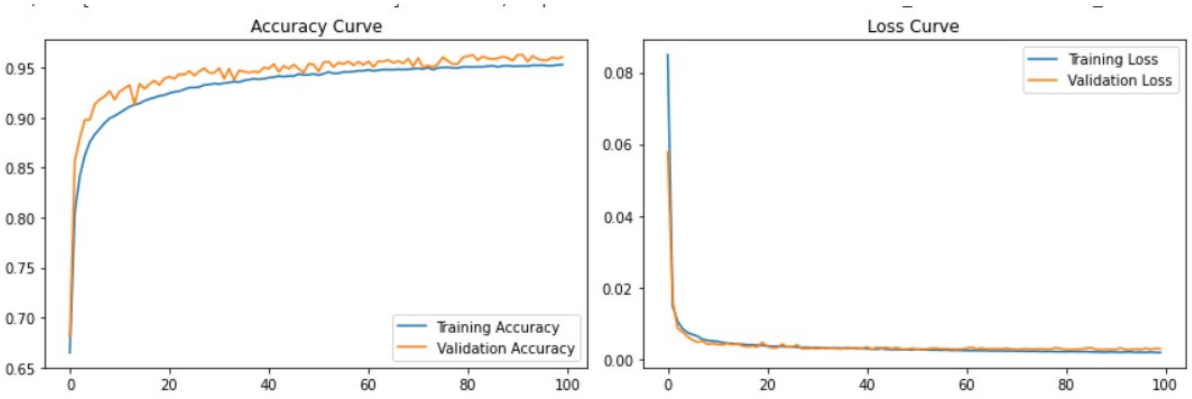

Figure 5.4: Graph accuracy and loss of autoencoder during Training

Confusion Matrix

A confusion matrix is a tabular representation used in classification tasks to evaluate the performance of a machine learning model. It helps to understand how well the model is classifying instances into different categories.

In Figure 5.5 and Figure 5.6 we can see the confusion matrix of siamese network using Xception and Mobilenetv3. where we can concluded True Positive = True Similar, True Negetive = True Different, False Positive = False Different and False Negetive = False Similar. by see this matrix we can easily calculate all the 4 matrics parameter accuracy, Precision, Recall and F1 score.

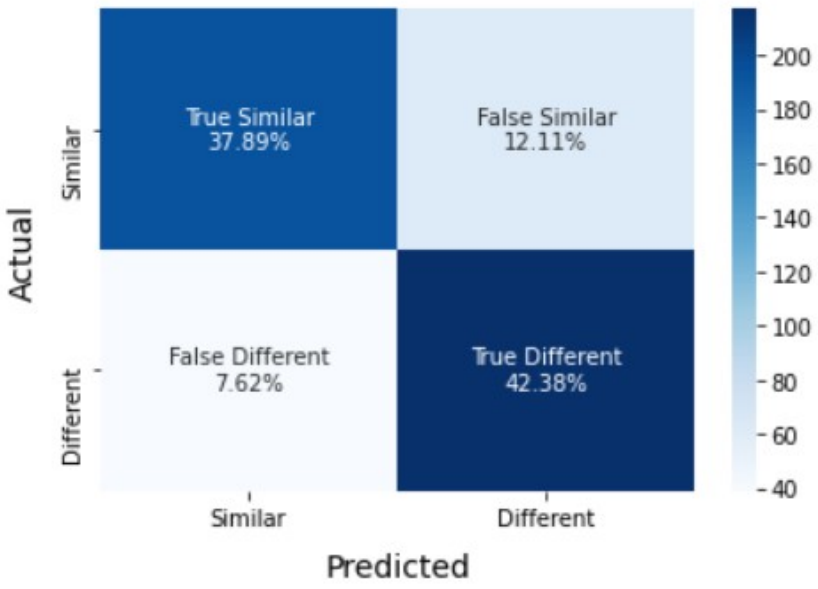

Figure 5.5: confusion matrix of Xception

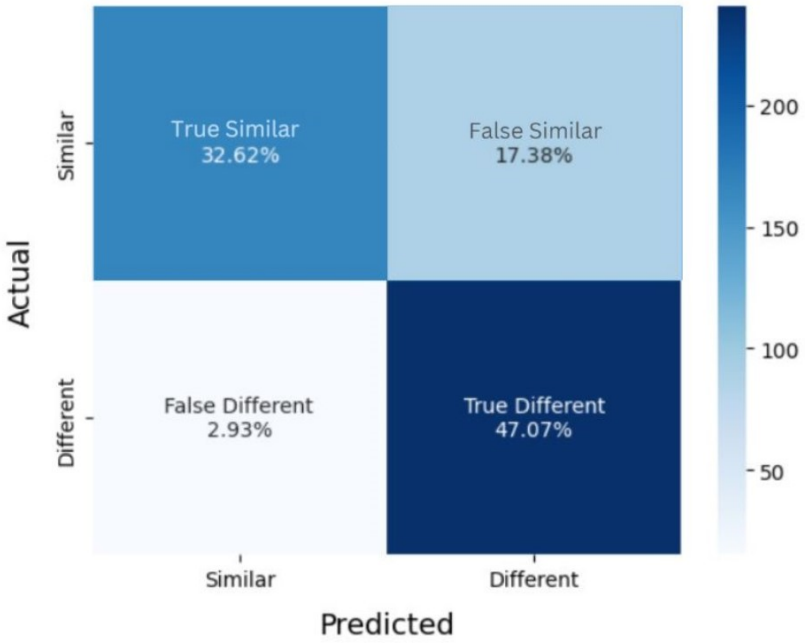

Figure 5.6: confusion matrix of Mobilenetv3

6

Conclusions and Future Scope

6.1 Conclusions

In conclusion, our project introduces a innovative approach to occluded face recognition by leveraging the power of a U-Net architecture-based autoencoder for occlusion removal and a Siamese network for accurate face recognition. Our proposed model outperforms existing methods like SSDA and DC-SSDA [22] in terms of both Peak Signal-to-Noise Ratio (PSNR) of 29% and Structural Similarity Index (SSIM)of 0.93. This demonstrates the effectiveness of our approach in addressing the challenge of recognizing occluded faces where our targeted occlusions were sunglasses and mask.

In the Siamese network for feature embedding, the integration of Xception and MobileNetV3 models yielded promising results, achieving accuracies of 83%and 79%, respectively. This outcome underscores the effectiveness of our de-occlusion process, as well as the model's ability to perform accurate face recognition on the de-occluded images.

6.2 Future Scope

Our proposed model has demonstrated promising results compared to other methods, but it currently operates in two separate stages: de-occlusion using the autoencoder and subsequent recognition with the Siamese network.

To enhance this process, we propose an innovative approach: integrating the autoencoder directly into the Siamese network. By doing so, we can simultaneously de-occlude the image while obtaining feature embeddings. This streamlined process ensures that the autoencoder's capabilities are utilized within the Siamese network, enabling end-to-end occluded face recognition. This innovative approach stands as a significant advancement in the field.

we've focused on two types of occlusions. However, we plan to expand our scope by incorporating a broader range of occlusion types in the future. It's important to note that our current training dataset is limited, but we have aspirations to enhance our model's performance by training it on a more extensive and diverse dataset in the future.

This expansion in occlusion types and dataset size will contribute to the robustness and effectiveness of our model.

Bibliography

[1] Vidyasheela, "Leaky relu activation function [with python code]," https://vidyasheela.com/post/leaky-relu-activation-function-with-python-code, Year, (accessed Aug. 21, 2023).

[2] L. Martín-Gómez, J. Pérez-Marcos, R. Cordero-Gutiérrez, and D. Hernández de la Iglesia, "Promoting social media dissemination of digital images through cbr-based tag recommendation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 45–53, 09 2022.

[3] A. Mahamudul Hashan, "Lung opacity identification using mathematical model based on deep learning," *International Journal of Engineering Applied Sciences and Technology*, vol. 5, pp. 25–29, 09 2020.

[4] S.-H. Tsang. (2022) Review — swish: Searching for activation functions. (accessed Aug. 21, 2023). [Online]. Available: https://sh-tsang.medium.com/review-swish-searching-for-activation-functions-993a9ef2b4b9

[5] (2023) Triplet loss. (accessed Jun. 09, 2023). [Online]. Available: https://en.wikipedia.org/wiki/Triplet_loss

[6] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991.

[7] M. Anggo and L. Arapu, "Face recognition using fisherface method," *Journal of Physics: Conference Series*, vol. 1028, p. 012119, 06 2018.

[8] A. Hadid, "The local binary pattern approach and its applications to face analysis," in *2008 First Workshops on Image Processing Theory, Tools and Applications*, 2008, pp. 1–9.

[9] M. Nakada, H. Wang, and D. Terzopoulos, "Acfr: Active face recognition using convolutional neural networks," 07 2017, pp. 35–40.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *Available: https://arxiv.org/pdf/1503.03832.pdf*, 2015.

[11] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 424–427.

[12] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[13] S. Ge, C. Li, S. Zhao, and D. Zeng, "Occluded face recognition in the wild by identity-diversity inpainting," vol. 30, no. 10, pp. 3387–3397, October 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=8963643

[14] Y. Li and J. Feng, "Reconstruction based face occlusion elimination for recognition," *Neurocomputing*, vol. 101, p. 68–72, 02 2013.

[15] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *Journal Name*, vol. Volume, p. Page, Year.

[16] J. H. Steven and D. E. Herwindiati, "Siamese network's performance for face recognition," *IEEE Xplore*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9557529

[17] Y. Su, Y. Yang, Z. Guo, and W. Yang, "Face recognition with occlusion," *IEEE Xplore*, 2015. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7486587

[18] H. Qiu, D. Gong, and D. Tao, "End2end occluded face recognition by masking corrupted features," Year. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=9495272

[19] Z.-M. Wang and J.-H. Tao, "Reconstruction of partially occluded face by fast recursive pca," 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4425497

[20] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," Year. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=9009826

[21] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust lstm-autoencoders for face de-occlusion in the wild," vol. 27, no. 2, pp. 778–790, February 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=8101544

[22] L. Cheng, J. Wang, Y. Gong, and Q. Hou, "Robust deep auto-encoder for occluded face recognition," October 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2733373.2806291

[23] P. S. Dinesh and M. Manikandan, "Fully convolutional deep stacked denoising sparse auto encoder network for partial face reconstruction," vol. 130, p. 108783, October 2022. [Online]. Available: FullyconvolutionalDeepStackedDenoisingSparseAutoencodernetworkforpartialfacereconstruction

[24] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Self-restrained triplet loss for accurate masked face recognition," p. 108473, December 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032100649X#

[25] Z. Zhang, X. Ji, X. Cui, and J. Ma, "A survey on occluded face recognition," December 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=&arnumber=9864502

[26] S. Shahsavarani, M. Analoui, and R. S. Ghiass, "Deep-id: A novel model for multi-view face identification using convolutional deep neural networks," *Accessed: Aug. 21, 2023. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2001/2001.07871.pdf*, Year.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Available:https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=9330594*, 2015.

[28] C. R. Kumar, S. N, M. Priyadharshini, D. G. E, and K. R. M, "Face recognition using cnn and siamese network," vol. 27, p. 100800, June 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917423001368#

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8099678*, 2016.

[30] A. Howard *et al.*, "Searching for mobilenetv3," *Available: https://ieeexplore.ieee.org/stamp/stamp.jsptp=arnumber=9008835*, 2019.

[31] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *Accessed: Aug. 21, 2023. [Online]. Available: https://openreview.net/pdf?id=Z7ALA0Ssza*, 2017.

[32] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.