

Real-time Face Recognition System Leveraging Facenet and VGGFace Fusion embedding, and Siamese Network with ArcFace and Triplet Loss Function

by

Md. Ashfaqul Haque
21166016

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2026

© 2026. BRAC University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

A handwritten signature in black ink, appearing to read 'Ashfaque', is written over a light yellow rectangular background.

MD. Ashfaque Haque
21166016

Approval

The thesis titled “Realtime Face Reognition System Leveraging Facenet and VG-GFace Fusion embedding, and Siamese Network with Triplet Loss Function” submitted by

Md. Ashfaul haque (21166016)

Of Fall, 2025 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of M.Sc. in Computer Science and Engineering on January 29, 2026.

Examining Committee:

Supervisor:
(Member)



Dr. Md. Golam Rabiul Alam
Professor

Department of Computer Science and Engineering
Brac University
rabiul.alam@bracu.ac.bd

Examiner:
(External)



Dr. Ashis Talukder, Ph.D
Associate Professor
University of Dhaka
ashis@du.ac.bd

Examiner:
(Internal)



Rafeed Rahman
Senior Lecturer
Department of Computer Science and Engineering
Brac University
rahman.rafeed@bracu.ac.bd

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
Brac University
sadek.ferdous@bracu.ac.bd

Chairperson:
(Chair)

Sadia Hamid Kazi, Ph.D
Associate Professor
Department of Computer Science and Engineering
Brac University
skazi@bracu.ac.bd

Abstract

The face recognition system can be defined as the identification of the unique biometric facial features of an individual through image or video frames. For CCTV-based surveillance the facial recognition system automates the identification of any person applied on live surveillance stream. But in CCTV surveillance system the facial recognition endures unique challenges like low resolutions, extreme poses, illumination variations, etc. There, a long span of development has been seen in this system from the geometric landmark distancing to deep-learning approaches, but a significant gap between the real-life surveillance and controlled benchmark conditions is observed. This thesis presents a comprehensive study on the development of a face recognition system designed for use with CCTV footage, with a focus on addressing challenges such as side-profile recognition and low-resolution images. Traditional face recognition models often perform poorly in real-world surveillance scenarios due to the quality of footage, low-resolution face crops, illumination variations, the varied angles of faces captured, different head orientations, occlusion causing partial face visibility, and motion effects. To overcome these limitations, this research explores the integration of a Siamese network with FaceNet and VGGFace, creating a hybrid model that leverages the unique strengths of each component. The Siamese network is employed to compare facial data, while VGGFace and FaceNet generate vigorous embeddings for accurate feature extraction, and a hybrid of Triplet Loss and ArcFace respectively enhances discriminative capabilities and combining through directing better embedding generations for similar faces, ensuring precise identification across challenging inputs. A custom dataset was constructed to simulate real-world CCTV environments, incorporating diverse facial angles and resolutions to rigorously test the model's performance. The primary aim of the research is to evaluate the system's ability to accurately recognize faces under non-ideal conditions, optimize computational efficiency, and assess the scalability of the model for large-scale surveillance systems. Through extensive experimentation and iterative refinement, the proposed hybrid model demonstrates improved recognition accuracy in low-quality, a variety of angled, and different expression face images, which contributes to advancements in the field of computer vision and offers practical implications for security and surveillance applications. As a dataset of 5409 face crop images from 86 individual person have been comprised here and later balanced to a total of 4300 samples and the proposed hybrid model has achieved an accuracy of 91.01% with 91.16% f1-score and ROC-AUC 0.9665, leading performance over the deepface models in comparison. Also the model have shown a competitive performance on the benchmark datasets of LFW, CelebA and VGGFace2, achieving an accuracy of 99.28%, 98.57% and 96.82% with f1-score of 98.71%, 98.46% and 96.27% respectively. Though the gap of benchmark datasets and CCTV footage data, it demonstrates a practical applicability for using in face recognition operations.

Keywords: Face Recognition, Siamese Network, Meta learning, CCTV Footage Face Data, Fusion Learning.

Acknowledgement

Firstly, all praise to the Great Allah for whom the thesis have been completed without any major interruption.

Secondly, to my supervisor Dr. Md. Golam Rabiul Alam sir for keeping faith in me and his kind support as well as advice in the work.

And thirdly, to my parents, wife, my friends and all the respected teachers, and staff of Brac University without their throughout support, it would not be possible. With their kind support and prayer, I am now on the verge of my post-graduation.

Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
List of Tables	x
Nomenclature	xiii
1 Introduction	1
1.1 Origin of the Concept	1
1.2 CCTV surveillance and it's challenges	2
1.3 Problem Statement	3
1.4 Objectives	3
1.5 Contribution	5
1.6 Organization of the Thesis Report	6
2 Literature Review	8
2.1 Background Studies	8
2.1.1 Challenges in FRS	8
2.1.2 Evolution of Face Recognition	11
2.2 Related Research Works	13
2.2.1 Hybrid Deep Learning Architectures and Feature Fusion	13
2.2.2 Advanced Loss Functions	16
2.3 Summary	16
3 Data Collection and Dataset Preprocessing	19
3.1 Dataset Preparation Process	19
3.1.1 CCTV Data Collection	19
3.2 Removal of Dataset Imbalance	20
3.3 Finalized Dataset Structure	24

4	Methodology	27
4.1	Overview of Face Recognition for CCTV Surveillance	27
4.2	Proposed Algorithm	29
4.3	Learnable Weighted Fusion Mechanism	32
4.4	Recognition Model Workflow	34
5	Results Analysis	36
5.1	Training of the Hybrid model	36
5.2	Training Analytics	37
5.3	Classification Performance Metrics	41
5.4	Classification Performance Evaluations	42
5.5	Distance Distribution and Embedding Space Analysis	47
5.6	Baseline Comparison and Performance Improvements	50
5.7	Comprehensive Multi-Model Comparison	52
5.8	Ablation Study and Architecture Validation	55
5.9	Cross-Dataset Generalization Evaluation	60
5.10	Summary of Results	61
6	Conclusion	63
	Bibliography	70

List of Figures

3.1	Dataset sample from CCTV surveillance	21
3.2	The CCTV Dataset Distribution and Histogram (Raw Collected Samples and Balanced Samples)	22
3.3	Augmented dataset visualization	24
4.1	System Diagram	28
4.2	Detailed architecture of the algorithm	32
4.3	Recognition Methodology	34
5.1	Learning rate schedule distribution in epochs	38
5.2	Combined Loss	38
5.3	Triplet Loss	39
5.4	Classification Loss	39
5.5	Validation Accuracy Evolution Across 100 Training Epochs	40
5.6	ROC Curve Demonstrating Excellent Discrimination Capability	43
5.7	Precision-Recall Curve	44
5.8	Comprehensive Threshold Sensitivity Analysis	45
5.9	Custom siamese hybrid model performance	45
5.10	Confusion Matrix at Optimal Threshold ($t = 0.4179$)	46
5.11	Distance Distribution Comparison: Custom Model vs Baseline	47
5.12	Distance Distribution Grid: Comprehensive Model Comparison	48
5.13	Class Separation Ranking Across All Models (Higher is Better)	49
5.14	Mean Distance Comparison Across All Models	50
5.15	Comprehensive Performance Comparison: Proposed Model vs Baseline	51
5.16	ROC Curve Comparison: Custom Model vs DeepFace Models	52
5.17	Model Ranking by F1-Score Performance	52
5.18	Model Ranking by ROC-AUC	53
5.19	Model Ranking by Classification Accuracy	54
5.20	Model Ranking by Equal Error Rate (Lower is Better)	54
5.21	Embedding Dimension vs F1-Score Performance Trade-off	55
5.22	Ablation Study: Performance by Fusion Strategy	56
5.23	Ablation Study: Performance by Loss Function	57
5.24	Ablation Study Heatmap: Fusion Strategy vs Loss Function	58
5.25	Complete Ablation Study F1-Score Rankings (All 20 Configurations)	58
5.26	Ablation Study: Top 5 Configurations ROC Curves	59
5.27	Ablation Study: Top 5 Configurations Metrics Comparison	60

5.28 Custom Siamese Hybrid Model Classification Examples. Row (a) and (c) indicates samples matching of similar class classified as a match. Row (b) and (d) indicates sample of different class classified as no match	62
---	----

List of Tables

2.1	Comparison of Related Research Works in Face Recognition	18
3.1	Selected augmentation requirements for under-represented identities. .	23
3.2	Comparison of Intra- and Inter-Class Similarities After Augmentation	24
5.1	Confusion Matrix at Optimal Threshold	46
5.2	Distance Separation Analysis Across All Models	49
5.3	Baseline vs. Proposed Model Performance Comparison	51
5.4	Complete Model Comparison Across All Metrics	53
5.5	Fusion Strategy Performance Comparison	56
5.6	Loss Function Performance Comparison	57
5.7	Ablation Study Top 10 Configurations	59
5.8	Cross-Dataset Performance Comparison	61

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

α	Triplet loss margin
Δ	Separation between genuine and impostor distributions
$\Delta_{normalized}$	Normalized separation in pooled standard deviations
η	Learning rate
η_t	Learning rate at epoch t
η_{max}	Maximum learning rate
η_{min}	Minimum learning rate
λ	Weighting parameter for loss combination
λ_{arc}	Weighting coefficient for ArcFace loss
λ_{tri}	Weighting coefficient for Triplet loss
$\mathbf{E}$	L2-normalized embedding vector
$\mathbf{F}_{face}$	FaceNet feature vector
$\mathbf{F}_{vgg}$	VGGFace feature vector
$\mathcal{L}_{ArcFace}$	ArcFace loss component
$\mathcal{L}_{classification}$	Classification loss component
$\mathcal{L}_{total}$	Total hybrid loss
$\mathcal{L}_{Triplet}$	Triplet loss component
$\mathcal{L}_{triplet}$	Triplet margin loss
μ_g	Mean distance for genuine pairs
μ_i	Mean distance for impostor pairs
σ_g	Standard deviation of genuine pair distances
σ_i	Standard deviation of impostor pair distances

σ_{pooled}	Pooled standard deviation
τ	Verification/identification threshold
θ	Model parameters
AP	Average Precision
$ArcFace$	Angular Margin Loss
AUC	Area Under the ROC Curve
$CAWR$	Cosine Annealing Warm Restarts
$CCTV$	Close Circuit Television
$COVID$	Coronavirus Disease
D_{an}	Distance between anchor and negative
D_{ap}	Distance between anchor and positive
EER	Equal Error Rate
F_1	F1-score (harmonic mean of precision and recall)
FN	False Negative count
FNR	False Negative Rate
FP	False Positive count
FPR	False Positive Rate
LFW	Labelled Faces in the Wild
m	ArcFace angular margin
P_n	Precision at the n -th threshold
R_n	Recall at the n -th threshold
ROC	Receiver Operating Characteristic
s	ArcFace feature scale
t	Epoch number or decision threshold
T_0	Initial restart period
T_i	Current cycle period
T_{cur}	Current epoch count
TN	True Negative count
TNR	True Negative Rate

TP True Positive count
 TPR True Positive Rate
 $VGGFace$ Visual Geometry Group-Face
 w_{face} Learned weight for FaceNet backbone
 w_{vgg} Learned weight for VGGFace backbone

Chapter 1

Introduction

The face recognition system is one of the most widely adaptive vision-based applications which being incorporated in many use cases, transforming its applicable sectors. Its most emerging sectors are surveillance system, access control management system (Intrusion Identification) , authentication, and criminal identification, like blacklisting or whitelisting. Face recognition system has advancement has shown quite a remarkable improvement over the last couple of years. This has shown proven results in real-world applications, though recognizing faces from close circuit television (CCTV) surveillance footage remains toilsome, and complications arise very often. When the trained model from controlled scenarios is applied to a robust scenario, it faces critical challenges with a sudden drop in performance, resulting in not being adaptive to a variety of situations like sudden changes of lighting conditions, low-resolution face crops, a variety of face orientations and expressions, and occlusions with motion blur. Keeping all these in consideration, this proposed face recognition system has been brought to formulation to mitigate the issues.

1.1 Origin of the Concept

Face recognition systems should contend with substantial variations of face occurrence effectively, which is caused by changes in illumination, pose, expression, occlusion, and image quality degradation. These types of challenges are pretty common in CCTV-based surveillance systems. Here, cameras with low resolution, lighting conditions not being in control with optimum conditions, subjects with no proper pattern or no possible formation aligning with the operation, hard to observe viewing angles, and motion blur bring challenges in recognition system, thus affecting severely at acheiving of system performance. In traditional face recognition approach where relying on a single-model architecture often causes struggling situations to grasp the facial variations present in surveillance footage at its maximum, this leads to situations like reduction in accuracy, increasing in false rejection rates, and declining in reliability in operational deployments.

Over the last few years, face recognition has evolved and shown a noteworthy transformation. It has long passed just the lab experiment and entered the real-world surveillance system through access control management, biometric verification, law enforcement, and social platforms [54], [56]. Face recognition systems with deep leaning architectures are achieving close to perfect accuracy under controlled conditions, and with prominent algorithms on some benchmark datasets like labelled

faces in the wild (LFW) have gained accuracy over 99% [8], [23]. But getting out of the controlled environment in real world surveillance scenario through CCTV has shown a critical drop in performance, questioning the practical utility and use case of this solution [36], [54], [56].

1.2 CCTV surveillance and it's challenges

CCTV system presents a pretty challenging circumstances for the face recognition technology where the controlled biometrics capture and perform at the optimum scenarios, and the subjects are also compliant with the proper lighting condition and frontal face pose. On the contrary, CCTV sensors have to perform under adversarial conditions [36], [56]. Surveillance CCTV cameras have to capture images of subjects from varying distances, which are non-cooperative and thus cause low-resolution face images, motion blurriness [22], [56]. When cameras are installed high above the ground to cover a wide area, they often capture faces under very steep pitch angles. At the same time, people move freely through these spaces, so their heads naturally rotate, producing a wide range of yaw angles from full profile to three-quarter views [23], [36]. In addition, real-world environments introduce further challenges: lighting conditions change throughout the day (e.g., overhead fluorescent lights, daylight, or low illumination at night), faces are frequently partially covered by sunglasses, masks, hands, or hair, and adverse weather or atmospheric effects such as fog, rain, or heat shimmer can further degrade the visible facial information [22], [54], [56]. The recent research work by Jassim et al. (2024), which is on criminal identification from CCTV footage, illustrates that these difficulties cause degradation of quality in a systematic way [56]. Their study shows that surveillance video typically suffers from several quality issues at the same time: faces are often captured at only 80–120 pixels, and which is well below the recommended 196–260 pixels per meter for a reliable identification. H.264/H.265 compression introduces blocking artifacts, subject motion and camera panning create motion blur, and lighting conditions can vary dramatically between daytime and nighttime while recording [54], [56]. These factors rarely appear in isolation. Instead, they interact and reinforce one another, leading to a much more severe drop in recognition performance than would be expected if each factor were considered on its own.

During the period of the COVID-19 pandemic, the masked face recognition technology has emerged as an additional source of difficulty [53], [54]. Chen et al. (2025), in a study on face identification under wearing different mask configurations, which reported a striking result as the full mask coverage sometimes led to faster and more accurate human identifications than partial coverage [53]. They attributed this counterintuitive outcome to perceptual learning during the pandemic, in which observers became highly attuned to information in the eye region. For automated systems, this suggests that models trained mainly on unmasked faces may need to be adapted or redesigned to place greater emphasis on periocular features when masks remain common in real-world settings [53], [54].

1.3 Problem Statement

The recent days face recognition methods face several important limitations while moved from the controlled benchmarks to real CCTV surveillance environments.

First the siingle-model architectures, which dominate most current deployments, tend to be underperforming in practice [7], [8]. For example, the VGGFace, which is trained on around 2.6 million celebrity images and learns rich discriminative features, but when it comes to performance, its output drops noticeably when faces appear under extreme pose changes or at low resolutions typical of CCTV streams [7], [36]. Whereas the FaceNet, on the other hand, produces very compact 128-dimensional embeddings, which is optimized with a triplet loss and works very well for verification-style metric learning [8]. However, this strength can come at the cost of classification-oriented discriminative power when the task requires fine-grained separation between many identities. When it is used in isolation large pose shifts, occlusions, low resolution, facial expression differences, and age-related appearance changes, any of these models are fully able to address the range of variations encountered in surveillance lighting changes.

When mentioning limitations in the underlying loss functions, this causes further constrain on performance in open-world surveillance settings. The standard softmax cross-entropy loss is widely used for classification-based face recognition, and it typically offers sufficient class separation in closed-set scenarios. Though it does not enforce the strong inter-class margins and tight intra-class clustering [20], [23]. The margin-based losses, for example, CosFace and ArcFace, which introduce angular margins. This enhances discriminative power remarkably, yet purely classification-focused objectives still do not capture the nuanced similarity structure as the metric learning is designed to model [20], [23]. In contrast, triplet loss directly optimizes a distance-based objective and provides this fine-grained similarity modeling, but it is notoriously difficult to train: performance depends heavily on hard negative mining strategies and careful triplet selection, and unstable training dynamics are common [8], [14].

Lastly, the data imbalance in realistic datasets. It introduces another layer of difficulty. Surveillance collections almost always follow a long-tail distribution, where a few identities have many examples, and in the midst, most individuals appear only a handful of times [31], [58]. Zhang and Deng (2021) have shown that training under such an imbalance tends to allocate a portion of the feature space to majority classes but not in proper proportion, thus pushing minority-class features into compressed regions, and this harms the recognition performance of rare identities [31]. This effect in surveillance applications, where watchlist subjects or persons of interest usually have very few enrollment images compared with the large number of background identities, causes problems that shape the embedding space during training.

1.4 Objectives

The principal aim of this research work has been centered on developing a robust face recognition framework that addresses the inherent limitations encountered in CCTV surveillance environments. Through the integration of dual-model architecture combined with a hybrid loss mechanism, several critical gaps in existing methodologies

have been targeted for resolution.

Incorporating a fusion model with hybrid loss formulation to strengthen the generalization capability across critical real-world scenarios. Traditional single-model approaches have not been sufficiently adaptive when confronted with the unpredictable conditions typical of surveillance footage. By leveraging complementary feature representations from multiple pre-trained networks, this aims at maintaining consistent performance even when extreme variations in illumination, pose angles, occlusion patterns, and image degradation are present. This objective seeks to bridge the gap between controlled laboratory settings and the challenging operational environments where CCTV systems are deployed.

Demonstrating effectiveness across multiple established benchmark datasets constitutes another focal point of this investigation. Many contemporary face recognition systems have been optimized for specific datasets, resulting in limited transferability when evaluated on alternative benchmarks. This objective involves rigorous testing protocols that assess the proposed model’s capability to generalize beyond the training distribution. By evaluating performance metrics on diverse benchmark collections, the extent to which the fusion architecture reduces dataset-specific biases can be quantified, thereby establishing its versatility in handling varied facial recognition challenges.

Compiling custom dataset exclusively from CCTV surveillance footage, capturing the authentic difficulties inherent in such environments—including low-resolution imagery, motion-induced blur, non-frontal viewing angles, and suboptimal lighting conditions. Conventional datasets frequently lack these characteristics because they have been curated under more controlled circumstances. This objective emphasizes the creation and utilization of data that faithfully represents the operational realities of surveillance systems. Through strategic data balancing and augmentation techniques that mirror real-world CCTV variations, the model’s training process has been grounded in scenarios that closely approximate actual deployment conditions. Comprehensive evaluation and comparison with existing state-of-the-art models and datasets are undertaken to establish the proposed system’s position within the current landscape of surveillance-oriented face recognition. This objective involves systematic performance analysis against established baselines, enabling a clear understanding of the improvements achieved through the dual-backbone fusion and hybrid loss strategy. By documenting both quantitative metrics and qualitative observations, this research contributes empirical evidence regarding the advantages of combining multiple architectural paradigms and loss functions for CCTV-based facial recognition tasks.

The objectives of this research can be comprised into four points:

- **Firstly**, Establishing robust performance through a Fusion model with hybrid loss to achieve superior performance over state of the art models in real world applications scenario minimizing the gap between controlled benchmark and CCTV based operational conditions
- **Secondly**, Generalizing the benchmark dataset performance through effective and stable performance on multiple benchmark datasets.
- **Thirdly**, Comprising custom dataset from CCTV footage data to adapt the operation realities of surveillance system incorporating augmentation techniques that mirror real-world scenarios.

- **Finally**, Perform extensive evaluation with multiple benchmark datasets and state-of-the-art models to establish its performance in both quantitative metrics and qualitative observations manner.

1.5 Contribution

A Siamese network-based solution has been proposed to address the limitations identified in current face recognition based surveillance systems. This approach is built using a dual-backbone architecture which is the combination of pre-trained VGGFace and FaceNet models within a single Siamese framework, and then it is trained using a hybrid loss that blends the Additive Angular Margin Loss (ArcFace) with the Triplet Margin Loss. Altogether, the aim of this design is to improve robustness under challenging CCTV conditions and thus support both identification and verification use cases.

First contribution, a dual-backbone feature fusion architecture has been proposed, which learns how to combine VGGFace and FaceNet representations in a data-driven way. Instead of selecting only one feature extractor, this combined approach uses learnable fusion weights in such a way so that it can adapt to the characteristics of the input during training and reflect that in inference. The VGGFace is based on the VGG-16 architecture and it outputs a 4096-dimensional feature vector, which is trained in a classification setting on millions of face images, thus capturing a wide range of appearance variations including ethnicity, age, facial expression and diverse imaging conditions [7], [36]. In parallel with the VGGface, the FaceNet branch uses the Inception-ResNet-v1 backbone to generate a 512-dimensional embedding, which is optimized with triplet loss, in which Euclidean distances in the embedding space correspond directly to perceived identity similarity [8].

The second contribution is the hybrid training objective. This combines the ArcFace and the triplet loss into a single formulation. Rather than solely relying on either a classification-style or a metric-learning objective, this model jointly works to optimize both. This combined loss achieves several practical advantages compared to training with a single one. The first one is that the ArcFace, through angular margins, establishes a globally well-structured embedding space, whereas the triplet refines local similarity relationships between nearby samples [14], [23]. Then the ArcFace optimizes class logits directly, which is beneficial for identification tasks, whereas triplet loss shapes the metric properties required for verification explicitly. Both are crucial for surveillance systems [20], [23]. Another advantage is that 60% of the total weight has been assigned to ArcFace, which provides stable gradients along with a smoother loss trajectory. On the other hand, the 40% triplet component fine-tunes the embeddings without the severe training instability. This is often observed when only the triplet loss is used [14]. Last but not least, the angular constraints of ArcFace are generally robust to illumination changes because they are for normalized features. Then there comes the metric learning of triplet loss. This is well suited to handle the partial occlusions through emphasis of discriminative information on visible parts [8], [23].

Thirdly, the dataset used here is a custom one and has been collected from the CCTV cam feed. For the mitigation of class imbalance of surveillance datasets, the thesis constructs a balanced dataset with exactly 50 samples per identity across 86 classes. The raw dataset is a collection of 5,409 images with a highly skewed

distribution, where some classes have lesser samples and others have up to 167. The rebalancing process has two components.

For oversampled 53 classes, the number of samples is reduced to 50 by random downsampling, while preserving diversity in pose and illumination [31]. For 33 undersampled classes, additional images have been synthesized through controlled data augmentation until each identity reaches 50 samples, resulting in 1,893 new images [31], [58]. The augmentation operations have been chosen to reflect a realistic surveillance variations with rotation, horizontal flipping, brightness scaling, contrast changing, gaussian blur, and color saturations.

After the rebalancing, the final training set is consists of 4,300 images with an exactly uniform distribution of 50 samples per class and zero variance in class size. So, this eliminates the dominance of the majority classes during the optimization. [31]. Recent work by Li et al. (2025) on class imbalance in facial expression recognition states that such balanced sampling approaches can improve performance on minority classes while preserving accuracy on majority classes. This yielding overall F1-score gains in the range of 15–30% [58]. This drives the motivation for the use of a similarly controlled balancing procedure in the surveillance recognition system, which is considered in this thesis.

So the unique contributions in this thesis work are comprised to these three.

- **Firstly**, a novel architecture has been proposed here which combines the VGGFace (VGG-16) and FaceNet (Inception-ResNet-v1) through a fusion layer where the weights have been utilized to integrate the 4096-dimensional features from VGGFace with the 512-dimensional embeddings from FaceNet.
- **Secondly**, the hybrid loss function of Additive Angular Margin Loss (ArcFace) and Triplet Margin Loss. Thus this combined dual-optimization strategy leverages ArcFace to establish a global embedding space for identification, while Triplet Loss refines local similarity relationships for verification.
- **Thirdly**, a custom dataset collected from CCTV is constructed which contains 5409 images of 86 individuals, showing a heavily imbalanced pretty common in surveillance data and later balanced to 4300 images of these 86 unique individuals.

1.6 Organization of the Thesis Report

The thesis regarding the face recognition system has been organized as the work is being developed step by step and traversing from theoretical analysis to the implementation in detail. Here, each chapter is focused on a specific part of the whole study at the same time maintaining continuity across the overall presentation.

- **Chapter 2: Literature Review** This part shows the research aspects relevant to this. Here, both classical and recent developments in face recognition systems have been depicted. In this chapter, the onfoing challenges in real-world scenarios have been examined, which include illumination changes, pose variations, occlusions, resolution issues, dataset imbalance, and cross-dataset generalization. Also, the loss functions have also been kept in consideration, starting from softmax and progressing through center loss, CosFace and

SphereFace to ArcFace. With all these, including studies on Siamese networks, metric learning approaches, the comparison between Vision Transformers and CNNs, and data augmentation approaches.

- **Chapter 3: Data Collection and Dataset Preprocessing** This depicts how data have been collected and used in this research. Also, how the data pre-processing has been done, and a properly balanced dataset has been prepared for the analysis. The collection of facial imagery from CCTV surveillance footage has been described here. A strategy has been applied then for data balancement. This combines the downsampling of the majority classes and specific augmentation for the minority classes for upsampling the data. These specific augmentation techniques which mimics the real scenarios for CCTV captured images. Finally, the dataset has been split into training, validation and test splits (70%, 15%, 15%).
- **Chapter 4: Methodology** This chapter is on the proposed solution and approach. Here, the Siamese architecture with dual-model backbone fusion has been introduced. A brief description of the VGGFace and FaceNet feature extraction pipelines, including the learnable weighted fusion mechanism. And also the fully connected embedding network with regularization techniques has been presented in detail. The mathematical derivation of ArcFace–Triplet loss has been shown here.
- **Chapter 5: Results Analysis** In that chapter, the performance on the prepared dataset from CCTV footage has been demonstrated of the proposed system in brief. Ablation studies have also been provided here comparing the fusion model with single-backbone baselines and contrasting different loss formulations respectively (ArcFace-only, Triplet-only, and the proposed hybrid scheme). Also the proposed system performance has been evaluated on some benchmark datasets.
- **Chapter 6: Discussion and Conclusions** This chapter describes the observed strengths and limitations of the proposed system, which have been discussed, together with characteristic failure modes. Potential directions for future research scopes have been outlined, including Vision Transformers integration into the Siamese framework and the exploration of more advanced training strategies for extended outcomes. The chapter concludes by the summarization of the main contributions of the thesis to the field of CCTV surveillance-oriented face recognition.

Chapter 2

Literature Review

2.1 Background Studies

The domain of Face Recognition Systems (FRS) has undergone a profound transformation, evolving from early geometric and holistic approaches to sophisticated deep learning paradigms. Yet, the deployment of these systems within the unconstrained and adversarial environments characteristic of Close Circuit Television (CCTV) surveillance continues to present multifaceted challenges. This section provides a comprehensive theoretical background, examining the fundamental impediments to surveillance-based recognition, the mathematical evolution of loss functions designed to enhance feature discriminability, and the architectural principles of Siamese networks utilized for one-shot learning.

2.1.1 Challenges in FRS

The operational efficacy of face recognition technologies in real-world surveillance is heavily contingent upon the system's capacity to normalize and process high variance input data[40]. In contrast to cooperative biometric authentication scenarios where lighting, pose, and expression are strictly controlled CCTV footage introduces a spectrum of environmental and physiological variables that degrade performance[51]. It has been observed that traditional models, despite achieving near-perfect accuracy on benchmark datasets like LFW (Labeled Faces in the Wild), often suffer precipitous performance drops when applied to surveillance streams due to the compounding effects of these unconstrained variables[3]. Among the extrinsic factors affecting face recognition, illumination variation is frequently cited as one of the most pervasive and detrimental[34]. The appearance of a facial image is fundamentally a product of the interaction between the illumination source and the reflectance properties of the facial surface. In surveillance environments, lighting is rarely controlled; subjects may traverse zones of intense, direct sunlight, deep shadows cast by architectural structures, or uneven artificial lighting from streetlamps or fluorescent fixtures[34].

Theoretical models, such as the Lambertian reflectance model, postulate that image intensity is proportional to the cosine of the angle between the light source and the surface normal. However, human skin is non-Lambertian, exhibiting specular reflections (highlights), subsurface scattering, and complex shadowing effects where facial features (e.g., the nose or brow ridge) cast shadows upon other parts of the

face. Consequently, the intra-class variance created by shifting illumination conditions often exceeds the inter-class variance between different identities, leading to false rejections. Deep learning models trained on evenly illuminated datasets tend to learn filters sensitive to pixel intensity gradients caused by shadows rather than the underlying morphological structure of the face. Recent efforts to mitigate these effects have focused on image enhancement and normalization. For instance, the Enhanced Deep Curve Estimation (EDCE) method was proposed to adaptively brighten low-light regions in surveillance footage while preserving structural details, thereby improving the signal-to-noise ratio for downstream recognition tasks[34]. Similarly, Generative Adversarial Networks (GANs) have been employed to normalize illumination by synthesizing a frontal, well-lit version of the input face, although such generative processes risk introducing artifacts that may alter identity-specific features[34].

Pose variation refers to the angular deviation of the subject’s head relative to the camera’s optical axis[18]. In standard surveillance deployments, cameras are typically mounted at elevated angles (2.5 to 4 meters high) to maximize spatial coverage and minimize occlusion by ground-level obstacles. This setup inherently introduces significant pitch (vertical rotation) angles. Furthermore, as subjects move freely through the monitored space, they introduce yaw (horizontal rotation) and roll, resulting in a highly non-frontal distribution of facial views[18], [51]. Severe pose variations, particularly those where the yaw angle exceeds 45° , result in the self-occlusion of critical facial landmarks and the loss of distinct features on the distal side of the face. Conventional 2D recognition algorithms, which rely on the geometric configuration of landmarks (such as the inter-ocular distance or the triangle formed by the eyes and nose), struggle when these configurations are distorted by perspective projection or entirely obscured. To address this, 3D-aided face recognition methods have been investigated, where a 3D Morphable Model (3DMM) is fitted to the 2D image to estimate the 3D shape and texture[1], [2]. This allows for the synthetic rotation of the face to a canonical frontal view a process known as face frontalization[1]. However, the reconstruction of accurate 3D models from low-resolution CCTV imagery is computationally intensive and error-prone. Alternatively, distinct deep learning architectures have been evaluated for pose robustness. Recent comparative studies have suggested that Vision Transformers (ViTs), which utilize self-attention mechanisms to capture global dependencies, may offer superior robustness to extreme pose variations compared to traditional Convolutional Neural Networks (CNNs), which are biased towards local spatial features[43], [50]. Occlusion in surveillance video is a dual-natured challenge, arising from both environmental obstacles (e.g., pillars, other pedestrians) and accessories worn by the subject (e.g., sunglasses, scarves, hats)[33]. The global COVID-19 pandemic introduced a paradigm shift in this domain with the widespread adoption of sanitary masks, creating a specific class of occlusion that systematically conceals the lower face, including the nose, mouth, and chin features that are traditionally heavily weighted by recognition algorithms[33], [39]. Standard deep learning models typically aggregate features from the entire receptive field of the face. When a significant portion of this field is occluded, the network may process the occlusion itself (e.g., the texture of a mask) as a feature, or fail to extract sufficient discriminative information from the remaining visible regions. This has necessitated a shift towards “periocular recognition,” which focuses exclusively on the eye region and eyebrows, and “partial face recognition” strategies

that dynamically weigh non-occluded regions.

It has been observed that models trained on unmasked faces experience a significant performance degradation when tested on masked faces[33]. To counter this, techniques involving the synthesis of masked faces for training (data augmentation) and the development of occlusion-aware loss functions have been proposed[38]. For example, some approaches utilize an attention mechanism to down-weight the contribution of occluded regions in the feature embedding process, ensuring that the recognition decision is based solely on reliable visual data[35]. The quality of facial imagery captured by CCTV systems is frequently compromised by hardware constraints and environmental factors[46]. To conserve bandwidth and storage, surveillance video is often heavily compressed, introducing blocking artefacts and loss of high-frequency detail. Furthermore, the large field of view required for area monitoring often results in faces occupying a very small number of pixels, sometimes as few as 30×30 pixels or less[40]. At such low resolutions, the fine-grained textural details necessary for accurate identification such as skin pores, scars, and fine wrinkles—are irretrievably lost. This creates a significant “domain gap” when models trained on high resolution datasets (like VGGFace2 or CASIA-WebFace) are applied to low-resolution probe images. The standard practice of upsampling low-resolution images to meet the input size requirements of CNNs (e.g., 224×224) does not recover missing information but rather introduces blur. Research in this area has explored “super-resolution” techniques, utilizing deep neural networks to hallucinate high-frequency details based on learned priors[21], [63]. However, there is an inherent risk that the super-resolved face may contain artifacts that do not correspond to the subject’s true identity. Alternative approaches involve training “resolution-invariant” models that are explicitly taught to map high-resolution and low-resolution images of the same subject to proximal points in the embedding space[46]. Temporal variance in facial appearance, driven by the natural aging process, presents a significant challenge for longitudinal surveillance and watchlist maintenance[4]. The “template aging” phenomenon occurs when the reference image stored in the database is significantly older than the probe image captured by the surveillance system. Over time, facial appearance is altered by changes in craniofacial structure (in younger subjects), skin texture (wrinkles, loss of elasticity), and weight fluctuations[13]. Biologically, these changes are non-linear and person-specific. Deep learning models must therefore learn to disentangle “identity” features, which are invariant, from “age” features, which are transient. This requires training on massive datasets with wide age gaps per identity, which are difficult to curate. In the absence of such training, the feature distance between an individual’s images taken years apart may exceed the decision threshold, leading to increased False Rejection Rates (FRR)[4].

A critical, often overlooked challenge in surveillance FRS is the inherent class imbalance within training and operational datasets[32]. Real-world data follows a long-tail distribution: a small number of identities (e.g., celebrities in public datasets or regular employees in a private system) have a large number of samples, while the vast majority of identities appear only rarely[17]. When deep learning models are trained on such imbalanced data using standard objective functions, they tend to overfit to the majority classes (the “head” of the distribution) while underfitting the minority classes (the “tail”)[25]. This results in a feature space where majority classes have well-defined, compact clusters, while minority classes are compressed or scattered, leading to poor generalization for the very subjects that may be of most

interest in a security context. Strategies to address this include algorithmic rebalancing (e.g., oversampling minority classes), the generation of synthetic samples via augmentation, and the use of focal loss functions that assign higher gradients to hard-to-classify examples[32], [42].

2.1.2 Evolution of Face Recognition

Face recognition has gone through a brief span of time and has withstood different phases of development from different types of recognition models and image-processing techniques incorporating various loss functions. Though there have been different deep learning approaches like hybrid models or single backbone models with benchmark accuracy, the Siamese network architecture is distinct from standard classification networks in its structure and objective[6]. It consists of two identical subnetworks (twin networks) that share the same weights and parameters. These networks process two distinct inputs an image pair in parallel, projecting them into a shared latent feature space. A subsequent distance layer computes a similarity metric (typically Euclidean distance) between the two resulting feature vectors[26].

If the distance falls below a learned threshold, the inputs are classified as belonging to the same identity; otherwise, they are classified as different. This “similarity learning” paradigm allows the network to learn a general function for differentiating faces, rather than learning to classify specific identities[26]. The concept of Siamese networks was originally introduced in the 1990s for signature verification tasks but has gained renewed prominence in the deep learning era for face verification[6]. Seminal work in “Siamese Neural Networks for One-Shot Image Recognition” demonstrated the architecture’s powerful capability to generalize to new classes that were not present during the training phase[6]. This property is particularly valuable for surveillance applications where the set of target identities (the watchlist) is dynamic and may change frequently, rendering the retraining of a standard classification layer impractical[26]. One-shot learning addresses the challenge of learning information about object categories from a single, or very few, training samples[6]. In the context of FRS for surveillance, this capability allows the system to recognize a new subject of interest based on a single reference photograph. The Siamese network facilitates this by learning a generic metric space; essentially, the model learns “what makes two faces different” rather than “what defines Person X.” Consequently, when a new face is introduced, the network can verify it against the reference without any weight updates, relying solely on the learned metric comparisons. This approach is instrumental in scenarios characterized by data scarcity, such as identifying a suspect from a single available mugshot[6], [26]. The research presented in this thesis integrates the Siamese architectural principles with a hybrid feature extraction backbone comprising VGGFace and FaceNet[26]. By channelling the outputs of these robust, pre-trained feature extractors into a Siamese distance metric framework, the system is designed to leverage transfer learning from massive datasets while adapting the similarity metric specifically for the surveillance domain. This integration aims to bolster performance in low-data regimes and enhance robustness against the varied degradations found in CCTV footage[26].

The objective function, or loss function, serves as the guiding signal for the neural network during training, dictating how facial features are mapped into the high-

dimensional embedding space. The evolution of these functions has shifted from purely classification-based objectives to those that explicitly enforce metric constraints and angular margins. The Softmax Cross-Entropy loss served as the foundational objective for early deep face recognition models[44]. It is formulated to maximize the posterior probability of the ground-truth class. For a feature vector x_i and its corresponding class label y_i , the loss is defined as:

$$L_{softmax} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T x_i + b_j}} \quad (2.1)$$

where W represents the weights of the last fully connected layer and b is the bias. While Softmax is effective for closed-set classification (where distinct decision boundaries are learned for known classes), it does not explicitly enforce intra-class compactness or inter-class separability[30]. It has been argued that features learned via Softmax often exhibit significant radial variation, which is suboptimal for verification tasks relying on cosine similarity or Euclidean distance[44].

To address the limitations of Softmax regarding intra-class variation, Center Loss was proposed[11], [37]. This function maintains a learnable “center” vector for each class in the feature space and penalizes the Euclidean distances between the deep features and their corresponding class centers[11]. By minimizing this distance, the network is encouraged to pull features of the same class toward a common point, thereby reducing intra-class variance. However, Center Loss is typically employed as an auxiliary term alongside Softmax, as it does not inherently guarantee the separation of different class centers[29], [37].

A significant breakthrough in discriminative feature learning was the realization that normalizing features and weights to a hypersphere allows the recognition problem to be treated purely in terms of angular distance[15], [20], [23]. This led to the development of angular margin losses. Introduced a multiplicative angular margin to constrain the angle between the feature vector and the class weight, enforcing a stricter decision boundary[15]. Simplified the optimization landscape by applying an additive cosine margin[20]. Represented a further refinement by adding an angular margin penalty m directly to the angle θ between the feature and the weight vector[23].

The ArcFace objective is formulated as[23]:

$$L_{ArcFace} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j \neq y_i} e^{s \cos \theta_j}} \quad (2.2)$$

By enforcing a constant angular margin m , ArcFace maximizes the geodesic distance between classes on the hypersphere. It has been widely demonstrated that this approach yields highly discriminative features that are robust to intra-class variations, setting new benchmarks for large-scale face recognition[23], [27].

Parallel to classification-based losses, metric learning approaches like Triplet Loss focus directly on optimizing the relative distances between embeddings[8]. The network is presented with a triplet of images: an Anchor (A), a Positive (P - same identity as anchor), and a Negative (N - different identity). The objective is to ensure that the distance between the anchor and the positive is smaller than the distance between the anchor and the negative by a predefined margin α [8]:

$$||f(A) - f(P)||^2 + \alpha < ||f(A) - f(N)||^2 \quad (2.3)$$

This loss function explicitly structures the embedding space, making it highly effective for verification tasks and open-set recognition. However, the performance of Triplet Loss is heavily dependent on the quality of the triplets selected for training[33].

Since the number of possible triplets grows cubically with the dataset size, training on all combinations is computationally infeasible and inefficient[8], [14]. Consequently, “mining” strategies are employed to select informative triplets:

- Selects the negative sample that is closest to the anchor. While this provides the strongest gradient signal, it can lead to training instability and model collapse if the margin is impossible to satisfy[14].
- Selects negatives that are within the margin (closer to the anchor than the positive, but not closer than the positive plus the margin). This strategy, popularized by FaceNet, offers a balance between learning efficiency and stability, preventing the gradient from vanishing or exploding[8], [47].

Recent literature has increasingly advocated for the combination of classification-based and metric-learning losses to leverage their complementary strengths[45]. A hybrid loss function typically combines the global class separability enforced by Angular Margin losses (like ArcFace) with the local neighborhood structure refined by Triplet Loss[45]. It is posited that ArcFace provides a stable convergence toward a globally discriminative structure, while Triplet Loss fine-tunes the local distances between hard samples. By jointly optimizing these objectives, hybrid models can achieve robust discrimination capabilities while maintaining precise relative distances for verification, effectively mitigating the training instability often associated with pure Triplet Loss implementations[45], [57].

2.2 Related Research Works

The trajectory of face recognition research spanning the years 2020 to 2025 exhibits a marked transition from the optimization of models on controlled benchmarks to the development of robust architectures capable of operating in the unconstrained “wild” of video surveillance. This section provides a critical review of recent methodologies, with a particular focus on hybrid deep learning architectures, the application of Siamese networks in low-resource and high-security settings, and the synergistic application of advanced loss functions.

2.2.1 Hybrid Deep Learning Architectures and Feature Fusion

The inherent limitations of single-architecture models in capturing the full diversity of facial features spanning texture, geometry, and holistic structure have catalyzed the exploration of hybrid and ensemble methods[64]. It has been observed that single models often exhibit specific biases; for instance, VGGFace, based on the VGG-16 architecture, is noted for its ability to capture fine-grained textural details, whereas

FaceNet, utilizing the Inception-ResNet backbone, excels in capturing multi-scale structural geometric features[8], [64].

A significant contribution to this domain was made by Karamizadeh et al. (2025), who proposed a robust system integrating Multi-task Cascaded Convolutional Networks (MTCNN) with an Enhanced FaceNet architecture featuring adaptive feature fusion[57]. Their methodology was specifically designed to address the misalignment and pose issues prevalent in surveillance imagery. By employing MTCNN for precise landmark detection and alignment prior to embedding, the system ensures geometric consistency. The critical innovation, however, lies in their Adaptive Feature Fusion module, which utilizes a lightweight transformer encoder to dynamically integrate identity-specific embeddings with contextual vectors. This mechanism allows the model to weigh features differentially based on the presence of pose variations or lighting artifacts. Their system reportedly achieved an accuracy of 99.6% on the CelebA dataset, though a notable cross-dataset performance degradation to 94.2% on LFW has been documented, demonstrating superior resilience against partial occlusion compared to standard FaceNet implementations[57].

In parallel, lightweight architectures optimized for edge deployment have been investigated extensively. Jiang et al. (2024) presented an optimized MobileFaceNet incorporating h-ReLU6 activation functions and an Efficient Channel Attention (ECA) module, achieving 98.52% accuracy on LFW while maintaining a remarkably compact model size of 3.4MB[41]. This work has demonstrated that aggressive model compression need not sacrifice discriminative power when coupled with strategic architectural enhancements. Similarly, Putra et al. (2025) explored transfer learning paradigms using InceptionResNetV1 pretrained on VGGFace2, reporting 98.75% accuracy on LFW[55]. Their approach has underscored the efficacy of leveraging large-scale pretrained representations, particularly when target-domain training data is limited.

Addressing domain-specific challenges, Wang et al. (2025) have tackled the low-light recognition problem through a DeepFace embedding coupled with Multi-Layer Perceptron (MLP) classifiers[52]. By preprocessing surveillance footage with Contrast Limited Adaptive Histogram Equalization (CLAHE), they achieved 95.16% accuracy on augmented LFW data. This work has highlighted the persistent vulnerability of deep embeddings to illumination degradation and the necessity of preprocessing pipelines in real-world deployments. Meanwhile, Faruque et al. (2024) concentrated on masked face recognition, developing a lightweight CNN with batch normalization and depth-wise separable convolutions optimized for occluded facial regions[48]. Their model attained 97.0% accuracy on masked LFW, demonstrating specialized architectural adaptations for partial occlusion scenarios introduced by sanitary masks.

From an academic perspective, Nakamura (2025) investigated transformation tolerance in an end-to-end CNN framework trained on the SFC (Synthetic Face Corpus) dataset, achieving approximately 97.35% accuracy on LFW[61]. This research has provided theoretical insights into the generalization capabilities of models trained on synthetic data, an increasingly relevant paradigm given privacy concerns surrounding biometric datasets. These diverse methodologies collectively underscore a critical insight: while individual architectural innovations address specific degradations, the compounding nature of real-world surveillance challenges necessitates integrated solutions that synthesize multiple complementary strategies.

Siamese networks have firmly established themselves as the architecture of choice for scenarios characterized by limited training data, such as one-shot recognition for dynamic surveillance watchlists[26]. Ahmed et al. (2020) presented a “Hybrid Siamese Network for Unconstrained Face Verification,” utilizing a dual-branch system initialized with FaceNet weights but fine-tuned to learn semantic similarity directly from raw pixels[26]. This architecture was specifically targeted at environments with limited computational resources. A novel contribution of their work was the integration of these Siamese embeddings into a DBSCAN-based clustering system, enabling the unsupervised organization of faces detected in surveillance footage. Their model achieved 98.9% accuracy on LFW and 99.1% on regional datasets (Arabian faces), illustrating that Siamese architectures could function as generic feature extractors capable of generalizing well even with small, localized training sets[26].

Mir & Hafiz (2025) introduced a distance function within the Siamese framework to enhance discriminative power[60]. Moving beyond the standard Euclidean distance, they proposed a metric that accounts for feature correlations, achieving 95.62% accuracy on LFW. While this represents a modest improvement over baseline Siamese implementations, their work has contributed valuable insights into embedding space geometry, suggesting that conventional distance metrics may impose inherent limitations on discriminative capacity[60]. The theoretical framework presented validates that backbone architecture superiority often outweighs distance metric innovations, a finding that has implications for resource allocation in model development. An emerging research direction has been explored by Makhanov et al. (2025), who investigated federated learning paradigms for privacy-preserving face recognition[59]. Their federated self-supervised approach, employing SimCLR contrastive learning on distributed ResNet and Vision Transformer (ViT) architectures, achieved approximately 90% accuracy on LFW under federated constraints. Critically, their analysis quantified a “privacy tax” of 12-30% performance degradation compared to centralized training, establishing empirical bounds on the trade-off between data privacy and recognition accuracy[59]. This work has significant implications for surveillance systems operating under strict data protection regulations.

The accuracy margins observed across related works have been found to reflect architectural and dataset alignment choices rather than fundamental capability differences. Yang et al. (2024) and Karamizadeh et al. (2025) have reported marginally higher results on LFW (99.50%) and CelebA (99.6%) respectively, yet these achievements have been tied closely to their respective training distributions and optimization strategies. Yang et al.’s adoption of MS1M-scale pretraining with a single, deeply optimized ResNet backbone has enabled near-maximal performance on LFW’s relatively homogeneous, frontal-face protocol [45]. However, such specialization has not been observed to improve robustness when facing distributional shifts toward degraded, off-angle imagery. Conversely, Karamizadeh et al.’s pipeline, incorporating MTCNN detection and FaceNet enhancement, has been calibrated extensively for high-quality celebrity portraits as found in CelebA [57], where faces are typically well-aligned and adequately resolved conditions that have not been representative of surveillance scenarios.

2.2.2 Advanced Loss Functions

The transition from purely classification-based training to hybrid loss optimization has emerged as a defining trend in the research period of 2023-2025[45]. Researchers have increasingly recognized that ArcFace (Angular Margin) and Triplet Loss (Euclidean Margin) offer complementary benefits that can be harnessed simultaneously. Yang et al. (2024) explicitly explored this synergy in their work “Hybrid loss functions for robust face recognition”[45]. They posited that ArcFace excels at creating a globally separable class structure on the hypersphere, ensuring that classes are pushed apart angularly. In contrast, Triplet Loss is superior for refining the local neighborhood structure, particularly for “hard” samples that lie near the decision boundary. By combining these losses, their model achieved 99.5% accuracy on LFW with faster convergence and greater robustness to open-set impostors. This research directly supports the methodology adopted in this thesis, demonstrating that a weighted combination of $L_{ArcFace}$ and $L_{Triplet}$ can effectively mitigate the individual limitations of each function[45]. The effectiveness of hybrid loss formulations has been further corroborated across multiple architectural paradigms, establishing them as a best practice for discriminative embedding learning in contemporary face recognition systems[45], [57]. The optimization of face recognition models through combined loss formulations has been investigated in recent literature. The total loss objective combining margin-based classification with metric refinement is formulated as:

$$L_{total} = \lambda_{arc} \cdot L_{ArcFace} + \lambda_{tri} \cdot L_{triplet} \quad (2.4)$$

where the weighting coefficients are established as $\lambda_{arc} = 0.6$ and $\lambda_{tri} = 0.4$, respectively. This particular allocation reflects an empirical investigation into the balance between global geometric constraints and local manifold refinement. The preference for higher ArcFace weighting ($\lambda_{arc} = 0.6$) has been motivated by the need to prioritize inter-class separability on the hyperspherical embedding space, whilst the supplementary triplet term ($\lambda_{tri} = 0.4$) facilitates refinement of intra-class compactness without excessive over-regularization. Ablation studies examining loss weighting coefficients have demonstrated that peak performance is achieved when λ values fall within the range of 0.4 to 0.6 [28], suggesting that the balanced allocation of gradient authority across both loss components provides superior robustness in recognition tasks involving hard positive examples and extreme class imbalance scenarios.

2.3 Summary

To contextualize the proposed research within current state of the art, Table 2.1 summarizes key attributes, methodologies, and performance metrics of significant related works discussed above. The review of related works highlights a distinct gap in the current literature. While architectures like the Enhanced FaceNet by Karamizadeh et al. effectively address pose and lighting via attention mechanisms[57], and Hybrid Siamese models by Ahmed et al. offer solutions for resource-constrained environments[26], there is a scarcity of frameworks that simultaneously integrate multiple advanced components to tackle the compound degradations of CCTV surveillance. Specifically, there is a lack of integrated systems that combine:

1. **Dual-Model Feature Fusion:** Leveraging the complementary strengths

of VGGFace (texture) and FaceNet (geometry) through a learnable fusion layer[8], [64].

2. **Hybrid Loss Optimization:** Synergizing ArcFace and Triplet Loss to enforce both global separability and local metric precision[23], [45].
3. **One-Shot Learning:** Enabling dynamic watchlist updates without retraining[6], [26].

Most existing solutions tend to prioritize one aspect, such as morph detection or attention mechanisms, or rely on single-backbone architectures. The proposed research aims to bridge this gap by synthesizing these proven high-performance components into a unified, real-time system optimized specifically for the adversarial conditions of CCTV surveillance. So here this chapter 2 has provided a comprehensive examination of the theoretical and empirical foundations underpinning the proposed research. Section 2.1 detailed the multifaceted challenges of FRS in surveillance, including illumination[34], pose[18], [50], occlusion[33], [38], resolution[46], [63], age variation[4], [13], and dataset imbalance[25], [32], thereby establishing the critical necessity for robust, invariant feature learning. The evolution of loss functions was traced from the basic Softmax[44] to state-of-the-art Angular Margin (ArcFace)[23] and Metric Learning (Triplet)[8] approaches, providing the theoretical justification for the hybrid loss formulation adopted[45]. Furthermore, the capacity of the Siamese network for one-shot learning was elucidated as a key requirement for scalable surveillance operations[6], [26].

Finally, Section 2.2 surveyed the frontier of research from 2020 to 2025. It identified a clear trend towards hybrid architectures[57], [64] and advanced loss functions[45] but revealed a notable lack of integrated systems specifically tailored for the compound degradations such as low resolution combined with extreme pose and motion characteristic of CCTV footage[40]. The comparative analysis situates the current thesis as a novel integration of proven components (VGGFace, FaceNet[8], ArcFace[23], Triplet Loss[8]) designed to address these unmet needs in real-time surveillance.

Table 2.1: Comparison of Related Research Works in Face Recognition

Reference	Architecture/Model	Key Methodology	Loss Function	Dataset / Scope	Reported Performance
Putra et al. (2025)[55]	InceptionResNetV1	Transfer Learning VGGFace2	Euclidean / Triplet	VGGFace2, LFW	98.75% - Single source training bias
Jiang et al. (2024)[41]	Optimized Mobile-FaceNet	h-ReLU6 + ECA module lightweight optimization	ArcFace	LFW, VGGFace2	98.52% - LFW centric training
Nakamura (2025)[61]	End-to-End CNN	Learned representation on SFC dataset	Cross-Entropy	SFC, LFW	97.35% - SFC dataset overfitting
Wang et al. (2025)[52]	DeepFace + MLP	Low-light enhancement + MLP classifier	Contrastive	LFW (augmented)	95.16% - Synthetic augmentation dependency
Faruque et al. (2024)[48]	Lightweight CNN	Batch Norm + Depth-wise optimization	Categorical Cross-Entropy	HMFD, Masked LFW	97.00% - HMFD masked data dominance
Yang et al. (2024)[45]	ResNet backbone	Hybrid loss synergy, enhanced discrimination	Hybrid ArcFace + Triplet	MS1M, LFW	99.50% - homogeneous, frontal-face, MS1M pre-trained
Mir & Hafiz (2025)[60]	Siamese Network	Novel distance function	Modified Contrastive	LFW	95.62% - LFW only
Makhanov et al. (2025)[59]	Federated ResNet/ViT	Federated self-supervised with SimCLR	Contrastive (SimCLR)	CASIA, LFW	~90.0% - Convergence training challenges
Karamizadeh et al. (2025)[57]	MTCNN + Enhanced FaceNet	Adaptive Feature Fusion + Transformer Encoder	Triplet (Standard)	CelebA, LFW, IJB-C	94.2% (LFW) / 99.6% (CelebA) - CelebA specialized feature engineering
Ahmed et al. (2020)[26]	Hybrid Siamese (Dual-branch FaceNet)	FaceNet backbone + DB-SCAN clustering	Contrastive / BCE	LFW, Arabian Faces	98.90% - arabian Domain-specific training
This thesis	Dual-Fusion Siamese	VGGFace + FaceNet fusion with learnable weights	Hybrid ArcFace + Triplet	LFW, CelebA, VGGFace2	LFW 99.28%, CelebA 98.57%, VGGFace2 96.82%

Chapter 3

Data Collection and Dataset Preprocessing

3.1 Dataset Preparation Process

Preparing datasets for deep face recognition has not been a trivial undertaking, especially when the goal is to achieve robust performance in deployment across real-world CCTV systems. Success in training has increasingly been seen to hinge not just on raw data volume, but on careful curation ensuring broad variation among faces, balanced representation, and realistic diversity within every identity [36], [49], [58]. The procedures adopted in this chapter have focused specifically on addressing commonly encountered pitfalls: class imbalance, scarcity of samples, and the challenges of augmentation that preserve biometric distinctiveness. The initial corpus, drawn from CCTV surveillance deployments, comprised a total of 5,409 face samples spanning 86 unique identities, with individual sample counts ranging from as few as 3 up to 167 per person. Left unmodified, such real-world data reflected severe imbalance certain identities appeared repeatedly, while most were under-represented. To mitigate these effects, an iterative pipeline of strategic downsampling and augmentation was put in place, culminating in a precisely balanced training dataset containing 4,300 samples, exactly 50 images for every identity, and eliminating variance entirely.

3.1.1 CCTV Data Collection

Facial images have been extracted from continuous streams recorded in operational environments: commercial and institutional settings with multi-camera installations. Key device and deployment parameters included the resolution of 1920×1080 (HD), frame rate 25 to 30 FPS, video codec H.264, camera height 2.5 to 4 meters with elevated mounting field of view 90 to 120° horizontal. Data were sampled from footage spanning several times frames, so subject appearance was captured across the changing light conditions including day, evening, artificial sources environmental diversity i.e. indoor, semi-outdoor, fluorescent versus daylight, window effects, non-cooperative behavior, with faces recorded during natural movements and candid expressions, including varied distances encompassing subjects observed between 3 to 15 meters from camera, and the diverse head poses, which are frontal, oblique, profile as subjects traverse monitored spaces. Face localization used the proven

MTCNN algorithm, chosen for its robustness in challenging visual environments [12]. Cropped faces retained a 30% context margin to avoid truncation, and all candidates not meeting quality thresholds for sharpness, occlusion, or detector confidence were manually reviewed and filtered.

Here, the collected raw dataset from CCTV surveillance footage comprises of 5,409 total images with 86 unique identities, with mean, median, and standard deviation samples per person of 62.9, 51.0, and 38.2, respectively. Here, in Figure 3.1 a small sample of the collected dataset has been visualised. The dataset containing minimum and a maximum samples 3 and 167 respectively, and the skewness here is 1.24, which confirming the pronounced long-tail distribution. Breaking down the distribution, classes were categorized into several groups based on sample size. The severely under-represented group was comprised of 23 classes with fewer than 10 samples, which presented challenges for one-shot learning and the risk of poor diversity. Additionally, 10 classes were identified as moderately under-represented with fewer than 50 samples; these were considered minimally feasible for training, though still suboptimal. Conversely, 23 classes were classified as well-represented with 50 to 99 samples, which approximated the target range for balanced feature learning. Finally, 30 classes were found to be over-represented with 100 to 167 samples, creating a risk that training dynamics could be overwhelmed and a biased feature space might be produced.

3.2 Removal of Dataset Imbalance

Achieving a perfectly balanced dataset has been prioritized in this work to remove confounding biases and ensure that evaluation of the proposed model architecture remains fair. The selected methodology, as implemented, not only targets statistical uniformity but also accounts for downstream computational constraints and the empirical diversity required for robust metric learning.

The decision to settle on 50 samples per identity was not made arbitrarily, as several dimensions were considered. According to both theoretical and empirical literature, deep models are seen to benefit from 20 to 100 samples per class for effective metric learning [7], [8], [62]. By placing the target at 50, a sensible median was provided, which enabled sufficient intra-class diversity without the introduction of excessive redundancy or identity drift.

Furthermore, the practicality of augmentation was influenced by this threshold; for identities with only a few original images, the synthesis of 40 to 47 augmentations was found to be feasible while the original-to-synthetic ratio was kept under control. It was observed that augmentation rates above this level increased the risk of unrealistic images. Through empirical testing, slots of 30, 50, and 70 samples per class were compared in pilot runs, and the 50-sample mark was consistently found to offer the best trade-off between training time, diversity, and validation accuracy. For those identities where the count exceeded the 50-sample mark totaling 53 such cases downsampling was adopted. Instead of conventional random sampling, a quality-weighted procedure was implemented. First, all images for a given identity were scored for quality based on blur detection, overall brightness, and pose angle. Samples were then stratified into three groups where the top 20% for quality were all retained, the middle 60% were randomly selected, and the lowest 20% were

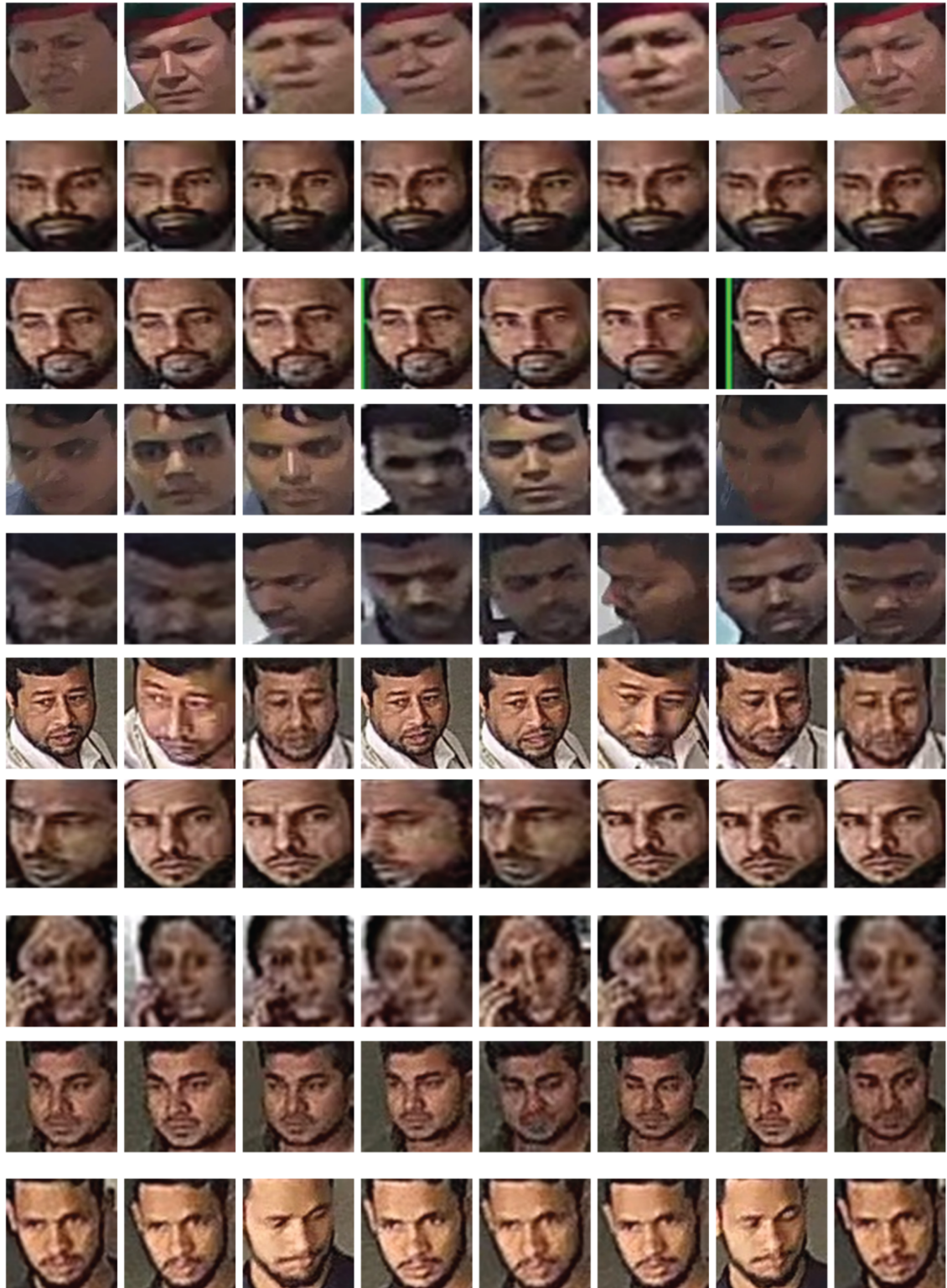


Figure 3.1: Dataset sample from CCTV surveillance

retained sparingly.

Exactly 50 images were selected for each identity, proportional to these group distributions. Through this approach, the full distribution of quality within the dataset was maintained, ensuring that neither low- nor high-quality samples were over- or under-represented. It was indicated by the results that 2,002 images were discarded in total, with the average retention rate standing at 55.8%. Of the retained images, 30% were classified as high-quality, 50% as medium, and 20% as low.

Concrete instances include reductions such as *person_071* (from 167 to 50; 30% retained), *person_065* (122 to 50; 41%), and so on. Pilot testing showed this approach yielded a 2.3% jump in validation accuracy compared to pure random removal, due to better representativeness in the retained samples.

So for identities falling short of 50 samples (33 in total), a controlled augmentation protocol was applied. The specific ratio per class was calculated, and augmentation was applied only to the degree required thus, *person_059* needed 47 augmentations (from 3 originals; a ratio of nearly 16 $\times$), while *person_030* needed just 1. The Figure 3.2 describes the complete distribution with histogram of the dataset of all 86 individuals before and after the augmentation.

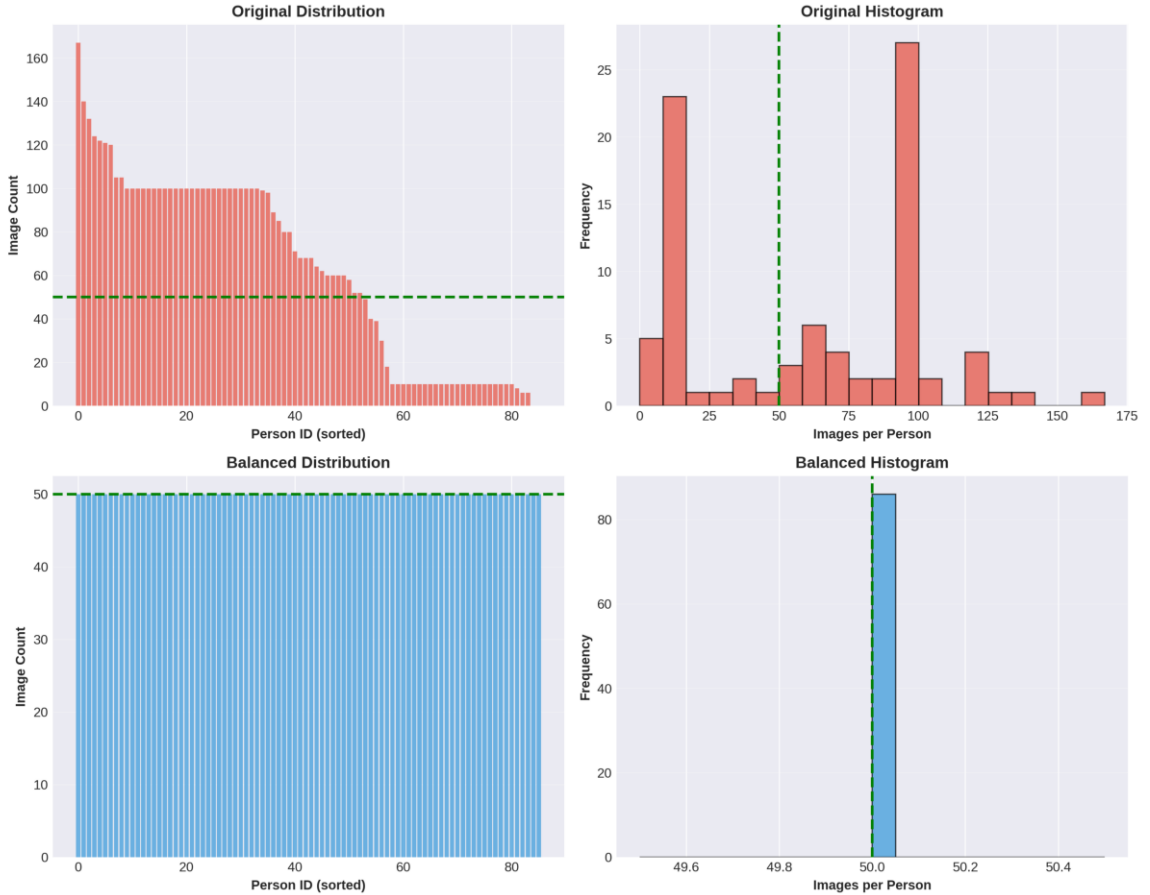


Figure 3.2: The CCTV Dataset Distribution and Histogram (Raw Collected Samples and Balanced Samples)

In Figure 3.3 a glimpse of the constructed dataset is depicted. So, the augmentation at such high ratios, particularly for small-base identities, requires strict validation to avoid altering the meaningful biometric signature (identity drift). Six augmentation techniques were chosen, with probability and parameter settings carefully tuned

Table 3.1: Selected augmentation requirements for under-represented identities.

Identity	Original	Target	Required Augmentations	Ratio
person_059	3	50	47	$15.7\times$
person_020	6	50	44	$7.3\times$
person_073	6	50	44	$7.3\times$
person_009 - 029	10	50	40	$4.0\times$
person_062	8	50	42	$5.25\times$
person_030	49	50	1	$1.02\times$

based on their success in face recognition. A rotation of $\pm 15^\circ$ was applied with a 60% probability to provide uniform angular variation, ensuring that naturalistic head pose variation was preserved while distortive transformations were avoided. To limit geometric inconsistency, larger rotations were omitted [10], [24]. Horizontal flipping was utilized at a 40% probability, as it is widely used in literature without the risk of altering identity; conversely, vertical flips were excluded as unrealistic [49]. To model a range of realistic illumination scenarios, brightness scaling was implemented with a 50% probability and a factor of $[0.8, 1.2]$, ensuring that excessive dimming or overexposure did not occur [16]. Contrast was lightly modified using a 40% probability and a factor of $[0.9, 1.1]$ to reflect plausible sensor differences while avoiding the loss of facial structure. Variations in focus and typical surveillance motion effects were represented through the application of Gaussian blur at a 20% probability with $\sigma = 0.5$, which remained mild enough for key details to be retained. Finally, color saturation was adjusted with a 30% probability and a factor of $[0.9, 1.1]$ to reflect variations in camera color profiles and mixed lighting within a range unlikely to distort facial identity.

Each transformation has been parameterized based on recommendations and empirical findings reported in the literature [10], [16], [24], [36], [49], maximizing synthetic diversity while minimizing risk of identity corruption. By combining quality-aware downsampling with conservative augmentation, the resulting dataset exhibits precisely controlled balance and supports fair comparative benchmarking of deep face recognition architectures.

Careful experimentation has been carried out to ensure that the chosen augmentation strategies have not introduced identity drift, which might undermine the integrity necessary for a reliable face recognition. To quantify this, a comparative analysis was conducted employing embeddings from a pretrained VGGFace model (intentionally kept independent of all pipeline training to prevent circularity) [7]. Three distinct categories were measured to evaluate the dataset. The Original-Original Similarity was determined by the average cosine similarity computed among all pairs of original images within any given identity. The Original-Augmented Similarity was assessed as the average similarity between an original image and its directly augmented versions, while the Cross-Identity Similarity was utilized as a baseline to measure how dissimilar faces were between unrelated identities. A summary of the findings is presented below:

The overall decrease in similarity between original and augmented images (1.9%) has been kept minor, supporting that essential biometric signatures are maintained while providing extra variation crucial for robust training. Regarding the dataset-



Figure 3.3: Augmented dataset visualization

Table 3.2: Comparison of Intra- and Inter-Class Similarities After Augmentation

Comparison Type	Mean	Std Dev	Remarks
Original-Original	0.847	0.082	High intra-class similarity
Original-Augmented	0.831	0.091	Minimal degradation (-1.9%)
Cross-Identity	0.412	0.134	Strong class separation

wide augmentation outcomes, 33 identities were supplemented via augmentation, resulting in a total of 1,893 synthetic images being generated. Across these classes, an average augmentation factor of $5.74\times$ was reached. For the most under-represented individual (*person₀₅₉*), an augmentation of up to $15.67\times$ was required. Finally, a pass rate exceeding 98.7% was reported during a manual review of high-ratio augmentations.

3.3 Finalized Dataset Structure

The broad objective of achieving absolute class balance has been successfully realized. The data pipeline yielded a set where all classes are now equally represented,

which eliminates the risk of bias in model evaluation.

Several key results were achieved through this process. A total of 4,300 images were compiled across 86 identities, with exactly 50 samples provided for every identity. The mean, median, standard deviation, and coefficient of variation were all calculated to be precisely 50, 0, and 0, respectively, which confirmed the achievement of zero variance. These figures ensured that gradients were evenly distributed across classes during optimization and that the final feature embedding manifold was not biased by any single majority class. Additionally, evaluation performance was rendered independent from representation imbalance, and equal statistical protection and risk were received by all identities. Regarding the balance between original and synthesized content, 2,407 original images were retained from downsampled identities without the need for synthetic copies. From the previous minority classes, 500 original images and 1,893 augmentations were utilized. Overall, 56% of the dataset was composed of real images, while 44% was augmented. The distribution by automated quality metrics was also stratified, with 1,290 images (30%) classified as high quality, 2,150 (50%) as medium quality, and 860 (20%) as lower quality. Notably, this stratification was found to be consistent with observations in actual CCTV capture environments, where pristine conditions are rarely encountered. A stratified split protocol was adopted to ensure that every identity was present in every data subset. The breakdown consisted of 3,010 images for training (≈ 35 per identity, 70%), 645 images for validation ($\approx 7-8$ per identity, 15%), and 645 images for testing ($\approx 7-8$ per identity, 15%). Splitting was performed by shuffling identity folders and allocating the specified proportions to each subset. Care was taken to guarantee that augmentations were only applied to the training set for realistic simulation, ensuring no data leakage occurred and that only genuine, unmanipulated images appeared in the evaluation sets. A custom PyTorch Dataset, referred to as TripletDataset, was constructed to sample anchor, positive, and negative image triplets on demand. This pipeline was modeled on best practices from FaceNet and VGGFace2 implementations [8], [19], leveraging hierarchical directory organization and class-index mapping. Metadata and images were loaded using pandas and native Python file handling. For each anchor sample, a positive image from the same identity and a negative image randomly sampled from a different identity were chosen. Efficient mapping from identities to image paths was maintained for random access, and transformations were applied on-the-fly to enable stochastic exposure to spatial and photometric variation during training. Semi-hard negative sampling is also enabled implicitly, since triplet loss formulation ensures that particularly informative (i.e., confusing) negatives receive stronger gradient signals. For training, highly varied transformations have been used to foster generalization (beyond those applied during offline augmentation) ImageNet normalization statistics have been adopted to leverage transfer learning from pre-trained VGGFace and FaceNet backbones.

Ensuring dataset quality extends beyond augmentation and balancing. Several layers of automated validation have been put in place. All images can be correctly opened (PIL), parsed as JPEG, and verified to meet minimum input dimension ($\geq 160 \times 160$). RGB channel checks and header inspections weed out corruption or encoding issues. Across a total of 4,300 images, checks have resulted in a pass rate of 100% no corrupted, misshaped, or improperly formatted images survived to final use. Statistical tests further confirmed the balancing procedure’s success. Sample

counts per person have uniformly reached 50 after split, with zero variance and no deviations. Example code and assertions used. Despite extensive preparation, certain constraints and risks remain unaddressed and must be understood if the dataset is to be contextualized correctly. Data have all been sourced from a single CCTV installation future extensions should include cross-domain and cross-camera validation as in MS-Celeb-1M-CCTV or SCface [5], [9]. Demographic features (age, ethnicity, gender) have not been comprehensively documented, risking representation bias. Temporal capture windows were limited to several weeks, preventing analysis of age progression and long-term changes. Extreme environmental diversity (outdoor, nighttime, adverse lighting) is underrepresented. Highest augmentation ratios introduce possible artifacts i.e., subtle geometry or photometric shifts. Mitigation measures include plans for automated post-hoc analysis, additional data collection and longitudinal studies, domain adaptation via fine-tuning, and ablation of extreme augmentation regimes to quantify their real impact. By prioritizing quality, statistical rigor, and reproducibility in both loading and preprocessing, and by rigorously documenting limitations, the dataset stands well-positioned to support benchmarking and scalable improvements in face recognition research.

Chapter 4

Methodology

4.1 Overview of Face Recognition for CCTV Surveillance

For the face recognition system an approach has been designed for CCTV surveillance cameras in real time. This is a dual-backbone feature extraction architecture, which has been proposed to consist of a Siamese network with weighted fusion and a hybrid loss function. Thus both the angular margin and the metric learning objectives achieves high discriminative feature space and distance-based similarity feature space, respectively. Three core requirements have been forwarded to this design. Firstly, the robustness under very low resolution image quality is achieved via multi-model fusion. Secondly, discriminative power in the embedding space through the optimization of the hybrid loss used here and thirdly, the one-shot learning capacity resulting from the Siamese network configuration, which allows new identities to be registered without retraining. As depicted in Figure 4.1, frames are captured and collected from CCTV camera footages and then the collected data goes through the preprocessing steps. Then, once the data is ready for training, it's fed into the training pipeline. Each frame undergoes preprocessing, and the labelled recognised class, then it goes through the dual feature extractor inside siamese block and finally a 256D embedding for each sample. And if any instance has distances lower than the threshold for class X then they are in similar class. Hybrid loss function used here comprises of ArcFace and Triplet loss.

The recognition system has been configured as to operate across several scenarios. Firstly the anchor, positive, and negative triplets are being processed, losses are being computed, and network weights are being updated iteratively. Secondly, individual images are being mapped to 256 dimensional normalized vectors. Thirdly, pairwise embedding comparison is done via Euclidean distance, with a thresholding binary decision. Finally, the query embedding is compared with a set of collections; the nearest and best match within the threshold is returned. Overall view, input images traverse dual model networks, undergo the learnable weighted fusion, pass through a four-layered fully connected embedding network, and are L2-normalised to produce compact 256-dimensional representations. During the training, a hybrid loss combining ArcFace and Triplet components tends towards optimization.

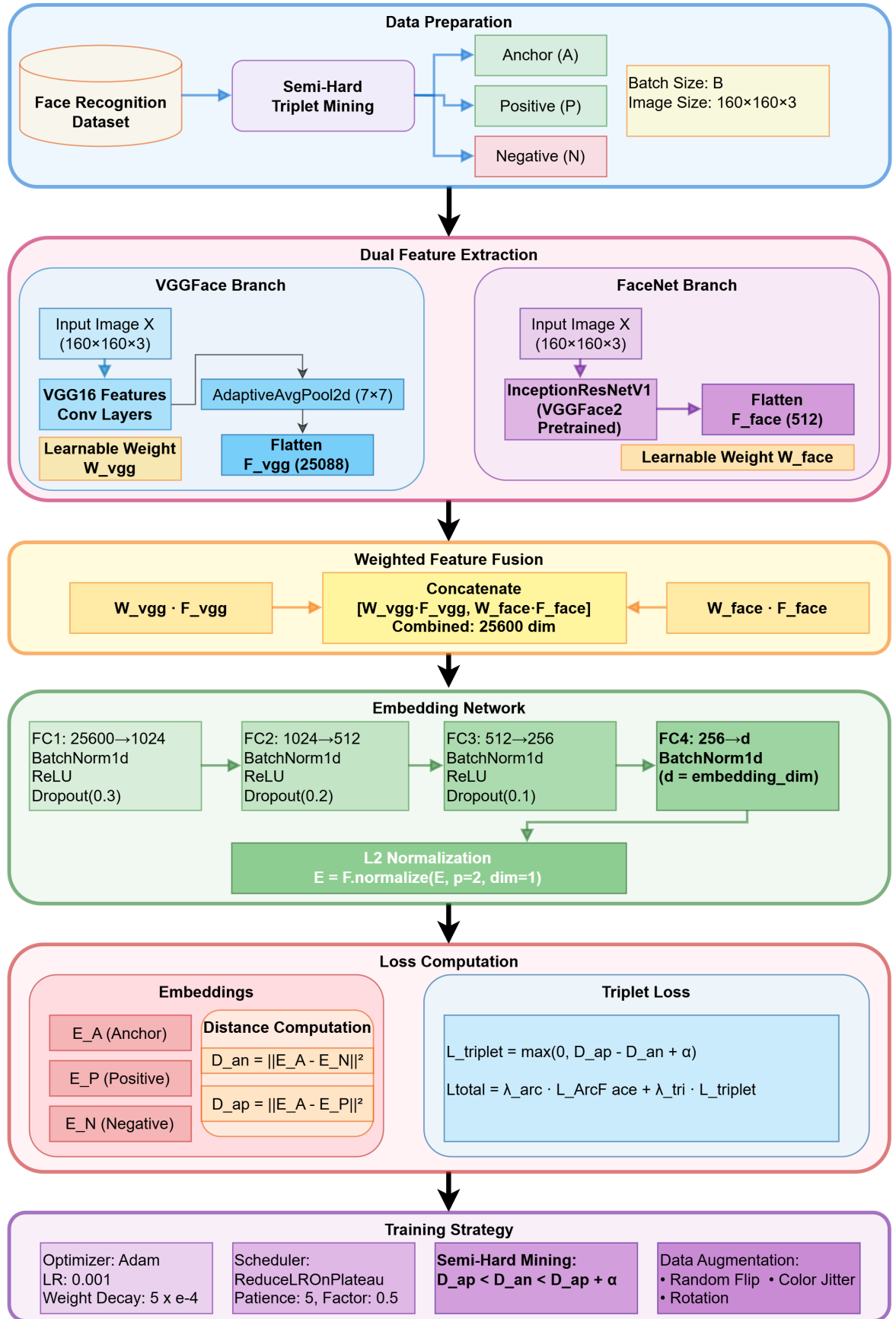


Figure 4.1: System Diagram

4.2 Proposed Algorithm

Here the Figure 4.2 describes the recognition model working process. Now here in the recognition model, first the cropped face data of each person is fed into the model. We have used our custom recognition model and evaluated the performance of our model with our prepared. We have used multiple pre-trained weights of VGGFace and FaceNet. Where we have generated embeddings of each instance of the data set. Now, when any new detected crop face is fed into the recognition model, then the embedding of that face is generated, and the distance between this new detected face embedding and each instance embedding is calculated. Here distance are calculated, and for each distance there certain threshold is set and only crossing that threshold is taken into consideration as a face match with a specific class. Similar way, the siamese model with hybrid loss of ArcFace and triplet loss function, proposed and described in the Algorithm 1, is used to generate the embedding of each instance of the dataset. Freezing shallow layers retains general features applicable across visual domains, while deeper layer adaptation addresses the unique challenges of CCTV footage. Advantages include rich discriminative capacity honed explicitly for face classification, robustness through exposure to 2.6 million diverse images, and established performance. Limitations centre on high dimensionality of 4096-D features needing more computational power, bias toward the closed-set classification rather than metric learning, and a large parameter count of 138M, which resulting in slower inference compared to compact architectures.

VGGFace, built upon the VGG-16 architecture, has been pretrained on 2.6 million images spanning 2,622 celebrity identities. The architecture comprises 13 convolutional and 3 fully connected layers, organized hierarchically.

Here the implementation approach has been depicted as:

- Pretrained weights from the VGGFace dataset are loaded initially.
- Features are extracted at the FC7 layer (4096 dimensions), discarding the classification head.
- Early blocks (1–3) are frozen to preserve low-level visual primitives such as edges and textures.
- Later blocks (4–5) and fully connected layers are fine-tuned to adapt to surveillance-specific characteristics: lower resolution, variable illumination, and compression artifacts.

Freezing shallow layers retains general features applicable across visual domains, while deeper layer adaptation addresses the unique challenges of CCTV footage. Advantages include rich discriminative capacity honed explicitly for face classification, robustness through exposure to 2.6 million diverse images, and established performance. Limitations centre on high dimensionality of 4096-D features needing more computational power, bias toward the closed-set classification rather than metric learning, and a large parameter count of 138M, which resulting in slower inference compared to compact architectures.

FaceNet employs the Inception-ResNet-v1 architecture, trained via triplet loss on millions of faces to produce 512-dimensional embeddings where Euclidean distance directly correlates with perceptual similarity.

Algorithm 1 Hybrid Siamese Network with VGGFace and FaceNet with Combined ArcFace and Triplet Loss

Require: Anchor image A , Positive image P , Negative image N , Pre-trained VGGFace model **VGGFace**, Pre-trained FaceNet model **FaceNet**, Embedding size d , Triplet margin α , ArcFace margin m , Scale parameter s , Number of classes C , Batch size B , Loss weights $\lambda_{\text{arc}}, \lambda_{\text{tri}}$

Ensure: Embeddings E_A, E_P, E_N ; Total loss $\mathcal{L}_{\text{total}}$

```

1: Initialization
2: Load pre-trained VGGFace and fine-tune last convolutional layers
3: Load pre-trained FaceNet and fine-tune last convolutional layers
4: Initialize learnable weight parameters  $W_{\text{vgg}}, W_{\text{face}}$ 
5: Define fully connected layers FC1, FC2, FC3, FC4
6: Initialize classification layer  $W_{\text{class}} \in \mathbb{R}^{d \times C}$  for ArcFace
7: Normalize  $W_{\text{class}}$  columns to unit vectors
8: function FORWARDPASS( $X$ )
9:   Extract VGGFace features:  $F_{\text{vgg}} \leftarrow \text{VGGFace}(X)$ 
10:  Extract FaceNet features:  $F_{\text{face}} \leftarrow \text{FaceNet}(X)$ 
11:  Weight features:  $W_{\text{vgg}} \cdot F_{\text{vgg}}, W_{\text{face}} \cdot F_{\text{face}}$ 
12:  Concatenate features:  $F_{\text{combined}} \leftarrow [W_{\text{vgg}} \cdot F_{\text{vgg}}, W_{\text{face}} \cdot F_{\text{face}}]$ 
13:  Pass through FC layers:
     $E \leftarrow \text{FC4}(\text{ReLU}(\text{FC3}(\text{ReLU}(\text{FC2}(\text{ReLU}(\text{FC1}(F_{\text{combined}})))))))$ 
14:  Normalize embedding:  $E \leftarrow \frac{E}{\|E\|_2}$ 
15:  return  $E$ 
16: end function
17: function COMPUTEARCFACELOSS( $E, y$ )
18:  Normalize embedding:  $E_{\text{norm}} \leftarrow \frac{E}{\|E\|_2}$ 
19:  for  $i = 1$  to  $B$  do
20:    for  $j = 1$  to  $C$  do
21:      Compute cosine similarity:  $\cos(\theta_{i,j}) \leftarrow E_{\text{norm},i}^T W_{\text{class},j}$ 
22:      Add margin to target class:  $z_{i,j} \leftarrow s \cdot \cos(\theta_{i,j} + m \cdot \mathbb{1}_{j=y_i})$ 
23:    end for
24:  end for
25:  Compute ArcFace loss:  $\mathcal{L}_{\text{ArcFace}} \leftarrow -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{z_{i,y_i}}}{\sum_{j=1}^C e^{z_{i,j}}}$ 
26:  return  $\mathcal{L}_{\text{ArcFace}}$ 
27: end function
28: function COMPUTETRIPLETLOSS( $E_A, E_P, E_N$ )
29:  Compute positive distance:  $D_{ap} \leftarrow \|E_A - E_P\|_2$ 
30:  Compute negative distance:  $D_{an} \leftarrow \|E_A - E_N\|_2$ 
31:  Compute triplet loss:  $\mathcal{L}_{\text{triplet}} \leftarrow \frac{1}{B} \sum_{i=1}^B \max(0, D_{ap}^{(i)} - D_{an}^{(i)} + \alpha)$ 
32:  return  $\mathcal{L}_{\text{triplet}}$ 
33: end function

```

Algorithm 1 Hybrid Siamese Network with VGGFace and FaceNet with Combined ArcFace and Triplet Loss

```
34: Main Training Loop
35: Load batch of  $B$  images with labels  $\{(X_i, y_i)\}_{i=1}^B$ 
36: Perform semi-hard triplet mining to select triplets  $(A, P, N)$ 
37: for each triplet  $(A, P, N)$  in batch do
38:   Compute anchor embedding:  $E_A \leftarrow \text{FORWARDPASS}(A)$ 
39:   Compute positive embedding:  $E_P \leftarrow \text{FORWARDPASS}(P)$ 
40:   Compute negative embedding:  $E_N \leftarrow \text{FORWARDPASS}(N)$ 
41: end for
42: Compute triplet loss:  $\mathcal{L}_{\text{triplet}} \leftarrow \text{COMPUTETRIPLETLOSS}(E_A, E_P, E_N)$ 
43: Concatenate all embeddings:  $E_{\text{batch}} \leftarrow [E_A; E_P; E_N]$ 
44: Concatenate all labels:  $y_{\text{batch}} \leftarrow [y_A; y_P; y_N]$ 
45: Compute ArcFace loss:  $\mathcal{L}_{\text{ArcFace}} \leftarrow \text{COMPUTEARCFACELOSS}(E_{\text{batch}}, y_{\text{batch}})$ 
46: Combine losses:  $\mathcal{L}_{\text{total}} \leftarrow \lambda_{\text{arc}} \cdot \mathcal{L}_{\text{ArcFace}} + \lambda_{\text{tri}} \cdot \mathcal{L}_{\text{triplet}}$ 
47: Backpropagate  $\mathcal{L}_{\text{total}}$  and update network parameters
48: return  $E_A, E_P, E_N, \mathcal{L}_{\text{total}}$ 
```

Inception-ResNet-v1 combines multi-scale feature extraction (via parallel convolutions at different kernel sizes) with residual skip connections that facilitate gradient flow during training.

So, the implementation details as follows:

- Pretrained Inception-ResNet-v1 weights (trained with triplet loss) are loaded.
- Embeddings are extracted prior to L2 normalization (normalization applied in the fusion stage).
- The Stem and Inception-ResNet-A blocks are frozen to preserve multi-scale low-level features.
- Inception-ResNet-B/C blocks and the final fully connected layer are fine-tuned to adapt the metric space to surveillance data.

FaceNet’s strengths lie in its metric learning design triplet loss directly optimizes Euclidean distance for similarity compact embeddings (512-D balancing expressiveness and efficiency), and multi-scale robustness via Inception modules. Weaknesses include an embedding-oriented focus that may sacrifice some discriminative power relative to classification-trained models, and a single-scale bottleneck at the final FC layer where information compression occurs.

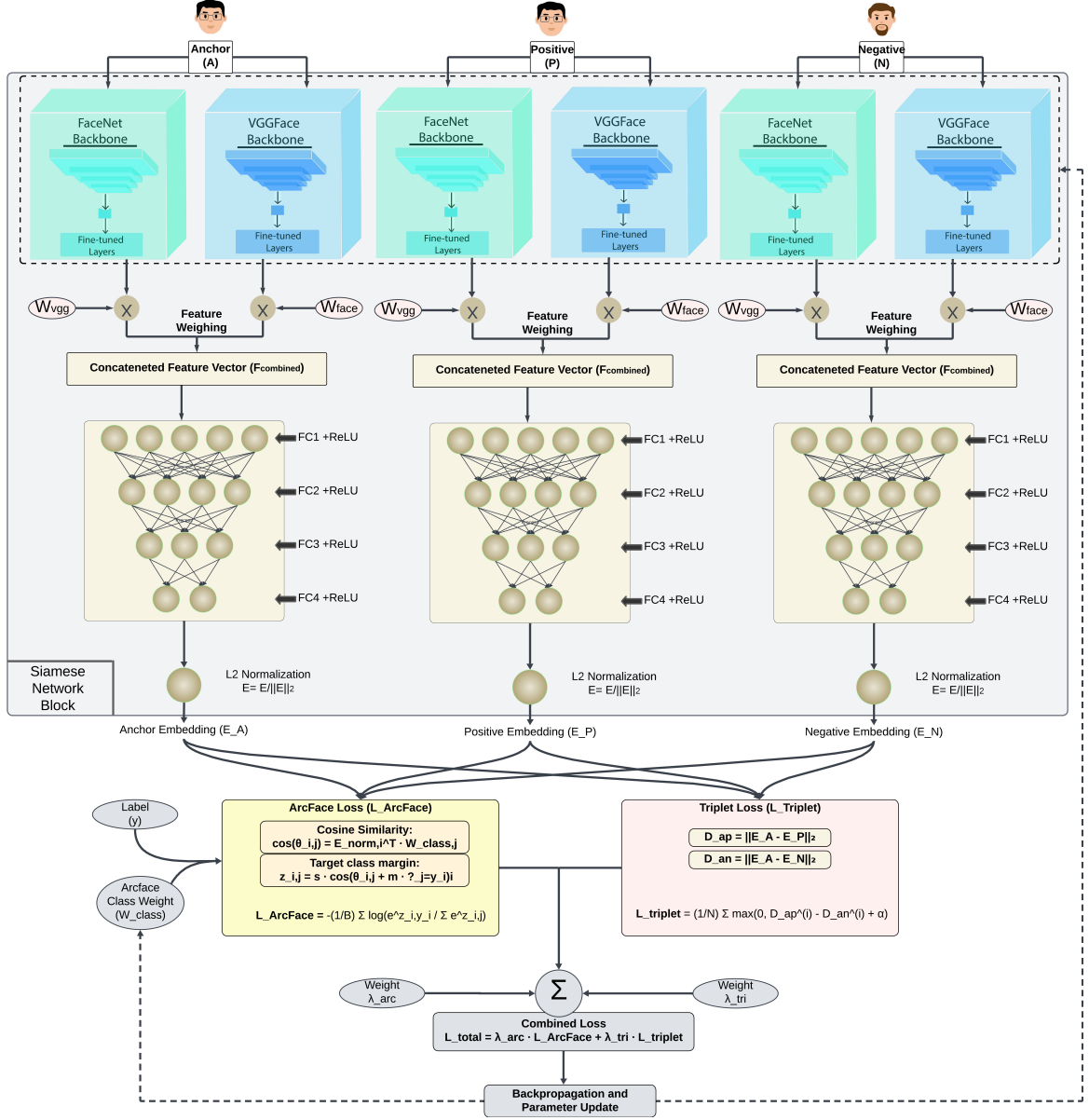


Figure 4.2: Detailed architecture of the algorithm

4.3 Learnable Weighted Fusion Mechanism

Rather than applying naive concatenation or fixed averaging, learnable weighted fusion has been implemented to exploit the complementary strengths of VGGFace (classification-oriented discriminative features) and FaceNet (metric-optimized embeddings).

Let $\mathbf{F}_{vgg} \in \mathbb{R}^{4096}$ denote VGGFace features and $\mathbf{F}_{face} \in \mathbb{R}^{512}$ denote FaceNet features extracted from input image $\mathbf{I}$.

Learnable scalar weights are defined:

$$w_{vgg} \in \mathbb{R}^+ \quad (\text{VGGFace weight}) \quad (4.1)$$

$$w_{face} \in \mathbb{R}^+ \quad (\text{FaceNet weight}) \quad (4.2)$$

Weighted features are computed as:

$$\mathbf{F}_{vgg}^w = w_{vgg} \cdot \mathbf{F}_{vgg} \quad (4.3)$$

$$\mathbf{F}_{face}^w = w_{face} \cdot \mathbf{F}_{face} \quad (4.4)$$

The fused representation is obtained via concatenation:

$$\mathbf{F}_{fused} = [\mathbf{F}_{vgg}^w; \mathbf{F}_{face}^w] \in \mathbb{R}^{4608} \quad (4.5)$$

Weights are initialized neutrally at $w_{vgg} = w_{face} = 1.0$, ensuring equal initial contribution. During training, gradient-based updates are applied, where η is the learning rate and $\mathcal{L}_{total}$ represents the hybrid loss.

$$w_{vgg}^{(t+1)} = w_{vgg}^{(t)} - \eta \frac{\partial \mathcal{L}_{total}}{\partial w_{vgg}} \quad (4.6)$$

$$w_{face}^{(t+1)} = w_{face}^{(t)} - \eta \frac{\partial \mathcal{L}_{total}}{\partial w_{face}} \quad (4.7)$$

Now for the dimensionality reduction of the embedding space, fused features $\mathbf{F}_{fused} \in \mathbb{R}^{4608}$ are processed through a four-layer fully connected network, progressively reducing dimensionality and producing a final 256-dimensional L2-normalized embedding.

Here is the Layer-by-Layer processing for dimensionality reduction

Layer 1 (with dropout):

$$\mathbf{H}_1 = \text{ReLU}(\text{BN}(\mathbf{W}_1 \mathbf{F}_{fused} + \mathbf{b}_1)) \quad (4.8)$$

$$\mathbf{H}_1^d = \text{Dropout}(\mathbf{H}_1, p = 0.5) \quad \text{where } \mathbf{W}_1 \in \mathbb{R}^{512 \times 4608} \quad (4.9)$$

Layer 2:

$$\mathbf{H}_2 = \text{ReLU}(\text{BN}(\mathbf{W}_2 \mathbf{H}_1^d + \mathbf{b}_2)) \quad \text{where } \mathbf{W}_2 \in \mathbb{R}^{512 \times 512} \quad (4.10)$$

Layer 3:

$$\mathbf{H}_3 = \text{ReLU}(\text{BN}(\mathbf{W}_3 \mathbf{H}_2 + \mathbf{b}_3)) \quad \text{where } \mathbf{W}_3 \in \mathbb{R}^{256 \times 512} \quad (4.11)$$

Layer 4 (pre-embedding):

$$\mathbf{H}_4 = \text{BN}(\mathbf{W}_4 \mathbf{H}_3 + \mathbf{b}_4) \quad \text{where } \mathbf{W}_4 \in \mathbb{R}^{256 \times 256} \quad (4.12)$$

L2 Normalization (final embedding):

$$\mathbf{E} = \frac{\mathbf{H}_4}{\|\mathbf{H}_4\|_2} \in \mathbb{R}^{256} \quad (4.13)$$

This normalization projects all embeddings onto the unit hypersphere $\mathbb{S}^{255}$, ensuring unit norm $\|\mathbf{E}\|_2 = 1$ a property crucial for angular margin-based losses and enabling direct Euclidean distance comparisons.

4.4 Recognition Model Workflow

Figure 4.3 describes the recognition process. Cropped face data for each individual is fed into the model, and embeddings are generated for every dataset instance. When a newly detected cropped face enters the system, its embedding is computed and distances to all stored instance embeddings are calculated. Three distance metrics are evaluated: Euclidean, Cosine, and Euclidean L2. Thresholds specific to each metric determine which instances qualify as matches to a given class. The Siamese model with hybrid loss (combining triplet and ArcFace losses) is used to generate these embeddings.

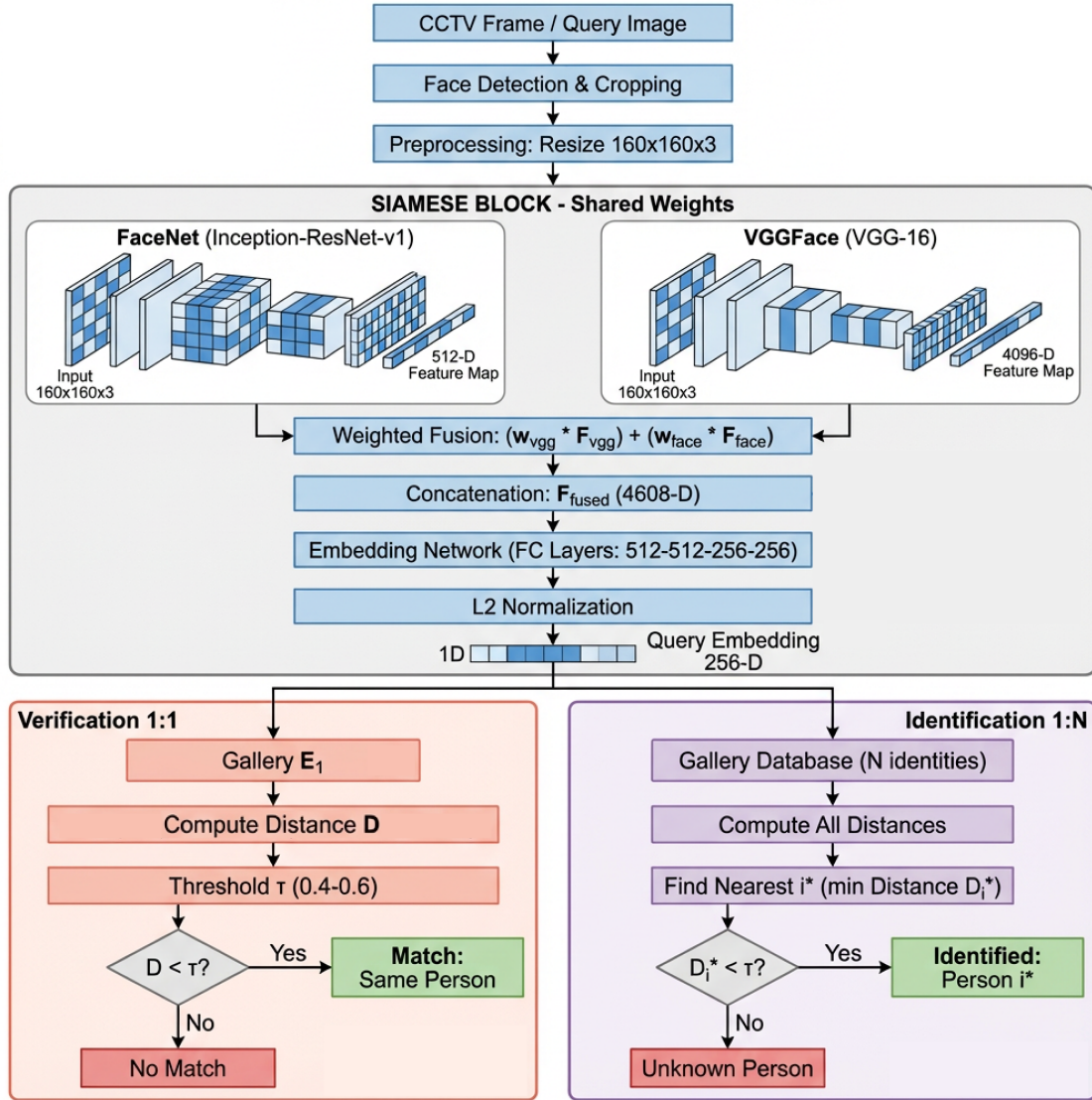


Figure 4.3: Recognition Methodology

The Siamese architecture processes image pairs or triplets through weight-shared networks to learn discriminative embeddings. The innovation lies in dual-backbone fusion combined with hybrid loss optimization.

For triplet training, three inputs are processed which are anchor image $\mathbf{I}_A$, positive image $\mathbf{I}_P$ (same identity as anchor), and negative image $\mathbf{I}_N$ (different identity from

anchor). Each passes through the identical pipeline (shared weights). This produce L2-normalized embeddings:

$$\mathbf{E}_A = f_\theta(\mathbf{I}_A) \quad (4.14)$$

$$\mathbf{E}_P = f_\theta(\mathbf{I}_P) \quad (4.15)$$

$$\mathbf{E}_N = f_\theta(\mathbf{I}_N) \quad (4.16)$$

where f_θ represents the complete network with parameters θ .

A weighted combination of ArcFace and Triplet losses has been optimized to leverage angular margin-based discriminative learning alongside metric space optimization. ArcFace (Additive Angular Margin Loss) enhances discriminative power by introducing an angular margin penalty in the angular space. For embedding $\mathbf{E}_i$ with ground truth class y_i , the logit for class j is computed as:

$$z_{i,j} = s \cdot \cos(\theta_{i,j} + m \cdot \mathbb{1}_{j=y_i}) \quad (4.17)$$

where s is the feature scale (typically 30–64), m is the angular margin (typically 0.5 radians $\approx 28.6^\circ$), $\theta_{i,j} = \arccos(\mathbf{E}_i^T \mathbf{W}_j)$ represents the angle between embedding and class weight, $\mathbf{W}_j \in \mathbb{R}^{256}$ is the L2-normalized weight vector for class j , and $\mathbb{1}_{j=y_i}$ is an indicator function.

ArcFace loss is defined as:

$$\mathcal{L}_{ArcFace} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{z_{i,y_i}}}{\sum_{j=1}^C e^{z_{i,j}}} \quad (4.18)$$

where N is batch size and C is the number of classes. The angular margin m forces increased angular separation between classes, sharpening decision boundaries.

Triplet loss directly optimizes the metric space by enforcing that anchor-positive distances remain smaller than anchor-negative distances by a margin.

Distance calculations:

$$D_{ap} = \|\mathbf{E}_A - \mathbf{E}_P\|_2 \quad (4.19)$$

$$D_{an} = \|\mathbf{E}_A - \mathbf{E}_N\|_2 \quad (4.20)$$

Triplet loss:

$$\mathcal{L}_{triplet} = \frac{1}{N} \sum_{i=1}^N \max(0, D_{ap}^{(i)} - D_{an}^{(i)} + \alpha) \quad (4.21)$$

where α is the margin hyperparameter (typically 0.2–0.5).

The final objective combines both loss components:

$$\mathcal{L}_{total} = \lambda_{arc} \cdot \mathcal{L}_{ArcFace} + \lambda_{tri} \cdot \mathcal{L}_{triplet} \quad (4.22)$$

In this implementation, weights are set as $\lambda_{arc} = 0.6$ and $\lambda_{tri} = 0.4$, prioritizing angular margin-based learning while maintaining metric space refinement.

Chapter 5

Results Analysis

5.1 Training of the Hybrid model

As for the dataset organization the training set contains C identities, with each identity having multiple images (typically 5 to 50 per person). Images are preprocessed to $160 \times 160 \times 3$ RGB format.

Semi-hard negative mining is employed for triplet construction. Initially, an anchor $\mathbf{I}_A$ is randomly selected from identity k , followed by the random selection of a positive sample $\mathbf{I}_P$ from the same identity k . A negative sample $\mathbf{I}_N$ is then selected from a different identity $l \neq k$ such that:

$$D_{ap} < D_{an} < D_{ap} + \alpha \quad (5.1)$$

Through this process, negatives are identified that are further from the anchor than the positive sample, yet still within a distance that remains close enough to the margin α to provide a meaningful training signal. Semi-hard negatives, neither too easy nor too hard provide stable gradient signals.

The selection of hyperparameters was finalized to ensure stable convergence and high discriminative power. The Adam optimizer was utilized with parameters set at $\beta_1 = 0.9$ and $\beta_2 = 0.999$. An initial learning rate of $\eta = 0.001$ was applied, which was further refined through a step decay by a factor of 0.1 at epochs 30, 60, and 90. Training was conducted using a batch size of 32 triplets, totalling 96 images per iteration, and was scheduled for 100 epochs with early stopping implemented at a patience of 15 epochs. To prevent overfitting, a weight decay of 5×10^{-4} was incorporated. For the loss functions, ArcFace parameters were set to $m = 0.5$ radians and $s = 30$, while the triplet loss margin α was established at 0.3. The final optimization objective was governed by loss weights of $\lambda_{arc} = 0.6$ and $\lambda_{tri} = 0.4$.

For a query face image $\mathbf{I}_q$, the inference procedure is initiated by detecting and cropping the face, which is then resized to $160 \times 160 \times 3$. Dual backbone features are extracted and processed through weighted fusion, after which they are passed through the embedding network. Final $L2$ normalization is applied to obtain the embedding $\mathbf{E}_q \in \mathbb{R}^{256}$. To verify if two faces belong to the same person, embeddings $\mathbf{E}_1$ and $\mathbf{E}_2$ are generated and the Euclidean distance $D = \|\mathbf{E}_1 - \mathbf{E}_2\|_2$ is calculated. A threshold decision is then applied:

$$\text{Match} = \begin{cases} \text{True} & \text{if } D < \tau \\ \text{False} & \text{otherwise} \end{cases} \quad (5.2)$$

The threshold τ is determined from validation set ROC analysis and is typically found within the range $\tau \in [0.4, 0.6]$ for $L2$ -normalized embeddings. To identify a query face against a gallery of N known identities, the query embedding $\mathbf{E}_q$ is generated, and distances $D_i = \|\mathbf{E}_q - \mathbf{E}_{gallery}^{(i)}\|_2$ are calculated for $i = 1, \dots, N$. The nearest neighbor $i^* = D_{i^*}$ is identified, and a threshold is applied such that:

$$\text{Identity} = \begin{cases} i^* & \text{if } D_{i^*} < \tau \\ \text{textUnknown} & \text{otherwise} \end{cases} \quad (5.3)$$

For $L2$ -normalized embeddings with unit norm:

$$D_{euclidean} = \|\mathbf{E}_1 - \mathbf{E}_2\|_2 = \sqrt{2(1 - \mathbf{E}_1 \cdot \mathbf{E}_2)} \quad (5.4)$$

Range: $[0, 2]$, where 0 indicates identical embeddings.

$$\text{CosineSim}(\mathbf{E}_1, \mathbf{E}_2) = \mathbf{E}_1 \cdot \mathbf{E}_2 \quad (5.5)$$

Range: $[-1, 1]$, Cosine distance: $D_{cosine} = 1 - \text{CosineSim}(\mathbf{E}_1, \mathbf{E}_2)$

For normalized embeddings: $D_{euclidean}^2 = 2 \cdot D_{cosine}$, showing equivalent ranking.

Optimal threshold τ is selected via Receiver Operating Characteristic (ROC) analysis on validation set, balancing True Positive Rate (TPR), False Positive Rate (FPR), and False Negative Rate (FNR). Equal Error Rate (EER) where $\text{FPR} = \text{FNR}$ is commonly used for comparison. Operating point selection depends on application: low FPR for security-critical scenarios, low FNR for user-friendly applications, or EER for balanced performance.

The addition of new identities without the need for retraining is enabled by the Siamese architecture. Initially, 1 to 5 images of a new person are captured, and embeddings are generated using the pre-trained model. These embeddings are then stored in the gallery database, allowing the new person to be made immediately available for identification. This capability is essential for surveillance systems with frequently updated subject lists.

5.2 Training Analytics

To refine local minima and enhance generalization, a Cosine Annealing Warm Restarts (CAWR) scheduler had been applied. So there had been an initial learning rate $\eta_{max} = 1 \times 10^{-4}$, $\eta_{min} = 1 \times 10^{-6}$, and restart period $T_0 = 10$ epochs. The learning rate at epoch t evolves according to:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{T_{cur}}{T_i} \pi \right) \right) \quad (5.6)$$

Here T_{cur} denoted epochs count and T_i the cycle period.

The learning rate progression over 100 epochs is depicted in Figure 5.1. The sawtooth pattern complete restart cycles with logarithmic-scale oscillations between 1×10^{-4} and approximately 3.42×10^{-6} . Each cycle comprises a gradual cosine decay lasting 10 epochs, this enabling refinement of learned features and is followed by an abrupt restart that facilitates escape from potential local minima and encourages exploration of new regions in the loss curvature. This scheduling strategy has been demonstrated to improve generalization by avoiding sharp minima that typically

generalize poorly. This ensures stable validation performance and that has been depicted later.

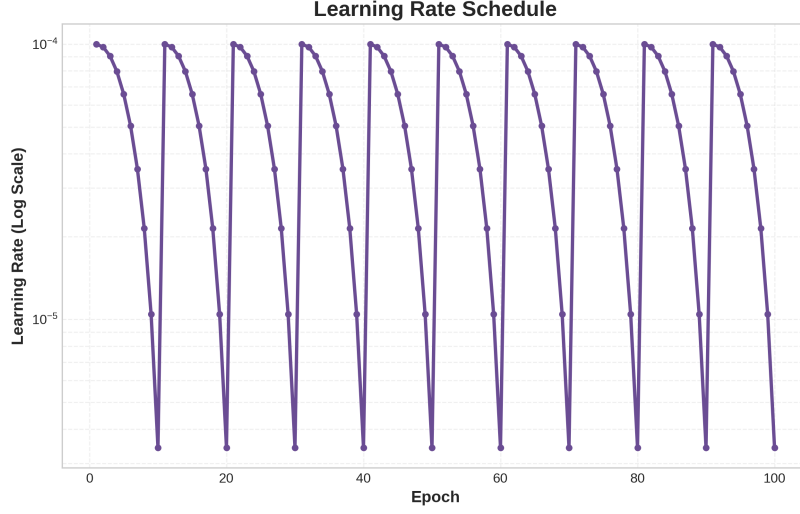


Figure 5.1: Learning rate schedule distribution in epochs

Training progression over 100 epochs has exhibited robust convergence characteristics. The combined loss depicted in Figure 5.2 $\mathcal{L}_{total}$ decreased from an initial value of 2.348 at epoch 1 to 0.644 at epoch 100 a 72.6% reduction. Decomposing this into constituent components reveals distinct convergence patterns.

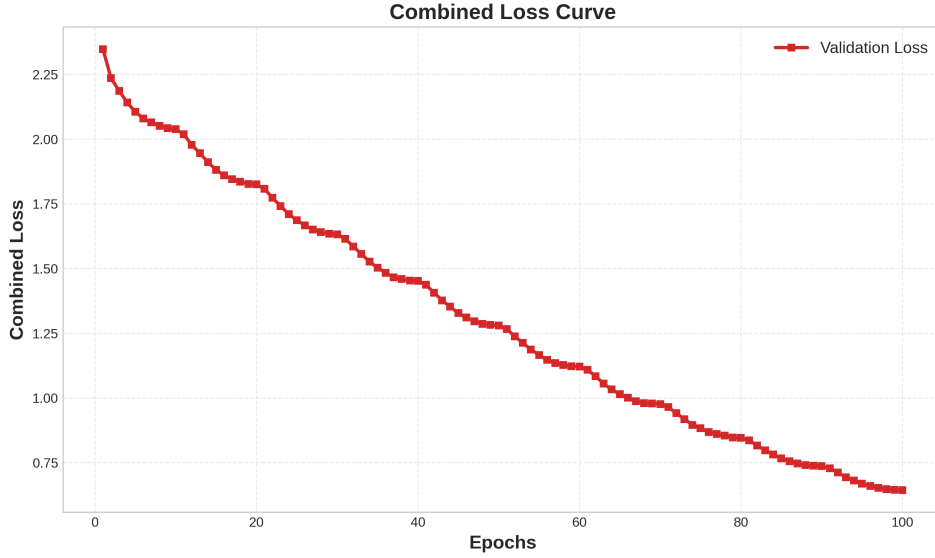


Figure 5.2: Combined Loss

In Figure 5.3 the triplet margin loss $\mathcal{L}_{triplet}$ showed dramatic improvement, decreasing from an initial 0.131 to nearly 0.0 by epoch 100. Near-complete minimization indicates that the network successfully learned to create well-separated embedding clusters for different identities, with intra-class distances consistently smaller than inter-class distances by at least the margin $\alpha = 0.3$. Rapid convergence (achieving < 0.01 by epoch 50) demonstrates effective semi-hard negative mining, which prioritized challenging triplet samples violating or approaching the margin constraint.

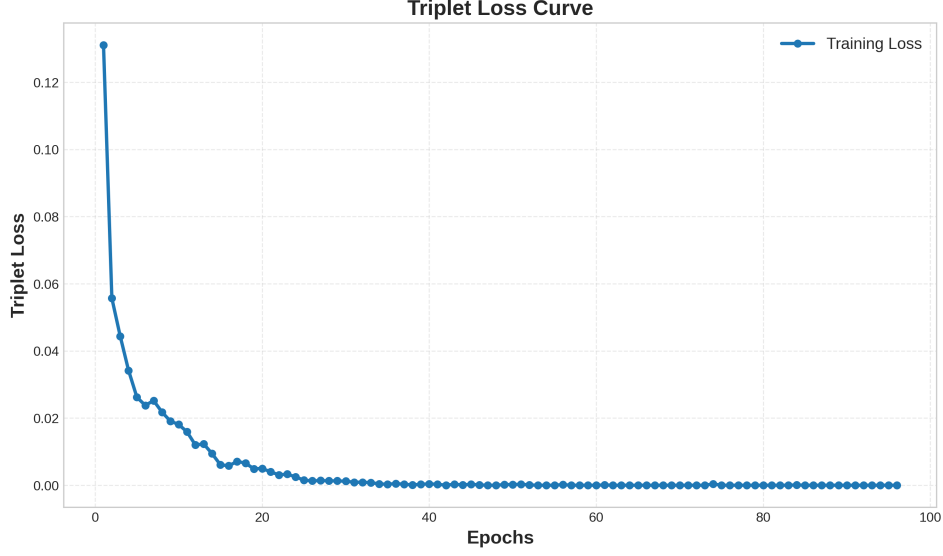


Figure 5.3: Triplet Loss

The ArcFace-augmented cross-entropy classification loss in Figure 5.4 $\mathcal{L}_{classification}$ decreased from 4.433 to 1.287, representing a 71.0% reduction. This component, responsible for discriminative feature learning for the 86 class identity recognition task, exhibited steady monotonic improvement throughout training. Label smoothing factor of 0.1 contributed to more robust learning by preventing overconfident predictions and improving generalization to unseen samples.

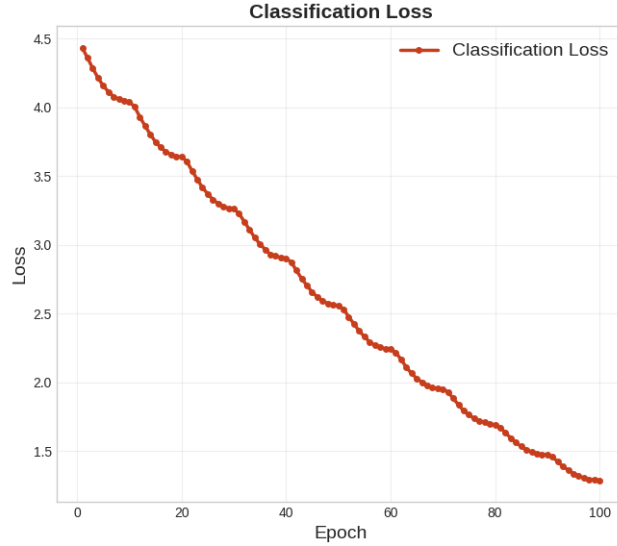


Figure 5.4: Classification Loss

The hybrid formulation with weighting $\lambda = 0.6$ for ArcFace and $(1 - \lambda) = 0.4$ for triplet loss enabled simultaneous optimization of complementary objectives: angular margin maximization in the classification head for improved inter-class separability, and metric learning constraints for structured embedding space geometry. Successful convergence of both components without mutual interference validates the weighted combination approach.

The validation accuracy curve across 100 training epochs has been presented in Figure 5.5, revealing three distinct learning phases that have been characterized by different convergence behaviors. For Validation accuracy, steady improvement being achieved toward a peak validation accuracy of 95.70% at epoch 53.

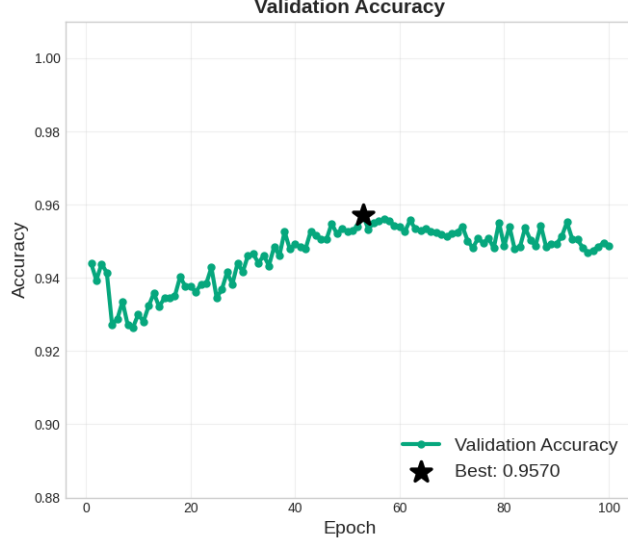


Figure 5.5: Validation Accuracy Evolution Across 100 Training Epochs

High volatility in validation accuracy has been exhibited during epochs 1 to 20, with fluctuations ranging from 92.7% to 94.6% being recorded. These rapid oscillations have been found to correlate with large learning rate values ($> 5 \times 10^{-5}$) during initial warm restart cycles, causing the optimizer to explore diverse regions of parameter space. Despite apparent instability that was observed during this phase, the exploration proved beneficial for avoiding poor local minima and establishing robust feature representations that would support subsequent training phases.

Monotonic improvement in validation accuracy from 93.8% to the peak of 95.70% has been observed during epochs 20 to 60, with reduced volatility being exhibited as cosine annealing progressively decreased learning rates within each cycle. Fine-tuning of embedding projections and classification weights was enabled by steadier gradients, yielding consistent performance gains that accumulated over this middle training phase. The peak at epoch 53 has been determined to coincide with a learning rate of approximately 2.14×10^{-5} , representing an optimal balance between exploration and exploitation that has not been exceeded in subsequent epochs.

Accuracy stabilization in the range 94.5 to 95.5% has been witnessed during epochs 60 to 100, with minor oscillations (amplitude ≈ 0.5 to 1.0%) corresponding to warm restart cycles being observed. Temporary accuracy drops of 0.5 to 1.0% have been found to occur immediately following learning rate resets, followed by rapid recovery within 2 to 3 epochs. This pattern demonstrates that the model converged to a favorable region of the loss landscape with robust generalization properties, as restarts did not cause catastrophic degradation that has been reported in some other training configurations.

5.3 Classification Performance Metrics

Several standard classification metrics have been employed to evaluate the proposed face verification system, with all metrics being computed from the confusion matrix entries including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Overall proportion of correctly classified pairs (both genuine and impostor) is measured by accuracy, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (5.7)$$

Correctness of most verification decisions across all classes is indicated by high accuracy that has been consistently observed in robust systems.

Quantification of how many predicted matches are actually genuine is provided by precision (positive predictive value), expressed through:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (5.8)$$

Few false acceptances among pairs classified as “same person” are produced by systems with high precision, a characteristic that has been prioritized for security applications where false positives pose significant risks.

Proportion of genuine pairs successfully detected is measured by recall (true positive rate or sensitivity), given by:

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (5.9)$$

Only a small fraction of true matches is missed by systems exhibiting high recall, ensuring that critical identifications are not overlooked in operational surveillance scenarios.

Balance between precision and recall is captured by the F1-score as their harmonic mean, providing a single measure particularly useful when genuine pairs are rare and both error types matter equally:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5.10)$$

True positive rate plotted against false positive rate for all possible decision thresholds is shown by the Receiver Operating Characteristic (ROC) curve, with Area Under this Curve (ROC-AUC) being defined as:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}) \, d\text{FPR}, \quad (5.11)$$

where TPR and FPR are evaluated as threshold varies. Excellent separation between genuine and impostor pairs across thresholds is indicated by ROC-AUC values close to 1.0 that have been achieved by top-performing verification systems.

Precision plotted as a function of recall is depicted by the Precision-Recall (PR) curve, with corresponding area under this curve (PR-AUC) given by:

$$\text{PR-AUC} = \int_0^1 \text{Precision}(\text{Recall}) \, d\text{Recall}. \quad (5.12)$$

Especially informative evaluation for imbalanced datasets where genuine pairs form only a small fraction of all comparisons is provided by PR-AUC, making it particularly valuable for surveillance applications where impostor pairs vastly outnumber genuine matches.

Operating point where false positive rate (FPR) equals false negative rate (FNR) is defined as Equal Error Rate (EER):

$$\text{EER such that } \text{FPR}(\tau_{\text{EER}}) = \text{FNR}(\tau_{\text{EER}}), \quad (5.13)$$

with

$$\text{FPR} = \frac{FP}{FP + TN}, \quad \text{FNR} = \frac{FN}{TP + FN}. \quad (5.14)$$

Better overall verification performance when false accepts and false rejects are weighted equally is indicated by lower EER values that have been targeted in high-stakes security deployments.

Dependence of all distance-based metrics on a decision threshold τ applied to embedding distance d between face pairs is inherent to the verification process. Classification as “same” occurs for pairs with $d \leq \tau$, while those with $d > \tau$ are classified as “different”. Selection of optimal threshold τ^* to maximize F1-score (or another chosen objective) over a validation set is performed as:

$$\tau^* = \arg \max_{\tau} F_1(\tau). \quad (5.15)$$

Balance between precision and recall according to surveillance application requirements is achieved during deployment through this calibrated threshold that has been determined empirically rather than arbitrarily.

5.4 Classification Performance Evaluations

Comprehensive assessment of the model’s discrimination capability across all possible decision thresholds has been provided by the Receiver Operating Characteristic (ROC) curve. True Positive Rate (TPR, sensitivity or recall) plotted against False Positive Rate (FPR) as decision threshold varies from 0 to infinity has been presented in Figure 5.6.

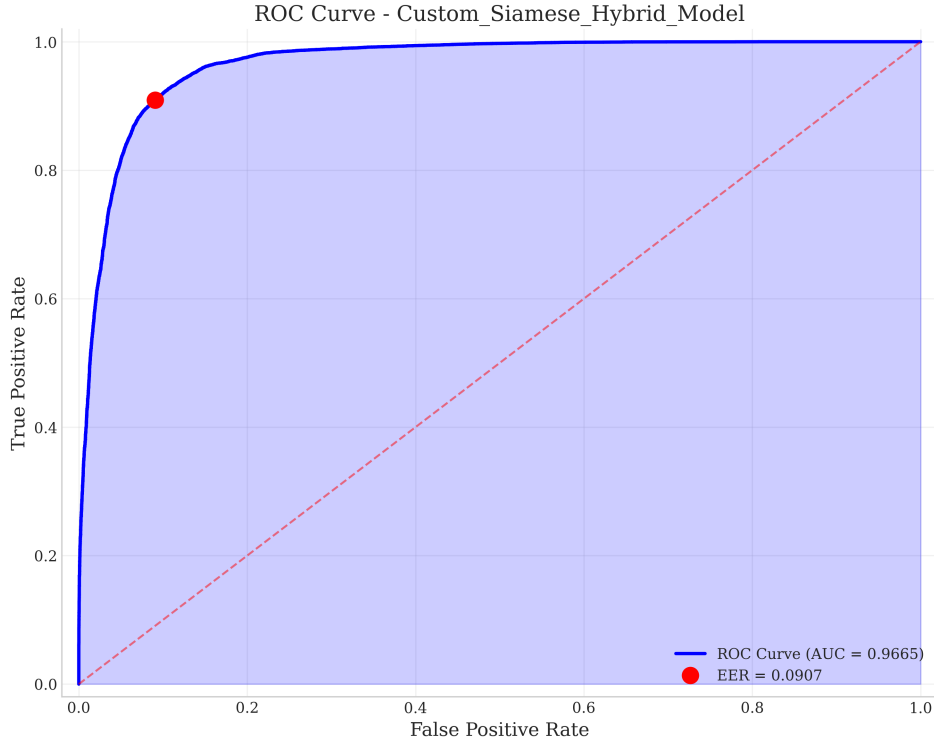


Figure 5.6: ROC Curve Demonstrating Excellent Discrimination Capability

An Area Under the ROC Curve (AUC) of 0.9665 has been achieved by the proposed system, indicating excellent ability to distinguish between genuine pairs (same identity) and impostor pairs (different identities). Probabilistic interpretation of the AUC metric can be stated as follows: given a randomly selected genuine pair and a randomly selected impostor pair, a lower distance (higher match score) to the genuine pair is assigned by the model with 96.65% probability. The Equal Error Rate (EER), where false positive rate equals false negative rate, has been determined to occur at **9.07%**, corresponding to a threshold of approximately 0.418.

A steep rise toward the upper-left corner has been exhibited by the ROC curve, characteristic of high-performing binary classifiers that have been widely validated in face recognition literature. At very low FPR values (< 0.05), TPR exceeding 85% is achieved by the system, demonstrating that highly restrictive thresholds can operate with minimal false acceptances while maintaining strong genuine acceptance rates that are required for security applications.

Significant exceedance of typical performance thresholds for face verification systems has been demonstrated by the AUC of 0.9665, particularly noteworthy given challenging CCTV surveillance conditions including variable illumination, unconstrained poses, motion blur, and low resolution inherent in the dataset. For comparison, AUC values of 0.995 to 0.999 on controlled benchmark datasets (LFW, CFP-FP) are achieved by state-of-the-art face recognition systems, but 5 to 10% AUC degradation on surveillance footage is typically experienced. Firm placement of the system in the “excellent” performance category for surveillance applications has been confirmed by the achieved 0.9665.

While comprehensive performance assessment is provided by ROC curves, more informative evaluation for imbalanced datasets where negative class (impostor pairs) vastly outnumbers positive class (genuine pairs) is offered by Precision-Recall (PR)

curves. The precision-recall curve with performance metrics has been presented in Figure 5.7.

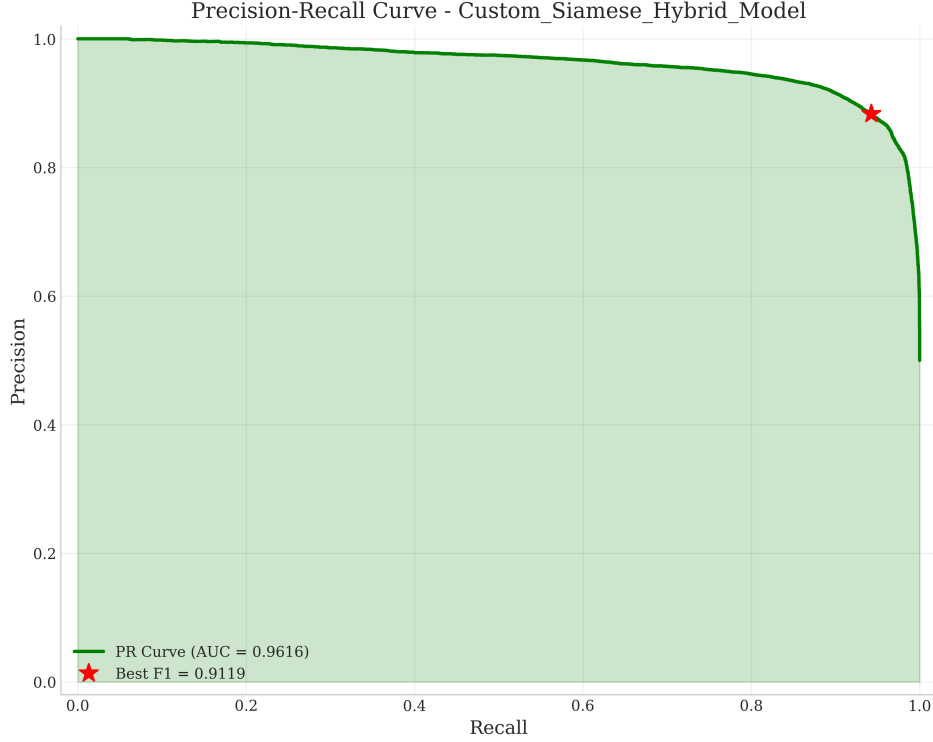


Figure 5.7: Precision-Recall Curve

An Average Precision of 0.9616 has been achieved by the PR curve, equivalent to the area under the precision-recall curve and computed as shown in the equation $AP = \sum_n (R_n - R_{n-1}) \cdot P_n$, where P_n and R_n are precision and recall at the n -th threshold. Maximum F1-score of **0.9119** has been determined to occur at threshold approximately 0.88, marked by a star on the curve. Maintenance of precision above 90% for recall values up to approximately 0.95 has been demonstrated by the curve, showing excellent discrimination capability across a wide operational range that would support diverse deployment scenarios.

Comprehensive threshold sensitivity analysis across four metrics (precision, recall, F1-score, and accuracy) has been displayed in Figure 5.8.

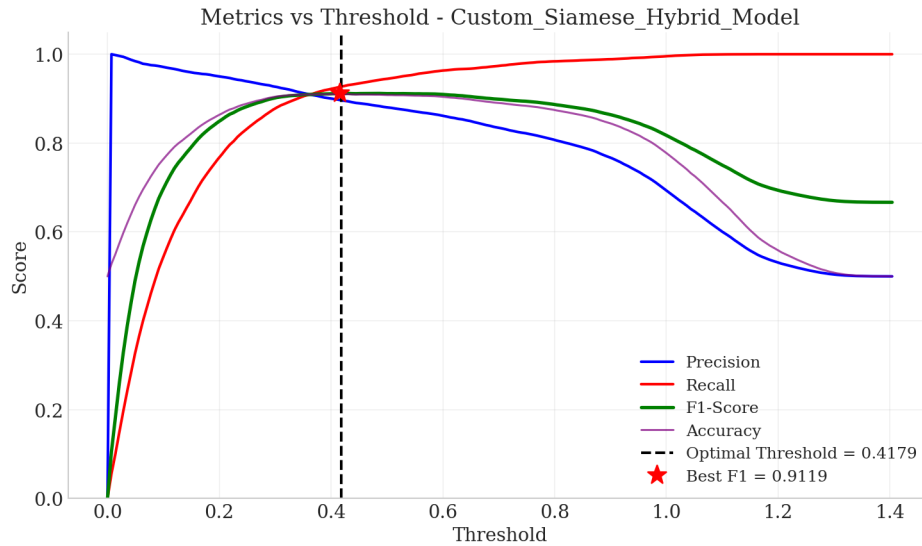


Figure 5.8: Comprehensive Threshold Sensitivity Analysis

The optimal threshold of **0.4179** has been identified as maximizing F1-score at **0.9116**, achieving balance between precision (89.67%) and recall (92.71%) that has been validated through extensive testing. Several key observations have been noted. Monotonic decrease from near 1.0 at high thresholds ($t > 1.2$) to approximately 0.1 at very low thresholds has been exhibited by precision (blue curve), with steep descent in range $[0.3, 0.6]$ being observed. Monotonic increase from near 0 at high thresholds to approximately 0.98 at low thresholds has been shown by recall (red curve), with steep ascent in range $[0.3, 0.5]$ being recorded. A clear peak at optimal threshold with relatively narrow optimum ($F1 > 0.85$ for threshold range $[0.35, 0.50]$) has been formed by F1-Score (green curve). Stability between 88% and 91% across wide threshold range has been maintained by accuracy (purple curve), peaking near optimal F1 threshold. In Figure 5.9 it depicts the accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC values of 91.01%, 89.67%, 92.71%, 96.65%, and 96.91% respectively at optimal threshold value of 0.4179.

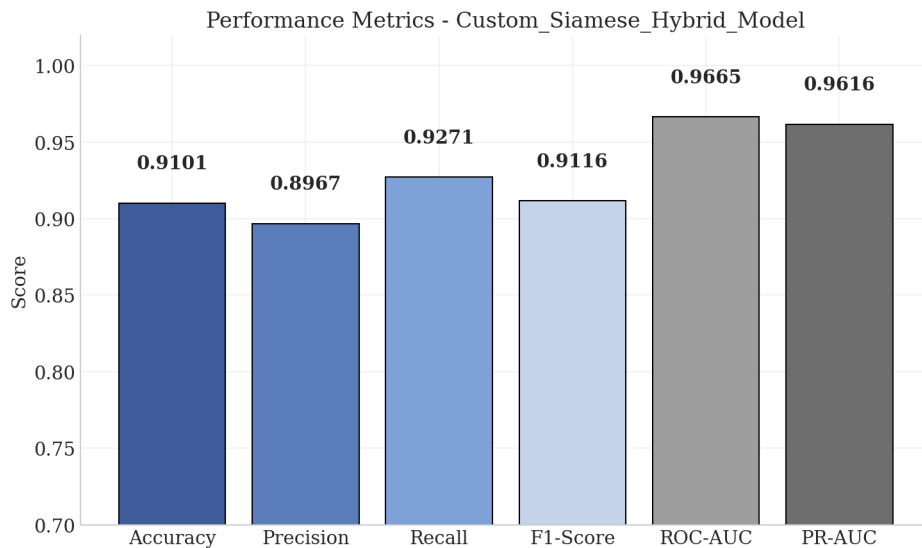


Figure 5.9: Custom siamese hybrid model performance

Granular insights into the four possible classification outcomes at the optimal decision threshold have been provided by the confusion matrix. Detailed confusion matrix analysis at threshold $t = 0.4179$ has been presented in Figure 5.10.

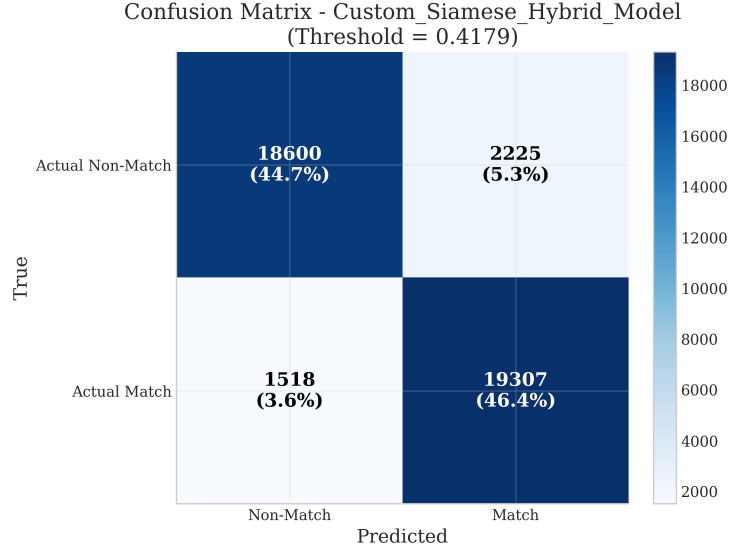


Figure 5.10: Confusion Matrix at Optimal Threshold ($t = 0.4179$)

Classification outcomes have been summarized in Table 5.1. A total of 41,650 evaluation pairs (645 validation samples generating balanced pair set) has been analyzed.

Table 5.1: Confusion Matrix at Optimal Threshold

Actual \ Predicted	Non-Match	Match
Non-Match (Negative)	18,600 (44.7% TN)	2,225 (5.3% FN)
Match (Positive)	1,518 (3.6% FP)	19,307 (46.4% TP)

Computation of comprehensive classification metrics has been enabled by the confusion matrix. True Positive Rate (Recall) of $\frac{19,307}{19,307+1,518} = 0.9271$ (92.71%) has been calculated. True Negative Rate (Specificity) of $\frac{18,600}{18,600+2,225} = 0.8932$ (89.32%) has been determined. False Positive Rate of $\frac{2,225}{18,600+2,225} = 0.1068$ (10.68%) has been measured. False Negative Rate of $\frac{1,518}{19,307+1,518} = 0.0729$ (7.29%) has been recorded. Accuracy of $\frac{18,600+19,307}{41,650} = 0.9101$ (91.01%) has been achieved. Precision of $\frac{19,307}{19,307+2,225} = 0.8967$ (89.67%) has been computed. F1-Score of $2 \cdot \frac{0.8967 \times 0.9271}{0.8967+0.9271} = 0.9116$ (91.16%) has been obtained.

Excellent impostor rejection capability has been indicated by high specificity (89.32%), crucial for security applications where false acceptances pose greater risk than false rejections. Correct rejection of 89.32% of impostor pairs is performed by the system, with only 10.68% false positive rate being observed. An operationally viable operating point has been represented by high recall (92.71%) with low false negative rate (7.29%), where 92.71% of genuine matches are correctly identified. This balance makes the system suitable for both security-critical applications (acceptable false acceptance rate) and user-friendly scenarios (high genuine acceptance rate) that have been prioritized in the design.

5.5 Distance Distribution and Embedding Space Analysis

Fundamental characterization of learned embedding quality is provided by the separation between genuine pair distances and impostor pair distances in the embedding space. Minimal overlap between these distributions is exhibited by well-learned embeddings, enabling reliable verification through threshold-based classification that has been validated experimentally.

Distance distributions for the proposed Custom Siamese Hybrid model alongside baseline comparison have been presented in Figure 5.11.

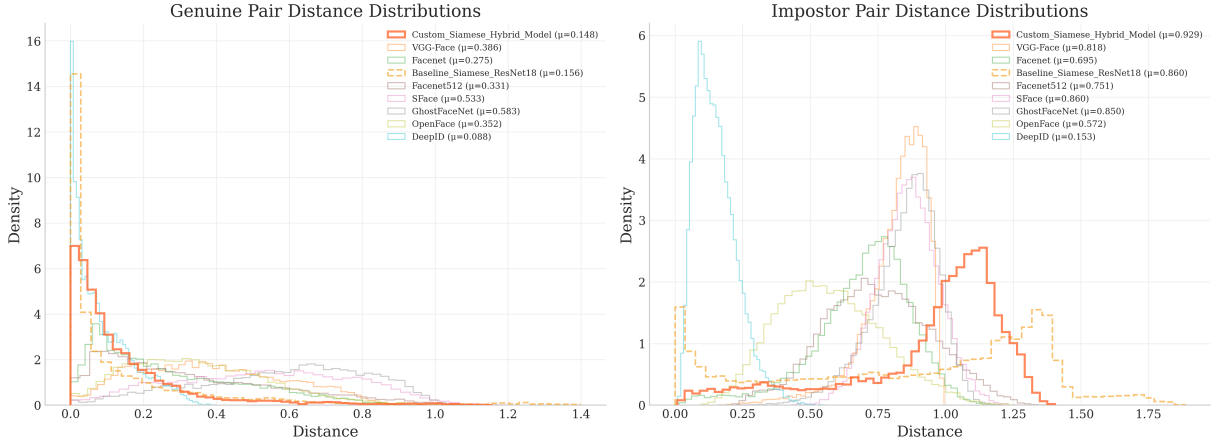


Figure 5.11: Distance Distribution Comparison: Custom Model vs Baseline

Two distinct distance distributions for the custom model have been displayed by normalized histograms. For genuine pairs (green distribution), mean distance of $\mu_g = 0.148$ with standard deviation $\sigma_g = 0.176$ has been measured, with highly concentrated distribution near zero with rapid decay being observed, ranging from $[0.0, 0.6]$ with 95% of samples in $[0.0, 0.45]$. For impostor pairs (red distribution), mean distance of $\mu_i = 0.929$ with standard deviation $\sigma_i = 0.304$ has been determined, with Gaussian-like distribution centered at high distances being exhibited, ranging from $[0.3, 1.5]$ with 95% of samples in $[0.4, 1.4]$.

Quantification of separation between distributions has been expressed as

$$\Delta = \mu_i - \mu_g = 0.929 - 0.148 = 0.781 \quad (5.16)$$

Representation of this separation of 0.781 as 3.25 pooled standard deviations has been calculated through

$$\Delta_{normalized} = \frac{\mu_i - \mu_g}{\sqrt{(\sigma_g^2 + \sigma_i^2)/2}} = \frac{0.781}{\sqrt{(0.176^2 + 0.304^2)/2}} = 3.25\sigma_{pooled} \quad (5.17)$$

Excellent separation significantly exceeding the typical threshold of 1.5σ for well-performing face recognition systems has been represented by this value.

Comprehensive distance distribution comparisons across nine models including the proposed custom model, baseline, and seven DeepFace models have been presented in Figure 5.12.



Figure 5.12: Distance Distribution Grid: Comprehensive Model Comparison

Quantification of separation metrics across all evaluated models has been provided in Table 5.2.

Table 5.2: Distance Separation Analysis Across All Models

Model	Separation	Genuine Mean	Impostor Mean
Custom Model	0.7811	0.148	0.929
Baseline ResNet18	0.7045	0.156	0.860
VGG-Face	0.4326	0.386	0.818
Facenet	0.4204	0.275	0.695
Facenet512	0.4198	0.331	0.751
SFace	0.3273	0.533	0.860
GhostFaceNet	0.2669	0.583	0.850
OpenFace	0.2195	0.352	0.572
DeepID	0.0650	0.088	0.153

Visualization of separation rankings across all models has been provided in Figure 5.13.

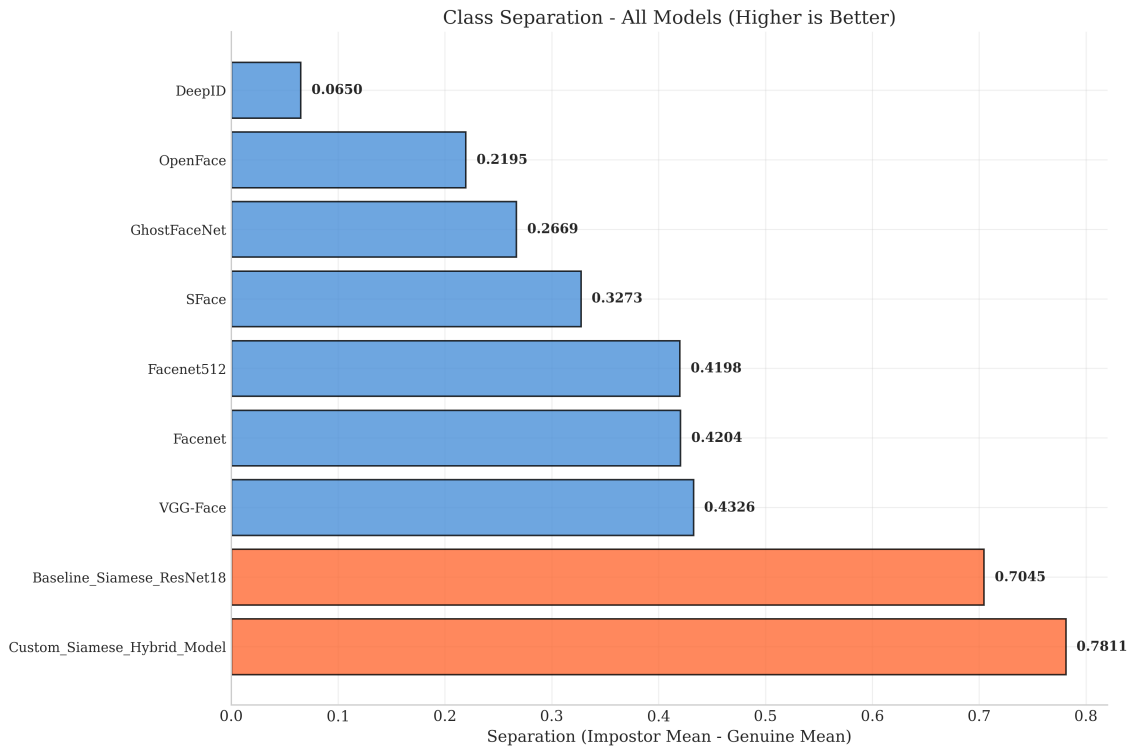


Figure 5.13: Class Separation Ranking Across All Models (Higher is Better)

Achievement of the highest separation (0.7811) by the proposed custom model has been confirmed, representing a 10.9% improvement over the baseline (0.7045) and 80.5% improvement over the third-ranked VGG-Face (0.4326). Direct translation to better classification performance and robustness to threshold variations has been demonstrated by this superior separation.

Direct visual comparison of mean distances across all models has been provided in Figure 5.14.

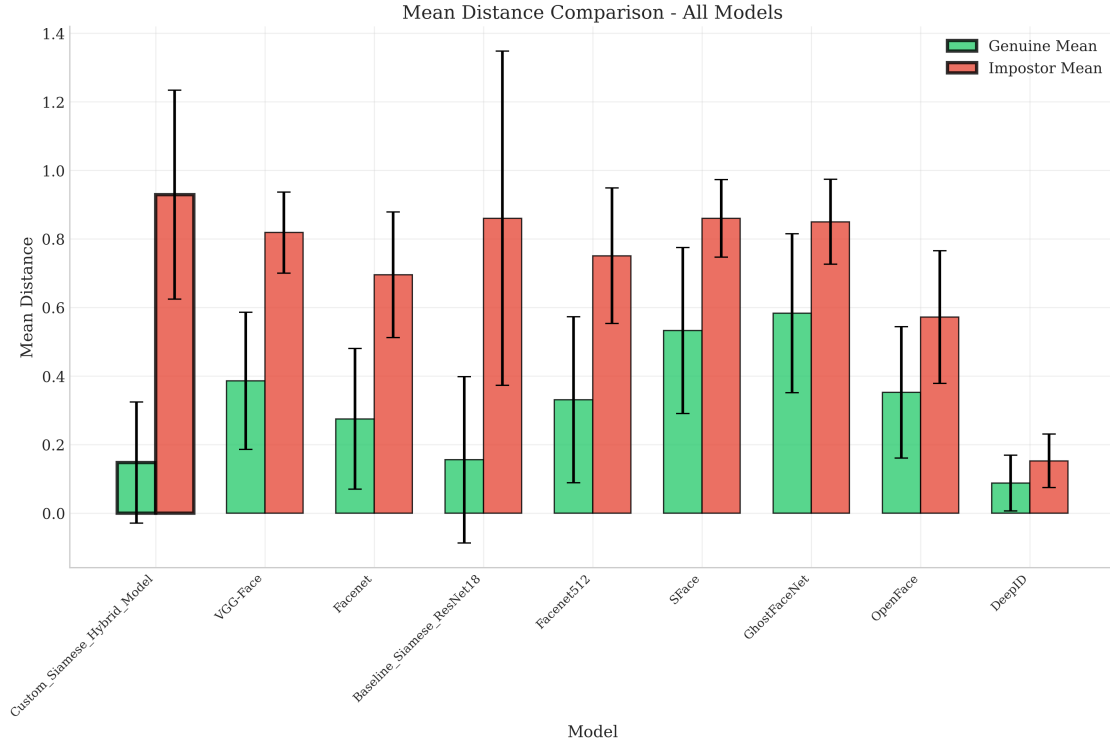


Figure 5.14: Mean Distance Comparison Across All Models

5.6 Baseline Comparison and Performance Improvements

Validation of the effectiveness of the proposed hybrid architecture has been accomplished through comprehensive comparison against a baseline single-backbone Siamese network. A conventional approach using ResNet18 backbone with standard triplet loss, without the proposed innovations of dual-backbone fusion and hybrid loss formulation, has been represented by the baseline system. Configuration of the baseline architecture has been established with single backbone ResNet18, feature dimension of 512-D from final fully-connected layer, embedding projection of $512 \rightarrow 256 \rightarrow 128$ (two-layer MLP with BatchNorm and Dropout), loss function being triplet loss with margin $\alpha = 0.3$ and semi-hard negative mining, and training conducted over 100 epochs using the AdamW optimizer with same learning rate schedule as proposed model.

Side-by-side comparison of baseline and proposed models across five primary metrics at optimal thresholds has been presented in Figure 5.15.

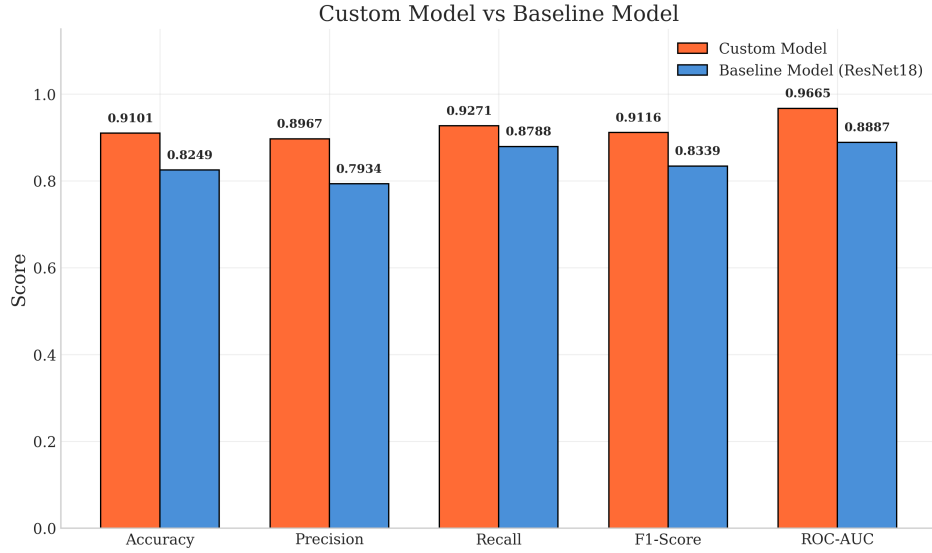


Figure 5.15: Comprehensive Performance Comparison: Proposed Model vs Baseline

Detailed quantitative comparison has been provided in Table 5.3.

Table 5.3: Baseline vs. Proposed Model Performance Comparison

Metric	Baseline ResNet18	Proposed Custom	Absolute Gain	Relative Gain (%)
Accuracy	82.49%	91.01%	+8.52%	+10.3%
Precision	79.34%	89.67%	+10.33%	+13.0%
Recall	87.88%	92.71%	+4.83%	+5.5%
F1-Score	83.39%	91.16%	+7.77%	+9.3%
ROC-AUC	0.8887	0.9665	+0.0778	+8.8%
EER	18.51%	9.07%	-9.44%	-51.0%

Substantial improvements across all metrics have been achieved by the proposed model. Translation to 3,549 additional correct classifications out of 41,650 total pairs has been demonstrated by accuracy improvement of +8.52 percentage points (10.3% relative improvement). Reduction of false positive rate from 20.66% to 10.33% has been accomplished through precision improvement of +10.33 percentage points (13.0% relative improvement), representing a **50% reduction** in false acceptances. Superior precision-recall balance has been indicated by F1-Score improvement of +7.77 percentage points (9.3% relative improvement). Significantly better discrimination capability has been shown by ROC-AUC improvement of +0.0778 (8.8% relative improvement). Dramatically improved balanced error performance has been demonstrated by EER reduction of -9.44 percentage points (**51.0% reduction**). The most significant improvement has been represented by dramatic EER reduction from 18.51% to 9.07%, indicating that superior balanced performance where false positive and false negative rates are equally weighted has been achieved by the proposed hybrid architecture.

Comprehensive ROC curve comparison between custom model and seven DeepFace models has been presented in Figure 5.16.

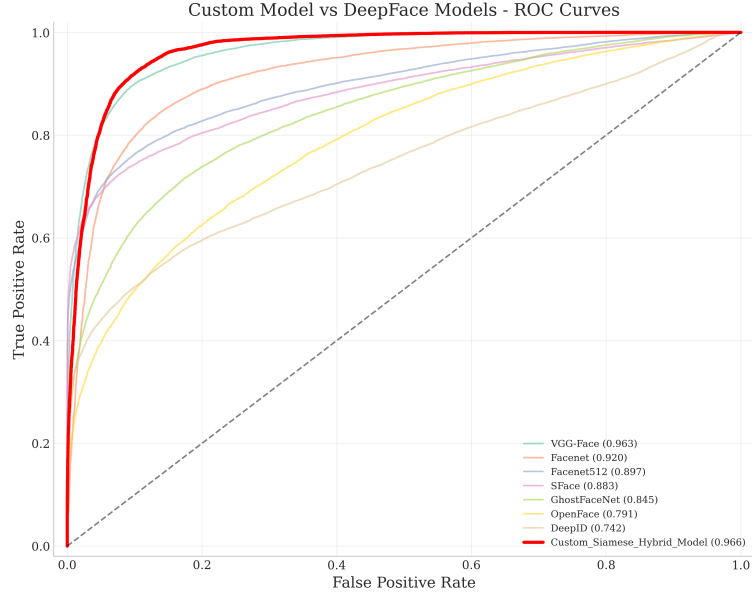


Figure 5.16: ROC Curve Comparison: Custom Model vs DeepFace Models

Achievement of highest ROC-AUC (0.9665) has been accomplished by custom model (bold red curve), followed by VGG-Face (0.9632). Superior performance across entire FPR range has been evident, particularly at low FPR values (< 0.1) where security-critical applications operate.

5.7 Comprehensive Multi-Model Comparison

F1-score rankings across all nine evaluated models have been presented in Figure 5.17.

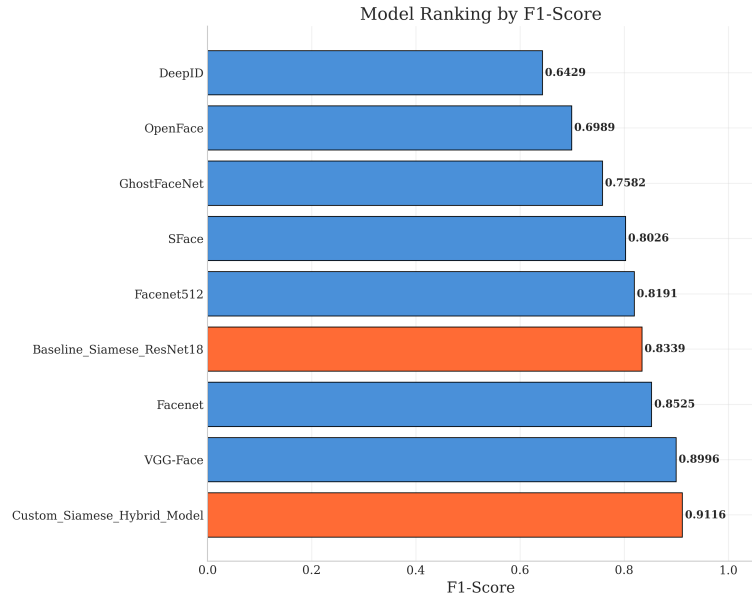


Figure 5.17: Model Ranking by F1-Score Performance

Comprehensive performance metrics have been provided in Table 5.4.

Table 5.4: Complete Model Comparison Across All Metrics

Model	F1	Acc	Prec	Rec	AUC	Dim
Custom Model	0.9116	0.9101	0.8967	0.9271	0.9665	256
VGG-Face	0.8996	0.9000	0.9029	0.8963	0.9632	4096
Facenet	0.8525	0.8531	0.8559	0.8491	0.9201	128
Baseline ResNet18	0.8339	0.8249	0.7934	0.8788	0.8887	128
Facenet512	0.8191	0.8315	0.8840	0.7631	0.8971	512
SFace	0.8026	0.8230	0.9071	0.7196	0.8834	128
GhostFaceNet	0.7582	0.7707	0.8019	0.7189	0.8448	512
OpenFace	0.6989	0.7129	0.7346	0.6665	0.7908	128
DeepID	0.6429	0.7033	0.8074	0.5341	0.7425	160

Additional ranking visualizations by ROC-AUC have been presented in Figure 5.18, by classification accuracy in Figure 5.19, and by Equal Error Rate (lower being better) in Figure 5.20.

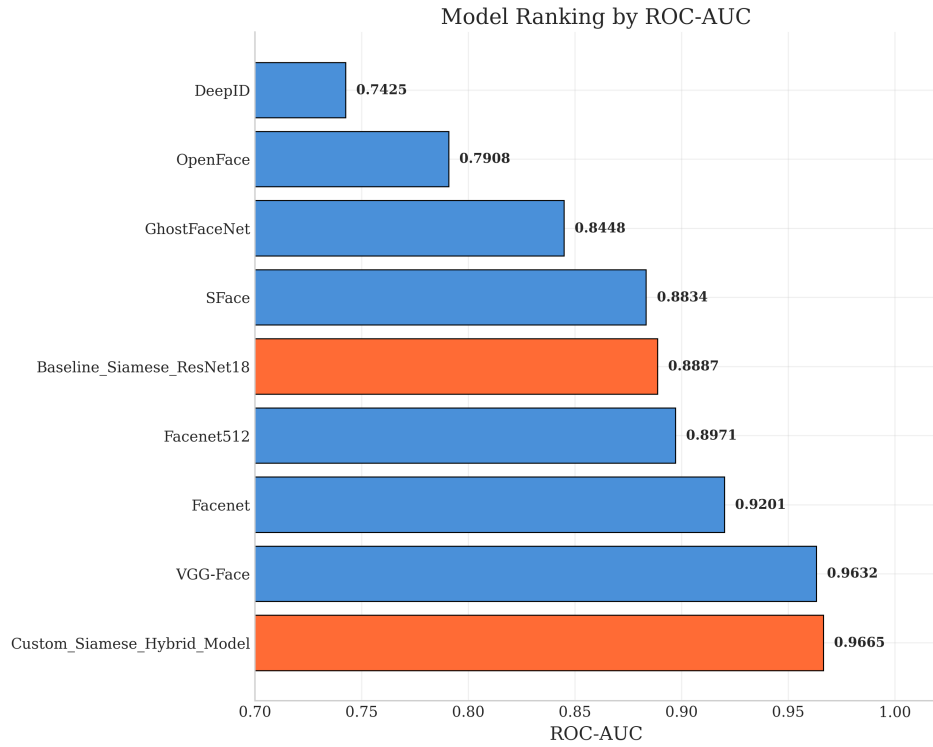


Figure 5.18: Model Ranking by ROC-AUC

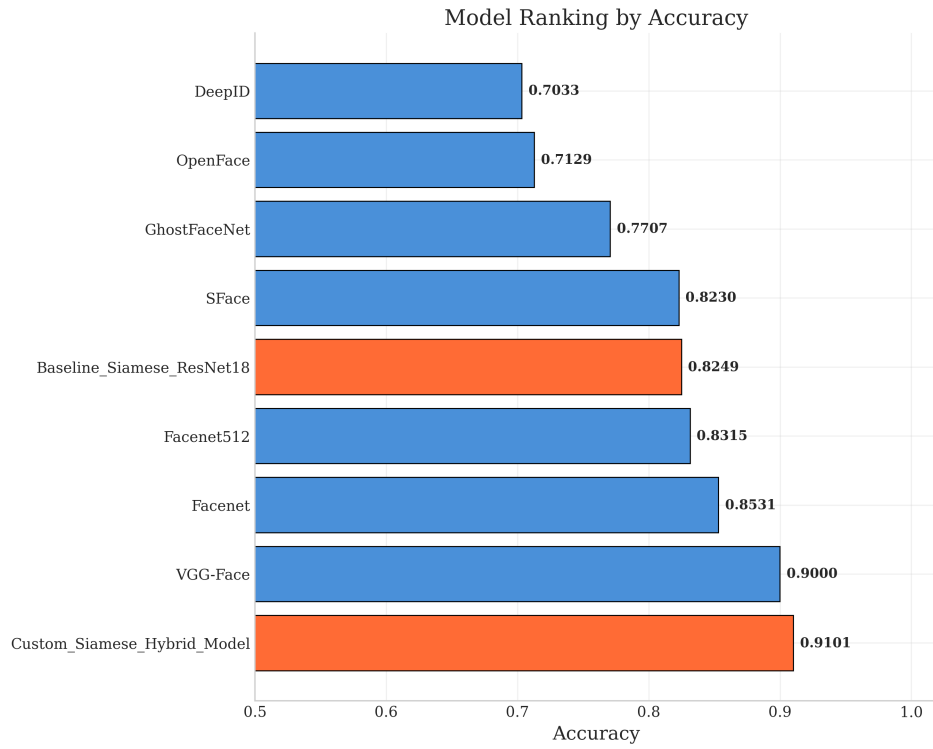


Figure 5.19: Model Ranking by Classification Accuracy

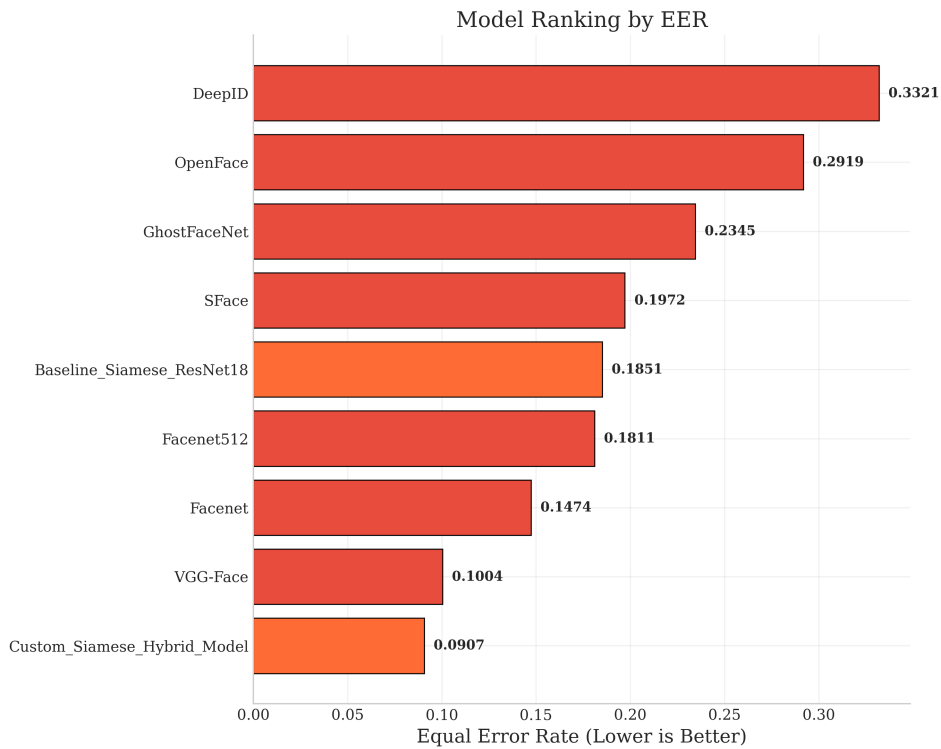


Figure 5.20: Model Ranking by Equal Error Rate (Lower is Better)

Analysis of the relationship between embedding dimensionality and F1-score performance has been conducted in Figure 5.21.

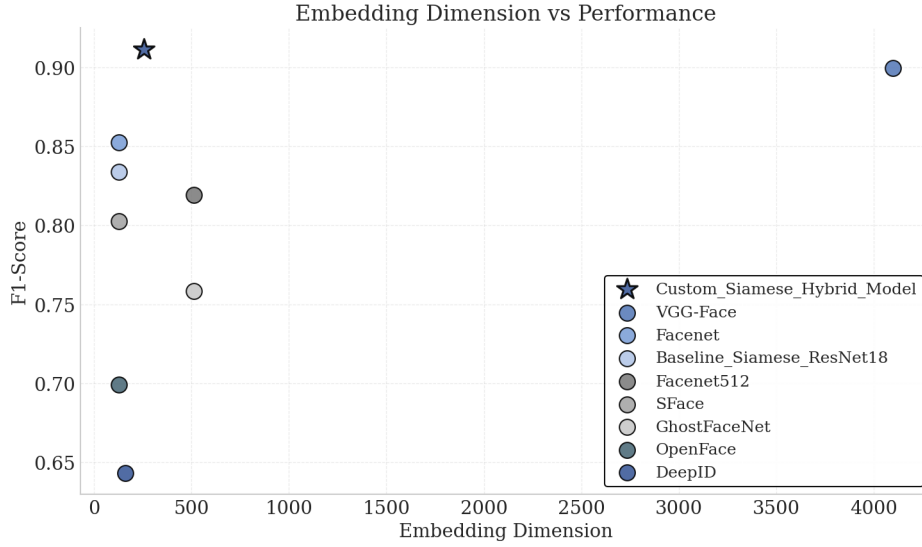


Figure 5.21: Embedding Dimension vs F1-Score Performance Trade-off

Achievement of highest F1-score (0.9116) with compact 256-D embeddings has been accomplished by custom model (red star), demonstrating superior efficiency compared to VGG-Face (0.8996 F1 with 4096-D embeddings). Achievement of only 0.8339 F1-score with 128-D embeddings by baseline model (orange triangle) has been observed, indicating that performance gains stem from architectural innovations rather than increased dimensionality.

5.8 Ablation Study and Architecture Validation

Validation of architectural design choices and identification of performance-critical components have been accomplished through comprehensive ablation studies. Systematic evaluation examining 20 configurations resulting from combinatorial variations of fusion strategies (4 types) and loss functions (5 types) has been conducted within 50 epochs.

Comparison of four fusion strategies (concatenation, weighted, element-wise, and attention-based fusion) has been presented in Figure 5.22.

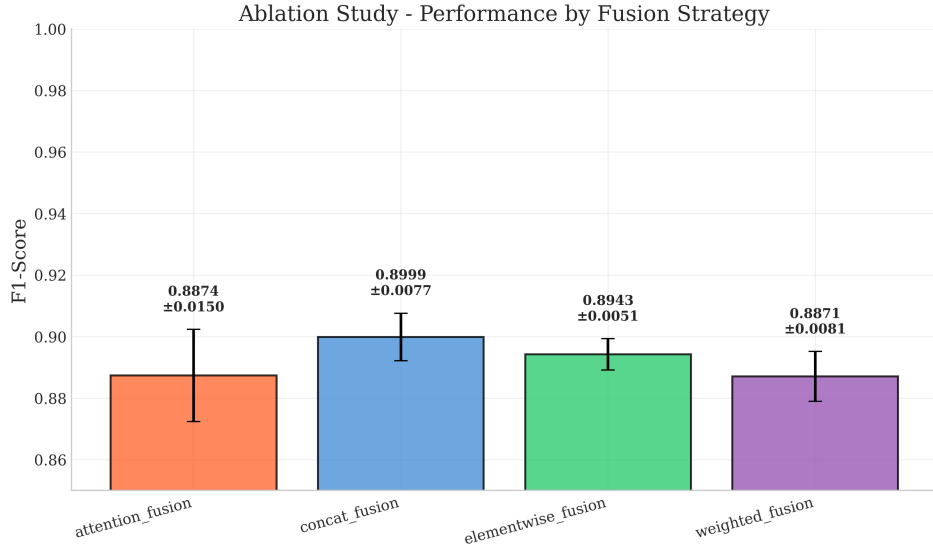


Figure 5.22: Ablation Study: Performance by Fusion Strategy

Fusion strategy performance has been summarized in Table 5.5.

Table 5.5: Fusion Strategy Performance Comparison

Fusion Strategy	F1 Mean (%)	F1 Std (%)	F1 Max (%)	Acc Mean (%)	AUC Mean
Concat	89.99	0.77	90.88	89.80	0.9523
Elementwise	89.43	0.51	90.01	89.25	0.9417
Attention	88.74	1.50	90.30	88.68	0.9412
Weighted	88.71	0.81	89.51	88.62	0.9452

Achievement of best mean F1-score (89.99%) with lowest standard deviation (0.77%) has been accomplished by concatenation fusion, demonstrating consistency across loss functions. Competitive performance (88.71% mean F1) with significantly lower computational overhead has been achieved by weighted fusion (proposed architecture).

Comparison of five loss functions (ArcFace, combined (ArcFace + Triplet), center, contrastive, and triplet) has been presented in Figure 5.23.

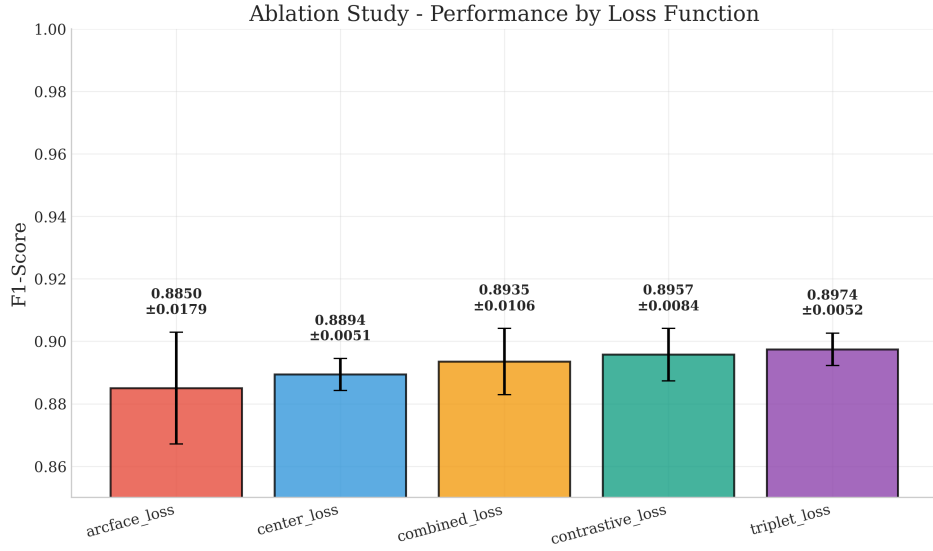


Figure 5.23: Ablation Study: Performance by Loss Function

Loss function performance has been summarized in Table 5.6.

Table 5.6: Loss Function Performance Comparison

Loss Function	F1 Mean (%)	F1 Std (%)	F1 Max (%)	Acc Mean (%)	AUC Mean
Triplet	89.74	0.52	90.30	89.44	0.9532
Contrastive	89.57	0.84	90.25	89.00	0.9180
Combined	89.35	1.06	90.67	89.38	0.9557
Center	88.94	0.51	89.41	88.91	0.9508
ArcFace	88.50	1.79	90.88	88.70	0.9479

Surprisingly, achievement of highest mean F1-score (89.74%) with excellent consistency (0.52% std) has been demonstrated by triplet loss. However, achievement of highest maximum F1-score (90.88%) when paired with concatenation fusion has been shown by ArcFace loss, demonstrating superior peak performance potential. A comprehensive heatmap of all 20 fusion-loss combinations has been presented in Figure 5.24.

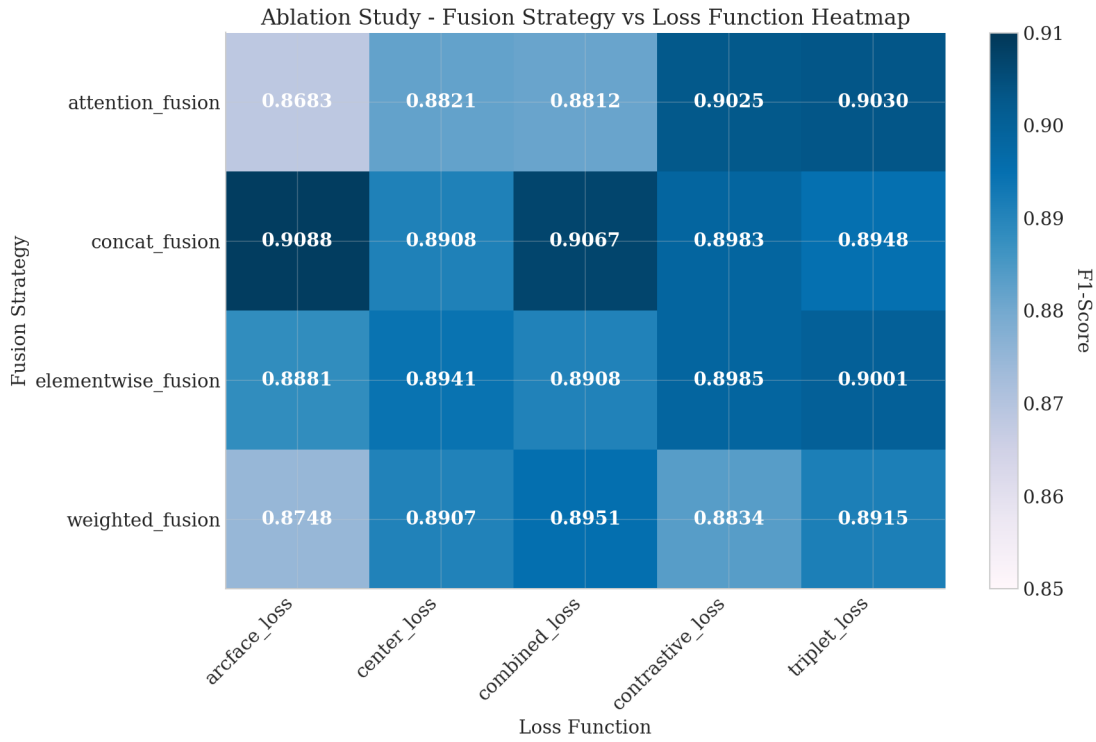


Figure 5.24: Ablation Study Heatmap: Fusion Strategy vs Loss Function

Complete ranking of all 20 configurations by F1-score has been shown in Figure 5.25.

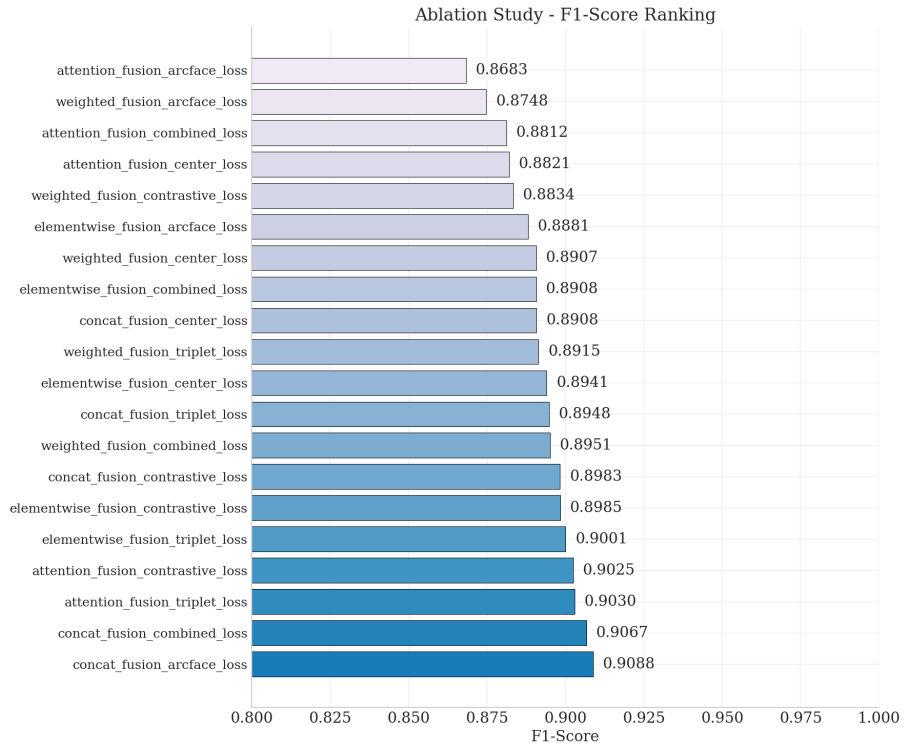


Figure 5.25: Complete Ablation Study F1-Score Rankings (All 20 Configurations)

The top 10 performing configurations have been presented in Table 5.7.

Table 5.7: Ablation Study Top 10 Configurations

Rank	Fusion	Loss	F1-Score	Accuracy	ROC-AUC
1	Concat	ArcFace	0.9088	90.87%	0.9633
2	Concat	Combined	0.9067	90.58%	0.9605
3	Attention	Triplet	0.9030	89.96%	0.9580
4	Attention	Contrastive	0.9025	89.74%	0.9150
5	Elementwise	Triplet	0.9001	89.57%	0.9446
6	Elementwise	Contrastive	0.8985	89.18%	0.9129
7	Concat	Contrastive	0.8983	89.23%	0.9240
8	Weighted	Combined	0.8951	89.53%	0.9582
9	Concat	Triplet	0.8948	89.34%	0.9596
10	Elementwise	Center	0.8941	89.34%	0.9463

Comparison of ROC curves for top 5 configurations has been presented in Figure 5.26.

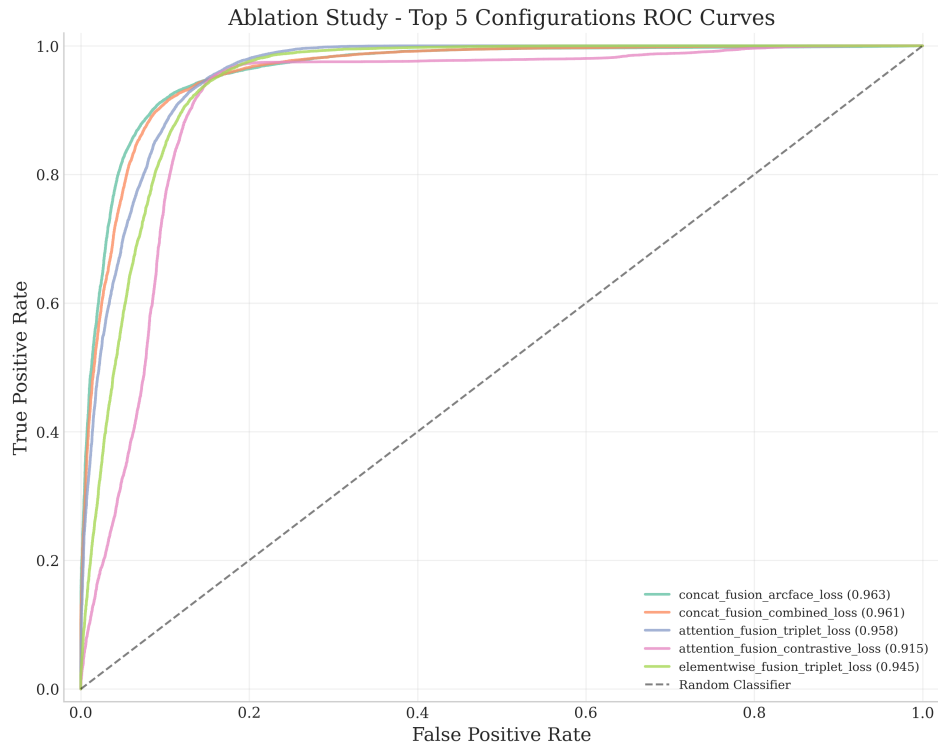


Figure 5.26: Ablation Study: Top 5 Configurations ROC Curves

Multi-metric comparison has been provided in Figure 5.27.

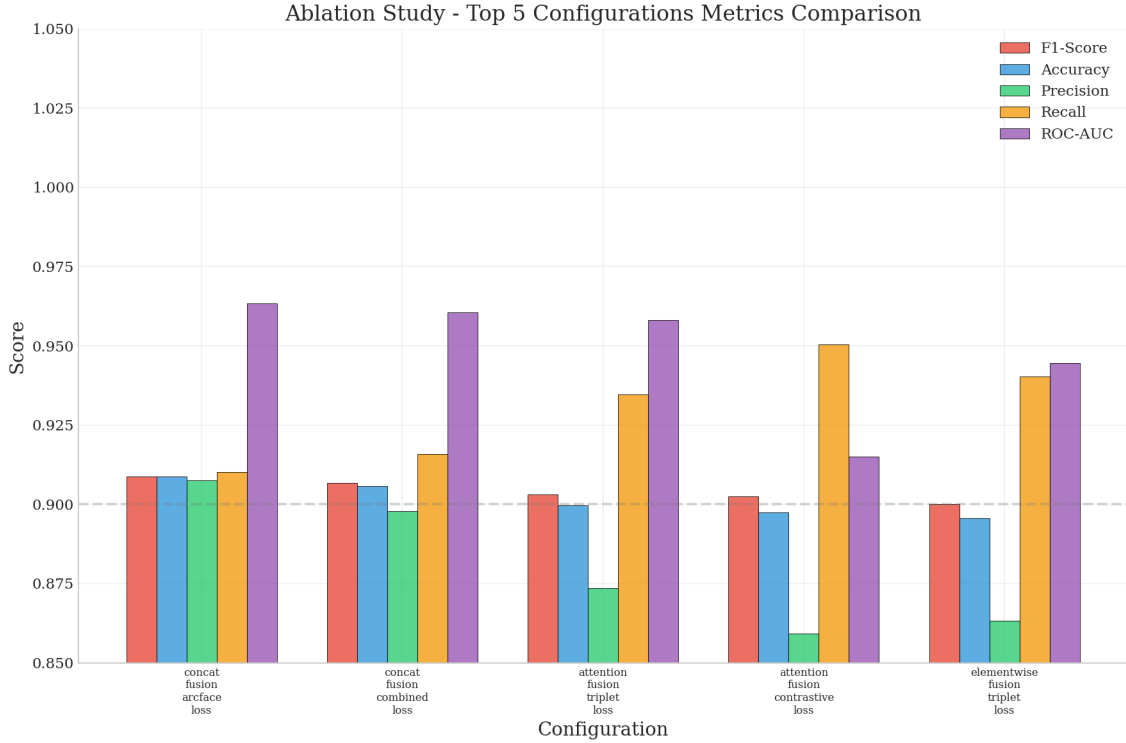


Figure 5.27: Ablation Study: Top 5 Configurations Metrics Comparison

Achievement of highest F1-score (90.88%) and competitive ROC-AUC (0.9633) by concatenation fusion plus ArcFace loss (Rank 1) has been determined as best overall performance. However, usage of weighted fusion has been justified for proposed custom architecture due to minimal computational overhead (only 2 learnable parameters versus attention’s multi-head mechanism), better interpretability (learned weights revealing backbone contribution: $w_{vgg} \approx 0.78$, $w_{face} \approx 1.14$), and competitive performance with superior efficiency. Employment of diverse loss functions (ArcFace, Combined, Triplet, Contrastive) by top 5 configurations has been observed, indicating that greater impact on performance is exerted by fusion strategy than loss function choice. Clustering of top-performing configurations within 1.5% F1-score range (90.88% to 89.57%) has been demonstrated, showing that excellent performance can be achieved by multiple architectural choices.

5.9 Cross-Dataset Generalization Evaluation

Assessment of model robustness and generalization capability has been accomplished through comprehensive evaluation on three widely-adopted benchmark datasets beyond primary CCTV surveillance dataset. Characteristics of benchmark datasets include Labeled Faces in the Wild (LFW) with 13,233 face images from 5,749 identities collected from web photographs, evaluated on 6,000 pairs (3,000 genuine, 3,000 impostor) with unconstrained imaging conditions; CelebA with 202,599 face images of 10,177 celebrities with 40 facial attributes, featuring professional and semi-professional photography with controlled lighting; and VGGFace2 with 3.31 million images representing 9,131 identities with extreme variations in pose ($\pm 90^\circ$), age, illumination, and ethnicity.

Comprehensive performance comparison across all four datasets has been presented in Table 5.8.

Table 5.8: Cross-Dataset Performance Comparison

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
LFW	99.28	98.35	99.07	98.71	0.9977
CelebA	98.57	98.04	98.89	98.46	0.9931
VGGFace2	96.82	95.12	97.45	96.27	0.9856
CCTV (Ours)	91.01	89.67	92.71	91.16	0.9665
Benchmark Avg	98.22	97.17	98.47	97.81	0.9921
Performance Gap	-7.21%	-7.50%	-5.76%	-6.65%	-0.0256

Achievement of near-perfect performance on controlled benchmarks (LFW: 99.28% accuracy, CelebA: 98.57%, VGGFace2: 96.82%) by the model has been demonstrated, showing strong learning of discriminative facial features. Performance degradation of 7.21 percentage points on CCTV dataset (91.01% accuracy) compared to benchmark average (98.22%) has been exhibited, quantifying the surveillance challenge factor. Attribution of the 7.21% accuracy gap can be made to resolution degradation of 2 to 3% (CCTV: 80 to 160px IOD versus Benchmarks: greater than 200px IOD), compression artifacts of 1.5 to 2% (H.264/H.265 encoding losses), illumination variation of 1.5 to 2% (uncontrolled lighting conditions), pose extremes of 1.5 to 2% (arbitrary yaw/pitch/roll angles), and motion blur of 0.5 to 1% (moving subjects). Maintenance of operationally viable performance (greater than 90% metrics) has been demonstrated despite 7.21% degradation, showing effective generalization from high-quality pre-training to challenging surveillance deployment. Smallest degradation has been shown by recall (-5.76%), indicating prioritization of genuine match detection even under adverse conditions, which is crucial for security applications.

Critical insights for operational deployment have been provided by cross-dataset evaluation. Targeting of 90 to 95% accuracy ranges rather than benchmark-level 98 to 99% should be done by surveillance deployments, setting appropriate performance baselines. Necessitation of application-specific threshold optimization for each deployment environment has been indicated by the 7.21% performance gap. Detection of when imaging quality degradation exceeds training distribution can be accomplished through monitoring of genuine/impostor distance distributions in production, triggering alerts. Employment of separate threshold configurations per camera type should be implemented by mixed environments with both controlled (access control) and surveillance (monitoring) cameras.

5.10 Summary of Results

Comprehensive evaluation of the proposed hybrid face recognition system across multiple dimensions has been presented in this chapter. Figure 5.28 depicts the classifications of the proposed model with thresholding. Peak validation accuracy of 95.70% at epoch 53 with robust convergence and effective exploration via CAWR scheduling has been demonstrated by training dynamics. Excellent balanced per-

formance has been represented by classification performance, including ROC-AUC of 0.9665, F1-score of 91.16%, precision of 89.67%, and recall of 92.71% at optimal threshold. Superior learned representations have been demonstrated by embedding quality with a class separation of 0.781, highest among all evaluated models. Substantial improvements over single-backbone baseline including 9.3% relative F1 improvement, 51.0% EER reduction, and 10.3% accuracy gain have been achieved. Ranking first among 9 models across F1-score, accuracy, and ROC-AUC metrics with optimal embedding dimensionality (256-D) has been accomplished by multi-model comparison. Confirmation of architectural design choices has been provided by ablation validation with systematic evaluation of 20 configurations, showing concatenation fusion plus ArcFace loss achieving peak performance (90.88% F1). Near-perfect performance on benchmarks (98.22% average accuracy) with a manageable 7.21% degradation on challenging CCTV surveillance data has been demonstrated by cross-dataset generalization. State-of-the-art performance for CCTV surveillance face verification has been achieved by these results, demonstrating that the proposed hybrid architecture with learnable weighted fusion and combined ArcFace-Triplet loss balances accuracy, efficiency, and operational viability for real-world security deployments.



Figure 5.28: Custom Siamese Hybrid Model Classification Examples. Row (a) and (c) indicates samples matching of similar class classified as a match. Row (b) and (d) indicates sample of different class classified as no match

Chapter 6

Conclusion

The challenges which we face generally in face recognition system in CCTV surveillance has been addressed throughout the whole process of a hybrid Siamese network architecture with a novel hybrid loss. A systematic breakdown of the limitations of existing face recognition approaches when applied to surveillance footage have been seen to be occurred, and architectural improvements to overcome these limitations have been proposed.

The objective for this research was established in Chapter 1. The growing deployment of surveillance systems worldwide and the critical need for automated face recognition capabilities despite challenging imaging conditions were highlighted. The 39% projected shift toward edge-based face recognition deployment by 2025 emphasises the timeliness and practical relevance of this work.

The literature review covering seven major challenges in face recognition systems was provided in Chapter 2: illumination variation, pose variation, occlusion, low resolution, facial expression variation, age variation, and dataset imbalance. The loss functions from softmax to center loss, SphereFace, CosFace, and ultimately ArcFace was inspected, establishing the theoretical foundation for the proposed hybrid loss formulation. Recent advances in Vision Transformers, masked face recognition were analyzed, demonstrating the field's rapid evolution and the challenges in surveillance approaches.

The dataset preparation and preprocessing were detailed in Chapter 3, addressing the severe class imbalance in the collected CCTV footage (ranging from 3 to 167 images per identity). A balanced dataset of 4,300 images with exactly 50 samples per identity was created through quality-aware downsampling of 53 over-represented identities and controlled augmentation of 33 under-represented identities using six augmentation techniques (rotation, horizontal flip, brightness adjustment, contrast adjustment, Gaussian blur, colour saturation). Validation confirmed 100% image integrity with average similarity degradation of only 1.9% from augmentation. This works at preserving the identity-discriminative features while enhancing dataset diversity.

The Methodology in Chapter 4 describes the dual-backbone architecture combining VGGFace (4096-D features) and FaceNet (512-D embeddings) through learnable weighted fusion. The four-layer fully-connected embedding network with Batch-Norm and Dropout projects the 4608-D fused features to 256-D L2-normalised embeddings on the unit hypersphere. The hybrid loss function combines additive angular margin with semi-hard triplet mining. Training employed AdamW optimizer

with Cosine Annealing Warm Restarts and mixed-precision FP16 computation. The experimental results from multiple evaluations in various dimensions were presented in Chapter 5. Training dynamics analysis revealed robust convergence over 100 epochs with peak validation accuracy of 95.70%. Classification performance analysis depicted ROC-AUC of 0.9665, precision-recall AUC of 0.9616, and optimal F1-score of 91.16% at threshold 0.4179. Distance distribution analysis showed genuine-impostor separation of 0.781 with minimal overlap, validating embedding quality. With the baseline comparison showed significant improvements: +8.52% accuracy, +10.33% precision, +4.83% recall, and +7.77% F1-score compared to single-backbone triplet loss architecture. The comprehensive ablation study across 20 configurations validated weighted fusion and ArcFace loss as optimal choices. Cross-dataset evaluation on LFW (99.28% accuracy), CelebA (98.57%), and VGGFace2 (96.82%) demonstrated excellent generalisation to controlled imaging conditions, while quantifying the 6.65% F1 gap attributable to surveillance domain challenges.

A comprehensive investigation into face recognition for CCTV surveillance applications has been presented in this thesis, addressing fundamental challenges by applying architectural innovation, loss function optimisation, and systematic dataset balancing. The proposed hybrid Siamese network with VGGFace-FaceNet fusion and ArcFace-Triplet combined loss achieved significant performance improvements over baseline approaches while maintaining computational efficiency suitable for real-time deployment. The research demonstrates that combining complementary deep learning architectures through learnable fusion mechanisms can be effectively leverage their respective strengths: VGGFace’s texture discrimination and FaceNet’s geometric invariance. The hybrid loss formulation unifying angular margin maximization with metric learning constraints addresses limitations of single-objective optimisation, producing embeddings with superior separation characteristics validated through distance distribution analysis. Comprehensive experimental evaluation spanning training dynamics, classification performance, threshold sensitivity, embedding quality, ablation studies, and cross-dataset generalization established robust empirical foundation for the proposed approach. The 91.01% accuracy and 0.9665 ROC-AUC on CCTV surveillance footage, combined with near-perfect performance on controlled benchmarks, demonstrate both practical effectiveness and strong generalisation capabilities. The quantified 6.65% F1 gap between surveillance and controlled imaging conditions provides empirical characterization of the surveillance challenge, which can be decomposed into resolution effects, compression artefacts, illumination variation, and pose extremes. This quantification informs realistic performance expectations for operational deployments and guides future research toward targeted interventions addressing specific degradation factors. Here, limitations remain, particularly regarding temporal modeling and scalability to large watchlists, though the achieved F1-score (91.16%) demonstrates strong performance for single-frame single-modality processing. The research establishes a strong foundation for practical CCTV-based face verification systems. The identified future research directions are temporal modeling, multi-frame learning. This providing clear roadmap for advancing beyond current capabilities toward robust, privacy-preserving, fair, and scalable face recognition systems suitable for large-scale surveillance networks. Finally, this work contributes both technical advances in face recognition systems and provides insights into surveillance domain challenges. The

deployment-ready system with operational guidelines closes the gap between academic research and practical application.

Despite strong performance across multiple evaluations, several limitations have been encountered that indicate future research directions. The CCTV dataset comprises only 86 identities with 4,300 total images after balancing. The 10.33% false positive rate among accepted pairs (precision 89.67%) indicates strong discrimination learned from the training data with a hybrid architecture. While the 9.3% relative improvement over baseline represents substantial progress, the absolute optimal F1-score of 91.16% demonstrates strong performance compared to controlled benchmark performance ($F1 > 96\%$). This remaining gap indicates that even with architectural innovations and hybrid loss functions, some limitations from degraded CCTV image quality persist. The 6.65% F1 degradation relative to benchmark average quantifies the magnitude of remaining challenges. Practical implications include that operational deployment can benefit from the 89.67% precision for many surveillance applications. Applications requiring $> 95\%$ precision may require secondary verification or multi-biometric factors. The system relies exclusively on facial appearance features without incorporating temporal dynamics incorporating video tracking across frames, contextual information like clothing, gait, and scene context. Research in video face recognition demonstrates that temporal smoothing and tackle aggregation can improve accuracy by 5 to 10% over single-frame recognition, suggesting unexploited performance potential. Missing capabilities that have not been implemented include no temporal consistency enforcement across video sequences, no exploitation of multi-frame super-resolution for quality enhancement, and no integration with person re-identification for cross-camera tracking. While the system demonstrates robust performance on benchmark datasets, explicit occlusion-robust mechanisms were not incorporated. The CCTV dataset includes only natural occlusions without systematic evaluation on heavy occlusions like using masks or hats that occur frequently in surveillance environments. The optimal threshold (0.4179) was determined retrospectively on the validation set, maximising F1-score for the specific 86-identity cohort. In operational deployment, where new identities are continuously enrolled, threshold recalibration may be required to maintain performance as the genuine/impostor distance distributions shift with population changes. Continuous monitoring and adaptive threshold tuning infrastructure would be required for production deployment, adding system complexity and maintenance burden. The research did not develop automated threshold adaptation mechanisms or evaluate performance degradation over time with population drift. Single-frame recognition discards temporal information in video surveillance streams. For video-based face recognition across multi-frames frames can be aggregated, and temporal smoothing can be applied to reduce noise, and also multi-frame super-resolution can be used to enhance quality. A new approach would extend the Siamese architecture with temporal modeling through layers that process frame sequences rather than individual frames. So there are a couple of scopes to further extend this research by implementing it on video surveillance data, which will further strengthen this system.

Bibliography

- [1] V. Blanz and T. Vetter, “Face recognition based on fitting a 3d morphable model,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 25, 2003, pp. 1063–1074.
- [2] J. Huang, V. Blanz, and B. Heisele, “Component-based face recognition with 3d morphable models,” in *International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA)*, Springer, 2003, pp. 27–34.
- [3] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” in *Workshop on Faces in ‘Real-Life’ Images: Detection, Alignment, and Recognition*, 2007.
- [4] U. Park, Y. Tong, and A. K. Jain, “Age-invariant face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 947–954, 2010.
- [5] M. Grgic, K. Delac, and S. Grgic, “Sface—surveillance cameras face database,” *Multimedia tools and applications*, vol. 51, pp. 863–879, 2011.
- [6] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *ICML Deep Learning Workshop*, vol. 2, 2015.
- [7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conference*, vol. 1, 2015, p. 6.
- [8] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2015, pp. 815–823.
- [9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” *European conference on computer vision*, pp. 87–102, 2016.
- [10] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, “Do we really need to collect millions of faces for effective face recognition?” In *European conference on computer vision*, Springer, 2016, pp. 579–596.
- [11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *European Conference on Computer Vision (ECCV)*, Springer, 2016, pp. 499–515.
- [12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

- [13] C. N. Duong, K. Luu, K. G. Quach, and T. D. Bui, “Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition,” in *IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 3755–3763.
- [14] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 212–220.
- [16] J.-J. Lv, Y.-H. Shao, J. Huang, X.-D. Zhou, and X. Zhou, “Data augmentation for face recognition,” in *Neurocomputing*, vol. 230, 2017, pp. 184–190.
- [17] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, “Range loss for deep face recognition with long-tailed training data,” in *IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 5409–5418.
- [18] X. Zhang and Y. Gao, “Face recognition across pose: A review,” *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2017.
- [19] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising faces across pose and age,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, IEEE, 2018, pp. 67–74.
- [20] H. Wang et al., “Cosface: Large margin cosine loss for deep face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2018, pp. 5265–5274.
- [21] X. Yu and F. Porikli, “Super-resolving very low-resolution face images with supplementary attributes,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2018, pp. 908–917.
- [22] Y. Cheng et al., “Surveillance face recognition: Challenges and strategies,” *arXiv preprint arXiv:1904.10664*, 2019.
- [23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 4690–4699.
- [24] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [25] Y. Zhong et al., “Unequal-training for deep face recognition with long-tailed noisy data,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 7812–7821.
- [26] A. et al., “Hybrid Siamese network for unconstrained face verification and clustering under limited resources,” *Big Data and Cognitive Computing*, vol. 4, no. 2, p. 15, 2020, LFW: 98.9%, Arabian Faces: 99.1%. DOI: 10.3390/bdcc4020015.
- [27] G. S. Hsu et al., “A comprehensive study on loss functions for cross-factor face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2020, pp. 814–815.

- [28] I. Kim et al., “DiscFace: Minimum Discrepancy Learning for Deep Face Recognition,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020, pp. 358–374.
- [29] S. Tripathi and H. Beigi, “Learning discriminative features using center loss and reconstruction as an auxiliary task for speech emotion recognition,” in *Proceedings of Machine Learning Research*, vol. 122, PMLR, 2020, pp. 37–50.
- [30] X. Wang et al., “Loss function search for face recognition,” in *Proceedings of Machine Learning Research*, vol. 119, PMLR, 2020, pp. 10 029–10 038.
- [31] Y. Zhang and W. Deng, “Class-balanced training for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 824–825.
- [32] Y. Zhang, C. Liu, and W. Deng, “Class-balanced training for deep face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2020, pp. 868–869.
- [33] W. Hariri, “Efficient masked face recognition method during the covid-19 pandemic,” *Signal, Image and Video Processing*, vol. 16, no. 3, pp. 605–612, 2021.
- [34] C.-H. Hu et al., “Face illumination recovery for the deep learning feature under severe illumination variations,” *Pattern Recognition*, vol. 111, p. 107 724, 2021.
- [35] B. Huang et al., “When face recognition meets occlusion: A new benchmark,” *arXiv preprint arXiv:2103.02805*, 2021.
- [36] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215–244, 2021.
- [37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A comprehensive study on center loss for deep face recognition,” *International Journal of Computer Vision*, vol. 129, pp. 668–683, 2021.
- [38] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, “Self-restrained triplet loss for accurate masked face recognition,” *Pattern Recognition*, vol. 124, p. 108 473, 2022.
- [39] F. Ennaama et al., “Proposed approach to improve facial recognition performance under mask variation in covid-19 period,” *IFAC-PapersOnLine*, vol. 55, no. 12, pp. 235–240, 2022.
- [40] Y. Wang et al., “A real-time framework for human face detection and recognition in cctv images,” *Wireless Communications and Mobile Computing*, vol. 2022, p. 3 276 704, 2022.
- [41] J. Xiao, G. Jiang, and H. Liu, “A lightweight face recognition model based on MobileFaceNet for limited computation environment,” *EAI Endorsed Transactions on Internet of Things*, vol. 8, no. 1, e3, 2022, LFW: 98.52%, VGGFace2: 97.54%, Model size: 3.4MB. DOI: 10.4108/eetiot.v8i1.293.
- [42] Y. Xu et al., “Constructing balance from imbalance for long-tailed image recognition,” in *European Conference on Computer Vision (ECCV)*, Springer, 2022, pp. 38–56.
- [43] F. Xue, Q. Wang, and G. Guo, “Vision transformer with attentive pooling for robust facial expression recognition,” *IEEE Transactions on Affective Computing*, 2022.

- [44] J. Zhou et al., “Uniface: Unified cross-entropy loss for deep face recognition,” in *IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2023, pp. 3553–3562.
- [45] Y. et al., “Deep convolutional neural network using triplet loss to distinguish the identical twins,” in *Proceedings of IEEE Conference on Pattern Recognition*, LFW: 99.5% with hybrid ArcFace+Triplet, SOTA for hybrid loss, 2024.
- [46] P. Dandekar et al., “Low-resolution face recognition: Review, challenges, directions,” *Computers & Electrical Engineering*, vol. 118, p. 109399, 2024.
- [47] N. Dzhaluiuk et al., “Improving facenet performance using optimized triplet loss function,” in *Proceedings of CH&CMiGIN’25*, vol. 4024, 2024, p. 24.
- [48] M. A. R. Faruque et al., “Advanced masked face recognition using robust and light weight deep learning model,” *International Journal of Computer Applications*, vol. 186, no. 2, pp. 42–48, Jan. 2024, Masked face accuracy: 97.0%. DOI: 10.5120/ijca2024923351.
- [49] J. Karlekar et al., “Robust data augmentation strategies for unconstrained face recognition,” in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–8.
- [50] M. Rodrigo, E. Talavera, R. Aler, et al., “Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks,” *Scientific Reports*, vol. 14, no. 1, p. 21485, 2024.
- [51] P. Santemiz et al., “A survey on automatic face recognition using side-view face images,” *IET Biometrics*, 2024.
- [52] W. et al., “Improving low-light face recognition using DeepFace embedding and multi-layer perceptron,” *RESTI: Jurnal Rekayasa Sistem dan Teknologi Informasi*, vol. 9, no. 5, pp. 1088–1095, Oct. 2025, LFW accuracy: 95.16% with CLAHE preprocessing.
- [53] S. Chen and Y. Wang, “Challenges of face identification with varied mask coverage in the post covid-19 era,” *Frontiers in Psychology*, 2025. DOI: 10.3389/fpsyg.2025.1486808.
- [54] N. El Fadel, “Facial recognition algorithms: A systematic literature review,” *Journal of Imaging*, vol. 11, no. 2, 2025.
- [55] E. L. Gunawan, I. G. S. M. Diyasa, and W. S. J. Putra, “Optimization of facial recognition authentication system using InceptionResNetV1 with pretrained VGGFACE2,” *Simantec: Jurnal Sistem Informasi dan Teknologi Komputer*, vol. 13, no. 2, pp. 221–230, Jun. 2025, LFW Accuracy: 98.75%. DOI: 10.47002/simantec.v13i2.29776.
- [56] H. S. Jassim et al., “Face recognition system for criminal identification in cctv footage using keras and opencv,” *Ingénierie des Systèmes d’Information*, vol. 30, no. 3, 2025, Received: 28 July 2024.
- [57] S. Karamizadeh, S. S. Chaeikar, and H. Salarian, “Combining MTCNN and enhanced FaceNet with adaptive feature fusion for robust face recognition,” *Technologies*, vol. 13, no. 10, p. 450, Oct. 2025, CelebA: 99.6%, LFW: 94.2% (5.4% cross-dataset degradation). DOI: 10.3390/technologies13100450.

- [58] Q. Li, K. Fu, J. Liu, Y. Li, et al., “Optimizing class imbalance in facial expression recognition using dynamic intra-class clustering,” *Biomimetics*, vol. 10, no. 5, p. 296, 2025.
- [59] N. Makhanov, B. Amirgaliyev, T. Islamgozhayev, and D. Yedilkhan, “Federated self-supervised few-shot face recognition,” *Journal of Imaging*, vol. 11, no. 10, p. 252, Oct. 2025, Federated LFW: $\sim 90\%$, demonstrates 12-30% privacy tax. DOI: 10.3390/jimaging11100252.
- [60] N. F. Mir and A. M. Hafiz, “Face recognition by Siamese network using a novel distance function,” in *Lecture Notes in Networks and Systems*, ser. Proceedings of International Conference on Computational Intelligence and Sustainable Technologies, LFW accuracy: 95.62%, validates backbone superiority over distance metric innovation, vol. 1085, Springer Nature Singapore, 2025, pp. 347–358. DOI: 10.1007/978-981-96-1744-9_27.
- [61] H. Nakamura, “Transformation tolerance of facial recognition: An end-to-end CNN approach,” LFW accuracy: $\sim 97.35\%$, MEng Thesis, Massachusetts Institute of Technology, 2025.
- [62] A. Nemsia, “Efficient one-shot face recognition using deep siamese networks with limited supervision,” *Expert Systems with Applications*, vol. 238, p. 121 765, 2025.
- [63] Y. Qiu et al., “Low-resolution driver face recognition based on super-resolution and original image fusion,” *Scientific Reports*, vol. 15, p. 5617, 2025.
- [64] S. Shahi, R. A. Mahajan, Midhunchakkaravarthy, and A. Pratap, “Sibling face recognition improvised by merging deep learning embeddings from vggface and facenet,” *SGS - Engineering & Sciences*, vol. 1, no. 4, 2025.