CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Enhanced Distillation based Deep Learning for Low Resolution Face Recognition

by

Mohsin Ullah

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Engineering
Department of Electrical & Computer Engineering

2025

Enhanced Distillation based Deep Learning for Low Resolution Face Recognition

By

Mohsin Ullah

(DEE193002)

Dr. Andrew Ware, Professor

University of South Wales, UK

(Foreign Evaluator 1)

Dr. José Valente de Oliveira, Senior Researcher

University of Lisboa, Portugal

(Foreign Evaluator 2)

Dr. Imtiaz Ahmad Taj

(Research Supervisor)

Dr. Noor Muhammad Khan

(Head, Department of Electrical & Computer Engineering)

Dr. Imtiaz Ahmad Taj

(Dean, Faculty of Engineering)

DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2025

Dedicated to my Teachers, Parents, Wife and Siblings

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Enhanced Distillation based Deep Learning for Low Resolution Face Recognition**" was conducted under the supervision of **Dr. Imtiaz Ahmed Taj**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Electrical & Computer Engineering, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Electrical Engineering.** The open defence of the dissertation was conducted on **January 03, 2025**.

Student Name : Mohsin Ullah (DEE193002)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Asifullah Khan
Professor
PIEAS, Islamabad

(b) External Examiner 2: Dr. Abdul Ghafoor
Professor
MCS, NUST, Rawalpindi

(c) Internal Examiner : Dr. Nadeem Anjum
Professor
CUST, Islamabad

Supervisor Name : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

Name of HoD : Dr. Noor Muhammad Khan
Professor
CUST, Islamabad

Name of Dean : Dr. Imtiaz Ahmad Taj
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Mohsin Ullah** (Registration No. **DEE193002**), hereby state that my dissertation titled, "**Enhanced Distillation based Deep Learning for Low Resolution Face Recognition**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(**Mohsin Ullah**)

Dated: 03, January, 2025

Registration No : DEE193002

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Enhanced Distillation based Deep Learning for Low Resolution Face Recognition**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled dissertation declare that no portion of my dissertation has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Mohsin Ullah**)

Dated: 03 January, 2025

Registration No : DEE193002

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **Ullah, M.**, Taj, I.A. and Raza, R.H., 2023. *Degradation Model and Attention Guided Distillation Approach for Low Resolution Face Recognition.* Expert Systems with Applications, p.122882

(Mohsin Ullah)
Registration No: DEE193002

Acknowledgement

All praise to the Almighty Allah, the Lord of the Universe, for His boundless grace, countless blessings, infinite mercy and granting us wealth of *Emaan*.

I am forever grateful to my supervisor Dr. Imtiaz Ahmad Taj, whose visionary guidance, tireless efforts, unwavering support and unrelenting encouragement have been the driving force behind this research journey. His exceptional expertise in computer vision and deep understanding of mathematics, which has been a constant source of inspiration, even I have only scratched the surface of his vast knowledge.

I extend my sincere gratitude to Dr. Rana Hammad Raza of PNEC, NUST Karachi, for the financial support provided through National Center of Big Data and Cloud Computing (NCBC) and Higher Education Commission (HEC) of Pakistan, which made this research possible. Also, thanks to Dr. Safwan Khalid of COMSATS University Islamabad, for his valuable suggestions.

I am also grateful to the members and colleagues of the Vision and Pattern Recognition System (VisPRS) research group for their support and guidance.

To my friends and colleagues, I offer my heartfelt admiration for rallying around me during my darkest moments. You refused to let me give up, even when I felt like collapsing. Muhammad Sufyan, you are one of those.

And finally, to my parents, wife, and siblings, your prayers, love, and affection have been my cheerleading, turning my dreams into reality.

(Mohsin Ullah)

Abstract

Face recognition is one of the most well-studied research topics in computer vision, having received significant attention in the last decades. This surge of interest is primarily driven by the effectiveness of deep convolutional neural networks (CNNs) in visual recognition tasks. CNNs excel at extracting highly distinctive facial features from facial images, making them ideal for facial recognition applications. Recent advancements in face recognition technology have achieved remarkable accuracy in high-resolution (HR) images, even with considerable variations in pose, illumination, and expression. However, a significant hurdle remains in identifying individuals from blurry, low-resolution (LR) images, such as those captured by surveillance cameras. A straightforward solution to this problem involves training with a combination of HR images and their corresponding down-sampled LR images. Although this strategy improves the performance of CNNs on LR images, it has some limitations. Firstly, the down-sampled images do not reflect the variations found in the real-world surveillance data. Secondly, capturing diverse representations of images across varying resolutions is challenging. Thirdly, the performance is biased toward LR images, leading to deteriorated performance on HR images. Fourthly, the existing LR testing benchmarks cannot highlight the efficacy and limitations of the LR face recognition model.

This dissertation proposes two novel schemes for LR face recognition and new protocols to provide a more rigorous testing environment. The first scheme targets face recognition in low-resolution images, especially for surveillance applications. It consists of a degradation model and attention-guided distillation. The degradation model simulates real-world degradation effects in the synthetic LR facial data using a combination of classical degradation techniques with a comprehensive evaluation of each degradation on the performance of widely used SCface and COXface datasets. The attention-guided distillation uses spatial attention maps to reduce the gap between HR and LR feature representations by transferring informative and discriminant features from the HR teacher network to the LR student network.

The second scheme also targets LR images while maintaining good recognition performance on HR images. The proposed scheme consists of sub-center learning and contrastive distillation loss. Sub-center learning captures diverse representations of images

across varying resolutions through multiple sub-centers defined for each class. Contrastive distillation loss enforces a strict constraint by pushing the LR and HR features closer together, in contrast to other non-corresponding features. This process effectively leads the model to learn compact and discriminant features.

The new protocols exploit the limitations found in the existing protocols for LR testing benchmarks and provide a comprehensive evaluation mechanism for both HR and LR images. It gets rid of fine-tuning that compromises the generalization capability of a face recognition model. A new combined evaluation metric (CEM) is also introduced to judge the performance simultaneously on HR and LR images. The first scheme is tested on existing protocols of LR testing benchmarks trained on the small-scale training dataset. The comparative analysis has demonstrated significant performance improvement with a margin of 6.95%, 4.97% and 9.38% on SCface, COXface and Tinyface, respectively, compared to the previous state-of-the-art (SOTA) methods. The second scheme is tested on existing and more rigorous new protocols trained on both small-scale and large-scale datasets. The proposed scheme outperformed all other schemes on the majority of benchmarks. Using CEM as the standard for the best compromise between HR and LR face recognition, the proposed scheme achieved higher scores, surpassing the previous SOTA by margins of 25.83 and 3.60 in small-scale and large-scale experiments, respectively.

Contents

Author's Declaration v

Plagiarism Undertaking vi

List of Publications vii

Acknowledgement viii

Abstract ix

List of Figures xv

List of Tables xvii

Abbreviations xviii

Symbols xix

1 Introduction 1
- 1.1 Overview . . . 1
- 1.2 History of Face Recognition . . . 2
  - 1.2.1 1960s . . . 2
  - 1.2.2 1970s . . . 2
  - 1.2.3 Late 1980s and Early 1990s . . . 2
  - 1.2.4 1990s . . . 3
  - 1.2.5 2000s . . . 4
  - 1.2.6 2010s . . . 4
  - 1.2.7 2020s . . . 6
- 1.3 Basics . . . 6
  - 1.3.1 Verification vs Identification . . . 6
  - 1.3.2 Probe Set vs Gallery Set . . . 6
  - 1.3.3 High Resolution vs Low Resolution . . . 6
  - 1.3.4 Separable vs Discriminant Features . . . 7
- 1.4 Challenges . . . 8
  - 1.4.1 Variations in Pose, Illumination and Expression . . . 8
  - 1.4.2 Image Resolution . . . 8

1.4.3 Occlusion . . . 8
1.4.4 Aging . . . 8
1.5 Motivation . . . 9
1.6 Face Recognition Pipeline . . . 9
1.6.1 Preprocessing . . . 9
1.6.2 Feature Extraction . . . 10
1.6.3 Matching . . . 10
1.7 Conventional Techniques to Deep Learning . . . 11
1.8 HR Face Recognition Problem . . . 11
1.9 LR Face Recognition Problem . . . 12
1.10 Face Detectors . . . 12
1.10.1 MTCNN . . . 12
1.10.2 Retinaface . . . 13
1.11 Backbone Architectures . . . 15
1.11.1 GoogleNet . . . 15
1.11.2 ResNet . . . 16
1.11.3 Squeeze-and-Excitation Networks (SENet) . . . 16
1.12 Thesis Contribution . . . 17
1.13 Thesis Organization . . . 18
1.14 Summary . . . 19

2 Literature Review **20**
2.1 Introduction . . . 20
2.2 Conventional Methods . . . 21
2.2.1 Geometric Approaches . . . 21
2.2.2 Holistic Approaches . . . 22
2.2.3 Feature-based Methods . . . 23
2.2.4 Hybrid Methods . . . 24
2.3 Deep Learning based Methods . . . 24
2.3.1 HR Face Recognition . . . 24
2.3.1.1 Softmax-based Methods . . . 25
2.3.1.2 Distance Metric Learning . . . 28
2.3.2 LR Face Recognition . . . 29
2.3.2.1 Super Resolution based Approaches . . . 29
2.3.2.2 Resolution Invariant Approaches . . . 32
2.3.2.3 Universal Learning . . . 33
2.3.2.4 Distillation Learning . . . 38
2.4 Emerging Trends . . . 45
2.5 Research Gap . . . 46
2.6 Problem Statement . . . 46
2.7 Research Objectives . . . 47
2.8 Summary . . . 47

3 Datasets and Evaluation Methods **49**
3.1 Outline . . . 49
3.2 Training Datasets . . . 50

3.2.1 CASIA Webface . . . 50
3.2.2 Mirosoft Celeb 1M . . . 50
3.2.3 Webface 4M . . . 50
3.2.4 VGGFace2 . . . 51
3.3 Testing Datasets . . . 51
3.3.1 LR Datasets . . . 51
3.3.1.1 SCface . . . 51
3.3.1.2 COXface . . . 53
3.3.1.3 Tinyface . . . 54
3.3.1.4 QMUL Survface . . . 56
3.3.2 HR Datasets . . . 57
3.3.2.1 Labeled Faces in the Wild (LFW) . . . 57
3.3.2.2 Celebrities in Frontal-Profile (CFP) . . . 57
3.3.2.3 AgeDB . . . 57
3.3.2.4 Cross-Age LFW (CALFW) . . . 58
3.3.2.5 Cross-Pose LFW (CPLFW) . . . 58
3.3.3 Mixed Resolution Datasets . . . 58
3.3.3.1 IJB-B Dataset . . . 58
3.3.3.2 IJB-C Dataset . . . 60
3.4 Evaluation Metrics . . . 60
3.4.1 Face Verification . . . 60
3.4.2 Close-Set Face Identification . . . 62
3.4.3 Open-Set Face Identification . . . 63
3.5 Summary . . . 63

4 Degradation Model and Attention-Guided Distillation 64
4.1 Outline . . . 64
4.2 Proposed Methodology . . . 65
4.2.1 Degradation Model . . . 65
4.2.1.1 Proposed Solution . . . 66
4.2.2 Attention Guided Distillation . . . 70
4.2.2.1 Attention Maps . . . 72
4.2.2.2 Proposed Solution . . . 73
4.3 Experimental Setup . . . 76
4.3.1 Implementation Details . . . 76
4.3.2 Degradation Model Setting . . . 76
4.4 Results and Discussion . . . 77
4.4.1 Ablation Study . . . 77
4.4.2 Comparison with SOTA Methods . . . 79
4.4.2.1 LR Datasets . . . 79
4.4.2.2 HR datasets . . . 84
4.4.3 Discussion . . . 85
4.5 Conclusion . . . 86

5 New Protocols 87
5.1 Outline . . . 87

5.2 Limitations in the Existing Protocols . . . 88
5.2.1 Fine-tuning . . . 88
5.2.2 Lack of Evaluation on HR Datasets . . . 89
5.2.3 Different Nature of LR Datasets . . . 89
5.2.4 Combined Evaluation Metric . . . 90
5.3 New Protocols . . . 91
5.4 Summary . . . 92

6 Sub-center Learning and Contrastive Distillation Loss 93
6.1 Outline . . . 93
6.2 Proposed Methodology . . . 95
6.2.1 Sub-center Learning . . . 96
6.2.2 Contrastive Distillation Loss . . . 98
6.2.2.1 Review of Other Distillation Losses . . . 99
6.2.2.2 Feature-Contrastive and Class-Contrastive Distillation Loss . . . 100
6.2.3 Distillation Mechanism . . . 103
6.2.4 Derivation of the Gradient on Contrastive Distillation Loss . . . 104
6.3 Results and Discussion . . . 107
6.3.1 Implementation Details . . . 107
6.3.1.1 Training Settings . . . 107
6.3.1.2 Augmentation Settings . . . 108
6.3.2 Ablation Analysis . . . 109
6.3.2.1 Effect of Each Component in Loss . . . 109
6.3.2.2 Effect of Augmentations . . . 110
6.3.3 Comparison with SOTA Methods . . . 110
6.3.3.1 LR Datasets . . . 111
6.3.3.2 HR and MR Datasets . . . 114
6.3.3.3 Combined Evaluation Metric (CEM) . . . 115
6.4 Summary . . . 116

7 Conclusion and Future Work 117
7.1 Conclusion . . . 117
7.2 Future Work . . . 119
7.3 Societal and Ethical Concerns . . . 120

Bibliography 122

Appendix A 135
A.1 Gradient of KL Divergence Loss . . . 135

List of Figures

1.1 History of Face Recognition 5
1.2 Image resolution worsens from left to right (Leftmost: HR, Rightmost: LR) 7
1.3 Separable Features (Left) and Discriminant Features (Right) 7
1.4 Face Recognition Pipeline 10
1.5 Pipeline of Cascaded Framework in MTCNN [19] 13
1.6 Pipeline of Retinaface Architecture [20] 14
1.7 Residual Block in ResNet Architecture 16
1.8 Squeeze and Excitation (SE) Block. [23] 17

2.1 Taxonomy of Face Recognition Techniques 21

3.1 Example images from the SCface dataset captured at three different distances. The first column shows high-quality images, while subsequent columns show images from the five surveillance cameras, respectively. . . 52
3.2 Example images from the COXface dataset captured by three different surveillance cameras. The first column shows high-quality images, while subsequent columns show images captured at different positions along the S-shaped pathway. 54
3.3 Example Images from Tinyface Dataset. (Left) Labeled Identities, (Right) Distractors 55
3.4 Example Images from QMUL Survface Dataset. (Left) Labeled Identities, (Right) Distractors 56
3.5 (Left) Positive Pairs, (Right) Negative Pairs. 59
3.6 Example Images from IJB-B and IJB-C datasets, having images of a single identity in a row 61

4.1 Overview of degradation model that is used to generate synthetic LR images. The input is an HR image and the coordinates of facial landmarks, while the output is a synthetic LR image. Each block represents a different degradation effect induced in the synthetic data. 67
4.2 Euclidean distance calculated between the coordinates of facial landmarks of HR image and its down-sampled versions. 68
4.3 A generic solution to abstain the network from discarding HR features. Distillation techniques shift the common-features subspace closer to the HR feature subspace. 71
4.4 HR and LR facial images with corresponding spatial attention maps are shown, highlighting where the teacher network has focused in low-, mid-, and high-level features. 72

4.5 Complete framework showing degradation model and attention-guided distillation. The basic module of the teacher and student network is shown at the top-right corner. The architecture is defined using the terminology $d \times n$ for each module, where d represents the depth of the output of the convolutional layer in each module and n represents the repetition of module. 73
4.6 Ablation study over each degradation induced in the degradation model. m: misalignment, b: blurring, n: noise, c: compression 78
4.7 Performance comparison of the proposed scheme against SR technique on SCface dataset 81
4.8 Performance Comparison of the Proposed scheme against SR Technique on COXface dataset 83

5.1 Response of LR datasets to different percentages of HR images in a batch during training 90

6.1 High-resolution (left) and low-resolution (right) samples of images visualized through t-SNE 94
6.2 Analysis of sub-center learning using feature distribution through tSNE. Different colours denote different classes. 98
6.3 The proposed methodology consists of a pre-trained HR teacher network and a student network. The student network is trained using the sub-center learning and contrastive distillation losses. (N: Number of classes, B: Batch Size) 103
6.4 The distribution of features visualized through t-SNE under Curricular-Face (Baseline) [15] **(Left)** and the proposed methodology **(Right)** is shown for 10 identities. Different colors denote different classes. 104

List of Tables

2.1 Comparative Analysis of SR based Approaches . . . 33
2.2 Comparative Analysis of Universal Learning Approaches . . . 37
2.3 Comparative Analysis of Distillation Learning Approaches . . . 44

3.1 Summary of Facial Datasets for Training and Testing (Images: I, Videos: V) . . . 62

4.1 Types of degradation used in the previous approaches and the proposed degradation model . . . 66
4.2 Ablation study over each contribution in the proposed approach. (DM: Degradation Model) . . . 77
4.3 Performance Comparison on SCface Dataset (Evaluation on the testing partition without fine-tuning) . . . 80
4.4 Performance Comparison on SCface dataset. **-FT** means fine-tuning with the SCface training set. . . . 80
4.5 Performance Comparison on COXface dataset: Video-to-Still face recognition. **-FT** means fine-tuning with the COXface training set. . . . 81
4.6 Performance Comparison on COXface Dataset: Still-to-Video face recognition. **-FT** means fine-tuning with the COXface training set. . . . 82
4.7 Performance Comparison on COXface Dataset: Video-to-Video scenario. 83
4.8 Performance Comparison on Tinyface Dataset. . . . 84
4.9 Performance Comparison on HR (1:1 Verification Rate) . . . 84

5.1 Performance comparison of ResNet-50 with REE [63] on fine-tuned protocols . . . 88
5.2 Performance evaluation of a face recognition model in HR and LR scenarios 89
5.3 Evaluation metrics for HR and LR testing benchmarks in Combined Evaluation Metric (CEM) . . . 91

6.1 Ablation Analysis of Proposed Loss . . . 109
6.2 Ablation Analysis of Augmentation . . . 110
6.3 Performance Comparison on SCface. (P: Protocol) . . . 111
6.4 Performance Comparison on Tinyface. (P: Protocol) . . . 112
6.5 Performance Comparison on QMUL Suvface. (P: Protocol) . . . 113
6.6 Performance Comparison on COXface. (P: Protocol) . . . 114
6.7 Performance Comparison on HR Datasets (1:1 Verification Rate) . . . 114
6.8 Performance Comparison on MR Datasets (TPR@FAR=1e-4) . . . 115
6.9 Performance Comparison on Combined Evaluation Metric (CEM) . . . 115

Abbreviations

AGD	Attention Guided Distillation
CEM	Combined Evaluation Metric
DFD	Deep Feature Distillation
DM	Degradation Model
HR	High Resolution
KD	Knowledge Distillation
KL	Kullback-Liebler
LR	Low Resolution
MSE	Mean Squared Error
RI	Resolution Invariant
SR	Super Resolution
SOTA	State-of-the-art

Symbols

x	Feature Vector / Embedding
z	Class Probabilities
w_t	Weights of the teacher network
w_s	Weights of the student network
$\mathcal{N}_t$	Teacher network
$\mathcal{N}_s$	Student network
$\mathcal{I}$	HR image
$\mathbb{I}_{\mathrm{H}}$	HR Dataset
$\mathcal{I}'$	Synthetic LR image
$\mathbb{I}_{\mathrm{S}}$	Synthetic LR Dataset
$\mathcal{L}_{cl}$	Classification loss
$\mathcal{L}_{dfd}$	Deep feature distillation loss
$\mathcal{L}_{ad}$	Attention-based distillation loss
$\mathcal{L}_{agd}$	Attention guided distillation loss
$\mathcal{L}_{f-cdl}$	Feature contrastive distillation loss
$\mathcal{L}_{c-cdl}$	Class contrastive distillation loss

Chapter 1

Introduction

1.1 Overview

Biometric refers to the automatic recognition of physiological and behavioral attributes in human beings. In past decades, there has been significant progress in biometric technology, resulting in the industrialization of systems based on various modalities such as face, iris, gait, fingerprint, and palmprint. Of these modalities, the human face is a widely used biometric due to its contactless acquisition, non-invasive nature, social acceptance, and suitability for non-cooperative scenarios.

Face recognition systems have evolved from a novelty to a powerful technology with a wide range of applications impacting our lives in many ways. In security and law enforcement, it can be used to identify suspects, verify identities at border crossings and even track missing persons, thus contributing to crime prevention, resolution and public safety. Beyond security, face recognition is used in everyday conveniences like user authentication in electronic devices, ensuring secure and convenient access, and automatic attendance systems in education and workplaces. The technology can also improve customer service by streamlining identification and personalizing experiences. Face recognition technology is also gaining importance in smart cities by enhancing public safety and security through surveillance and monitoring. Overall, face recognition systems offer a powerful tool for identification, surveillance, personalized experiences, and automation across various industries.

1.2 History of Face Recognition

Facial recognition systems have a fascinating history that spans several decades, evolving from early conceptualizations to advanced technologies used today. The journey of face recognition systems can be outlined in key milestones:

1.2.1 1960s

Woody Bledsoe is considered as the father of face recognition. He started work with Helen Chan Wolf and Charles Bisson in 1964 and 1965 [1]. Initially, their efforts involved manually marking various facial landmarks. These landmarks were then computationally adjusted to accommodate pose variations. To establish identity, distances were calculated between corresponding landmarks across images. This project was termed "man-machine" as it involved a human in extracting coordinates of facial features from photographs, which were then utilized by the computer for recognition. Graphics tablets such as GRAFACON or RAND TABLET were used to extract facial features. These features included pupil centers, inner and outer eye corners and the widow's peak, etc.

1.2.2 1970s

In the mid-1970s, Goldstein, Harmon, and Lesk made significant advancements in face recognition [2]. They improved the process by utilizing a subset of 22 critical features pruned from an initial set of 34. These features included more detailed attributes like eyebrow separation, lip thickness, ear length, hair texture, etc.

1.2.3 Late 1980s and Early 1990s

The development of Eigenfaces marked a turning point in the face recognition problem. Sirovich and Kirby first introduced the concept of Eigenfaces in the late 1980s [3]. Later, in 1991, Matthew Turk and Alex Pentland employed Eigenfaces for face classification [4]. Eigenfaces are generated using a mathematical process called Principal Component Analysis (PCA). PCA allows for data representation with fewer variables while retaining

the most crucial information. This makes the data easier to visualize, analyze, and use for classification problems.

1.2.4 1990s

1993 — Lades et al. proposed the Dynamic Link Architecture (DLA), an extension to conventional artificial neural networks for face discrimination [5]. Additionally, they introduced the innovative use of Gabor-type wavelets for feature extraction from face images, which significantly improved the accuracy and efficiency of face recognition systems.

1994 — Local Binary Patterns (LBP) is a visual descriptor commonly used for classification in computer vision. It was first described in 1994 by Ojala et al. [6] and was a specific case of the texture spectrum model proposed in 1990. It compares the intensity of pixels in a local neighborhood to generate binary codes. LBP is robust to illumination changes and has been successfully combined with other descriptors, such as the Histogram of Oriented Gradients (HOG), for improved performance.

1997 — Elastic Bunch Graph Model (EBGM) is a technique used for recognizing human faces in single images out of a large database that contains one image per person [7]. In EBGM, faces are represented as model graphs, where nodes correspond to local features or key points, and edges represent the relationships between these features.

1998 — Linear Discriminant Analysis (LDA) is a dimensionality reduction technique widely used in machine learning and pattern recognition. Unlike PCA, which focuses on maximizing variance, LDA aims to find the feature subspace that optimally separates different classes or groups in the data. It maximizes the between-class variance while minimizing the within-class variance.

1998 — To encourage collaboration between industry and academia in facial recognition research, the Defense Advanced Research Projects Agency (DARPA) launched the Face Recognition Technology (FERET) program [8]. This program offered the research community a valuable resource: a large and challenging database containing facial images of 850 individuals, totalling 2,400 images.

1999 — The Scale-Invariant Feature Transform (SIFT), developed by David Lowe in 1999, extracted robust features that are largely invariant to changes in scale, illumination, and local affine distortions [9].

1.2.5 2000s

2001 — Viola-Jones face detection [10], named after its inventors Paul Viola and Michael Jones, was a foundational algorithm for real-time face detection. Their approach is based on Haar-like features and the AdaBoost algorithm, enabling real-time detection of frontal-view faces.

2005 — The Face Recognition Grand Challenge (FRGC) [11] was a U.S. government-backed initiative (2004-2006) to encourage and develop face recognition technology. The FRGC provided a platform for researchers and developers to benchmark their algorithms against standardized datasets and evaluation protocols.

1.2.6 2010s

2012 — The seminal paper by Alex Krizhevsky, titled "ImageNet Classification with Deep Convolutional Neural Networks", introduced the groundbreaking Convolutional Neural Network (CNN) architecture known as AlexNet. Published in 2012, this paper revolutionized the field of computer vision and laid the foundation for Deep Learning.

2014 — DeepFace [12] is a deep learning-based face recognition model created by a Facebook research group. They trained a CNN model with 120 million parameters on 4,000 identities across 4 million facial images. The success of DeepFace has catalyzed widespread adoption of facial recognition technology, enabling its integration into diverse applications such as security, surveillance, social media, and biometric authentication.

2015 — FaceNet [13] is also a deep learning based face recognition system developed by a group of researchers affiliated with Google. They Proposed Triplet Loss to train CNN.

2019 — The loss in Arcface [14] is a variant of margin based softmax loss. It is proposed to increase inter-class variance and minimize intra-class variance among the extracted

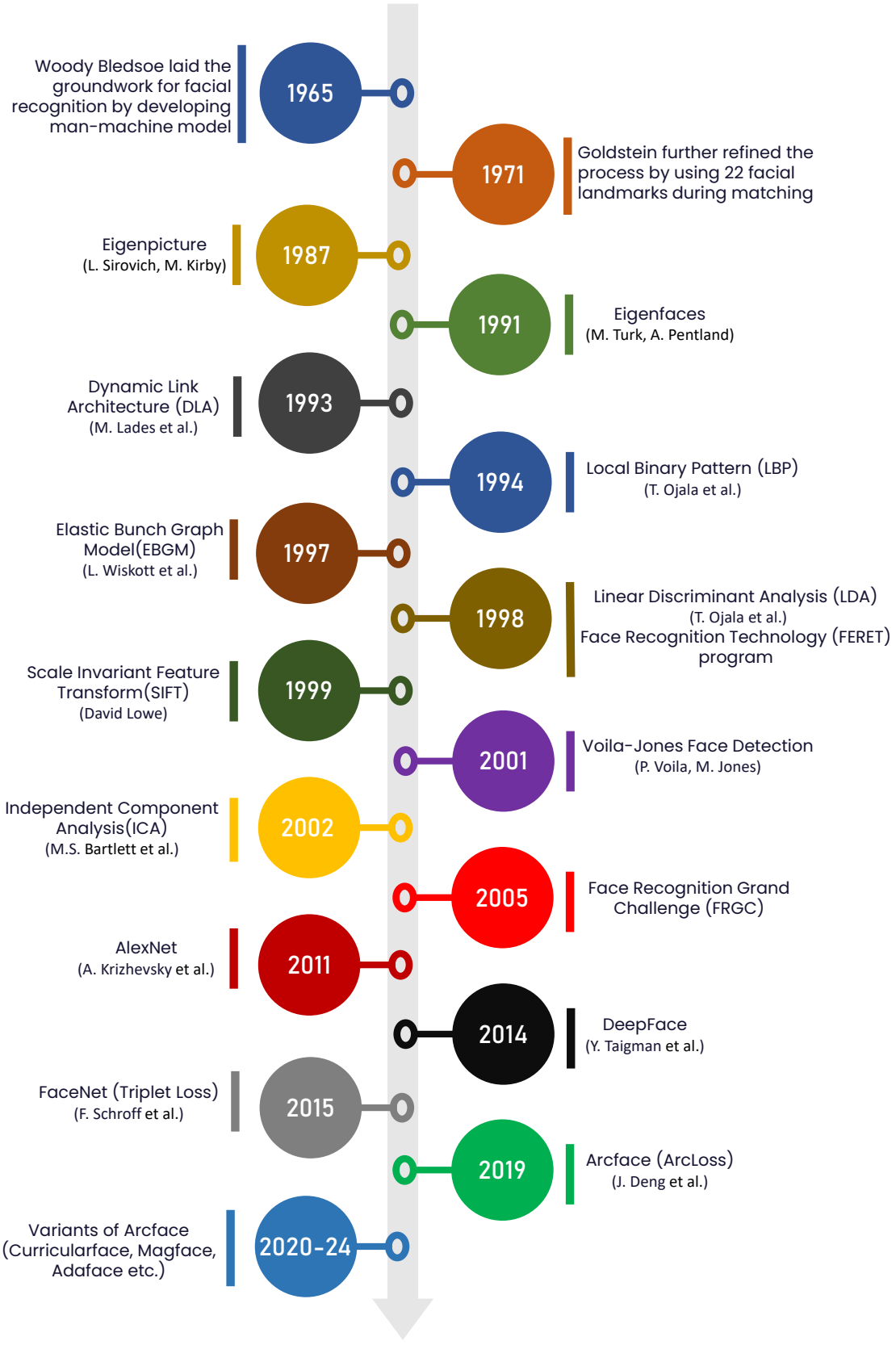


FIGURE 1.1: History of Face Recognition

features. Despite its many variants, Arcface remains a state-of-the-art (SOTA) approach in HR face recognition.

1.2.7 2020s

2020-24 — Various variants of Arcface have been proposed, including Curricularface[15], Magface[16], and Adaface [17] to name a few.

1.3 Basics

1.3.1 Verification vs Identification

Face verification is a one-to-one matching problem. Given two images, we have to determine whether they match or mismatch. In contrast, face identification is a one-to-many matching problem. In this case, a query image is compared with the entire database to find its identity.

1.3.2 Probe Set vs Gallery Set

In a typical face recognition scenario, the probe set, consisting of query images, is compared against the gallery set, which usually contains HR frontal images of known individuals. This comparison aims to find the identity of the individual in the probe image by matching it to the most similar entry in the gallery set.

1.3.3 High Resolution vs Low Resolution

High-resolution and low-resolution images are two terms used to describe the level of detail in a digital image. HR images have a higher number of pixels and contain more detailed information, resulting in sharper and visually richer representations. LR images, however, have fewer pixels and are less detailed, appearing blurry or pixelated. This also applies to facial images. HR facial images capture more fine details about facial features,

FIGURE 1.2: Image resolution worsens from left to right (Leftmost: HR, Rightmost: LR)

expressions, and textures. LR images, with less information about facial features, appear blurry and noisy. In the LR face recognition literature, faces appearing in less than 32×32 pixels are considered LR facial images. Images with different resolutions are shown in Fig. 1.2.

1.3.4 Separable vs Discriminant Features

Separable features are those that are separated by a boundary, while the discriminant features are compactly clustered for each class and separated at some margin from other classes. When learning discriminant features, the goal is to maximize the variance between classes (inter-class variance) while minimizing the variance within each class (intra-class variance). The difference between separable and discriminant features is illustrated in Fig. 1.3. In classification problems, separable features are typically learned through the softmax loss function. However, in face recognition problem, discriminant features are learned through specialized losses like NPT loss and Arcface loss.

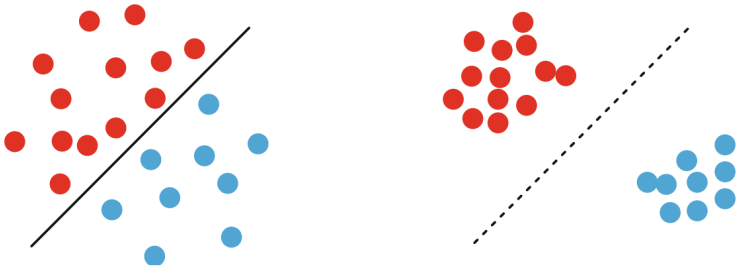


FIGURE 1.3: Separable Features (Left) and Discriminant Features (Right)

1.4 Challenges

Face recognition technology has achieved remarkable progress, but several key challenges limit its performance. These challenges are described below.

1.4.1 Variations in Pose, Illumination and Expression

These are the longstanding challenges in face recognition, commonly known as PIE (Pose, Illumination, and Expression) in the face recognition literature. These factors can significantly alter the appearance of a face, making it difficult for algorithms to match it accurately against reference images. Pose variations can cause different parts of the face to be visible, illumination changes can create shadows and highlights that obscure facial features, and expressions can distort the face's geometry.

1.4.2 Image Resolution

LR images or images captured from a distance can lack the necessary detail to extract discriminative features. Additional factors associated with these images include motion blur, out-of-focus blur, and camera noise. All these factors contribute to poor recognition performance.

1.4.3 Occlusion

Facial occlusion is caused when the part of a person's face is obstructed by objects such as eyeglasses, sunglasses, hats, scarves, or other items placed in front of the face. It disrupts the visibility of facial features that are essential for accurate identification.

1.4.4 Aging

Aging introduces a myriad of changes to human faces that can complicate facial recognition. Deep wrinkles, sagging skin, and alterations in facial features and skin texture are just a few of the transformations that occur over time. These changes can significantly

challenge algorithms, especially when trying to match current images with those from childhood or young adulthood.

1.5 Motivation

Our world is increasingly monitored by surveillance cameras, silently recording our movements and interactions, which results in the capture of a vast amount of visual data. This ever-expanding web of watchful eyes presents both opportunities and challenges in our pursuit of enhanced security and public safety. The quality of faces in these surveillance footage varies greatly. While some footage may capture clear and detailed faces, others may be pixelated, blurry, or lack sufficient resolution for accurate identification. Additionally, factors like varying pose, illumination, and expressions further complicate the recognition process. Face recognition systems designed for HR images struggle to handle these real-world complexities. In this context, integrating an accurate LR face recognition system with surveillance applications could be a game-changer. It has the potential to revolutionize surveillance capabilities, making rapid identification of persons of interest, missing individuals, or suspects.

1.6 Face Recognition Pipeline

The model development for face recognition via deep learning algorithms comprises two phases: training and inference. Training involves preprocessing and model training, during which the network learns to extract discriminative features from facial images. In contrast, inference encompasses preprocessing, feature extraction, and matching, utilizing the trained model to identify faces in new images. The face recognition pipeline is visualized in Fig. 1.4.

1.6.1 Preprocessing

To prepare an image before feeding it to a face recognition model, it undergoes a crucial preprocessing step that involves detecting and aligning the face. Face detection

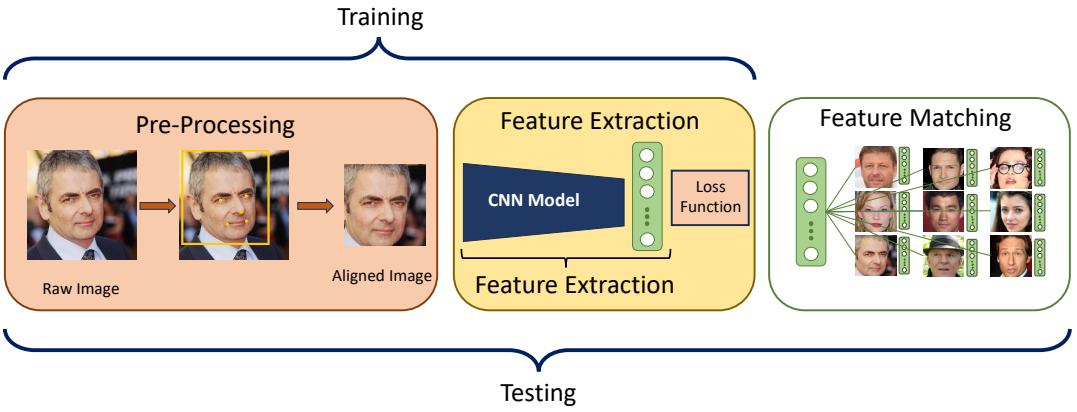


FIGURE 1.4: Face Recognition Pipeline

is performed by face localization algorithms such as MTCNN and RetinaFace, which usually output coordinates of a bounding box and landmarks associated with the face along with the confidence value. Following detection, alignment ensures that the face is positioned consistently and involves rotating the image so that both eyes rest on a horizontal line. To achieve this alignment, the estimated landmarks of the detected face are matched with a set of reference landmarks, guiding the necessary rotation.

1.6.2 Feature Extraction

Feature extraction is pivotal for the efficacy of face recognition systems. During the training stage, the face recognition model, usually a CNN, is trained to learn discriminative features. An extra classification layer and a loss function are employed to train the model. In the testing stage, the classification layer is omitted, and the output features are utilized for matching.

1.6.3 Matching

During the matching stage, the features extracted from probe images are compared with those extracted from gallery set images. This comparison uses cosine similarity, which evaluates the angular distance between feature vectors in the feature space, or Euclidean distance, which computes the straight-line distance between feature vectors in the feature space.

1.7 Conventional Techniques to Deep Learning

The earlier techniques for face recognition relied on hand-crafted features or manual feature engineering techniques. These techniques were suitable for recognizing frontal face images but struggled with changes in pose, illumination, and facial expressions, making them less robust in real-world scenarios. Subspace methods, such as Eigenfaces and Fisherfaces, emerged as an improvement over hand-crafted features. These methods project facial images into a lower-dimensional subspace that highlights the key differences between individuals. While subspace methods achieved some level of success, they still needed to improve their ability to handle significant variations. The introduction of deep learning, particularly convolutional neural networks (CNNs), marked a turning point in the performance of face recognition. The multi-layer structure of deep CNNs excels at automatically extracting intricate features from images, making them ideal for face recognition tasks. The images in the training data are repetitively passed through the CNN, loss is calculated and parameters are updated until the initial layers of the model learn to focus on low-level features, intermediate layers on mid-level features, and the top layer extracts high-level discriminant features. While CNNs are considered data-hungry, the longstanding issues of variations in pose, illumination, and expression can be mitigated to some extent with a large amount of training data. The more the data is enriched with these variations, the more efficient the face recognition model will be under challenging conditions.

1.8 HR Face Recognition Problem

HR face recognition has attracted significant interest from researchers in the past decade and has undergone considerable development in the literature. This can be attributed to advancements in deep learning algorithms, the availability of large-scale training datasets, and challenging testing scenarios for evaluation. Since training datasets for HR face recognition problem consist solely of HR images, the objective becomes maximizing inter-class variance to distinguish between individuals and minimizing intra-class variance to account for variations within a single person's face. Algorithms like Arcface [14] and NPT loss [18] employed this principle and have demonstrated remarkable

performance in recognizing faces from HR images, even with considerable variations in pose, illumination, and expression.

1.9 LR Face Recognition Problem

LR face recognition is a more complex problem compared to HR face recognition. It involves matching images of varying resolutions. Both the probe and gallery sets can potentially contain images of any resolution. LR face recognition is an understudied topic in the literature due to the lack of availability of LR training datasets and the absence of challenging testing scenarios. The training dataset for LR face recognition is generated by augmenting HR datasets with down-sampled LR images. This affects the distribution of training data and poses several challenges. Firstly, the augmentation makes it difficult to distinguish between very LR images of different individuals, leading to high interclass similarity. Second, the varying resolutions within a single person's images contribute to high intraclass variance. These factors cause the losses for HR face recognition problems to struggle in learning discriminative features. They cannot effectively map HR and LR features close together, resulting in degraded performance. Therefore, the primary objective in LR face recognition problem is learning corresponding HR and LR features of the same subject to be close while discriminative from others.

1.10 Face Detectors

1.10.1 MTCNN

MTCNN, or Multi-task Cascaded Convolutional Neural Networks [19], is a face localization algorithm known for its lightweight CNN architecture, enabling real-time performance. It starts with resizing the input image to various scales, creating an image pyramid. This pyramid is then fed into a three-stage cascaded framework for extracting the bounding box and facial landmarks. The overview of the three-stage cascaded network is shown in Fig. 1.5.

Stage 1: In the first stage, Proposal Net (P-Net) generates candidate bounding boxes that might contain faces. It also predicts bounding box adjustments (regression vectors) to improve accuracy. Highly overlapping candidates are merged using a technique called non-maximum suppression (NMS).

Stage 2: In the second stage, all candidate boxes from the previous stage are passed through the Refinement Net (R-Net). The R-Net further eliminates a significant portion of false detections. It refines the remaining proposals through bounding box regression and NMS merging.

Stage 3: The final stage utilizes a more powerful CNN, the Output Net (O-Net), to achieve even greater accuracy. It refines the bounding boxes from the previous stage and outputs the positions of key facial landmarks along with confidence value.

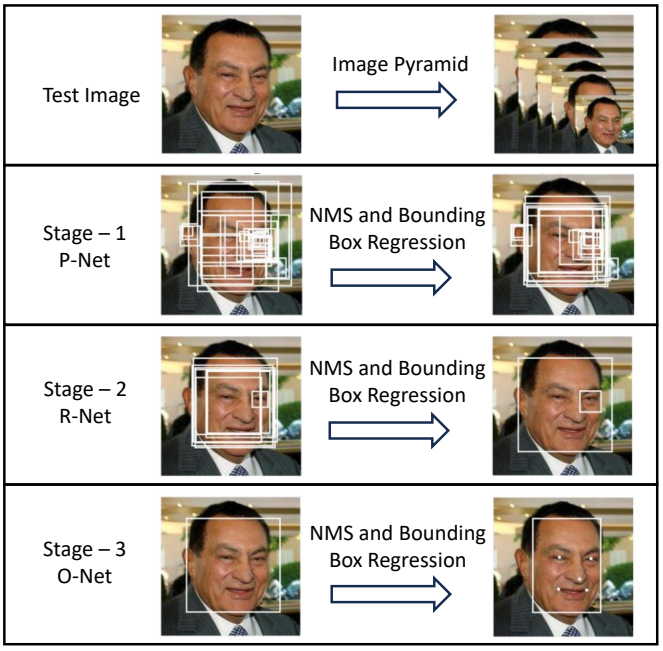


FIGURE 1.5: Pipeline of Cascaded Framework in MTCNN [19]

1.10.2 Retinaface

Retinaface [20] is a robust single-stage face detection algorithm designed for accurate and efficient face localization in challenging real-world scenarios. It can simultaneously predict the face score, bounding box around the face, five facial landmarks and even 3D

shape information. Retinaface has demonstrated the impact of face detection and alignment on the performance of deep face recognition. The results show that faces aligned by Retinaface exhibit improved recognition performance compared to those aligned by MTCNN. The backbone architecture of Retinaface is available in both ResNet and MobileNet architectures. Retinaface leverages three main components visualized in Fig. 1.6.

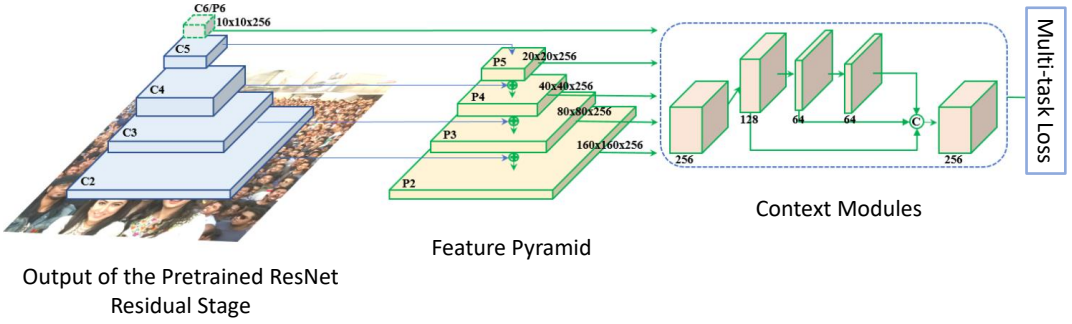


FIGURE 1.6: Pipeline of Retinaface Architecture [20]

Feature Pyramid: This component takes the input image and generates five feature maps, each representing a different scale. This allows Retinaface to detect faces of varying sizes within the image.

Context Modules: Independent context modules are employed on each of the five feature pyramid levels. These modules increase the receptive field, allowing the network to see a larger area around a specific point in the image. This enhances the ability of the model to capture contextual information and improve the accuracy of rigid face detection (e.g., non-deformed faces).

Loss Head: For negative anchors, only classification loss is applied. For positive anchors, the proposed multi-task loss is calculated. The multi-task loss consists of following components:

1. Classification loss, i.e., Cross entropy loss for predicting face and not face
2. Bounding box regression loss
3. Facial landmarks regression loss
4. Dense regression loss for predicting 3D information.

1.11 Backbone Architectures

This section discusses the following widely used CNNs as backbone architectures for face recognition:

1. GoogleNet
2. ResNet
3. SENet

1.11.1 GoogleNet

In 2014, a research team at Google introduced a groundbreaking deep CNN architecture known as GoogleNet, or Inception v1 [21]. This architecture achieved top results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. GoogleNet deviates from conventional CNN architectures by incorporating a novel building block termed the Inception module. This module facilitates the efficient allocation of computational resources while simultaneously achieving exceptional performance. The key features that distinguish GoogleNet are:

Multiple Paths with Varying Filter Sizes: The Inception module utilizes several parallel convolutional layers with different filter sizes (1x1, 3x3, 5x5) within a single block. This allows the network to capture features of various scales and complexities simultaneously.

1x1 Convolution for Dimensionality Reduction: The Inception module also incorporate 1x1 convolution layers. These layers reduce the dimensionality of the input data before feeding it to the next convolutional layers while also reducing the number of parameters and computational costs without sacrificing accuracy.

Following the groundbreaking success of GoogleNet (Inception v1), researchers continued to refine the Inception architecture and developed Inception v2, v3, and v4. These subsequent versions introduced several innovative techniques, including factorized convolution, batch normalization, and residual connections within the inception module. These enhancements significantly boosted the performance, enabling the models to achieve even higher accuracy and robustness in various computer vision tasks.

1.11.2 ResNet

ResNet, short for Residual Network, is a deep learning architecture designed for computer vision tasks by a research group at Microsoft [22] in 2015. They addressed the challenge of vanishing gradients, which hindered training very deep CNNs. Unlike traditional CNNs that stack convolutional layers directly, ResNet introduces a concept called residual blocks. These blocks contain skip connections that bypass the convolutional layers within the block, as shown in Fig. 1.7. The output of these bypass connections is added element-wise to the output of the convolutional layers. This design allows the network to learn the identity function (simply copying the input) along with more complex transformations. ResNet comes in various configurations, with ResNet-50 and ResNet-101 being popular choices in face recognition task. These variants differ only in the number of convolutional layers.

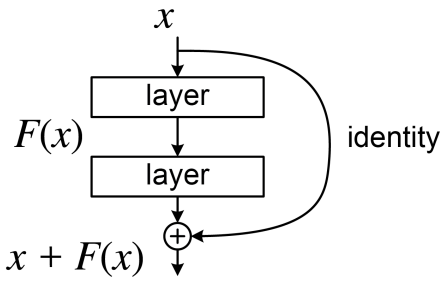


FIGURE 1.7: Residual Block in ResNet Architecture

1.11.3 Squeeze-and-Excitation Networks (SENet)

Squeeze-and-excitation Network (SENet) [23] is another advancement in CNN architecture. Introduced in 2017, it addressed the challenge of effectively utilizing feature channels within CNNs. SENet introduced a novel building block called the Squeeze-and-Excitation (SE) block. This block is inserted within the existing CNN architecture, typically after each block of convolutional layers. The SE block performs two fundamental operations:

Squeeze: This operation aims to capture global feature information across each channel. It typically uses spatial pooling (like average pooling) to squeeze the feature maps into a single value per channel.

Excitation: This operation refines the importance of each channel based on the information gathered by the squeeze operation. It often involves a small neural network that analyzes the channel-wise information and generates per-channel scale factors. These factors then modulate the original feature maps, emphasizing informative channels and suppressing less important ones.

SE blocks enable the network to adaptively learn feature importance, leading to a more robust representation of the input data. SENet has demonstrated significant accuracy improvements in various computer vision tasks like image classification and object detection. SE blocks are lightweight and easily integrated into existing CNN architectures with minimal computational overhead. The SE block is shown in Fig. 1.8.

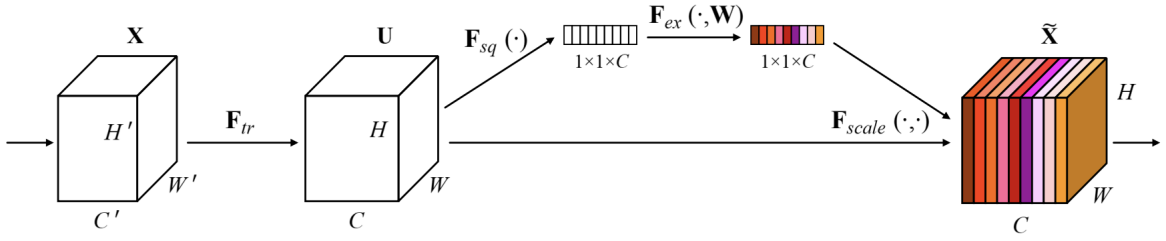


FIGURE 1.8: Squeeze and Excitation (SE) Block. [23]

1.12 Thesis Contribution

The following contributions are made in this thesis:

1. A LR face recognition system is developed that targets very LR images, especially from surveillance scenarios. It comprised two key components: a Degradation Model and Attention Guided Distillation. Due to the lack of real-world LR facial data, the Degradation Model simulates real-world surveillance effects in downsampled LR images, generating synthetic LR facial data. The Attention-Guided Distillation approach transfers informative HR features from an HR teacher network to an LR student network. This is achieved by leveraging not only the final layer but also the intermediate convolutional layers, utilizing spatial attention maps to guide the process.

2. A LR face recognition system is developed that can handle both HR and LR images. The system utilizes two key components: Sub-center Learning and Contrastive Distillation Loss. Sub-center Learning captures diverse image representations across varying resolutions through multiple sub-centers defined for each class. The Contrastive Distillation Loss pushes the LR and HR features close to each other in contrast to other non-corresponding features to learn compact and discriminant features.

3. Existing testing protocols are analyzed, and their limitations are highlighted with empirical evidence. New protocols are then developed to more effectively assess both the efficacy and limitations of the LR face recognition model.

1.13 Thesis Organization

The thesis is organized as follows:

Chapter 2 presents a detailed survey of face recognition algorithms. It covers both conventional methods and deep learning approaches. Deep learning algorithms are further categorized into those designed for HR and LR images. Margin-based softmax losses are the prominent methods for HR face recognition, while knowledge distillation techniques are successful in LR face recognition.

Chapter 3 presents a discussion of the datasets and evaluation methods. The datasets are categorized into those used for training the face recognition model and those used for testing its performance. The testing datasets are further divided into LR, HR, and Mixed Resolution (MR) datasets.

Chapter 4 presents the proposed scheme for LR face recognition that tackles the challenges of real-world surveillance scenarios. The system exhibits robustness against various types of degradation commonly encountered in surveillance footage, including blurriness, noise, and compression artifacts. It effectively handles identification for both near and distant images captured by surveillance cameras.

Chapter 5 presents the new protocols for LR face recognition. Since LR face recognition involves matching images of varying resolutions, including both HR and LR, it

is necessary to revise the previous protocols and assess their performance on both HR and LR images.

Chapter 6 presents the proposed LR face recognition scheme that effectively handles variations within LR images while maintaining performance on HR images. The scheme utilizes multiple sub-centers defined for each class, allowing them to learn the inherent variations in input images. Additionally, contrastive learning plays a key role in mapping HR and LR features closer together. To assess the effectiveness and limitations of this proposed scheme, it is tested on multiple LR and HR benchmark datasets.

Chapter 7 presents the conclusion and future work.

1.14 Summary

This chapter provides an overview of face biometric systems. It begins with the history of face recognition systems, followed by a discussion of basic terminologies. It then explores the challenges in this field and the motivation for studying this topic. Since face recognition encompasses two parallel research areas – HR and LR face recognition – both problems are described. The chapter also discusses famous backbone architectures and widely used face detectors.

Chapter 2

Literature Review

2.1 Introduction

Face recognition has emerged as one of the most active and significant research areas in computer vision and pattern recognition. The journey of face recognition technology began several decades ago, with early efforts primarily focused on measuring distances between key facial landmarks. As technology progressed, the field shifted towards appearance-based methods, which leveraged statistical techniques to represent and analyze facial images. The introduction of Principal Component Analysis (PCA) and the development of the Eigenface method in the early 1990s represented a major breakthrough. These methods reduced the dimensionality of facial images and allowed for more efficient and robust recognition.

The advent of machine learning and, more recently, deep learning has propelled face recognition into a new era. Neural networks, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in handling the complexities of face recognition, achieving high accuracy even in challenging scenarios involving variations in pose, lighting, and expression.

This chapter provides a comprehensive literature survey of face recognition techniques, specifically focusing on LR face recognition. The categorization of face recognition techniques is shown in Fig. 2.1.

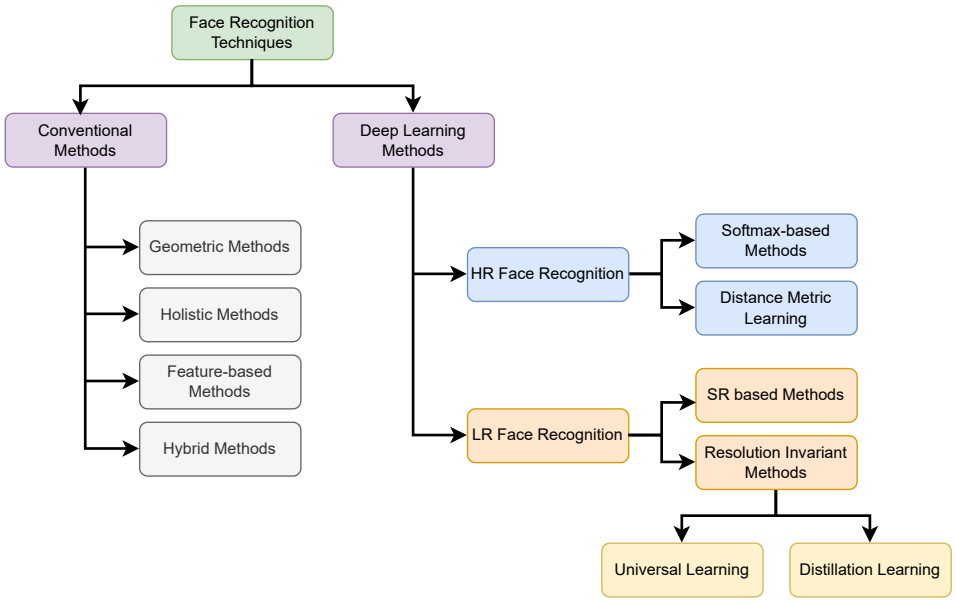


FIGURE 2.1: Taxonomy of Face Recognition Techniques

2.2 Conventional Methods

2.2.1 Geometric Approaches

Geometry-based methods are one of the foundational approaches in face recognition. These methods focused on analyzing the spatial relationships and distances between key facial features like eyes, nose, mouth, and chin. One of the earliest attempts was performed by Bledsoe [1], who developed a semi-automated system to classify photos of faces by computing distances between key facial landmarks. In the 1970s, Goldstein, Harmon, and Lesk [2] extended the range of facial features used for identification by describing a system that measured 21 specific features, such as the width of the mouth and the distance between the eyes. These comprehensive measurements were then utilized to classify and recognize faces.

Kelly's PhD thesis [24] is also considered a pioneering effort to explore the feasibility of using computers to automatically identify individuals based on their facial features. Kelly also focused on extracting geometric features from facial images to automatically identify individuals with a reasonable degree of accuracy, given the computational limitations of that time.

2.2.2 Holistic Approaches

In the holistic approach, the face is projected onto a low-dimensional space, discarding superfluous detail. This allows for the consideration of the entire face for matching purposes. The main challenge in this approach was the complexity and multi-dimensionality of faces.

Turk and Pentland [4] proposed a face recognition method using Principal Component Analysis (PCA). In PCA, eigenvectors are calculated to account for the largest variance in data distribution. Eigenvectors associated with the largest few eigenvalues have a face-like image; therefore, they are also known as eigenfaces. Traditional PCA worked in the original input space, which may not have been ideal for capturing complex relationships in the data. Kernel PCA [25], an extension of traditional PCA, operated in a high-dimensional feature space. Here, kernel functions are employed to project the data onto a higher dimensional feature space, enabling linear analysis using PCA. Independent Component Analysis (ICA) extended this concept by considering higher-order dependencies, thereby facilitating the capture of more complex relationships [26]. One of the disadvantages of PCA was that it maximized variance across all the images, including intra-person variations that were not relevant for recognition tasks.

Fisherfaces [27] aimed to find a low-dimensional subspace that separated different face classes effectively, even under severe variations in lighting and facial expressions. The Linear Discriminant Analysis (LDA) technique in Fisherfaces used class targets to find a projection matrix that maximized intra-class variance and minimized inter-class variance. Therefore, LDA was considered a better technique than PCA. However, LDA was prone to overfitting when the input data had limited samples per class, as it did not exploit inter-class variance. Support Vector Machine (SVM) in face recognition was also considered a holistic approach. While face recognition is inherently a multi-class problem (identifying a specific person from a set of individuals), SVM is typically a binary classifier. To address this, P. J. Phillips [28] reformulated the representation of facial images as a two-class problem. Instead of treating each image independently, the author worked in a "difference space". Here, the focus was on two classes: (1) dissimilarities between images of the same person and (2) dissimilarities between images of different people.

2.2.3 Feature-based Methods

Feature-based methods refer to methods that leverage local features extracted at different locations in a face image. An extension of the original eigenface technique employed a modular approach [29] that focused on specific facial features (eyes, nose, mouth) rather than the entire face. In this approach, Principal Component Analysis (PCA) was independently applied to different local regions in the face image to produce sets of eigenfeatures.

Another prominent feature-based method for recognizing human faces from single images within a large database was elastic bunch graph matching (EBGM) [7]. This technique builds upon the concept of dynamic link architecture introduced in [5]. In EBGM, faces are represented as model graphs, where nodes correspond to local features or keypoints, and edges represent the relationships between these features. The nodes contain Gabor wavelet coefficients extracted around a set of predefined facial landmarks. A variant of this method employs Histograms of Oriented Gradients (HOG) [30, 31] instead of Gabor wavelet features.

The Local Binary Pattern (LBP) [32] is a simple yet effective texture descriptor that summarizes the local structure in an image by comparing each pixel with its neighbors. The technique has become popular due to its robustness to changes in illumination and its computational efficiency. The method involves dividing the face image into multiple regions, from which LBP feature distributions are extracted. These distributions are then concatenated to form a global feature vector representing the entire face image. Many variations of this method have been proposed to improve face recognition accuracy, such as LBP descriptors extracted from Gabor feature maps [33], rotation-invariant LBP descriptors [34] and local derivative patterns (LDP) to extract high-order local information by encoding directional pattern features [35].

Scale-invariant feature transform (SIFT) [9] is a powerful feature matching technique that has found widespread use in computer vision, including face recognition. It identifies and describes local features in images that are invariant to scale, rotation, and affine transformations. These features are highly distinctive and can be reliably matched between different images of the same object or scene.

2.2.4 Hybrid Methods

Hybrid methods combined the strengths of holistic techniques, which consider the entire face image, with feature-based techniques, which focus on specific facial features. Researchers have proposed various hybrid methods that leverage Gabor wavelets alongside subspace methods. In [36], an enhanced Fisher linear discriminant model is proposed. It combines the discriminative power of Fisher's criterion with the robustness of Gabor features. The use of Independent Component Analysis (ICA) with Gabor features is investigated in [37]. ICA-based Gabor features improved the robustness of face recognition systems. The benefits of combining Gabor features with kernel-based techniques are demonstrated in [38]. By enhancing the kernel function with fractional power polynomial models, nonlinear relationships between features were captured. Similarly, the LBP descriptor also proved to be very successful in hybrid models.

In [39], the author applied LDA to multi-scale LBP histograms. The goal was to enhance face recognition performance by capturing both local texture information and global shape attributes. The application of this technique to color images was explored in [40], where LBP histograms were separately computed for each color channel. These multi-spectral LBP captured both texture and color information. Laplacian PCA proposed in [41] extends the traditional PCA by incorporating local neighborhood information. It aims to capture both global and local structures in high-dimensional data. Hybrid methods offer the best of holistic and feature-based methods. Before the advent of deep learning, most SOTA face recognition systems were based on hybrid methods.

2.3 Deep Learning based Methods

Deep learning has revolutionized face recognition, achieving superior performance compared to conventional methods. Broadly, deep learning approaches for face recognition are classified into HR and LR face recognition.

2.3.1 HR Face Recognition

There are two main research approaches for training deep CNNs for HR face recognition. In the first approach, a multi-class classifier is trained to learn a representation of faces.

This involves categorizing faces into multiple classes, allowing the network to distinguish between different individuals effectively. The second approach involves distance metrics to learn embedding directly from the data.

2.3.1.1 Softmax-based Methods

The softmax loss function is the most widely used loss function in multi-class classification problems. It consists of two main components: the softmax probability layer and the cross-entropy loss function. The softmax probability layer converts the raw output scores (logits) from the last layer into probabilities by applying the softmax activation function. Following the softmax layer, the cross-entropy loss function measures the difference between the predicted probability distribution and the true distribution. Deepface [12] is one of the initial attempts to solve the face recognition problem using deep learning by a research group at Facebook. A 9-layer deep CNN having locally connected layers is trained on 4 million images with 4000 identities using the softmax loss function. Mathematically, the softmax loss function is represented by:

$$\mathcal{L}_{softmax} = -\log\left[\frac{\exp(W_{y_i}^T x_i + b_i)}{\sum_{j=1}^{N}\exp(W_j^T x_i) + b_j}\right], \tag{2.1}$$

where $x_i \in \mathbb{R}^d$ denotes the features of the $i-th$ sample, belonging to the $y_i - th$ class. $W_j \in \mathbb{R}^d$ denotes the $j-th$ column of the weight $W \in \mathbb{R}^{d\times N}$, $b_j \in \mathbb{R}^N$ is the bias term, and N is the total number of classes.

After that, a series of papers known as DeepID were proposed and enhanced the face recognition performance on several benchmarks i.e., LFW and Youtubefaces. In DeepID [42], the representation is learned through a 4-layer multi-scale CNN on the Celebface dataset in a supervised fashion. DeepID2 [43] introduced identification and verification signals in the loss function. Identification is achieved using softmax loss function to increase inter-personal variations, while verification is a contrastive loss function to reduce the intra-personal variations and is represented by:

$$\mathcal{L}_{verification} = \begin{cases} \frac{1}{2}\|x_i - x_j\|_2^2 & y_{ij} = 1 \\ \frac{1}{2}\max(0, m - \|x_i - x_j\|_2^2) & y_{ij} = -1 \end{cases}, \tag{2.2}$$

where x_i and x_j are the feature vectors extracted from the two face images in comparison. $y_{ij} = 1$ means that x_i and x_j are from the same identity and $y_{ij} = -1$ means different identities. m is the margin parameter, and the distance between images should be larger than it. DeepId2+ [44] introduced a fully connected layer to intermediate convolutional layers of the network. Loss is then calculated from the intermediate layers in addition to the last layer.

Standard softmax loss does not induce any explicit margin between multi-class features. Therefore, the learned features are separable only but not discriminative enough. As discussed in Chapter 1 (Section 1.3.4), discriminative features have large intra-class variance and small inter-class variance. Introducing the margin parameter in the softmax loss function enhances the discriminative power of the learned features. The primary motivation behind margin-based softmax loss functions is to improve the model's ability to distinguish between different classes by increasing the margin between the learned features of different classes. This leads to more robust and reliable face recognition performance.

The Large Margin Softmax Loss (LMCL) [45] was the first to induce margin in the softmax loss function. The features and weights are normalized to lie in the cosine space, then a multiplicative margin of m is introduced. Sphereface [46] further demonstrates the problem and conducted experiments on facial databases. The derivation of Sphereface loss is presented below. Starting from the classification layer, which is represented as:

$$z_i = W^T x_i, \tag{2.3}$$

where $x_i \in \mathbb{R}^d$ denotes the $i{-}th$ the feature vector of dimension d. $W \in \mathbb{R}^{d \times N}$ is the weight matrix while N is number of classes. Normalizing the weights and features results in:

$$z_i = \hat{W}^T \hat{x}_i, \tag{2.4}$$

$$z_i = \cos\theta_i, \tag{2.5}$$

where $\cos\theta_i$ consists of positive and negative cosine similarities. $\cos\theta_{y_i}$ is the positive cosine similarity of $y_i - th$ class while all others are negative cosine similarities represented by $\cos\theta_j$ and $j \neq y_i$. The final equation for loss in Sphereface is given by:

$$\mathcal{L}_{sphereface} = -\log\left[\frac{\exp(s\cos(m\theta_{y_i}))}{\exp(s\cos(m\theta_{y_i})) + \sum_{j=1,j\neq y_i}^{N}\exp(s\cos\theta_j)}\right]. \tag{2.6}$$

In Cosface [47], the authors claimed that the margin in the Sphereface was not consistent over all the value of the angle; therefore, an additive angular margin is introduced, which is more robust and consistently applied to all the samples. The loss function in Cosface is given by:

$$\mathcal{L}_{cosface} = -\log\left[\frac{\exp(s(\cos\theta_{y_i} - m))}{\exp(s(\cos\theta_{y_i} - m)) + \sum_{j=1,j\neq y_i}^{N}\exp(s\cos\theta_j)}\right]. \tag{2.7}$$

Arcface [14] introduced additive cosine margin and claimed that it was a constant linear angular margin in contrast to Sphereface and Cosface, which were non-linear angular margins. The loss function in Arcface is given by:

$$\mathcal{L}_{arcface} = -\log\left[\frac{\exp(s\cos(\theta_{y_i} + m))}{\exp(s\cos(\theta_{y_i} + m)) + \sum_{j=1,j\neq y_i}^{N}\exp(s\cos\theta_j)}\right]. \tag{2.8}$$

Margin-based softmax losses discussed above used fixed margins for all samples. It does not explicitly emphasize each sample according to its importance. In other words, both easy and hard samples are weighted equally. This can lead to convergence issues during training, especially when using small backbone architectures like MobileFaceNet. To address this, the concept of hard mining was developed, which led to the proposal of mining-based loss functions. The basic principle of mining-based losses was to modulate the negative cosine similarities, giving higher weight to hard samples. MKV-arc-Softmax [48] modulated the negative cosine similarity by giving higher emphasis to the hard samples using a manually defined constant. In contrast, Curricularface [15] proposed an adaptive curriculum learning strategy that emphasized easy samples relative to the hard samples in the initial stage of training. Once easy samples are correctly classified, it assigns higher weightage to the hard samples according to their difficulty. The loss function in the Curricularface is given by:

$$\mathcal{L}_{curricularface} = -\log\left[\frac{\exp(s\cos(\theta_{y_i} + m))}{\exp(s\cos(\theta_{y_i} + m)) + \sum_{j=1,j\neq y_i}^{N}\exp(sN(t,\cos\theta_j))}\right]. \tag{2.9}$$

where

$$N(t, cos\theta_j) = \begin{cases} \cos\theta_j & \cos(\theta_{y_i} + m) \geq \cos\theta_j \\ \cos\theta_j(t + \cos\theta_j) & \cos(\theta_{y_i} + m) < \cos\theta_j \end{cases}.$$

2.3.1.2 Distance Metric Learning

Softmax-based loss functions necessitate the inclusion of a classification layer atop the representation layer, which is trained on known identities. This additional classification layer is often viewed as an extra overhead and can be eliminated through the use of distance metric losses. Distance Metric Learning aims to create an object embedding space that aligns with semantic similarity, ensuring that objects of similar classes are positioned closer together, while objects of different classes are spaced further apart. This approach enhances the model's ability to distinguish between different classes without the need for an extra classification layer.

In 2015, researchers affiliated with Google proposed a face recognition model known as FaceNet [13] that directly learns a mapping from input data in an Euclidean space. In FaceNet, Triplet loss was proposed, which input a pair of matching face images and a non-matching one. The loss function enforced the minimization of the distance between matching pairs and the maximization of the distance between non-matching pairs. Although the loss function could directly learn embedding, it was difficult to optimize the parameter of the network due to the high dependence on mining useful triplets from the massive training data. Mining hard triplets is computationally expensive and their absence can lead to convergence issues. The Triplet loss is represented by:

$$\mathcal{L}_{triplet} = \max\left[d(x_i^a, x_i^p) - d(x_i^a, x_i^n) + m, 0\right] \tag{2.10}$$

where x_i^a is the feature vector of any input image (i.e., anchor image), x_i^p is the feature vector of the anchor positive image (i.e., matching image with input), x_i^n is the feature vector of the anchor negative image (i.e., non-matching image with the input) and m denotes the margin. $d(.)$ corresponds to the Euclidean distance.

To accelerate convergence speed and relax the requirement for mining hard triplets, Movshovitz-Attias et al. [49] proposed a distance metric using proxies. This Proxy-based loss compute class representations during training. These class representations

serve as a proxy for the mean representation of each class. Every sample is pushed towards the corresponding class proxy and away from the non-corresponding class proxy. However, the Proxy-based loss solves the convergence issue, it can only leverage data-proxy relations instead of rich data-to-data relations. This vanilla proxy loss is modified in [50], and proxy-anchor loss is proposed. This loss utilizes both data-to-data and data-to-proxy relationships during training. Margin-based softmax losses dominated distance metrics until the Nearest Proxies Triplet (NPT) loss [18] was proposed. NPT loss maximizes information between each sample and its corresponding class proxy, in contrast to the nearest negative proxy, and has achieved comparable results to margin-based softmax loss. NPT loss is simple to implement and doesn't require any hyper-parameter tuning like margin-based softmax loss. The NPT Loss is given by:

$$\mathcal{L}_{NPT} = \sum_{k} \max\left[d(x_i, W_i) - d(x_i, W_k^{(i)}) + m, 0\right] \tag{2.11}$$

where W_i is the corresponding class proxy of $i{-}th$ feature vector x_i, $W_k^{(i)}$ denotes the nearest negative proxy against the $i{-}th$ feature vector. k corresponds to top-k nearest negative proxies.

2.3.2 LR Face Recognition

Compared to HR face recognition, LR face recognition has received less attention in the literature. The methods and techniques proposed for LR face recognition can be broadly categorized into Super Resolution (SR) based approaches and Resolution Invariant (RI) approaches.

2.3.2.1 Super Resolution based Approaches

In super-resolution-based approaches, HR images are synthesized from LR images and then recognition is performed on the estimated HR images. The challenging aspect of SR-based approaches is retaining the identity features necessary for accurate identification in the estimated HR images. The authors in [51] addressed this issue by proposing a framework that utilized a super-identity loss and a domain-integrated training approach known as Super-Identity Convolutional Neural Network (SICNN). This framework employed two cascaded networks: CNN_H for hallucination and CNN_R for identity

recognition. The super-identity loss minimized the distance between the original and estimated HR images in the feature space, ensuring that identity features were retained. The domain-integrated training approach addressed the challenge caused by the significant gap between the HR and the hallucination domain that can lead to a domain divergence problem. In this approach, CNN_R is first trained on HR images using recognition loss, while CNN_H is trained on HR and corresponding LR images using pixel-wise Euclidean loss. Then, CNN_R and CNN_H are alternatively fine-tuned in each iteration. First, CNN_R is updated using the recognition loss. Then, CNN_H is jointly updated using the pixel-wise Euclidean loss and super-identity loss. The loss function in sphereface [46] is employed for identity recognition.

SR-based face recognition relies on paired data (HR and corresponding LR images) to estimate HR images from LR ones. This paired data is typically generated by artificially creating LR counterparts from HR images. However, the corresponding HR images are unavailable in case of native LR images. Consequently, the training stage often involves only synthetically generated paired data, neglecting the valuable information present in the native LR images. To address this challenge and enhance both visual quality and identity consistency, Zheng et al. [52] proposed the Complement Super-Resolution and Identity (CSRI) learning mechanism. This framework consisted of two separate cascaded networks that shared parameters. Each cascaded network comprised a super-resolution sub-network and a face recognition sub-network. In the first cascaded network, the authors adopted joint learning of super-resolution and identity recognition using synthetically generated LR and HR data. The second cascaded network focused on complement-super-resolution learning. Since native LR data lacked corresponding HR images, the network leveraged the parameters inherited from the joint learning stage on synthetic data and performed super-resolution and identity recognition on native LR data.

Synthetically generated paired data has perfect pixel-to-pixel correspondence between the LR and HR images. However, for native LR data, the corresponding HR images may be available in some cases, it lacks perfect pixel-to-pixel correspondence. Real-world factors like pose variations, lighting changes, and occlusions further cause discrepancies between the LR and HR versions. The Feature Adaptation Network (FAN) [53] addressed this challenge and performs surveillance face recognition along with normalizing

the faces. This was achieved by employing a feature disentanglement technique along with feature and image-level supervision. In the first step, two encoders were trained using adversarial training. These encoders were trained to separate the identity features from the non-identity features. Specifically, *Enc_H* extracted identity features from HR images, while *Enc_Z* extracted non-identity features from HR images. A decoder, denoted as *Dec*, then reconstructed the original input image by combining these representations. The decoder network was trained to ensure the reconstructed image closely resembled the original. Next, feature adaptation took place. A separate LR encoder, *Enc_L*, was trained for unpaired/paired data. This *Enc_L* leveraged the feature and image-level supervision signals obtained from the previously trained encoders and decoders. In addition, FAN utilized a random scale augmentation (RSA) technique. This technique involved generating down-sampled LR images of random sizes to learn resolution-robust identity features.

Most SR-based techniques for LR face recognition typically address enhancing quality of LR image or directly normalizing the faces. VividGAN [54], on the other hand, proposed a two-step process for generating HR frontal faces from LR images. This process involved super-resolution of the input image followed by face frontalization. This sequential approach significantly reduced unwanted blur and artifacts. Within the VividGAN framework, two key components were crucial: Coarse-level and Fine-level Face Hallucination (FH) Networks, along with their corresponding discriminators. The coarse-level network performed both face super-resolution and frontalization in a unified manner. Subsequently, the fine-level network further refined the initial, coarsely generated HR frontal face images. By employing two-level discriminators, VividGAN enforced photo-realism by focusing on both local and global details of the image. This strategy led to visually superior results.

A significant challenge for super-resolution models lies in the vast difference between the source domain data (down-sampled LR images) and the target domain data (native LR images). The down-sampled LR data lacks the natural variations in pose, noise, and blurring present in native LR data. To address this issue, the Dual Domain Adaptive Translation Network (DDAT) [55] utilized a novel dual domain adaptive translation structure. This structure comprised two key modules: an adaptive adversarial module and an anti-perturbation classifier module. The adaptive adversarial module, essentially

a Generative Adversarial Network (GAN), aimed to bridge the gap between down-sampled and native LR images. Its discriminator network attempted to distinguish between real and generated images, guiding the generator to produce target domain images that closely resembled those in the source domain. The anti-perturbation classifier module served the purpose of improving accuracy not only on the target domain but also on the source domain. To achieve this, DDAT employed three types of losses: 1) Consistency loss preserved the identity information between the super-resolution images and the HR images in the source domain 2) Anti-perturbation loss enhanced the overall robustness of the module 3) Classification loss ensured that the identity information remained consistent between the input fed into the generator and the output produced by the classifier.

The main advantage of SR techniques lies in their ability to estimate HR images from LR ones. However, these techniques struggle to generate HR images and retain identity information when the pose variations become large. Moreover, from a recognition perspective, SR-based approaches are computationally expensive compared to resolution-invariant methods. SR methods first need to upsample the image to achieve high quality for recognition. Beyond that, some methods first extract discriminant features from LR images and then generate HR images. In both the techniques, the identification and verification results reported in the literature indicate that SR-based approaches underperform compared to resolution-invariant approaches. The possible reason is that the loss function in SR-based approaches optimizes the model not only for extracting discriminative features but also for extracting features that are further used for super-resolution. While in resolution-invariant approaches, the model is optimized only for extracting discriminative features. A comparative analysis of SR approaches is presented in Table 2.1.

2.3.2.2 Resolution Invariant Approaches

In resolution-invariant methods, the distance between the features extracted from HR and corresponding LR images is minimized in a lower-dimensional feature space. As a result, features are learned in a common feature subspace. This can be achieved in two ways. The first method involves simultaneously training a CNN on both HR and

TABLE 2.1: Comparative Analysis of SR based Approaches

Methods	**Losses**	**Limitations**
SICNN [51]	• Super Identity Loss • Sphereface • Pixel-wise Euclidean Loss	• Identity features assessment on LR benchmarks is not performed. • SR results are shown on down-sampled LR images
CSRI [52]	• Cross Entropy Loss • Pixel-level MSE Loss	• The model has only been validated on a single LR dataset, i.e, Tiny-face. • SR results are evaluated only on frontal-pose images.
FAN [53]	• CE Loss • Disentanglement Loss • Adversarial Loss • MSE for feature-level and image-level super-vision	• Complex training methodology. • Identity changes during normalization in the case of extreme pose images. • Low recognition performance.
Vivid GAN [54]	• Mirror Symmetry Loss • Pixel-wise Similarity Loss • Feature-wise Identity Similarity Loss • Structure-wise Similarity Loss • Cross Entropy Loss • Adversarial Loss	• SR results deteriorate with very LR images and varying pose images. • Identification or verification performance on LR datasets is not evaluated.
DDAT [55]	• KL divergence Loss • Adversarial Loss	• SR results are evaluated only on down-sampled LR images

LR images, which we call Universal Learning. The second method, called Distillation Learning, utilizes knowledge distillation techniques.

2.3.2.3 Universal Learning

A study by Zeng et al. [56] trained a 10-layer deep CNN simultaneously on HR and down-sampled LR images in a supervised manner. This approach aimed to learn resolution-robust features within a unified feature space. The authors validated their method on datasets containing uncontrolled scenarios and LR images, such as SCface

and COXface. This validation demonstrated significant improvements over conventional methods. Deep-coupled Resnet [57] introduced branch networks to explicitly minimize the distance between HR and LR features. This approach consisted of one trunk network and two branch networks. The trunk network is a 27-layer ResNet architecture trained on three different image resolutions in a supervised manner. Once the trunk network was trained, two branch networks were attached: one for HR images and another for LR images. These branches aimed to minimize the difference between HR and LR image features in the feature space using a coupled-mapping loss function. This couple-mapping loss function consisted of a Softmax loss and Center loss for each branch, and an Euclidean loss between the branches.

In [58], an augmentation technique for generating synthetic LR images, ensembled CNN, and regularized Triplet loss is proposed to learn discriminative embeddings simultaneously from HR and LR images. Synthetic LR images were generated by introducing degradations through averaging, out-of-focus, and motion blur filters. Ensembled CNN consisted of a trunk and branch network whose output was concatenated to form the final representation. Trunk Network was trained on whole images, while branches were trained on patches taken from the images. The Triplet loss was regularized by introducing a constraint on the distance between mean representations of each class.

Another ensembled CNN architecture and regularized Triplet loss function are also proposed in [59]. The trunk network in the ensemble CNN extracted holistic representations, while the branch networks extracted local representations based on Haar-like features. The regularized Triplet loss function not only imposed constraints on the distance between mean representations but also on the standard deviation of each class.

Much of the work in LR face recognition relies on down-sampled LR images. Generating realistic LR images remains a significant challenge. The Resolution Adaptive Network (RAN) tackles this issue by introducing a Multi-Resolution Generative Adversarial Network (MR-GAN) [60] to generate a realistic LR face and then recognition is performed through a Feature Adaptation Network (FAN). The MR-GAN learns multi-resolution representations of LR faces. These representations are then combined and fed into the lowest resolution stream to refine the LR faces. This process enhances

the local-global information, mitigating artifacts commonly found in down-sampled images. Subsequently, the FAN utilizes a translation gate to integrate the discriminative information extracted from realistic LR faces into the backbone network responsible for HR representation. Notably, this integration occurs while preserving the discriminative ability of the original HR face representations.

The training data for the LR face recognition problem is generated by down-sampling the HR data. This data cannot capture all the variations found in the test data. To address this challenge, a single universal deep feature representation [61] was learned that handles the variations in face recognition without requiring access to test data distribution. This was achieved by introducing several changes in the training methodology and loss function. A confidence-aware identification loss was proposed that utilizes sample confidence during training to leverage hard samples. Each embedding was partitioned into sub-embeddings with an independent confidence value during training to maximize the representational power of embedding. The sub-embedding was further penalized with different regularizations to reduce the correlation between them. For better generalization ability, more variations were mined with semantic meaning. Finally, a probabilistic aggregation method was used to combine the sub-embeddings. This method accounts for the uncertainties associated with each sub-embedding, recognizing that the discriminative power of each sub-embedding varies depending on the specific type of variation.

In LR face recognition, there have been very few attempts to validate models on both HR and LR testing benchmarks. Adaface [17] addressed this gap by being designed to improve face recognition performance on low-quality datasets while maintaining good performance on HR datasets. Adaface introduced adaptiveness in the margin-based softmax loss. The adaptiveness was based on the feature norm, used as a proxy for image quality. The core idea was that the importance of misclassified (hard) examples should be adjusted based on their image quality. In this way, the issue of unidentifiable images that may arise from augmenting the dataset with down-sampled LR images is addressed. ARoFace [62] is based on the idea of introducing face alignment errors as an image degradation technique for augmenting data with synthetic LR images. To achieve this, adversarial data augmentation is combined with differentiable spatial transformations. The model is trained with synthetic LR images using the Adaface

loss function and surprisingly achieved improved performance compared to the baseline Adaface model The NPT loss [18] discussed earlier was also evaluated on the SCface dataset, a well-known benchmark for LR face recognition. The results demonstrated the loss function's ability to learn discriminative features simultaneously from both HR and LR images.

In [63], the author proposed a novel approach Recognizability Embedding Enhancement (REE) to improve the performance of very LR images by focusing on enhancing the recognizability of faces in the embedding space rather than directly improving image quality. A recognizability index (RI) is introduced that evaluates how well a face embedding can be recognized. This index is calculated based on the distance of the embedding to cluster of unrecognizable faces and its similarity to positive and negative class prototypes. To improve recognizability, the author proposed an index diversion loss that pushes hard-to-recognize faces with low RI away from the unrecognizable faces cluster. Additionally, a perceptibility attention mechanism is also introduced to focus on the most informative face regions for better embedding learning. The proposed model is trained in an end-to-end manner using Arcface loss. Experimental results on multiple LR datasets have demonstrated superior performance compared to SOTA methods. However, it is important to note that all the results presented are based on fine-tuned models, which may not effectively assess the model's true generalization capabilities.

The literature contains very few attempts at designing lightweight architectures for face recognition. In [64], MobileFaceNet and ShuffleFaceNet architectures are trained in a supervised manner using different interpolation techniques to generate down-sampled LR data. These approaches have yielded promising results across various LR testing benchmarks. MixFaceNet [65], another lightweight CNN architecture specifically designed for LR face recognition, is also trained in a supervised manner. Drawing inspiration from the MixNet architecture [66], MixFaceNet has surpassed the widely used MobileFaceNet architecture in both verification and identification tasks, demonstrating its superior performance and efficiency

In the universal learning approach, face representations of both LR and HR images are learned within a shared feature subspace. While this approach is straightforward, its performance tends to be biased towards LR images. This bias arises because multiple

TABLE 2.2: Comparative Analysis of Universal Learning Approaches

Methods	**Losses**	**Backbone**	**Limitations**
Zeng et al. [56]	• Softmax Loss	10-layer CNN	• Softmax loss can learn separable features but may not be discriminative enough
DCR [57]	• Softmax Loss • MSE • Center Loss	ResNet-27	• The model has only been validated on a single LR dataset, i.e, SCface
TBE-CNN [58]	• Regularized-Triplet Loss	GoogleNet	• Difficulty in mining hard triplets • Multiple hyper-parameters in the loss function needs proper tuning.
Haar [59]	• Regularized-Triplet Loss	GoogleNet	• Difficulty in mining hard triplets • Multiple hyper-parameters in the loss function needs proper tuning.
TURL [61]	• Confidence-aware Sub-embedding Identification Loss • Sub-embedding Decorrelation Loss	ResNet-100	• Does not account the domain gap between the LR and HR images
Adaface [17]	• Adaptive Angular Margin Loss	ResNet-100	• Performance is biased towards HR images. • Performance in small-scale Experiments is underrated
NPT [55]	• Nearest Proxy Triplet Loss	ResNet-50,100	• The NPT Loss has only been validated on a single LR dataset, i.e, SCface • Different models are used for LR and HR evaluation
REE [63]	• Arcface • MSE • Index Diversion Loss • L1 loss	ResNet-50	• The performance is evaluated on various LR datasets using finetuned protocols.

LR versions of a single HR image are used during training. Adaface [17], although achieving acceptable performance on both HR and LR images within this category, has certain limitations. Adaface is trained on a large proportion of HR images within each batch during training, resulting in a performance bias towards HR images. This bias is more evident in the results of small-scale experiments. A comparative analysis of universal learning approaches is presented in Table 2.2.

2.3.2.4 Distillation Learning

Among RI approaches, knowledge distillation techniques have recently become prevalent in LR face recognition due to their promising results. Knowledge distillation techniques explicitly bridge the gap between HR and LR features by transferring informative HR features from the HR teacher network to the LR student network.

Recognizing LR objects is always a great challenge in computer vision. Zhu et al. [67] proposed Deep Feature Distillation (DFD) to address this challenge. In DFD, informative HR features were transferred from the HR teacher network to the LR student network. In the first stage, the teacher network is trained on HR images. Then, a student network was established for recognizing LR images by minimizing two objectives, i.e., the Euclidean distance between the last layer features of the teacher and student network and the cross-entropy loss function. Though the experiments were performed on CIFAR-10 and SHVN datasets, it laid the foundation of distillation learning for LR face recognition. Massoli et al. [68] also used the idea of DFD for LR face recognition with some modifications. The teacher and student model was selected to be SOTA SeNet-50 architecture. The weights of the teacher were frozen and used for feature extraction only, while the student model was trained under curriculum learning strategy by minimizing two objectives as in DFD. The curriculum learning strategy involved training the model in a way that gradually increased the complexity of the data inputs as the learning process progressed. Furthermore, the images were randomly resized to a resolution between 8 and 256 pixels. This resizing was achieved by choosing a random exponent between 3 and 8 of a base 2.

In [69], the author used the same teacher-student paradigm to resolve resolution invariant face recognition with a slightly different distillation technique. The distillation

technique comprised KL divergence loss and parameter sharing approach. In the first stage, the teacher network was trained on HR images. Then, a student network was trained by minimizing three objectives: 1) Minimizing the KL divergence loss between the pre-trained softmax probability layer of the teacher network and the trainable softmax probability layer of the student network 2) Minimizing the KL divergence loss between the trainable softmax probability layer of the teacher and student networks 3) Minimizing the cross-entropy loss function for the student network. After each iteration, the parameters of the trainable softmax probability layer of the teacher were shared with the softmax probability layer of the student network. This approach achieved a similar generalization performance for the student network to that of the teacher network and obtained SOTA results on LR testing benchmarks

LR face recognition methods often struggle with images containing significant variations, such as extreme poses or low quality. While HR methods like Arcface perform well on images with slight variations (easy samples), their performance drops significantly on these harder samples. To address this challenge, the Distribution Distillation Loss (DDL) [70] focused on narrowing the performance gap between easy and hard samples. First, two similarity distributions were constructed, i.e., teacher distribution from easy samples and student distribution from hard samples, using the loss in Arcface [14]. Then, student distribution was forced to approximate the teacher distribution using distribution distillation loss. The DDL consisted of two terms: KL divergence loss and order loss. The KL divergence loss constrained the similarity between the student and teacher distributions, while the order loss minimized the distances between the expectations of similarity distributions.

The large resolution gap makes it challenging for the network to learn discriminative features simultaneously from HR and LR images. Based on this idea, Transferable Couple Network (TCN) [71] used transferable Triplet loss as a distillation technique to reduce the resolution mismatch. Transferable Triplet loss was essentially a Triplet loss that was used to minimize the resolution mismatch between the HR and LR features. HR features were extracted from the pre-trained teacher network, while LR features were extracted from the student network. The TTL ensured that the LR images of a specific person should be closer to all the HR images of the same person than to other identities in the HR domain, and an HR image of a specific person should be closer

to all the LR images of the same person than to other identities in LR domain. This type of cross-resolution matching minimized the resolution mismatch and allowed the network to learn resolution invariant features.

Deep Rival Penalized Competitive Learning (D-RPCL) [72] was based on the idea of Rival Penalized Competitive Learning (RPCL). RPCL introduced the concept of a rival. The rival was the second-closest node to the data point. During the update process, RPCL not only strengthens the winning node's similarity to the data point but also penalizes the rival node, pushing it away from the data point. Based on this idea, a modified version of the Arcface [14] and Cosface [47] loss function was developed to encourage the separation from the hardest non-target logit. As a result, the deeply learned face features became more discriminant. A D-RPCL version of DCR [57] was also developed, and improved performance was noted.

Existing approaches often train face recognition models simultaneously on both HR and LR images. However, treating HR and LR samples equally hinders learning discriminative features due to the large intra-class variance and inter-class similarity. DeriveNet [73] proposed a solution to this problem using class-specific margins. This approach leveraged two novel loss functions: Derived-margin Softmax Loss and Reconstruction-Center (ReCent) Loss. Derived-margin Softmax Loss was developed from traditional softmax loss but incorporated class-specific margins for non-corresponding classes. The goal is to encourage the model to differentiate between similar classes with a larger margin and separate highly discriminative classes with a smaller margin. ReCent Loss utilized a reconstruction network. This network projected the LR embedding and its corresponding HR class center onto a reconstruction space. It minimized two objectives: 1) Minimizing the distance between the reconstructed HR image and the original HR image. 2) Minimizing the distance between the learned reconstructed center of the HR image and the original HR images. The distance between the reconstructed HR class centers was then used to determine the class-specific margin. Furthermore, DeriveNet employed a data augmentation technique called multi-resolution pyramid augmentation. This technique exposed the model to images of various resolutions during training, enhancing its ability to learn robust features that are less sensitive to variations in image resolution.

Low et al. [74] proposed a dual-stream mutual information distillation network (MIND-Net) to address the challenge of realistic LR images captured by real-world surveillance cameras at extreme standoff distances. These images were very poor quality and significantly differed from synthetic LR images. The idea was to distill the non-identity-specific mutual information characterized by generic facial features. Mind-Net technique involved simultaneous training of the cross-target network and target network while sharing parameters. The cross-target network was trained on the HR and corresponding down-sampled LR images, while the target network was trained on native LR images. The loss function comprised of large margin cosine loss for the teacher and student networks, and cosine Triplet loss was used for learning non-identity-specific mutual information from both native LR and down-sampled LR face images. Moreover the degree of mutual information between cross-target and target network was quantified through normalized mutual information index.

A Deep Siamese Network (DSN) proposed in [75] reduced the domain gap between realistic LR and HR images. In a typical face recognition scenario, the probe consists of LR images, while the gallery set contains known HR images. The DSN extracted deep features across different resolutions using multiple networks and used a shared classifier to classify these features using the AM-Softmax loss function [76]. This enabled the network to learn the features in a unified feature space. Additionally, a cross-resolution Triplet loss was also employed to effectively pull matching pairs closer and push non-matching pairs further apart across different resolutions.

The authors in [77] addressed the challenge of limited LR query face datasets for LR face recognition. They proposed a novel data augmentation (DA) strategy called identity-extended DA and the corresponding network for its implementation, named the Identity-Extended Augmentation Network (IDEA-Net). The DA strategy aimed to fulfill both affinity and diversity requirements, which are essential for effective data augmentation. IDEA-Net consisted of two CNNs trained with a softmax-based loss function while sharing parameters. One CNN processed the LR query face dataset, while the other handled an auxiliary HR face dataset. The auxiliary dataset extended the query dataset regarding identity labels, also called the identity-extended dataset. A calibrator employing Triplet loss is introduced to regulate the resulting representation space. This calibrator refined the intra-class compactness and inter-class separation. This approach minimized

the distribution shift between the query and the identity-extended examples, quantifying their affinity. At the same time, it maximized the training complexity, quantifying diversity.

Unlike deep face datasets that provide both breadth (a large number of identities) and depth (sufficient samples per identity), shallow face data typically has only two face images available for each identity. This lack of intra-class diversity can lead to feature dimension collapse and network degeneration. To tackle this problem, the authors proposed a novel training method called Semi-Siamese Training (SST). SST involved a pair of Semi-Siamese networks to extract features from gallery and probe images. The training loss was computed using an updating gallery queue, which effectively optimizes the shallow training data. This training scheme can be integrated with any form of existing loss function (no matter classification loss or embedding loss) and network architectures.

In recent research, the focus is on distillation loss to effectively transfer informative HR features from the teacher network to the student network and reduce the disparity between HR and LR features. In CCFace [78], the distillation loss consisted of a contrastive loss, which minimized the distance between HR features from the teacher network and their LR counterparts from the student network while simultaneously penalizing the similarity. The classification loss for the student network incorporated an adaptive angular margin, where the adaptiveness of the margin was achieved by learning it from the HR teacher network. Furthermore, the weights learned by the HR teacher network in the angular margin loss were shared with the LR student network. QGface [79], also utilized distillation loss similar to that in CCFace, while the classification loss is Adaface [17]. The overall paradigm is that if the feature quality exceeds a threshold, it is fed into the classification loss; otherwise, the contrastive loss is applied. The feature quality decides the difficulty of the sample.

CATface [80] introduced a vision transformer for low-resolution face recognition. The overall paradigm consists of two stages of training. In the first stage, a dual-branch CNN with a self-attention distillation module is trained on both HR and LR images. In the second stage, a cross-attribute-guided transformer is used to fuse the discriminative facial information between the last layers of the dual-branch CNN. The two-stage

dual-resolution face network was proposed in [81]. A dual-branch CNN with bilateral connections between them is first trained on HR and LR images using the Arcface loss function. Bilateral connections are used to fuse HR features into the LR branch. In the second stage, the network is fine-tuned with the triplet loss function using competence-based curriculum learning. In [82], a method motivated by the Siamese network architecture is proposed. Two CNN networks, sharing parameters with each other, are simultaneously trained on HR and LR images, respectively. The loss function in Cosface is used as classification loss for each network. To minimize the resolution gap between HR and LR features, a feature constraint loss based on the idea of contrastive learning is proposed.

Lightweight architectures are crucial for LR face recognition due to their efficiency and effectiveness in resource-constrained environments such as mobile devices and surveillance systems. In [83], a lightweight model is proposed for deployment on low-end devices. The authors introduce a bridge distillation technique to transfer knowledge from a pre-trained complex model to a relatively simpler model with fewer parameters. Bridge distillation comprised two phases: Cross-dataset distillation and Resolution-adapted distillation. In cross-dataset distillation, an adaptation module was attached to the pre-trained teacher model, assumed to be trained on any private dataset. The adaptation module transformed the teacher's knowledge from the private domain into the public domain with a reduced feature space, enabling the student model to easily mimic it with reduced computational resources. Resolution-adapted distillation involved training the student network on LR images under the supervision of the teacher network. The supervision employs a deep feature distillation loss (regression loss).

Distillation Learning has achieved prominent performance in the domain of LR face recognition. Unlike universal learning, it employs a two-stage training process, effectively leveraging HR information from a HR teacher network to learn both HR and LR features through a student network. Another significant advantage is the training of a relatively simpler LR student model compared to the more complex HR teacher network. Apart from its efficacy in the results, the two-stage training process makes the training methodology complex. The number of terms in the loss function is increased, and careful tuning of the auxiliary weights in the loss function is required. A comparative analysis of Distillation Learning approaches is presented in Table 2.3.

TABLE 2.3: Comparative Analysis of Distillation Learning Approaches

Methods	**Losses**	**Backbone**	**Limitations**
DFD [67]	• Softmax Loss • Squared Euclidean Distance between features	VGG-16, ResNet-18	• Validation has not been performed on facial datasets
T-C [68]	• Softmax Loss • Squared Euclidean Distance between features	SeNet-50	• Domain Adaptation issue due to curriculum learning • Performance is biased towards HR datasets
RIFR [69]	• Softmax Loss • KL Divergence Loss	ResNet-50	• Separate models are evaluated on SCface and Tinyface datasets
DDL [70]	• KL Divergence Loss • Order Loss	ResNet-50, 100	• The model has only been validated on a single LR dataset, i.e, SCface
TCN [71]	• Softmax Loss • Center Loss • Triplet Loss	VGGFace, ResNet	• The model has only been validated on a single LR dataset, i.e, SCface. • Convergence issues associated with the Triple loss
D-RPCL [72]	• Adaptive Angular Margin Loss	ResNet-18,50	• Does not account for the intra-class variance of varying resolution images
Derive Net [73]	• Modified Softmax Loss • ReCent Loss	Light CNN	• The model has only been validated on a single LR dataset, i.e., SCface.
DSN [75]	• Softmax Loss • Triplet Loss	ResNet-21	• The approach outperforms previous methods on fine-tuned results only.

Methods	Losses	Backbone	Limitations
IDEA-Net [77]	• Normface • Triplet Loss	ResNet-50, MobileFaceNet	• While impressive on the SCface dataset, the method underperforms compared to previous approaches on Tinyface and QMUL Surv-face datasets.
SST [115]	• SST scheme	Attention-56, MobileFaceNet	• Fine-tuning is necessary for evaluation on LR datasets. • Baseline is kept Softmax loss, that can learn separable features only

2.4 Emerging Trends

Transformer models have revolutionized natural language processing and are now being applied to various computer vision tasks, including face recognition. While traditional CNNs have been the dominant approach for face recognition, transformers offer several advantages, such as their ability to capture global context more effectively. In [84, 85], transformer-based deep neural network algorithms are employed to accomplish low-resolution face recognition tasks. The attention module within these transformers is particularly powerful, as it effectively captures and incorporates long-range dependencies between image regions. This enables the model to identify and reconstruct important facial features, proving robust even against heavily degraded or LR images.

According to available datasets and results on face recognition, transformer models [80] that are equivalent in architecture size to CNNs have not yet achieved significant improvements. Their performance remains comparable to that of CNNs. However, due to their weaker inductive bias compared to CNNs, transformers require a large amount of data. In the future, with the availability of large-scale datasets, there is a strong possibility that transformers will surpass CNN architectures in the domain of face recognition.

2.5 Research Gap

LR face recognition remains an understudied topic in the literature. The following points are identified as research gaps in the problem of LR face recognition.

1. Using down-sampled LR images is a common practice for training LR face recognition models. However, these down-sampled LR images lack real-world degradation effects. A detailed analysis of developing a degradation model to simulate real-world degradation effects is lacking.

2. The data generated for training LR face recognition models typically consists of HR and LR images. However, treating all HR and LR images equally fails to capture the diversity in variations within the training data. It limits the ability of the model to generalize to real-world scenarios.

3. Face recognition models in the literature are typically designed for HR or LR faces. HR models are evaluated on HR benchmarks, while LR models are tested on LR benchmarks. Notably, there have been no attempts to develop a single model that can effectively handle both HR and LR images, with a comprehensive evaluation of both HR and LR testing benchmarks.

4. LR face recognition encounters both HR and LR images. However, testing benchmarks in LR face recognition are often small-scale and have limited variations in the resolution. It makes it challenging to thoroughly assess the effectiveness and limitations of LR face recognition techniques.

2.6 Problem Statement

Face recognition techniques developed for HR images have demonstrated impressive performance on HR testing benchmarks, even with significant variations in pose, illumination, and expression, and have reached a saturation point. However, face recognition in LR images from surveillance applications poses more challenges due to the presence of both HR and LR images. Probe images of varying resolutions, including both HR and LR, are typically matched against HR gallery images. To achieve acceptable

performance on both HR and LR images, the LR face recognition model must map corresponding HR and LR features closer together. The significant domain gap between HR and LR images makes this mapping challenging and can lead to deteriorating performance on HR images.

2.7 Research Objectives

The primary goal in LR face recognition is to learn discriminative features from HR and LR images to achieve acceptable performance on both. The following efforts have been made to accomplish this.

1. To develop a degradation model that emulates real-world degradation effects in the synthetic LR images.
2. To develop an attention-guided distillation technique that transfers informative HR features from a powerful teacher network to a student network.
3. To develop a sub-center learning approach that captures diverse representations across varying image resolutions.
4. To develop contrastive distillation loss that reduces the gap between HR and LR features and facilitates learning compact and discriminative features.
5. To develop new protocols to more effectively assess efficacy and identify limitations in LR face recognition on HR and LR images.

2.8 Summary

This chapter delves into face recognition literature, encompassing both HR and LR face recognition approaches. In HR face recognition, margin-based softmax loss functions are prominent methods that maximize the inter-class variance and minimize the intra-class variance among the extracted features. In contrast, two main approaches dominate LR face recognition: super-resolution-based and resolution-invariant methods. Super-resolution-based approaches first enhance the quality of LR images before extracting

features for identification. Resolution-invariant methods, on the other hand, learn both LR and HR features in a unified feature space. Resolution-invariant methods are further categorized into universal learning and distillation learning. All these methods are comparatively analyzed, and a research gap is identified. Furthermore, a problem statement is formulated, and research objectives are outlined.

Chapter 3

Datasets and Evaluation Methods

3.1 Outline

Datasets play a pivotal role in the development, training, and evaluation of face recognition algorithms. They can be categorized into training and testing datasets. Training datasets, which are typically larger, are used exclusively for training the algorithms. Without large and diverse training datasets, the algorithms would lack the variability needed to generalize well to new, unseen data. Testing datasets, separate from the training data, allow researchers to objectively evaluate the algorithm's performance. This helps identify strengths and weaknesses, guiding further improvements.

Evaluation methods provide the tools to assess how well a face recognition algorithm performs after being trained. Traditional metrics like accuracy might not be sufficient for evaluating these algorithms. Specialized metrics, such as the Rank-1 Recognition Rate, measure the probability of the correct identity being ranked first in a search. Additionally, verification and identification rates at specific error thresholds provide insights into the algorithm's ability to distinguish between genuine and impostor access attempts. These metrics offer a more comprehensive understanding of an algorithm's performance in real-world scenarios.

The interplay between datasets and evaluation methods is crucial for advancing face recognition technology. Training on diverse datasets enables researchers to develop alg-

orithms that are robust to variations in real-world images. Utilizing well-defined evaluation methods allows researchers to objectively compare algorithms and identify areas for improvement. This continuous cycle of training, evaluation and refinement pushes the boundaries of face recognition capabilities, paving the way for even more reliable and secure applications in the future.

3.2 Training Datasets

3.2.1 CASIA Webface

CASIA Webface [86] is a small-scale dataset developed by the Institute of Automation, Chinese Academy of Sciences (CASIA). The CASIA Webface dataset contains over 10,575 identities and approximately 494,414 images. These images were collected from the internet, resulting in a diverse set of facial images with variations in pose, expression, illumination, and occlusion. The images were then manually filtered and annotated to ensure dataset quality and accuracy. Famous subjects have more images, while others are less represented.

3.2.2 Mirosoft Celeb 1M

The Microsoft Celeb 1M (MS Celeb 1M) dataset [87] is a large-scale collection of facial images designed to advance face recognition research. Developed by Microsoft, the dataset is publicly available to researchers, enabling them to advance their research. It contained approximately 10 million images of 1 million celebrities collected from the internet. The dataset has uniform distribution of age, gender, nationality. There are three clean versions available at the InsightFace repository [88].

3.2.3 Webface 4M

WebFace 4M is one of the smaller subsets of the WebFace 260M dataset [89]. It contains 4.2 million facial images of 2 million individuals. The original version comprises 4 million

identities and 260 million images with varying noise levels. Its cleaned version consists of 42 million images of 2 million individuals. WebFace42M covers most major races in the world including Caucasian, Middle East, East Asian, African, Latino, Indian and South East Asian. The dataset is publicly available for research purpose.

3.2.4 VGGFace2

VGGFace2 [90] is a large-scale face recognition dataset comprising approximately 3.31 million images of 9,131 individuals. It is a significant improvement over its predecessor, VGGFace, with a more diverse range of images in terms of pose, age, illumination, ethnicity, and profession. The dataset is approximately gender-balanced and spanning wide range of ethnicity.

3.3 Testing Datasets

The testing datasets for evaluating LR face recognition models are categorized into three types: LR, Mixed Resolution (MR), and HR. These datasets are chosen based on their wider acceptance within the research community for testing HR and LR face recognition.

3.3.1 LR Datasets

3.3.1.1 SCface

The SCface dataset is an indoor surveillance dataset. It is widely known for testing a resolution invariant face recognition. The dataset has 130 subjects having visible and infrared images.

Acquisition Scenario: The videos were recorded at the Video Communications Laboratory, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. The dataset includes a total of 130 subjects, with 114 males and 16 females. Six cameras were used for data acquisition: five surveillance cameras and one high-quality photo

camera. Surveillance cameras were installed at a height of 2.25 meters. Sunlight streaming through a window on one side provided the primary illumination source. Two of these five surveillance cameras were also capable of recording in infrared night vision mode. The high-quality photo camera was located in a separate room for capturing controlled-scenario mugshots in both infrared and visible light. This room was kept dark for infrared captures, while standard indoor lighting was used for the high-quality photo camera. Subjects were instructed to walk in front of the cameras and pause at three designated points, allowing for image capture at varying distances (d1=4.2m, d2=2.6m, and d3=1m) from the cameras.

The data collection spanned five days. Each subject contributes a variety of images to the dataset. In a controlled environment, one high-quality RGB image and one infrared image are captured for each subject. Additionally, fifteen probe images are taken in visible light from five different surveillance cameras, capturing the subject at three distinct distances. Finally, six probe images are captured by the two dedicated infrared surveillance cameras (cameras 6 and 7 in the database). This comprehensive

(A) distance (d1)

(B) distance (d2)

(C) distance (d3)

FIGURE 3.1: Example images from the SCface dataset captured at three different distances. The first column shows high-quality images, while subsequent columns show images from the five surveillance cameras, respectively.

collection of images ensures the dataset includes a diverse range of lighting conditions and distances. Example Images from the dataset are shown in Fig. 3.1.

Protocols: The evaluation protocol includes identification performance on daytime and night time images separately. The partition of subjects for training and testing are 50 and 80 respectively.

Comments: The SCface dataset exhibits negligible pose variation, as all images captured from surveillance cameras depict frontal faces. Additionally, the dataset may be biased towards specific lighting conditions and indoor settings, potentially limiting its generalizability to outdoor or diverse lighting scenarios. These biases could lead to models that excel in controlled indoor surveillance contexts but struggle in real-world environments, such as outdoor or crowded settings. Furthermore, the subjects of the dataset is limited and lack diversity. All subjects are white individuals, which may introduce demographic biases.

3.3.1.2 COXface

COXface is a video-based indoor surveillance database. There are 1000 subjects, with 435 males and 565 females. Half of them are Mongolian and the other half are Caucasian.

Acquisition Scenario: The videos are recorded at the large gym at Xinjiang University. Three SONY HDR-CX350E DV camcorders (surveillance cameras) and one Canon EOS 500D DC (high-quality digital camera) were used. The surveillance cameras are mounted on a two meter high tripod. The source of illumination was half indoor i.e., one side wall of transparent glass with a high ceiling and half outdoor lighting during day time. Each Subject was allowed to walk on an S-shaped path to emulate different facial expressions and poses to record a video. The walking speeds were different from subject to subject, therefore, the duration of the video clips of different subjects were different even for the same camera. There are 3000 videos of 1000 subjects. The mugshot images are captured by a high-quality digital camera mounted on a tripod about 3 meters away from the subjects. Example images from the COXface dataset are shown in Fig. 3.2.

Protocols: The evaluation protocol includes the mean identification accuracy and standard deviation on three video-based surveillance scenarios. i.e., video-to-still (V2S),

(A) Video 1 (Cam 1)

(B) Video 2 (Cam 2)

(C) Video 3 (Cam 3)

FIGURE 3.2: Example images from the COXface dataset captured by three different surveillance cameras. The first column shows high-quality images, while subsequent columns show images captured at different positions along the S-shaped pathway.

still-to-video (S2V) and video-to-video (V2V). Ten sets of random 300/700 partitions of subjects for training and testing are provided officially.

Comments: The COXface dataset contains videos with pose variations. However, the majority of images within these videos exhibit a frontal pose. Therefore, pose variation has a minimal impact on the overall results. This dataset may also be biased towards specific lighting conditions and indoor settings, potentially limiting its generalizability to outdoor or diverse lighting scenarios. Furthermore, the subjects in this dataset are predominantly of Chinese and Mongolian ethnicity.

3.3.1.3 Tinyface

Tinyface dataset is specifically designed for comprehensive evaluation of LR face recognition.

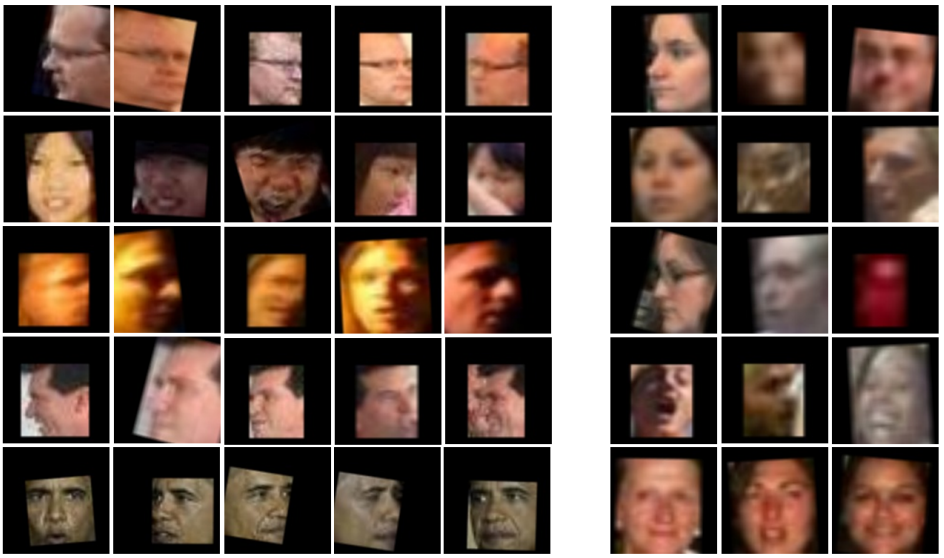


FIGURE 3.3: Example Images from Tinyface Dataset. (Left) Labeled Identities, (Right) Distractors

Image Collection: The Tinyface dataset [52] consists of labeled and unlabeled identities. The images of labeled identities were collected from publicly available datasets, PIPA [91] and MegaFace2 [92]. Both datasets contain unconstrained social-media web face images with a wide variety of facial expressions, poses, and imaging conditions. To ensure all selected faces are LR, the authors removed faces with a size exceeding 32×32 pixels based on detection results. The final image height ranges from 6 to 32 pixels with an average of 20 pixels. The dataset contains 15,975 labeled LR faces belonging to 5,139 unique identities (IDs). These identities are further split for training (2,570 IDs) and testing (2,569 IDs). Unlabeled faces (153,428) are used as distractors during testing. Finally, the test set images are further divided into probe and gallery sets. Example Images from the Tinyface dataset are shown in Fig. 3.3.

Protocols: The evaluation protocol includes the Cumulative Matching Characteristic (CMC) curve and mean Average Precision (mAP).

Comments: The Tinyface dataset consists of tightly cropped, LR images. When these images are aligned, artifacts are generated at the corners, which degrade the performance of face recognition models. However, these artifacts are not typically generated in real-world scenarios when using surveillance cameras.

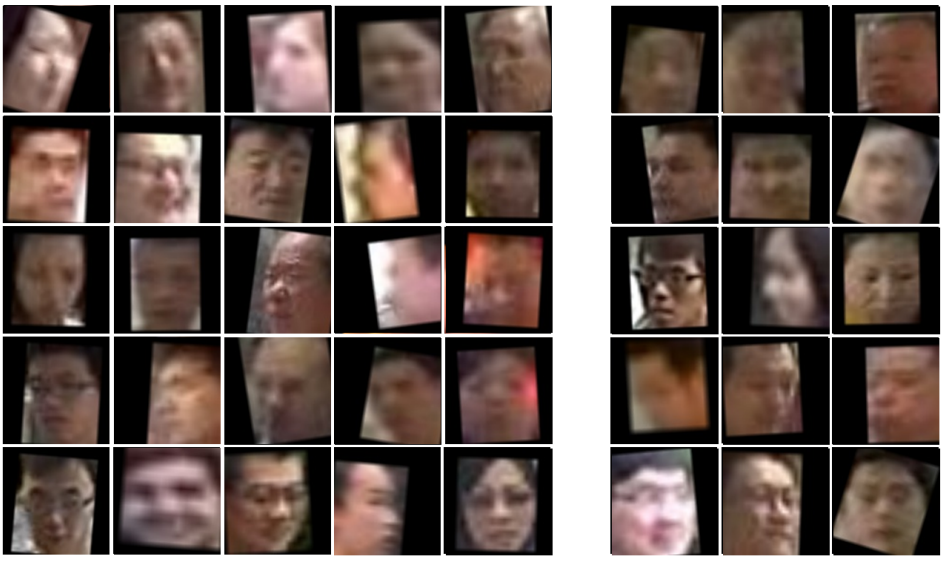


FIGURE 3.4: Example Images from QMUL Survface Dataset. (Left) Labeled Identities, (Right) Distractors

3.3.1.4 QMUL Survface

QMUL Survface is a native LR surveillance dataset

Image Collection: The QMUL Survface Challenge [93] utilizes real-world surveillance images collected from 17 different person re-identification benchmarks. These benchmarks encompass diverse locations and capture subjects in various surveillance scenarios across multiple countries. The dataset consists of 463,507 LR face images representing 15,573 unique individuals. The images depict uncontrolled variations in factors like pose, illumination, motion blur, occlusion, and background clutter. Notably, 68.3% (10,638 people) have multiple face images, while the remaining 4,935 individuals have only one image (single-shot IDs). For evaluation purposes, the dataset splits the 10,638 multi-shot IDs into two halves: 5,319 IDs for training and another 5,319 IDs for testing. The testing set is further supplemented with the 4,935 single-shot IDs, resulting in a total of 10,254 IDs. Furthermore, it also contains 153,428 unlabeled LR face images that are used as distractors. Example images from QMUL Survface dataset are shown in Fig. 3.4.

Protocols: The evaluation protocol includes the TPIR@FAR=[0.1, 0.2, 0.3] and TAR@FAR=[0.1, 0.01, 0.001, 0.0001].

Comments: The QMUL Survface dataset also comprises tightly cropped, LR images. During alignment, artifacts are generated at the corners.

3.3.2 HR Datasets

3.3.2.1 Labeled Faces in the Wild (LFW)

Labeled Faces in the Wild (LFW) [94] is a publicly available benchmark dataset used for face verification. The LFW dataset, first published in 2007, consists of 13,323 web photos of 5,749 individuals. These images are divided into 6,000 face pairs across 10 splits, representing a collection captured in uncontrolled environments (often referred to as "in the wild"). At the time of its publication, LFW was considered a challenging dataset due to the unconstrained nature of the images. However, with the advancements in deep learning algorithms, face verification accuracy on LFW has surpassed 99% [14, 18]. Example images are shown in Fig. 3.5(a).

3.3.2.2 Celebrities in Frontal-Profile (CFP)

The performance of face recognition algorithms degrades by more than 10% when transitioning from frontal-to-frontal to frontal-to-profile face verification tasks. To address this challenge, the Celebrities in Frontal-to-Profile (CFP) [95] dataset was developed to facilitate research in frontal-to-profile face verification. It comprises 7,000 face pairs for both frontal-to-frontal and frontal-to-profile scenarios. The CFP-FP dataset specifically corresponds to frontal-to-profile face pairs, providing a valuable resource for improving the accuracy and robustness of face recognition algorithms in diverse viewing angles. Example images are shown in Fig. 3.5(b).

3.3.2.3 AgeDB

The AgeDB dataset [96] is specifically designed for age-invariant face verification. It holds the distinction of being the first manually collected (in the wild) age database. The dataset ensures clean and accurate age labels, unlike other databases that rely on semi-automatic collection methods using crawlers, which can introduce noise into

the age data. Notably, within AgeDB-30, each face pair has an age difference of 30 years. This substantial age gap introduces variability in facial features, making it a more demanding dataset for face recognition algorithms. AgeDB-30 is divided into 10 folds, with each fold containing 300 intra-class pairs (same person) and 300 inter-class pairs (different people). Example images from AgeDB-30 are shown in Fig. 3.5(c).

3.3.2.4 Cross-Age LFW (CALFW)

The Labeled Faces in the Wild (LFW) dataset has long served as the benchmark for unconstrained face verification. However, deep learning algorithms have achieved near-perfect accuracy, approaching 100%. To address the issue of performance saturation, the Cross-Age LFW (CALFW) dataset [97] was introduced to further increase intra-class variance, particularly through age differences. CALFW includes both the large intra-class variance and the tiny inter-class variance simultaneously. It comprises 7,000 face pairs across 10 splits. Notably, compared to the accuracy achieved on LFW, performance on CALFW typically drops by 10% to 17%. Example images are shown in Fig. 3.5(d).

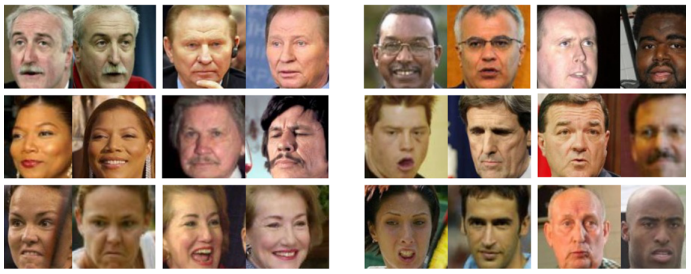
3.3.2.5 Cross-Pose LFW (CPLFW)

Similar to CALFW, the motivation behind the Cross-Pose LFW (CPLFW) dataset [98] is to further explore the challenge of intra-class variance. However, CPLFW focuses specifically on pose variations, aiming to increase the difficulty compared to LFW. Like CALFW, CPLFW exhibits both the large intra-class variance and the tiny inter-class variance simultaneously. It comprises 7,000 face pairs across 10 splits. It is worth noting that, similar to CALFW, the expected accuracy on CPLFW is significantly lower than LFW, dropping by 10% to 17%. Example images are shown in Fig. 3.5(e).

3.3.3 Mixed Resolution Datasets

3.3.3.1 IJB-B Dataset

The IJB-B (IARPA Janus Benchmark B) dataset is a widely recognized and challenging benchmark for face recognition and verification tasks in unconstrained environments.

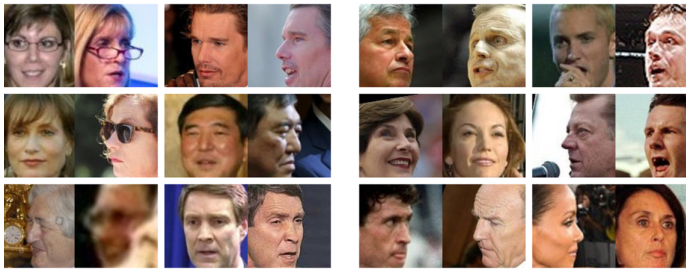


(A) LFW

(B) CFP-FP

(C) AgeDB-30

(D) CALFW



(E) CPLFW

FIGURE 3.5: (Left) Positive Pairs, (Right) Negative Pairs.

Developed by the Intelligence Advanced Research Projects Activity (IARPA) as an extension of IJB-A, it offers a larger and more diverse dataset with 1,845 subjects having 21,798 still images, and 55,026 video frames. This makes it a cornerstone for evaluating face recognition systems. IJB-B comprises ten protocols, among which the widely used 1:1 verification protocol was employed. This protocol involves 8,010,270 comparisons, consisting of 10,270 genuine and 8,000,000 impostor pairs. Example images are shown in Fig. 3.6(a).

3.3.3.2 IJB-C Dataset

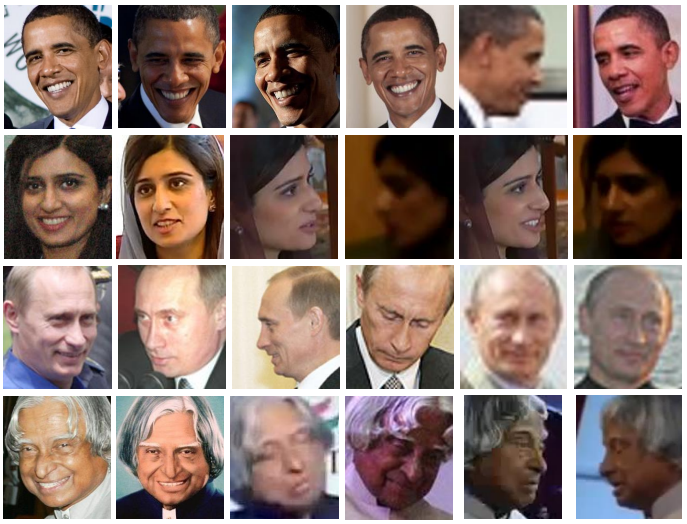
The IJB-C (IARPA Janus Benchmark C) dataset is the third iteration in the IARPA Janus program, designed to rigorously evaluate face recognition systems under challenging, real-world conditions. Building upon the foundation of IJB-A and IJB-B, IJB-C provides a significantly larger and more complex dataset featuring 3,531 subjects having 31,334 still images, and 117,542 video frames. This makes it a leading benchmark for assessing the capabilities of modern face recognition algorithms. The widely used 1:1 verification protocol within the IJB-C dataset was utilized for the experiments. This protocol comprises 15,658,489 comparisons, including 19,557 genuine and 15,638,932 impostor pairs. Example images are shown in Fig. 3.6(b).

3.4 Evaluation Metrics

Face recognition performance is evaluated through three standard tasks: face verification, closed-set face identification, and open-set face identification. Each of these tasks employs specific evaluation metrics to measure the effectiveness and accuracy of face recognition algorithms. A concise overview of the evaluation metrics are provided below.

3.4.1 Face Verification

In face verification, the matching process calculates the distance between a pair of facial images. A predefined threshold determines if the images belong to the same



(A) IJB-B

(B) IJB-C

FIGURE 3.6: Example Images from IJB-B and IJB-C datasets, having images of a single identity in a row

person. Images with a distance below the threshold are considered a match, while those exceeding it are not. The system's performance is evaluated based on correctly classified pairs (true positives/negatives) and errors (false positives/negatives). False positives occur when different people are mistakenly identified as the same, while false negatives happen when the same person is identified as different individuals. These error types form the basis for various metrics used to assess face verification performance.

Accuracy: The percentage of image pairs the system correctly classifies, encompassing both true positives (same person) and true negatives (different persons).

TABLE 3.1: Summary of Facial Datasets for Training and Testing (Images: I, Videos: V)

Type	**Datasets**	**#Identity**	**# I/V**
Training	CASIA Webface [86]	10,575	494,414 (I)
	Webface4M [89]	205,990	4,235,242 (I)
	MS1M [88]	85K	5.8M (I)
	VGGFace2 [90]	9,131	3.8M (I)
LR (Testing)	SCface [99]	130	1,950 (I)
	COXface [100]	1,000	3,000 (V)
	Tinyface [52]	5,139	15,975 (I)
	Survface [93]	15,573	463,507 (I)
HR (Testing)	LFW [94]	5,749	13,233 (I)
	CFP-FP [95]	500	7,000 (I)
	CPLFW [98]	5,749	11,652 (I)
	AgeDB [96]	568	16,488 (I)
	CALFW [97]	5,749	12,174 (I)
MR (Testing)	IJB-B [101]	1,845	66,780 (I & V)
	IJB-C [102]	3,531	138,836 (I & V)

False Positive Rate (FPR): Rate of incorrect matches (different people classified as same).

False Negative Rate (FNR): Rate of missed matches (same people classified as different).

Equal Error Rate (EER): The point at which FPR and FNR are equal.

Receiver Operating Characteristic (ROC): The curve that plots the true positive rate (TPR) against the false positive rate (FPR) calculated by varying the threshold.

Area Under Curve (AUC): The percentage of the area under the ROC curve.

3.4.2 Close-Set Face Identification

In close-set face identification, the query face image is compared with each image in the gallery set. The resulting distances are sorted and ranked, The top n subjects

with the closest distances are retrieved. A true match occurs when the true identity is observed within the top n ranks. True Positive Identification Rate (TPIR) refers to the probability of observing the true identity in the top n ranks. Evaluation metrics for close-set face identification based on TPIR are defined as follows:

Rank-n Accuracy: TPIR at the rank of n. Typical values for n are 1, 5, 10.

Cumulative Match Characteristic (CMC): The curve that plots TPIR against ranks.

3.4.3 Open-Set Face Identification

Open-set face identification is more complex than closed-set identification. It adds another essential step: comparing the closest distance between the query face and gallery-set face to a predefined threshold. This comparison determines whether the query belongs to a known person (someone already in the gallery) or an unknown person. If the distance is less than the threshold, the system considers it a potential match. Otherwise, it recognizes the face as unknown.

3.5 Summary

This chapter provides a detailed discussion of datasets used by the face recognition community for training and testing algorithms. Testing datasets are categorized into LR, HR, and MR. The LR datasets include SCface, COXface, Tinyface, and QMUL Survface. Similarly, the HR dataset includes LFW and its variants, and the MR datasets include IJB-B and IJB-C. Furthermore, the chapter explores evaluation methods for assessing the performance of face recognition algorithms.

Chapter 4

Degradation Model and Attention-Guided Distillation

4.1 Outline

Deep convolution neural networks (CNN) have shown their efficacy in face recognition tasks because they extract highly discriminant face representations from face images. On HR benchmark datasets, outstanding identification and verification results have been achieved, even with considerable variations in pose illumination and expression. However, the performance of these networks is significantly degraded when tested on LR images. Such scenarios usually arise in surveillance applications that capture tiny faces due to coverage of a large view of the scene. These face images exhibit different degradation, like blurring and noise. In LR face recognition, a straightforward solution to this problem is to train a deep CNN simultaneously on HR images and their corresponding LR versions. Mostly, the LR versions are created by down-sampling the HR images and then resizing them to the required size by the network. In this way, the image loses most of the HR information and appears as blurry LR image. Although this strategy improves the performance of deep CNNs on LR images, it has some limitations. First, a significant difference exists between down-sampled LR images and LR images from surveillance cameras. The absence of different types of degradations in the down-sampled LR images leads to less variation within the data, eventually causing pe-

rformance saturation at an earlier stage. Another limitation is the deterioration in the performance of HR images. Once the network is trained on a combination of HR and LR images, it becomes biased towards the LR images. In this work, solutions to both these limitations are proposed. A proposed degradation model synthesizes LR images from the corresponding HR, emulating the real-world degradation effects in synthetic data, thus enabling the face recognition model to tolerate blurry and noisy effects. The proposed attention-guided distillation leverages attention maps from intermediate convolutional layers, along with deep features, to transfer informative HR features from the teacher network to the student network. In this way, HR features of the teacher network guide the student network to a better optimum and facilitate the learning of resolution robust face representations.

4.2 Proposed Methodology

The proposed methodology is based on a degradation model and an attention-guided distillation. The degradation model generates synthetic LR images for training that mimic real-world surveillance data. In contrast, the attention-guided distillation technique transfers informative HR features from the teacher network to the student network.

4.2.1 Degradation Model

LR face recognition requires a combination of HR and LR images to train a deep CNN. Enormous HR datasets are publicly available, i.e., CASIA Webface [86] with 0.5M images, MS1M (v2 by Insightface) [87, 88] with 5.8M images and Webface260M [89] with 42M images. In contrast, real-world LR datasets with sufficient training images are lacking. Additionally, the available datasets have a single HR image per identity. In that case, these LR datasets cannot be used for training. This problem is usually addressed by down-sampling the HR images to get their LR versions. The downside of this approach is that the performance reaches a saturation point prematurely due to the lack of inherent variations found in the real-world LR images. Therefore, a degradation model is needed to simulate real-world surveillance and degradation effects in synthetically generated LR images. The aim of this degradation model is to improve

the performance of LR face recognition model by generating a more precise representation of real-world surveillance effects. The proposed degradation model incorporates five different degradation effects: down-sampling, blurring (including out-of-focus and motion blur), noise, and JPEG compression. The types of degradation effects used in the previous approaches and proposed degradation model are presented in Table 4.1.

TABLE 4.1: Types of degradation used in the previous approaches and the proposed degradation model

Methods	**Down-sampling**	**Miss-alignment**	**Blurring**	**Noise**	**JPEG**
TBE-CNN [58]	✓	×	✓	✓	×
FAN [53]	✓	×	×	×	×
Gao et al. [103]	✓	×	×	×	×
DDL [70]	✓	×	×	×	×
Khalid et al. [69]	✓	×	×	×	×
Proposed	✓	✓	✓	✓	✓

4.2.1.1 Proposed Solution

The proposed degradation model introduces close-to-real-world degradation effects in synthetic LR images by employing a combination of classical degradation techniques. These classical techniques, such as down-sampling, motion blur, out-of-focus blur, noise, and JPEG compression, are selected due to their frequent occurrence in surveillance images.

The degradation model first performs affine transformation according to the coordinates of five reference facial landmarks (i.e., eye centers, nose tip and mouth corners) while adding a few miss-alignment pixels in the extracted facial landmarks. A motion blur or Gaussian kernel is randomly selected and convolved with the misaligned image. Then, a down-sampling operation is performed and Gaussian noise is added. Finally, the image is resized, and JPEG compression is also applied, as it is commonly used in image acquisition systems The degradation model can be represented as:

$$\mathcal{I}' = [(([\,\mathcal{I}\,]_{misalign} * k) \downarrow_{p\times p} + n) \uparrow_{m\times m}]_{JPEG_q} \tag{4.1}$$

where $\mathcal{I} \in \mathbb{I}_{\mathrm{H}}$ and $\mathcal{I}' \in \mathbb{I}_{\mathrm{S}}$

FIGURE 4.1: Overview of degradation model that is used to generate synthetic LR images. The input is an HR image and the coordinates of facial landmarks, while the output is a synthetic LR image. Each block represents a different degradation effect induced in the synthetic data.

$\mathcal{I}$ and $\mathcal{I}'$ denotes the HR and the synthetic LR image, respectively, while $\mathbb{I}_{\mathrm{H}}$ and $\mathbb{I}_{\mathrm{S}}$ denote the HR and synthetic LR dataset, respectively. k signifies the kernel of the Gaussian or motion blur filter, and n represents the Gaussian noise. $p \times p$ denotes the down-sampling size while $m \times m$ is the required size for the CNN network. The subscript in $JPEG_q$ represents the compression ratio. The effect of each degradation in the degradation model is visualized in Fig. 4.1. The degradation techniques are briefly revisited below

Misalignment: Facial landmarks like eyes, nose, and mouth are the most distinctive parts of a face. Alignment ensures these features are always in the same relative positions, making them easier for the face recognition model to analyze. Face alignment is necessary for HR face recognition because it helps the face recognition model to focus on the most relevant parts of the face.

In LR face recognition, face detectors are unable to locate landmarks perfectly, which leads to imperfect alignment. In Fig. 4.2, the difference between the landmark locations is calculated for different resolutions of the same image. It is evident that the landmark localization error increases as the image resolution degrades. In real-world scenarios, this increased localization error will also adversely affect the performance of face recognition systems. Therefore, developing LR face recognition system that are invariant to misaligned facial images is crucial. To address this, the detected landmarks are randomly perturbed before an affine transformation is applied for alignment. Retinaface [20] is used for the extraction of facial landmarks.

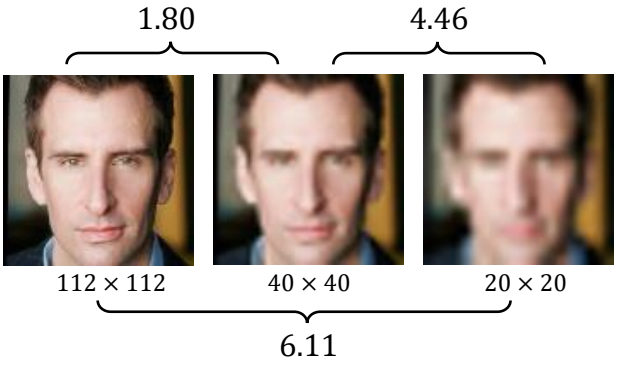


FIGURE 4.2: Euclidean distance calculated between the coordinates of facial landmarks of HR image and its down-sampled versions.

Motion-Blur: Motion blur in surveillance cameras is primarily caused by the movement of the subject being recorded during the camera's exposure time. Motion blur can significantly degrade the quality of surveillance footage, making it difficult for recognition systems to identify individuals or objects of interest. By simulating motion blur during training, the model is exposed to a more realistic distribution of LR images, improving its ability to generalize to real-world scenarios. Assuming that during the shutter exposure time, the relative motion occurs along a single direction, the motion blur can be represented by the following 2D kernel [104].

$$k(i, j, L, \theta) = \begin{cases} \frac{1}{L}, & \text{if } \sqrt{i^2 + j^2} < \frac{L}{2} \text{ and } \frac{i}{j} = -\tan\theta. \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

where (i, j) represents the pixel coordinate starting from the center of the image, L denotes the size of the kernel and specifies the motion distance during exposure, and θ is the motion direction.

Out-of-focus Blur: Out-of-focus blur is a frequent occurrence in surveillance camera footage, especially when cameras are not perfectly focused or when subjects are at varying distances. It is typically caused by external vibrations, such as those caused by wind or nearby traffic, which can make the camera shake. Lens deterioration over time can also contribute to out-of-focus blur. Training the model with out-of-focus blurred images forces it to learn more robust and discriminative features that are less sensitive to this type of degradation. This enhances the model's ability to accurately identify individuals even when their images are blurred. A Gaussian kernel is typically used to

model the out-of-focus blur.

$$k(i,j) = \frac{1}{N}\exp(-\frac{1}{2}C^T\Sigma^{-1}C), \tag{4.3}$$

$$C = [i,j]^T$$

where Σ shows the covariance matrix representing the spread of blur, N is the normalization constant, while C denotes spatial coordinates from the center. The covariance matrix can be further represented as follows:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

where σ_1 and σ_2 are the standard deviation in x and y directions.

Down-sampling: Down-sampling is a fundamental operation in synthesizing LR images from HR ones in LR face recognition problem. HR images are down-sampled to various resolutions and subsequently up-sampled to the required size by the network. This process inherently leads to a loss of detail in the HR images, resulting in a blurry and pixelated appearance. Surveillance cameras often capture a wide field of view, leading to faces appearing very small and consequently blurry or pixelated. To effectively identify individuals with limited information available in these LR images, down-sampling is intentionally introduced into the training data. Down-sampling and up-sampling are both image resizing operation. Common techniques include bi-cubic, bi-linear, and nearest-neighbor interpolation. These methods exhibit varying visual effects:

- Bi-linear interpolation can produce the most pixelated and blocky images.
- Nearest-neighbor interpolation may result in smoother images than nearest-neighbor, but can introduce some blurriness, especially around edges.
- Bi-cubic interpolation generally produces the highest quality results, with the smoothest transitions and the least noticeable artifacts.

In the proposed degradation model, bi-cubic interpolation is employed for the resizing operation.

Noise: In addition to motion blur and out-of-focus blur, real-world surveillance images are often captured with acquisition noise. Additive Gaussian noise is frequently used to simulate the noise characteristics commonly observed in real-world surveillance camera footage. Introducing Gaussian noise to the training data ensures that the model is exposed to more realistic LR images, enabling the model to perform effectively in practical settings. The noise intensity is determined by the standard deviation (i.e., σ) of the Gaussian distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{4.4}$$

JPEG Compression: Surveillance cameras generate vast amounts of video data, which poses significant challenges in terms of storage and transmission. Efficient management of bandwidth and storage resources is essential for the practical deployment of such systems, particularly in large-scale surveillance networks. Compression techniques are usually employed to reduce the data size without severely compromising visual quality. Among the various compression methods for images, JPEG is one of the most prevalent. JPEG compression achieves a balance between reducing file size and retaining critical image information. By applying JPEG compression to training data, models are better equipped to handle the challenges posed by real-world surveillance scenarios, resulting in improved recognition accuracy and reliability.

JPEG compression works by dividing an image into small blocks, and processing each block individually using a combination of discrete cosine transform, quantization, and entropy encoding. This block-based approach allows for efficient compression but can sometimes introduce visible artifacts, especially around edges or in areas with sharp contrasts, often referred to as "blockiness". The quality of a JPEG image is influenced by the compression level, controlled by a quality factor $q \in [0, 100]$. A lower q-value results in higher compression, reducing file size significantly but at the cost of image quality.

4.2.2 Attention Guided Distillation

Generally, knowledge distillation approaches are used for network complexity reduction, whereby information is distilled from a complex network to a simpler one. However,

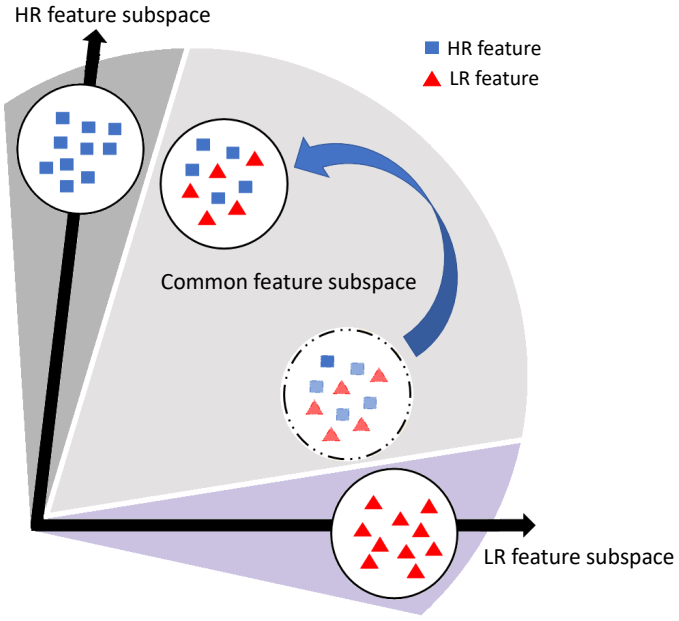


FIGURE 4.3: A generic solution to abstain the network from discarding HR features. Distillation techniques shift the common-features subspace closer to the HR feature subspace.

in the context of the LR face recognition problem, the main objective is to transfer informative HR features from the teacher network to the student network. It ensures that the student network can preserve both HR and LR features while maintaining a lower complexity than the teacher network. The importance of preserving HR knowledge comes in cross-resolution scenarios. In cross-resolution scenarios, an LR query image is usually compared to HR images of the gallery set. Thus, it is necessary to map both HR and LR features in the same (common) feature subspace. Simultaneously, training the model on HR and LR images is a straightforward solution, but it leads to mapping HR and LR features close to the LR feature subspace, thus discarding informative HR features. This is due to the requirement of multiple LR images of a single HR image during training. To overcome this problem, a generic solution is to shift the common feature subspace towards the HR feature subspace, as shown in Fig. 4.3. This domain shifting enables the network to preserve HR features, which is accomplished using knowledge distillation techniques.

In the LR face recognition problem, knowledge distillation is carried out in two phases. During the first phase, the teacher network is trained on HR images. In the second

phase, the weights of the teacher network are fixed, and the student network is trained simultaneously on HR and LR images under the guidance of the HR teacher network. In this way, the gap between the HR and LR features subspace is explicitly minimized. Previous research only used the last-layer features or the softmax probability layer for guidance. In the proposed methodology, attention maps are utilized to distill informative HR features from intermediate convolutional layers in combination with the last-layer features.

4.2.2.1 Attention Maps

Attention maps [105] are spatial maps that try to encode those spatial areas of the input image, which are given more emphasis by the network for its output decision. Consider a CNN network and a resulting 3D tensor map from one of its convolutional layers is represented by $F \in \mathbb{R}^{C\times H\times W}$, which is composed of C channels with spatial dimensions $H \times W$. Taking the above 3D tensor F as input, the mapping function $\mathcal{A}$ generates a spatial attention map, which is a flattened 2D tensor defined over the spatial dimensions.

$$\mathcal{A} : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{H\times W}$$

Mathematically, it is represented as:

$$\mathcal{A}(F) = \sum_{i=1}^{C} |F_i|^p \quad \text{where} \;\; p > 1 \tag{4.5}$$

FIGURE 4.4: HR and LR facial images with corresponding spatial attention maps are shown, highlighting where the teacher network has focused in low-, mid-, and high-level features.

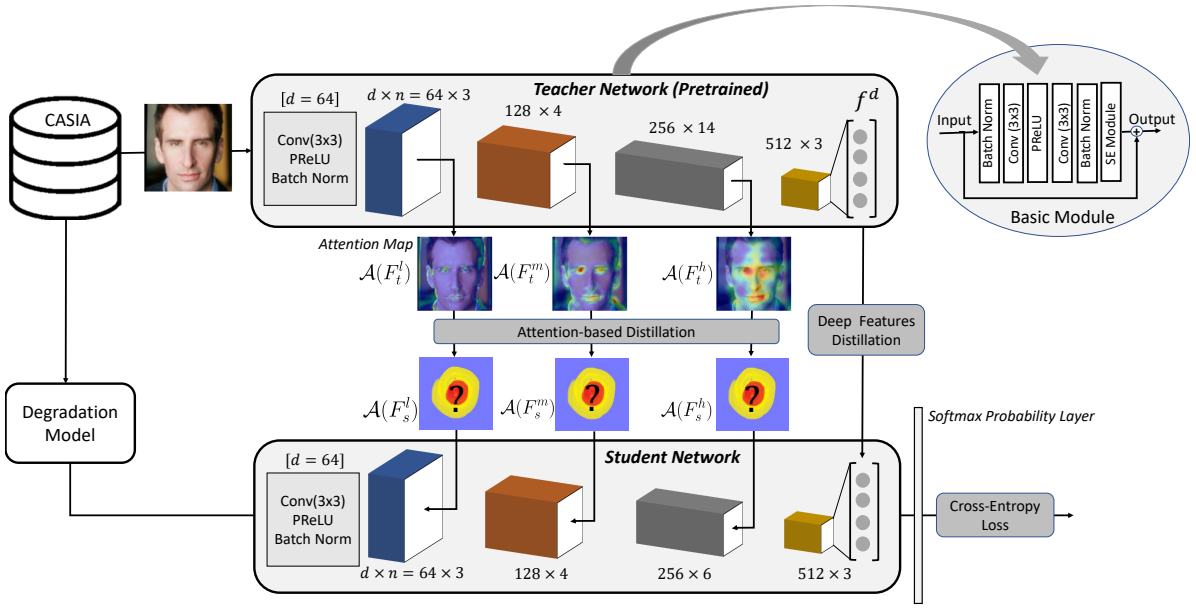


FIGURE 4.5: Complete framework showing degradation model and attention-guided distillation. The basic module of the teacher and student network is shown at the top-right corner. The architecture is defined using the terminology $d \times n$ for each module, where d represents the depth of the output of the convolutional layer in each module and n represents the repetition of module.

In the HR images, the attention maps highlight the information that is essential for CNNs in learning discriminative features. In Fig. 4.4, attention maps are calculated for HR and corresponding LR images from various layers of the HR teacher network. The HR network can clearly focus on the eyes, nose, and mouth regions in the HR image. However, details in these facial areas become unclear and difficult for the network to focus on in the LR image. It results in reduced efficiency in learning discriminative features. Therefore, the guidance of spatial attention maps was employed in the proposed mechanism to distill informative HR features from the HR teacher network to the student network.

4.2.2.2 Proposed Solution

The ideal solution for the cross-resolution scenario is to shift the common feature subspace to the HR feature subspace. The proposed solution shifts the common feature subspace as close as possible to the HR feature subspace. The proposed strategy is shown in Fig. 4.5.

Teacher and Student Network: In this work, the teacher network is assumed to be represented by:

$$\mathcal{N}_t = (\mathcal{F}_t, \mathcal{S}_t) \tag{4.6}$$

It consists of feature extraction layers $\mathcal{F}_t$ and softmax classification layer $\mathcal{S}_t$. The feature extraction layers can be further breakdown into:

$$\mathcal{F}_t = [\mathcal{F}_t^l; \mathcal{F}_t^m; \mathcal{F}_t^h; f_t^d] \tag{4.7}$$

Equation 4.7 denotes low-level, mid-level, high-level and deep feature layers. The output of low-level, mid-level and high-level feature layers is a 3D tensor, while the output of the deep feature layer is a 1D vector, also known as representation or embedding. Thus, given an HR image $\mathcal{I} \in \mathbb{I}_H$, first, low-level features are computed, then mid-level, high-level and at last deep features, given by:

$$F_t^l(\mathcal{I}) = \mathcal{F}_t^l(\mathcal{I}; w_t^l) \tag{4.8}$$

$$F_t^m(\mathcal{I}) = \mathcal{F}_t^m(F_t^l(\mathcal{I}); w_t^m) \tag{4.9}$$

$$F_t^h(\mathcal{I}) = \mathcal{F}_t^h(F_t^m(\mathcal{I}); w_t^h) \tag{4.10}$$

$$f_t^d(\mathcal{I}) = \mathcal{F}_t^d(F_t^h(\mathcal{I}); w_t^d) \tag{4.11}$$

Deep features are then processed through softmax classification layer $\mathcal{S}_t(f_t^d(\mathcal{I}); w_t^p)$ in order to obtain classification scores where w_t^p denotes weights of the softmax probability layer. $w_t = [w_t^l; w_t^m; w_t^h; w_t^d; w_t^p]$ denote the weights of the teacher network. For the student network, which is comparatively smaller than the teacher network, the same nomenclature will be followed for network representation, with subscript s instead of t.

Distillation Mechanism: The teacher network is initially trained on HR dataset $\mathbb{I}_{\mathrm{H}}$ and its weights are then frozen. The student network is then trained on synthetic LR dataset $\mathbb{I}_{\mathrm{S}}$ while being guided by attention maps and deep features from the teacher network. The guidance through an attention map is provided through attention-based distillation by employing squared Euclidean loss between the attention maps of the student and teacher network. The deep feature distillation is employed by calculating squared Euclidean loss between the last layer features of the teacher and student networks. Cross entropy loss function is used as classification loss for the student network

to learn both LR and HR features in the common feature subspace. The guidance provided by attention maps, coupled with deep feature distillation and cross-entropy loss, allows the student network to retain both HR and LR features.

Distillation Losses: The deep feature distillation loss encouraged the student network to extract the features that are aligned closely with those of the HR teacher network. The teacher network acts as a guide, transferring knowledge learned from HR data to the student network. The deep feature distillation loss is defined as:

$$\mathcal{L}_{dfd} = \sum_{\mathcal{I}' \in \mathbb{I}_S} \| f_t^d(\mathcal{I}) - f_s^d(\mathcal{I}') \|_2^2. \tag{4.12}$$

The subscript in $\mathcal{L}_{dfd}$ corresponds to the deep feature distillation.

The attention-based distillation loss guides the student network to focus on those areas of the input images that the teacher network emphasizes in its feature maps at low, mid, and high-level convolutional layers. This leads to more effective and informative transfer of HR knowledge and improved generalization ability in the student network. The attention-based distillation loss is given by:

$$\mathcal{L}_{ad} = \sum_{\mathcal{I}' \in \mathbb{I}_S} \sum_{f \in [l,m,h]} \| \frac{\mathcal{A}(F_t^f(\mathcal{I}))}{\|\mathcal{A}(F_t^f(\mathcal{I}))\|_2} - \frac{\mathcal{A}(F_s^f(\mathcal{I}'))}{\|\mathcal{A}(F_s^f(\mathcal{I}'))\|_2} \|_2^2. \tag{4.13}$$

The subscript in $\mathcal{L}_{ad}$ corresponds to the attention-based distillation. It is important to note that the attention map undergoes a transformation into a vector, followed by l_2 normalization.

The overall attention-guided distillation loss combines attention-based distillation loss, deep feature distillation loss, and cross-entropy loss function, given by:

$$\mathcal{L}_{agd} = \mathcal{L}_{cl} + \lambda_1 \mathcal{L}_{dfd} + \lambda_2 \mathcal{L}_{ad}. \tag{4.14}$$

The cross-entropy loss function, denoted by $\mathcal{L}_{cl}$, was adopted as classification loss for the student network. The subscript in $\mathcal{L}_{agd}$ and $\mathcal{L}_{cl}$ corresponds to the attention-guided distillation and classification, respectively. λ_1 and λ_2 are auxiliary weights of the loss function.

4.3 Experimental Setup

4.3.1 Implementation Details

Experiments are conducted using the Casia Webface dataset for training. The teacher and student networks are based on SENet-50 and SENet-34 [23], respectively. The teacher network is trained for 30 epochs with a batch size of 64, while the student network is trained for 35 epochs using batch size 32. Initially, the learning rate of 0.1 is used and decreased by a factor of 10 at 17 and 27 epochs. Weight decay and momentum are set to 0.0005 and 0.9, respectively. The values of λ_1 and λ_2 are set to 0.1 and 0.01, respectively, and the value of p in calculating attention maps is set to 2. These values were selected empirically after several experiments. After the training phase, the softmax layer is removed, and for each test image, the network outputs a feature vector of 512 dimensions. To recognize the identity in the test image, its Euclidean distance is then calculated with the feature vectors of gallery images.

4.3.2 Degradation Model Setting

In the implementation of the degradation model, a random perturbation of 0 to 3 pixels is introduced to the coordinates of the facial landmarks detected by the face detector. During misalignment, images are resized 112×112. Gaussian and Motion blur kernels are selected with two-thirds and one-third probabilities, respectively. To induce a significant blurring effect in the image, a Gaussian kernel size of 41 is selected, and its standard deviation σ is randomly sampled from $[0.1, 10]$. In motion blur, the value of L is kept at 7 and the value of θ is randomly chosen from $\{0, \pi/2\}$. The images are randomly down-sampled to various resolutions of size $p \times p$ where $p \in \{112, 112, 110, 100, 80, 60\}$. The noise sigma range is set to $[0, 20]$. After adding noise, images are resized to size $m \times m$ where $m = 112$. The quality factor q of the JPEG compression technique is randomly selected from $\{20, 40, 60, 80\}$. If the size for down-sampling is randomly selected to 112, then no degradation effects are applied, and the degradation model outputs an HR image. In this way, the synthetic LR dataset contains a combination of HR and synthetic LR images.

4.4 Results and Discussion

4.4.1 Ablation Study

In the ablation study, the impact of each contribution to the proposed approach is analyzed. Experiments are performed on all the images of the SCface and COXface datasets without splitting it into training and testing. The results are tabulated in Table 4.2. Our baseline involved simultaneous training on HR and down-sampled LR images using Softmax loss.

The degradation model achieved the highest performance on distance d1 of the SCface dataset and video V1 of the COXface dataset. This was due to its ability to accurately model the degradation effects observed in these datasets within the synthetically generated training data. However, this advantage came at the cost of reduced performance on HR images, particularly those captured at distance d3 of the SCface dataset.

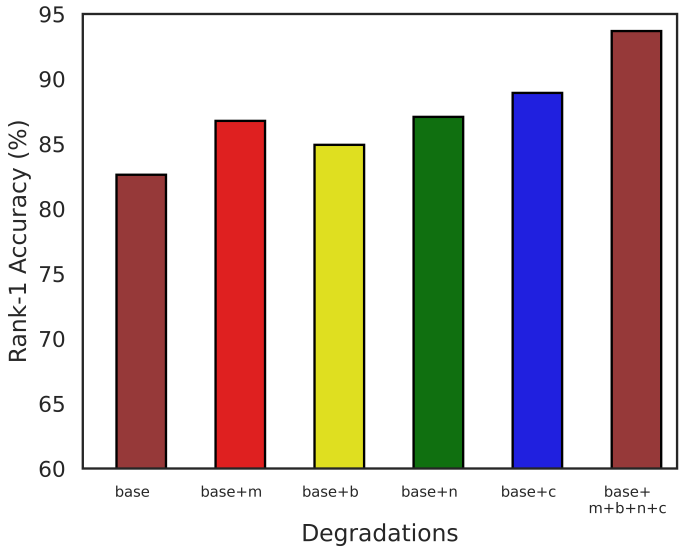
The combination of attention-based distillation with deep feature distillation led to superior performance on distance d3 of SCface and video V3 of COXface compared to deep feature distillation used individually. This demonstrates the effectiveness of attention-based distillation in preserving HR features.

The proposed scheme, employing attention-guided distillation, outperformed individual improvements in most scenarios, offering a good trade-off between HR and LR face recognition. These results confirm the efficacy of both the degradation model and attention-based distillation for LR face recognition.

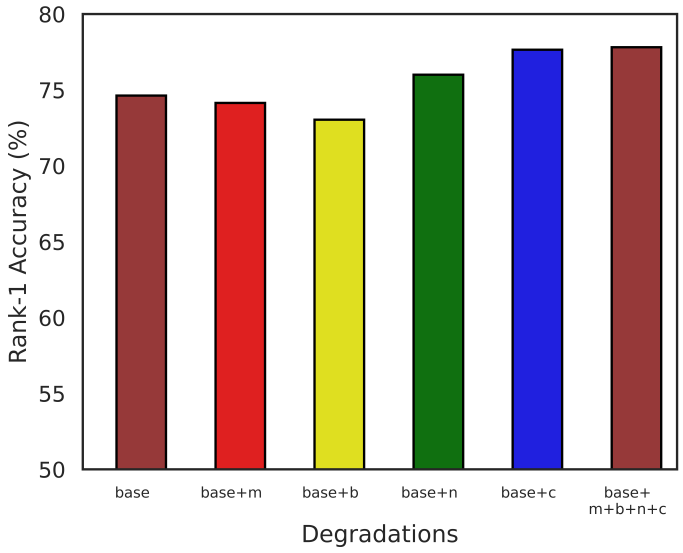
TABLE 4.2: Ablation study over each contribution in the proposed approach. (DM: Degradation Model)

Methods	$\mathcal{L}_{cl}$	**DM**	$\mathcal{L}_{dfd}$	$\mathcal{L}_{ad}$	**SCface**			**COXface**		
					d1	**d2**	**d3**	**V1**	**V2**	**V2**
Baseline	✓				82.62	96.62	94.92	74.63	81.86	94.18
I	✓	✓			**93.69**	**98.92**	97.50	**77.81**	**85.64**	95.44
II	✓		✓		85.30	98.20	98.40	75.83	83.61	95.38
III	✓		✓	✓	87.11	98.40	**99.00**	76.84	85.31	**95.84**
Proposed	✓	✓	✓	✓	92.92	**99.54**	**99.38**	**80.37**	**87.96**	**95.84**

In Fig. 4.6, the impact of each type of degradation on the recognition performance is analyzed. Experiments are performed on d1 and V1 of the SCface and COXface datasets, respectively. It is noted that each degradation has individually improved the performance as compared to the baseline. On SCface, the introduction of JPEG compression significantly improves the performance by 8.30%. Similarly, misalignment, blurring, and noise enhanced the performance by 4.15%, 2.3%, and 4.46%, respectively. The complete degradation model achieves the highest accuracy of 93.69%, improving the performance by 11.07% and confirming its effectiveness.



(A) SCface dataset (d1)



(B) COXface dataset (V1)

FIGURE 4.6: Ablation study over each degradation induced in the degradation model. m: misalignment, b: blurring, n: noise, c: compression

Unlike SCface, the trend observed in COXface is different. Misalignment and blurring have slightly decreased the performance due to the absence of such degradations in the COXface dataset. In contrast, noise and JPEG compression have individually improved the performance by 0.49% and 1.59%, respectively, compared to the baseline. The complete degradation model has demonstrated its effectiveness on the COXface dataset and achieved the highest accuracy of 77.81%.

The main aim of degradations is to lose information in HR images and make them representative of LR images. That's why degradations have improved performance on LR images. It is noted that in both datasets, JPEG compression significantly enhances performance on LR images. This is due to the fact that JPEG generates artifacts in synthetically generated LR images that are inherently found in real-world surveillance imagery.

4.4.2 Comparison with SOTA Methods

The proposed method is evaluated on LR datasets against SOTA techniques using two distinct protocols. The first protocol involves fine-tuning, meaning the model has been fine-tuned on the training partition of the dataset and evaluated on the testing partition. Fine-tuning is a common practice mentioned in the evaluation criteria of testing benchmarks. The second protocol excludes fine-tuning, which is used by some previous approaches as a more rigorous testing criterion. This protocol assesses the model's ability to generalize and perform well without additional adjustments, providing a more rigorous evaluation of its inherent capabilities.

4.4.2.1 LR Datasets

SCface: The rank-1 identification performance of the proposed scheme is rigorously evaluated against several state-of-the-art (SOTA) algorithms. Tables 4.3 and 4.4 present the results with and without fine-tuning, respectively. Focusing on the results without fine-tuning in Table 4.3, our proposed scheme consistently outperforms the previously proposed knowledge distillation technique from [69] across all three distances (d1, d2,

and d3) corresponding to different image resolutions. Notably, our scheme achieves an impressive accuracy of 95.25%, surpassing the previous best of 88.3% reported in [69]. This translates to a significant performance improvement of 6.95% for images captured at distance d1. Even after fine-tuning, the proposed approach continues to surpass the previous best [103], which utilized an adaptive down-sampling strategy and area-attention pooling.

TABLE 4.3: Performance Comparison on SCface Dataset (Evaluation on the testing partition without fine-tuning)

Methods	**Rank-1 IR**			**Average**
	4.2m	**2.6m**	**1.0m**	
Arcface [14, 53]	48.00	92.00	99.30	79.80
FAN [53]	62.00	90.00	94.80	82.30
DDL [70]	86.80	98.30	98.30	94.40
ARDAA & AAP [103]	87.23	98.46	98.92	94.87
RIFR [69]	88.30	98.30	98.60	95.00
Proposed	**95.25**	**99.50**	**99.00**	**97.91**

TABLE 4.4: Performance Comparison on SCface dataset. **-FT** means fine-tuning with the SCface training set.

Methods	**Rank-1 IR**			**Average**
	4.2m	**2.6m**	**1.0m**	
Arcface-FT [14, 53]	67.30	93.50	98.00	86.30
DCR-FT [57]	73.50	93.50	98.00	88.30
FAN-FT [53]	77.50	95.00	98.30	90.30
ARDAA & AAP -FT [103]	99.75	100.0	99.00	99.58
Proposed-FT	**99.75**	**100.0**	**99.50**	**99.75**

The effectiveness of the proposed scheme is further evaluated against a Super-Resolution (SR) technique. Generative Facial Prior (GFP) GAN [106], a well-known SR technique, is employed for facial image restoration. Initially, the image quality of the SCface dataset was enhanced using GFP-GAN, and subsequently, recognition was performed using our proposed scheme. Fig. 4.7 illustrates the comparison in Rank-1 identification

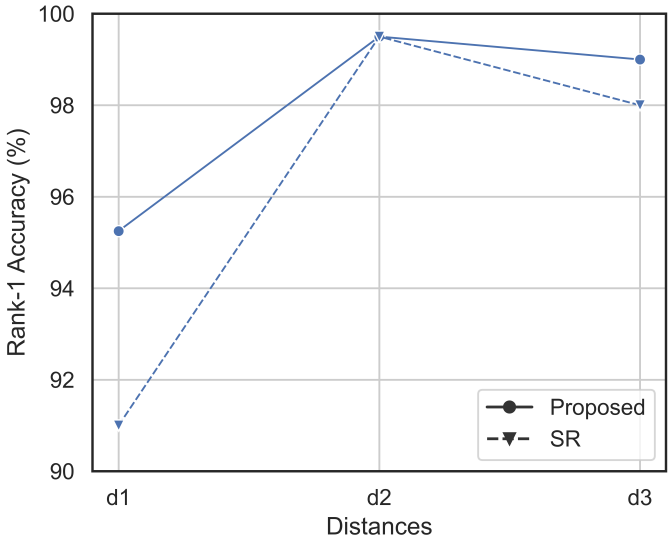


FIGURE 4.7: Performance comparison of the proposed scheme against SR technique on SCface dataset

performance. Interestingly, the SR technique significantly degrades recognition performance on LR images, while the impact is minimal for HR images. This is because SR techniques, while enhancing visual quality, can inadvertently discard the subtle features crucial for accurate facial recognition.

COXface: Tables 4.5, 4.6, and 4.7 summarize the rank-1 identification performance of our proposed scheme on the COXface dataset under different scenarios, compared to various SOTA algorithms. Notably, the algorithms used for comparison differ between COXface and SCface because none have been evaluated on both datasets. Table 4.5 presents the results for a video-to-still face recognition scenario. Here, each LR image captured from the videos is compared against all the HR images. Our proposed scheme

TABLE 4.5: Performance Comparison on COXface dataset: Video-to-Still face recognition. **-FT** means fine-tuning with the COXface training set.

Methods	**Mean IR ± Standard Deviation**		
	V1-S	**V2-S**	**V3-S**
TBE-CNN-FT [58]	88.2±0.45	87.86±0.85	95.74±0.67
CCM-FT [107]	88.65±1.1	87.82±0.8	92.13±0.9
HaarNet-FT [59]	89.31±0.94	87.90±0.60	97.01±1.65
Proposed-FT	**94.28 ±0.14**	**95.06±0.25**	**98.00±0.05**
Proposed	82.23 ± 0.34	89.80±0.41	96.82±0.07

significantly outperforms previous techniques based on the distance metric learning approach on all three videos (V1, V2, and V3). Without fine-tuning, our scheme achieves performance comparable to HaarNet [59] on videos V2 and V3, indicating the superior generalization capability of our model. When fine-tuning is applied, the proposed scheme achieves an accuracy of 94.28% on video V1, which contains severely degraded images of the COXface dataset. This exceeds the previous best of 89.31% [59] by a significant margin of 4.97%.

Table 4.6 presents the results of the still-to-video face recognition scenario. Here, each HR still image is compared against all LR images in the dataset. As expected, identification performance in this scenario is comparatively higher than the video-to-still scenario due to the presence of multiple LR images of the same subject in the gallery set. Notably, the proposed scheme with fine-tuning achieves an IR of 97.62% on video V1, outperforming HaarNet [59] i.e., 92.73%.

TABLE 4.6: Performance Comparison on COXface Dataset: Still-to-Video face recognition. **-FT** means fine-tuning with the COXface training set.

Methods	**Mean IR ± Standard Deviation**		
	S-V1	**S-V2**	**S-V3**
TBE-CNN-FT [58]	93.57±0.65	93.69±0.51	98.96±0.17
HaarNet-FT [59]	92.73±1.93	93.57±1.62	97.48±1.54
Proposed-FT	**97.62±0.24**	**96.20±0.26**	**98.42±0.17**
Proposed	88.81±.29	89.43±0.69	97.06±0.16

Table 4.7 showcases the result of the video-to-video face recognition. Here, each LR image from one video is compared against all LR images in another video. With three videos in the dataset, this leads to six different matching combinations. The proposed scheme outperforms existing methods in all these combinations, demonstrating its effectiveness for LR-to-LR face matching.

The effectiveness of the proposed scheme is compared with the application of the SR technique on the V2S scenario of the COXface dataset. As shown in Fig. 4.8, the SR technique degrades performance at lower resolutions. However, as the resolution improves, the difference in performance becomes minimal.

TABLE 4.7: Performance Comparison on COXface Dataset: Video-to-Video scenario.

Methods	**Mean IR ± Standard Deviation**	
	V1-V2	**V1-V3**
VGGFace-FT	93.39±0.56	96.10±0.27
TBE-CNN-FT	97.20±0.26	99.30±0.16
Proposed-FT	**99.29±0.004**	**99.46±0.005**
Proposed	91.19±0.06	92.13±0.18
Methods	**V2-V1**	**V2-V3**
VGGFace-FT	99.30±0.16	96.60±0.52
TBE-CNN-FT	98.07±0.32	99.33±0.19
Proposed-FT	**99.50±0.007**	**99.77±0.001**
Proposed	97.14±0.04	97.39±0.03
Methods	**V3-V1**	**V3-V2**
VGGFace-FT	99.33±0.19	96.39±0.42
TBE-CNN-FT	98.16±0.23	96.39±0.42
Proposed-FT	**99.77±0.006**	**99.92±0.0003**
Proposed	98.32±0.01	98.51±0.01

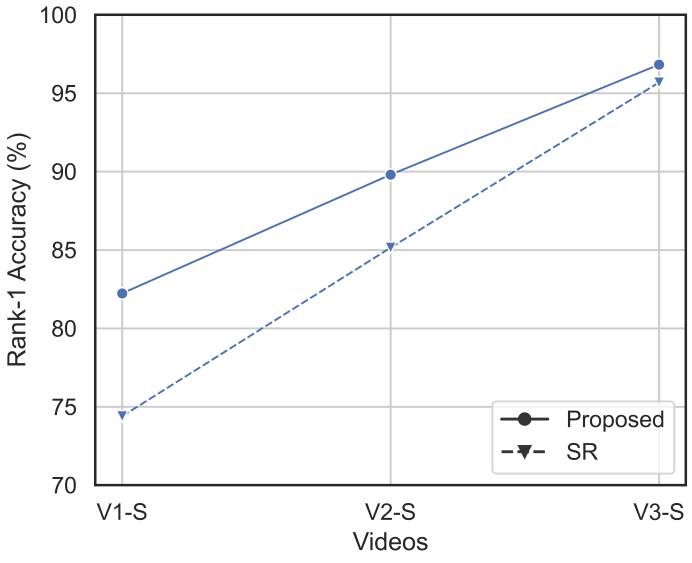


FIGURE 4.8: Performance Comparison of the Proposed scheme against SR Technique on COXface dataset

Tinyface: Table 4.8 provides a comprehensive summary of the identification performance of our proposed scheme on the Tinyface dataset. Both versions of our scheme, with and without fine-tuning, significantly outperform previous SOTA approaches. The Tinyface dataset comprises tiny and tightly cropped facial images, which can introduce artifacts at the boundaries following the alignment process. Despite these potential degradations, our proposed scheme demonstrates a remarkable ability to tolerate and effectively manage these imperfections, ensuring robust performance.

TABLE 4.8: Performance Comparison on Tinyface Dataset.

Methods	**IR**			**mAP**
	Rank-1	**Rank-20**	**Rank-50**	
DeepID2 [43, 52]	17.4	25.2	28.3	12.1
Sphereface [46, 52]	22.3	35.5	40.5	16.2
VGGface [52, 108]	30.4	40.4	42.7	23.1
Centerface [52, 109]	32.1	44.5	48.4	24.6
CSRI [52]	44.8	60.4	65.2	36.2
Proposed-FT	**60.78**	**70.54**	**74.66**	**50.14**
Proposed	54.18	66.44	70.30	45.14

4.4.2.2 HR datasets

The proposed scheme was evaluated on HR benchmark datasets (Table 4.9) and compared it with various HR algorithms, including Arcface [14] and NPT Loss [18]. The

TABLE 4.9: Performance Comparison on HR (1:1 Verification Rate)

Method	**LFW**	**AgeDB**	**CFP-FP**	**CALFW**
Norm-Softmax [18, 110]	97.55	87.14	87.15	88.46
Proxy-Triplet [18, 49]	97.48	84.15	90.90	85.31
Arcface [14]	99.30	94.23	95.30	93.34
Curricularface [15, 18]	99.36	94.18	95.61	93.34
NPT loss [18]	99.40	95.38	96.81	93.46
Proposed	**97.70**	**85.60**	**89.84**	**87.46**

proposed scheme exhibits lower performance on this task. This highlights that HR images from the SCface dataset do not necessarily translate to optimal performance on HR images captured by digital cameras. Furthermore, the results suggest that evaluating LR face recognition on HR benchmarks can provide valuable insights for comprehensive analysis.

4.4.3 Discussion

The ablation study confirms the effectiveness of the degradation model and attention-guided distillation within our proposed approach. The degradation model, which introduces real-world surveillance and degradation effects into synthetic LR images, proves to be more efficient than simple down-sampling techniques commonly used in previous work. This reduces the domain gap between training and testing data, leading to significant improvements in LR face recognition performance. Attention-based distillation combined with deep feature distillation also demonstrates its significance by minimizing the gap between LR and HR features as compared to deep feature distillation alone.

The SCface dataset has different resolutions of images and is perfect for testing cross-resolution face recognition problems. In previous work, distillation techniques have utilized only the last layer features, and training has been carried out on down-sampled LR images. That's why the results are limited. Our proposed technique utilizes a degradation model that can make the training data more actual representative of the conditions in which the system will be used, and incorporating spatial attention maps from convolutional layers in combination with last layer features can preserve both HR and LR features more efficiently.

The COXface dataset also consisted of images in different resolutions. The previous methods consist of distance metric learning approaches and have used a modified form of Triplet loss. Due to heavy dependency on mining good triplets, its accuracy is limited. Our proposed approach consisted of two stages that can better leverage the information in HR and synthetic LR images.

The LFW and its updated versions are commonly used as HR benchmark datasets. Although our proposed approach has achieved slightly lower performance, it highlights

the impact of using synthetic LR images for training on HR performance. This challenge will be further investigated in the next chapter to develop systems that can achieve optimal results on both LR and HR benchmark datasets.

4.5 Conclusion

In this chapter, a novel training strategy has been proposed for highly accurate LR face recognition. Two major limitations of existing LR face recognition techniques are addressed by focusing on performance improvement in real-world surveillance scenarios. The limitations include early performance saturation on down-sampled LR images and unsatisfactory performance on HR images. To address the early performance saturation problem on down-sampled LR images, a degradation model that simulates real-world degradations in synthetic LR images was proposed. The face recognition system trained on those synthetic LR images can tolerate natural blurring and noisy effects, resulting in improved performance. To simultaneously achieve better results on LR images as well as HR images, an attention-guided distillation scheme is proposed that transfers informative HR features from the HR teacher network to the LR student network. The proposed scheme combining both strategies is tested on popular real-world surveillance/LR datasets, i.e., SCface, COXface and Tinyface. The comparative analysis has demonstrated significant performance improvement on LR images in these datasets with a margin of 6.95%, 4.97% and 9.38%, respectively, as compared to the previous SOTA methods, which confirms the effectiveness of the proposed degradation model and attention-guided distillation mechanism.

Chapter 5

New Protocols

5.1 Outline

The training-testing dichotomy remains a fundamental aspect of dataset protocols in machine learning. This approach is crucial for assessing the model's generalization ability on unseen data. With the advent of deep learning, which is considered data-hungry, enormous large-scale datasets have been proposed for visual recognition problems. In the case of face recognition problem, millions of facial images are collected from the web, cleaned, and made publicly available for training. This abundance of data has also impacted testing procedures in HR face recognition. Previously, HR face recognition models were validated by training on one portion of the HR benchmark dataset and testing the remaining portion. The recent methods involve training the HR model on either large-scale or small-scale training datasets entirely separate from the benchmarks that are used for testing. In LR face recognition problems, the previous approach is followed. The LR face recognition model is trained on one of the large-scale or small-scale datasets that is augmented with LR images, fine-tuned on the training partition of LR benchmarks, and tested on the remaining portion. The first downside of this approach is that the model might memorize patterns specific to that dataset instead of learning discriminative features. Secondly, even with a strict train-test split, there might be subtle overlaps or similarities within the dataset that the model can exploit, leading to unrealistic performance metrics. Thirdly, the LR testing benchmarks have

limited variation in resolution, making it challenging to assess the effectiveness of LR face recognition techniques and analyze their limitations.

5.2 Limitations in the Existing Protocols

Rigorous experimentation is conducted to highlight the limitations of the existing testing benchmarks and their associated protocols. This thorough analysis enables us to develop new protocols that not only highlight efficacy but also analyze limitations more effectively. By identifying and addressing these shortcomings, we can significantly improve the accuracy and reliability of our evaluation methods.

5.2.1 Fine-tuning

The drawback of fine-tuning contributes to the potential loss of the model's generalizability. A model fine-tuned on one dataset might not perform well on others, making it challenging to assess the effectiveness of the approach in learning discriminative features. To demonstrate this issue, the ResNet-50 model was implemented with fine-tuning and its results were compared with a more advanced method REE [63] on the SCface and Tinyface datasets. As evident from the results in Table 5.1, the ResNet-50 model outperformed REE [63] on SCface despite being an inferior model because of straightforward training on HR and LR images. For Tinyface, although REE [63] outperforms our ResNet-50 model, it is unclear whether they used only the training partition or incorporated additional LR images during fine-tuning. This ambiguity and lack of details about fine-tuning raises questions about its validity for evaluation. Therefore, to more

TABLE 5.1: Performance comparison of ResNet-50 with REE [63] on fine-tuned protocols

Methods	**FT**	**Rank-1 IR**	
		SCface (avg)	**Tinyface**
REE (R-50, Vggface2) [63]	✓	98.70	77.22
Curricularface (R-50, vggface2)	✓	99.75	70.98

effectively evaluate the model, it should be evaluated on the complete dataset without dividing it into training and testing.

5.2.2 Lack of Evaluation on HR Datasets

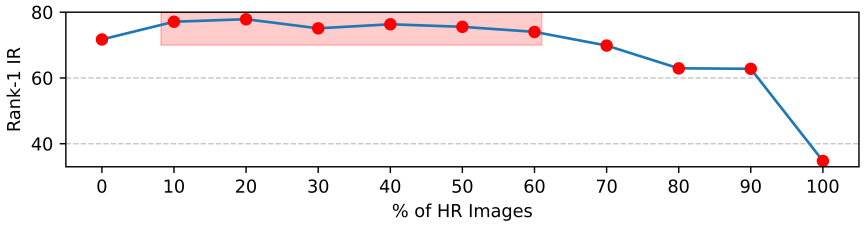
The joint training on HR and LR images for LR face recognition can lead the model to discard informative HR features. The effect becomes more pronounced as the proportion of LR images increases and the degradation of these images worsens. This results in improved performance on LR images but at the cost of a significant drop in performance for HR images. Existing testing scenarios for LR face recognition are often small-scaled and lack sufficient variation to highlight these issues effectively. Even the SCface dataset, known for testing resolution-invariant face recognition, exhibits this limitation. Interestingly, as more severe degradations like JPEG artifacts are introduced, the accuracy of d3 (presumably HR images) of SCface also increases. However, the performance on the HR dataset decreases with the introduction of JPEG artifacts, as shown in Table 5.2. Therefore, to more effectively validate resolution invariance in the LR face recognition, it is necessary to evaluate the model both on HR and LR datasets.

TABLE 5.2: Performance evaluation of a face recognition model in HR and LR scenarios

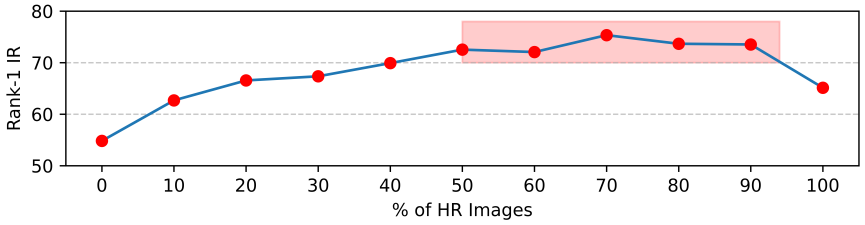
Degradations (R-18, casia)	**SCface**			**HR**
	4.2m	**2.6m**	**1m**	
HR	34.77	83.23	96.00	92.59
HR + LR (Down-sampled)	75.54	94.62	94.92	89.36
HR + LR (JPEG)	81.23	97.38	97.54	88.30

5.2.3 Different Nature of LR Datasets

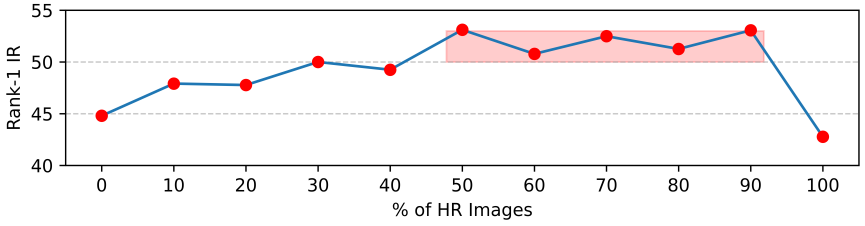
In Fig. 5.1, the performance response of four widely used LR datasets to different percentages of HR images in a batch during training, is plotted. Among them, SCface is biased towards a high percentage of LR images, while COXface and Tinyface are biased towards a high percentage of HR images. QMUL Survface exhibits fluctuations and is biased towards a high percentage of HR images. For a fair analysis, evaluating the LR



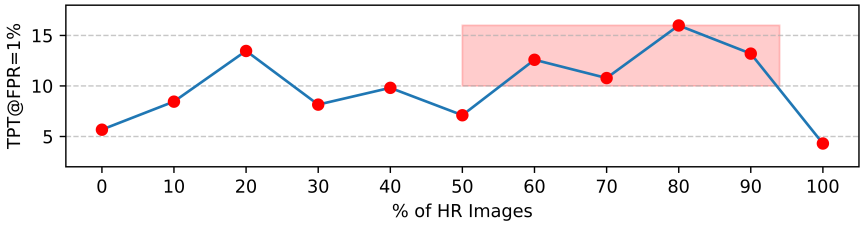
(A) SCface (d1)



(B) COXface (cam 1)



(C) Tinyface



(D) QMUL Survface

FIGURE 5.1: Response of LR datasets to different percentages of HR images in a batch during training

face recognition model on all the datasets with different natures is necessary to highlight its efficacy and limitations.

5.2.4 Combined Evaluation Metric

The Combined Evaluation Metric (CEM) is introduced to evaluate performance across various image resolutions, including HR, MR, and LR. This metric is designed to maintain high recognition accuracy regardless of image quality. Inspired by the F1-score's capability to balance precision and recall, the harmonic mean assigns greater weight

to the lower recognition rate, whether it be HR or LR. Essentially, CEM is based on the harmonic mean, which is particularly well-suited for this task due to its ability to balance different performance measures. CEM ensures that models experiencing a significant drop in accuracy on any resolution are penalized in the overall metric, thereby promoting robustness across varying image qualities.

5.3 New Protocols

In light of the aforementioned empirical evidence, the following testing protocols for LR face recognition are proposed:

1. Train the model on one dataset (either small-scale or large-scale) and test completely on different datasets without any fine-tuning process.

2. Use SCface, Coxface, Tinyface and QMUL Survface dataset for LR testing.

3. For HR testing, calculate the average verification accuracy for LFW, Agedb-30, CALFW, CPLFW, and CFP-FP, along with individual accuracy.

4. For Mixed Resolution (MR) testing, calculate the TAR at FAR=1e-4 for IJB-B and IJB-C datasets.

5. Use CEM to judge overall performance across different datasets. The performance metrics for each dataset is listed in Table 5.3.

TABLE 5.3: Evaluation metrics for HR and LR testing benchmarks in Combined Evaluation Metric (CEM)

Datasets	**Metric**
SCface	IR@d1
Tinyafce	IR(Rank-1)
COXface	IR@V1 (V2S)
QMUL	Average(TAR@FAR=[0.3, 1e-1,1e-2,1e-3])
HR	Average
IJB-B	TAR@FAR=1e-4
IJB-C	TAR@FAR=1e-4

5.4 Summary

Testing protocols are fundamental for evaluating the effectiveness and limitations of any machine learning or deep learning approach, particularly overfitting to the training data. In particular, adopting separate datasets for training and testing, rather than splitting a single dataset, is more suitable for robust model evaluation. Extensive experimentation was performed to analyze the limitations in LR testing benchmark datasets and their protocols. Finally, new protocols were proposed to effectively assess efficacy as well as limitations of LR face recognition model.

Chapter 6

Sub-center Learning and Contrastive Distillation Loss

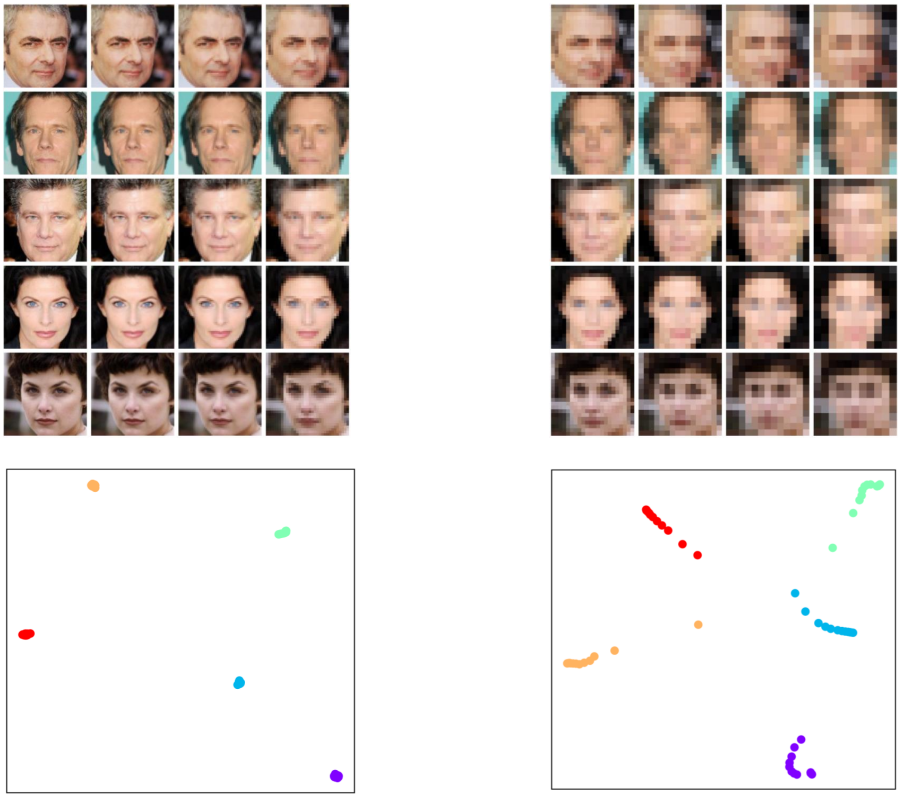
6.1 Outline

The convergence of three key advancements – large-scale cleaned facial datasets, increased computational power, and advanced deep CNN architectures – has significantly propelled progress in face recognition research. Initially focused solely on recognizing LR images, LR face recognition systems are now expanding their capabilities to handle images of varying resolutions, including HR ones. This integration of diverse image resolutions aims to improve the accuracy and robustness of face recognition technology, making it more adaptable to real-world scenarios where image quality can vary significantly.

In the HR face recognition problem, the training data consists of HR images. The primary goal is to increase the inter-class distance while decreasing the intra-class distance among the extracted features during training. Arcface [14] and NPT Loss [18] employed the same principle and demonstrated remarkable performance on HR images, even with considerable variations in pose, illumination, and expression. In contrast, the training dataset for the LR face recognition problem is augmented with LR images. The proposed losses for the HR face recognition problem are unable to map both HR and LR features close to each other [111], and the objective of achieving high inter-class distance

and low intra-class distance is not met. It also leads to a slight decline in the performance of HR images [69]. Hence, an explicit compulsion is required to minimize the gap between HR and LR features. Therefore, the objective of the LR face recognition problem is to reduce the gap between HR and LR features.

Augmenting the dataset with LR images in the LR face recognition problem poses a significant challenge while learning discriminant features from HR and LR images. In Fig. 6.1, HR images are down-scaled to 30% of their original size and are visualized using t-SNE. It shows that there is no significant effect on the distribution of the dataset. In contrast, when HR images are downscale to the range between 30% and 10% of the original size, i.e., close to real world representation of LR images and visualized using t-SNE, it is evident that the inter-class similarity and intra-class variance both have increased. The high inter-class similarity corresponds to the challenge of distinguishing between very LR images of distinct identities, while high intra-class variance corresponds to variation between the varying resolution images of a single identity.



(A) High Resolution (B) Low Resolution

FIGURE 6.1: High-resolution (left) and low-resolution (right) samples of images visualized through t-SNE

Knowledge distillation techniques are the most prevalent techniques in the LR face recognition problem. In Knowledge distillation techniques, distillation loss plays a major role in minimizing the gap between LR and HR features. Commonly employed distillation losses include Mean Squared Error (MSE) [68] and Kullback-Leibler (KL) divergence loss [69, 112]. These losses somewhat tackle the issue arising from high inter-class similarity and high intra-class variance in the dataset when learning discriminative features. However, the problem remains due to the inherent challenges in the data.

In the proposed methodology, the issues in the LR face recognition problem are categorized into high intra-class variance and high inter-class similarity in the LR dataset. High intra-class variance is tackled by introducing sub-center learning in LR face recognition. Sub-center learning uses multiple sub-centers for each class and captures the variations between the varying LR images of a single identity more efficiently as compared to the single center used previously. The second issue, high inter-class similarity, is tackled through the proposed contrastive distillation loss. The contrastive distillation loss minimizes the distance between corresponding LR and HR features while at the same time maximizing the distance between non-corresponding features, leading to the learning of more compact and discriminative features.

6.2 Proposed Methodology

The training data comprises images with varying resolutions in LR face recognition. This presents two significant challenges:

1. **High Intra-Class Variance**: Images of the same identity captured at different resolutions can exhibit substantial visual differences, making recognition challenging.
2. **High Inter-Class Similarity**: It is challenging to differentiate very low-resolution images of distinct identities due to the loss of fine-grained details.

The introduction of sub-center learning in LR face recognition tackles the problem of high intra-class variance. At the same time, the proposed contrastive distillation loss addresses the challenge of high inter-class similarity.

6.2.1 Sub-center Learning

In sub-center learning, multiple centers are learned for each class during training, rather than a single center. The distance between the features is minimized according to the closest sub-center. Sub-center Arcface [14], a notable example of sub-center learning, has been effectively employed for noise removal from datasets by associating clean and noisy samples with different sub-centers. In LR face recognition, high intra-class variance in the training data poses challenges for parameter optimization, where samples are centered on a single class center. This results in capturing limited variations in the learned features. Sub-center learning offers a more suitable strategy for addressing this challenge, enabling the model to capture the diverse representations of individuals across varying resolutions using multiple sub-centers. However, multiple sub-centers can result in multiple clusters within a class, increasing intra-class variance in the learned features and impacting performance in HR scenarios. This issue is addressed by the proposed contrastive distillation loss in Section 6.2.2.

In sub-center learning, the weights associated with the classification layer are defined according to the sub-centers defined for each class. The classification layer can be represented as:

$$z_{j_k,i} = W_{j_k}^T x_i, \qquad z_{j_k,i} \in \mathbb{R}^{k \times N}, \tag{6.1}$$

where $x_i \in \mathbb{R}^d$ denotes the i^{th} feature vector of dimension d. $W \in \mathbb{R}^{d \times (k \times N)}$ is the weight matrix whereas $W_{j_k} \in \mathbb{R}^{d \times (k \times 1)}$ is the weight matrix for each class, k is the number of sub-centers while N is number of classes. The feature vectors and the weights are both normalized on a unit hyper-sphere and a max-pooling step is employed with respect to the sub-centers for each class:

$$z_{j_{max},i} = \max_k(\hat{W}_{j_k}^T \hat{x}_i), \;\; z_{j_{max},i} \in \mathbb{R}^{1 \times N}, \tag{6.2}$$

$$z_{j_{max},i} = \max_k(\cos\theta_{j_k,i}), \tag{6.3}$$

$$z_{j_{max},i} = \cos\theta_{j_{max},i}, \tag{6.4}$$

where $\theta_{j_{max},i}$ represents the angle between the i^{th} feature vector and the weight vectors associated with class j with the highest cosine similarity among the sub-centers k. Defining y_{max} as the class label of the feature vector with highest cosine similarity

among the sub-centers, $\cos\theta^{+}_{y_{max},i}$ is the cosine similarity of the target class of the i^{th} sample and $\cos\theta^{-}_{j_{max},i}$ are the cosine similarities of remaining classes, i.e., $j=1:N$ and $j \neq y$. Additionally, introducing m as an additive angular margin parameter, the loss function for the sub-center learning for the i^{th} sample becomes:

$$\mathcal{L}_{sl} = -\log\left[\frac{\exp(s\cos(\theta^{+}_{y_{max},i}+m))}{\exp(s\cos(\theta^{+}_{y_{max},i}+m)) + \sum\limits_{\substack{j=1,\\ j\neq y}}^{N}\exp(s\cos\theta^{-}_{j_{max},i})}\right], \tag{6.5}$$

where s is the scaling parameter, its effectiveness is proved in [110]. The subscript in $\mathcal{L}_{sl}$ corresponds to sub-center learning.

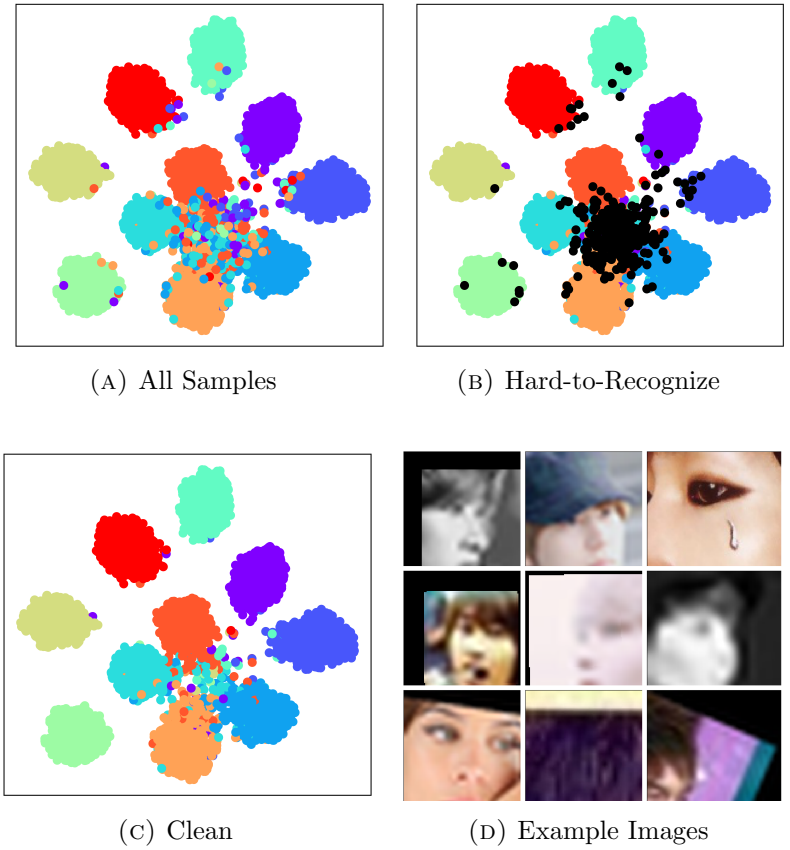
Sub-center learning loss is employed as a classification loss for the student network. Unlike traditional classification losses that rely on a single class center for each category, sub-center learning introduces multiple sub-centers per class. These sub-centers represent different variations within the same class, allowing the network to capture diverse feature distributions more effectively. In the context of LR datasets, images with varying levels of resolution can be mapped to one of the sub-centers for their corresponding class. This flexibility significantly eases the optimization process by accommodating the inherent variability in LR data. Instead of forcing all samples to cluster around a single rigid class center, the network learns to associate similar variations with the closest sub-center, improving its ability to generalize.

Feature Analysis in 2D: Face images from 10 different identities with enough samples (around 2,100 images) are selected to obtain a 2-D feature embedding space using sub-center learning. The distribution of features is shown in Fig. 6.2, where different clusters are visualized in distinct colours. The parameters are optimized according to the closest sub-center within a set of multiple sub-centers defined for each class. Consequently, one sub-center tends to dominate, attracting most data points, while the others become non-dominant sub-centers. Fig. 6.2(a) shows the features' distribution of all samples, while Fig. 6.2(b) highlights the features' distribution of samples attached to the non-dominant center. Notably, most samples associated with the non-dominant center form a cluster of hard-to-recognize samples known as unidentified identities (UIs) [113]. Removing these hard-to-recognize samples results in well-defined clusters, as shown in Fig. 6.2(c). This finding supports our hypothesis that optimizing parameters with a single center

becomes challenging due to the UIs when dealing with LR images. Therefore, the sub-center learning mechanism, which distributes samples among dominant and non-dominant sub-centers, is employed, leading to more efficient optimization of parameters. Example images of the hard-to-recognize samples are shown in Fig. 6.2(d).

6.2.2 Contrastive Distillation Loss

To overcome the issue of high inter-class similarity in the training dataset of LR face recognition and performance deterioration on HR images, two variants of contrastive distillation loss, namely feature-contrastive and class-contrastive distillation loss, are proposed. It is important to note that the student network outputs an LR feature vector, and the teacher network outputs an HR feature vector. In this regard, the feature-contrastive distillation loss minimizes the distance between the LR feature vector and its corresponding HR feature vector of the same class in contrast to other HR feature



(A) All Samples (B) Hard-to-Recognize

(C) Clean (D) Example Images

FIGURE 6.2: Analysis of sub-center learning using feature distribution through tSNE. Different colours denote different classes.

vectors in a batch. The class-contrastive distillation loss, on the other hand, minimizes the distance between the LR feature vector and its corresponding HR mean vector (class center) of the same class in contrast to the HR mean vectors of other classes.

6.2.2.1 Review of Other Distillation Losses

The purpose of distillation is to minimize the distance between the LR and corresponding HR features in LR face recognition. This results in transferring HR information from the HR to the LR model. The most widely used distillation losses are mean-squared error (MSE) and KL divergence loss.

Let x^t and x^s denote the output feature vectors of the pre-trained HR teacher and student network, respectively. Similarly, let $p(z^t)$ and $p(z^s)$ represent the outputs of the softmax probability layers of the pre-trained HR teacher and student network, respectively, where $z = W^T x$. The losses are given by:

$$\mathcal{L}_{mse} = \frac{1}{2}\|x^t - x^s\|_2^2. \tag{6.6}$$

$$\mathcal{L}_{kl} = \sum_{i=1}^{N} p(z_i^t) \log \frac{p(z_i^t)}{p(z_i^s)}. \tag{6.7}$$

The MSE calculates the squared Euclidean distance between the LR and corresponding HR features of the student and teacher network, respectively. The KL divergence minimizes the error between the distributions of the teacher and student network. The gradients of these losses are given by:

$$\frac{\partial \mathcal{L}_{mse}}{\partial x^s} = -(x^t - x^s). \tag{6.8}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_i^s} = p(z_i^s) - p(z_i^t)\,. \tag{6.9}$$

The gradient equations in equations 6.8 and 6.9 show that minimizing these losses ensures that both HR and LR features are pushed closer together. In contrast, the proposed contrastive distillation loss is based on the idea of contrastive estimation.

It pushes the corresponding samples towards each other while simultaneously pushing apart non-corresponding samples. The derivation of Equation 6.9 can be found in Appendix A.1.

6.2.2.2 Feature-Contrastive and Class-Contrastive Distillation Loss

The proposed distillation loss is based on the idea of noise contrastive estimation framework [114] and additive angular margin [14]. The loss consisted of query, positive and negative samples that are mapped to d-dimensional vector, $v, v^+ \in \mathbb{R}^d$, and $v^- \in \mathbb{R}^{d\times(N-1)}$, respectively, where $v_j^- \in \mathbb{R}^d$ denotes the j-th negative vector. The idea is to associate the query with its positive sample in contrast to $N-1$ negative samples. The vectors are normalized on a unit sphere to prevent space from collapsing or expanding, and the distance between the samples is scaled by a parameter s. The $N-way$ classification problem is set up and is represented by:

$$\mathcal{L}_{cdl} = -\log\left[\frac{\exp(s(\hat{v}\cdot\hat{v}^+))}{\exp(s(\hat{v}\cdot\hat{v}^+)) + \sum_{j=1}^{N-1}\exp(s(\hat{v}\cdot\hat{v}_n^-))}\right]. \tag{6.10}$$

The subscript in $\mathcal{L}_{cdl}$ corresponds to the contrastive distillation loss.

Feature-Constrastive: In feature-contrastive distillation loss, the information is maximized between the i^{th} feature vector of the student network, denoted by x_i^s and the corresponding HR feature vector of the teacher network, denoted by x_i^t. This is achieved in contrast to HR feature vectors of other classes in a batch, denoted by x_j^t. The loss is represented by:

$$\mathcal{L}_{f-cdl} = -\log\left[\frac{\exp(s(\hat{x}_i^s\cdot\hat{x}_i^t))}{\exp(s(\hat{x}_i^s\cdot\hat{x}_i^t)) + \sum_{j=1}^{B-1}\exp(s(\hat{x}_i^s\cdot\hat{x}_j^t))}\right], \tag{6.11}$$

where B denotes the batch size.

The equation 6.11 in terms of cosine angle is given by:

$$\mathcal{L}_{f-cdl} = -\log\left[\frac{\exp(s\cos\phi_{ii}^+)}{\exp(s\cos\phi_{ii}^+) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^-)}\right], \tag{6.12}$$

where ϕ_{ii}^{+} and ϕ_{ij}^{-} denote the angles between the corresponding and non-corresponding feature vectors of the teacher and student network, respectively. With an additive angular margin m incorporated into the cosine function to achieve both intra-class compactness and inter-class diversity, the loss is represented as:

$$\mathcal{L}_{f-cdl} = -\log \underbrace{\left[\frac{\exp(s\cos(\phi_{ii}^{+}+m))}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})} \right]}_{p_i^{+}}. \tag{6.13}$$

The subscript in $\mathcal{L}_{f-cdl}$ corresponds to the feature-contrastive distillation loss.

Feature contrastive distillation loss improved the knowledge distillation process between a teacher and student network by maximizing the cosine similarity between the corresponding feature representations from both networks. The introduction of an additive angular margin m into the cosine similarity enhances the discriminative power of the loss function. It increases the angular separation between features from different classes (non-corresponding features) in the angular space. This helps alleviate the problem of high inter-class similarity in HR datasets augmented with LR images.

Class-Contrastive: In class-contrastive distillation loss, the information is maximized between the i^{th} feature vector of the student network, denoted by x_i^s, and the corresponding HR class center of the teacher network, denoted by $\bar{x}_i^t$. This is achieved in contrast to HR class centers of other classes, denoted by $\bar{x}_j^t$. The $N-way$ classification problem is setup where N denotes the number of classes and is represented by:

$$\mathcal{L}_{c-cdl} = -\log \left[\frac{\exp(s(\hat{x}_i^s \cdot \hat{\bar{x}}_i^t))}{\exp(s(\hat{x}_i^s \cdot \hat{\bar{x}}_i^t)) + \sum_{j=1}^{N-1}\exp(s(\hat{x}_i^s \cdot \hat{\bar{x}}_j^t))} \right]. \tag{6.14}$$

Rewriting the equation 6.14 in terms of cosine and adding additive angular margin m, the class-contrastive distillation loss is represented as:

$$\mathcal{L}_{c-cdl} = -\log \underbrace{\left[\frac{\exp(s\cos(\varphi_{ii}^{+}+m))}{\exp(s\cos(\varphi_{ii}^{+}+m)) + \sum_{j=1}^{N-1}\exp(s\cos\varphi_{ij}^{-})} \right]}_{p_i^{+}}, \tag{6.15}$$

where φ_{ii}^{+} and φ_{ij}^{-} denote the angles between the feature vector and the corresponding / non-corresponding HR class centers, respectively. The subscript in $\mathcal{L}_{c-cdl}$ corresponds to the class-contrastive distillation loss.

Class contrastive distillation loss introduces a stricter constraint compared to feature contrastive distillation. Instead of aligning features directly between the teacher and student networks, it maximizes the cosine similarity between the feature representation of the student network and its corresponding class center as learned by the teacher network. This approach enforces the student to learn features that are not only aligned with the teacher's representation but are also tightly clustered around the corresponding class center.

Derivative Analysis: The gradient of equation 6.13 can be represented as:

$$\frac{\partial \mathcal{L}_{f-cdl}}{\partial \hat{x}_i^s} = -s(1-p_i^{+})\frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}}\hat{x}_i^t + s\sum_{j=1}^{B-1} p_j^{-}\hat{x}_j^t, \tag{6.16}$$

$$\text{where} \quad p_i^{+} = \frac{\exp(s\cos(\phi_{ii}^{+}+m))}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})},$$
$$p_j^{-} = \frac{\exp(s\cos\phi_{ij}^{-})}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}.$$

The derivation of equation 6.16 can be found in Section 6.2.4. The loss will be minimized if its gradient approaches zero.

$$-s(1-p_i^{+})\frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}}\hat{x}_i^t + s\sum_{j=1}^{B-1} p_j^{-}\hat{x}_j^t = 0, \tag{6.17}$$

$$p_i^{+}\frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}}\hat{x}_i^t + \sum_{j=1}^{B-1} p_j^{-}\hat{x}_j^t = \frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}}\hat{x}_i^t. \tag{6.18}$$

Since the margin m have a small value, the term $\frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}} \approx 1$.

$$p_i^{+}\hat{x}_i^t + \sum_{j=1}^{B-1} p_j^{-}\hat{x}_j^t \approx \hat{x}_i^t. \tag{6.19}$$

It can be seen in the above equation that the terms p_i^{+} and p_j^{-} are contrastive. Since our objective is to maximize the similarity between features from corresponding samples

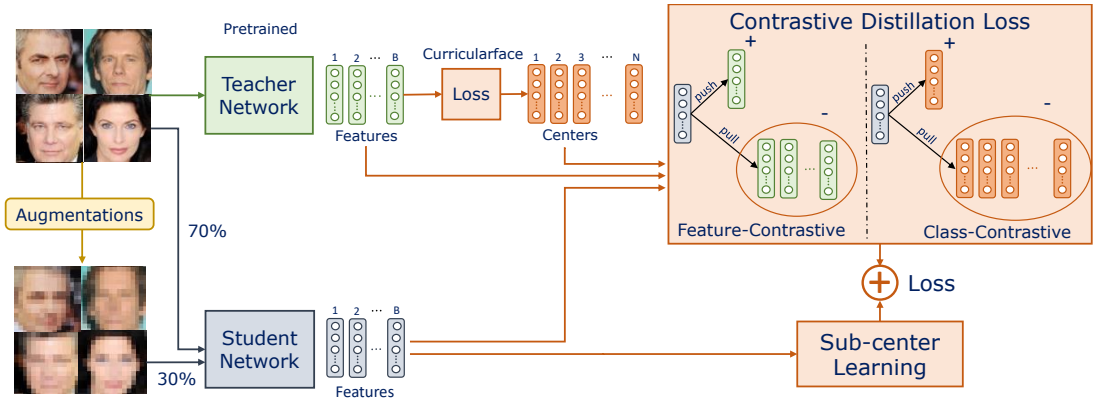


FIGURE 6.3: The proposed methodology consists of a pre-trained HR teacher network and a student network. The student network is trained using the sub-center learning and contrastive distillation losses. (N: Number of classes, B: Batch Size)

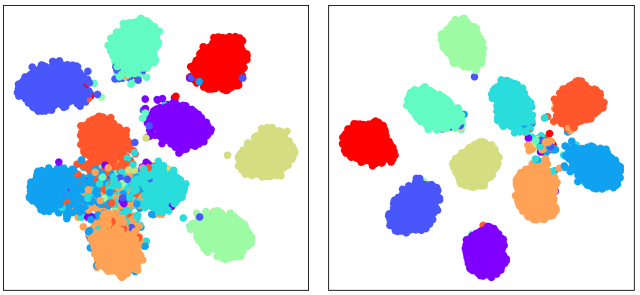
while minimizing similarity between those from non-corresponding samples, this translates to maximizing p_i^+ and minimizing all p_j^-. Consequently, the overall contrastive distillation loss is minimized.

6.2.3 Distillation Mechanism

The proposed distillation mechanism consists of two steps. In the first step, the teacher network is trained on one of the small-scale or large-scale datasets using Curricularface loss [15]. The student network is then initialized with the same weights as those of the teacher network, and is further trained simultaneously on HR and augmented LR images. The methodology of the proposed scheme is shown in Fig. 6.3. The combined loss function is represented by:

$$\mathcal{L} = \mathcal{L}_{sl} + \lambda_1 \mathcal{L}_{f-cdl} + \lambda_2 * \mathcal{L}_{c-cdl}. \tag{6.20}$$

Face images from 10 different identities containing enough samples (around 2,100 images) are selected to obtain 2-D feature embedding space using both schemes i.e., the Curricularface [15] as our baseline and the proposed methodology. In the visualization shown in Fig. 6.4, it can be observed that the baseline struggles to optimize parameters with respect to a single center due to the presence of hard-to-recognize samples. However, our proposed methodology has gathered the hard-to-recognize samples in a separate cluster and successfully learned discriminative features resulting in compact clusters of classes.



(A) Baseline (B) Proposed

FIGURE 6.4: The distribution of features visualized through t-SNE under Curricular-Face (Baseline) [15] **(Left)** and the proposed methodology **(Right)** is shown for 10 identities. Different colors denote different classes.

6.2.4 Derivation of the Gradient on Contrastive Distillation Loss

The feature-contrastive distillation loss can be represented by:

$$\mathcal{L}_{f-cdl} = -\log \underbrace{\left[\frac{\exp(s\cos(\phi_{ii}^{+}+m))}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})} \right]}_{p_i^{+}}. \tag{6.21}$$

Rewriting the above equation:

$$\mathcal{L}_{f-cdl} = -s\cos(\phi_{ii}^{+}+m) + \log\left[\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\right]. \tag{6.22}$$

Taking partial derivative with respect to $\hat{x}_i^s$ results in:

$$\frac{\partial\mathcal{L}_{f-cdl}}{\partial\hat{x}_i^s} = \frac{\partial}{\partial\hat{x}_i^s}\left(\underbrace{-s\cos(\phi_{ii}^{+}+m)}_{\text{Ⓐ}} + \underbrace{\log\left[\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\right]}_{\text{Ⓑ}} \right). \tag{6.23}$$

To express $\cos(\phi_{ii}^{+}+m)$ in terms of sines and cosines of individual angles, following trigonometric identities are used.

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta. \tag{6.24}$$

$$\sin\alpha = \sqrt{1-\cos^2\alpha}. \tag{6.25}$$

Hence $\cos(\phi_{ii}^{+} + m)$ becomes:

$$\cos(\phi_{ii}^{+} + m) = \cos\phi_{ii}^{+}\cos m - \sin\phi_{ii}^{+}\sin m, \tag{6.26}$$

$$= \cos\phi_{ii}^{+}\cos m - \sqrt{1-\cos^2\phi_{ii}^{+}}\sin m, \tag{6.27}$$

Since $\cos\phi_{ii}^{+}$ is the dot product between the corresponding feature vectors from the student and teacher network i.e., $\cos\phi_{ii}^{+} = \hat{x}_i^s \cdot \hat{x}_i^t$. Its partial deriative with respect to $\hat{x}_i^s$ is given by:

$$\frac{\partial\cos\phi_{ii}^{+}}{\partial\hat{x}_i^s} = \frac{\partial}{\partial\hat{x}_i^s}(\hat{x}_i^s \cdot \hat{x}_i^t), \tag{6.28}$$

$$= \hat{x}_i^t. \tag{6.29}$$

Solving for the partial derivative in expression Ⓐ of the equation 6.23:

$$\frac{\partial}{\partial\hat{x}_i^s}(-s\cos(\phi_{ii}^{+} + m)) = -s\frac{\partial}{\partial\hat{x}_i^s}(\cos\phi_{ii}^{+}\cos m - \sqrt{1-\cos^2\phi_{ii}^{+}}\sin m), \tag{6.30}$$

$$= -s(\cos(m)\ \hat{x}_i^t - \frac{1}{2}\frac{\sin(m)}{\sqrt{1-\cos^2\phi_{ii}^{+}}}\ \frac{\partial}{\partial\hat{x}_i^s}(1-\cos^2\phi_{ii}^{+})), \tag{6.31}$$

$$= -s(\cos(m)\ \hat{x}_i^t + \frac{\sin(m)\cos\phi_{ii}^{+}}{\sqrt{1-\cos^2\phi_{ii}^{+}}}\ \hat{x}_i^t), \tag{6.32}$$

$$= -s(\cos(m)\ + \sin(m)\frac{\cos\phi_{ii}^{+}}{\sin\phi_{ii}^{+}})\ \hat{x}_i^t, \tag{6.33}$$

$$= -s(\frac{\cos(m)\sin\phi_{ii}^{+}\ + \sin(m)\cos\phi_{ii}^{+}}{\sin\phi_{ii}^{+}})\ \hat{x}_i^t, \tag{6.34}$$

$$= -s(\frac{\sin(\phi_{ii}^{+} + m)}{\sin\phi_{ii}^{+}})\ \hat{x}_i^t. \tag{6.35}$$

Solving for the partial derivative in expression Ⓑ of the equation 6.23:

$$\frac{\partial}{\partial\hat{x}_i^s}\left(\log\left[\exp(s\cos(\phi_{ii}^{+} + m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\right]\right) \tag{6.36}$$

$$= \frac{1}{\exp(s\cos(\phi_{ii}^{+} + m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}\frac{\partial}{\partial\hat{x}_i^s}(\exp(s\cos(\phi_{ii}^{+} + m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})),$$

$$= \frac{1}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}(s\exp(s\cos(\phi_{ii}^{+}+m))\frac{\partial}{\partial\hat{x}_i^s}\cos(\phi_{ii}^{+}+m) \tag{6.37}$$

$$+ s\sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\frac{\partial}{\partial\hat{x}_i^s}\cos\phi_{ij}^{-}),$$

Since $\cos\phi_{ij}^{-}$ is the dot product between the $i-th$ feature vector of the student network and $j-th$ feature vector of the teacher network, its partial derivative with respect to $\hat{x}_i^s$ is given by:

$$\frac{\partial\cos\phi_{ij}^{-}}{\partial\hat{x}_i^s} = \frac{\partial}{\partial\hat{x}_i^s}\sum_{j=1}^{B-1}\hat{x}_i^s\cdot\hat{x}_j^t, \tag{6.38}$$

$$= \sum_{j=1}^{B-1}\hat{x}_j^t. \tag{6.39}$$

The partial derivative of $\cos(\phi_{ii}^{+}+m)$ can be found from equation 6.35. Hence the equation 6.38 can be written as:

$$\frac{\partial}{\partial\hat{x}_i^s}\left(\log\left[\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\right]\right)$$

$$= s\frac{\exp(s\cos(\phi_{ii}^{+}+m))}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}(\frac{\sin(\phi_{ii}^{+}+m)}{\sin\phi_{ii}^{+}})\hat{x}_i^t \tag{6.40}$$

$$+ s\frac{\sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})\hat{x}_j^t}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})},$$

let consider

$$p_i^{+} = \frac{\exp(s\cos(\phi_{ii}^{+}+m))}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}.$$

$$p_j^{-} = \frac{\exp(s\cos\phi_{ij}^{-})}{\exp(s\cos(\phi_{ii}^{+}+m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^{-})}.$$

The equation 6.41 becomes:

$$\frac{\partial}{\partial \hat{x}_i^s}\left(\log\left[\exp(s\cos(\phi_{ii}^+ + m)) + \sum_{j=1}^{B-1}\exp(s\cos\phi_{ij}^-)\right]\right) = sp_i^+(\frac{\sin(\phi_{ii}^+ + m)}{\sin\phi_{ii}^+})\hat{x}_i^t + s\sum_{j=1}^{B-1}p_j^-\hat{x}_j^t. \tag{6.41}$$

The equation 6.23 can be written as:

$$\frac{\partial \mathcal{L}_{f-cdl}}{\partial \hat{x}_i^s} = -s(\frac{\sin(\phi_{ii}^+ + m)}{\sin\phi_{ii}^+})\hat{x}_i^t + sp_i^+(\frac{\sin(\phi_{ii}^+ + m)}{\sin\phi_{ii}^+})\hat{x}_i^t + s\sum_{j=1}^{B-1}p_j^-\hat{x}_j^t, \tag{6.42}$$

$$= -s(1-p_i^+)(\frac{\sin(\phi_{ii}^+ + m)}{\sin\phi_{ii}^+})\hat{x}_i^t + s\sum_{j=1}^{B-1}p_j^-\hat{x}_j^t. \tag{6.43}$$

6.3 Results and Discussion

6.3.1 Implementation Details

6.3.1.1 Training Settings

Training is performed separately on both small-scale (e.g., Casia Webface) and large-scale (e.g., WebFace4M) datasets. The SENet-50 architecture [23] is employed for both teacher and student networks. For the small-scale experiment i.e., Casia Webface, both the teacher and student network is trained for 40 epochs with the learning rate decreased by a factor of 10 at the 20th and 30th epochs. For the large-scale experiment i.e., Webface4M, both teacher and student network are trained for 21 epochs with the learning rate reduced by a factor of 10 at the 10th and 18th epochs. In both experiments, the base learning rate is chosen to be 0.1 and the batch size as 32. The hyper-parameters in the loss function are empirically chosen after several experiments. The scaling term s in sub-center learning and contrastive distillation losses is chosen as 64 and 128, respectively while the auxiliary term in the loss function λ_1 and λ_2 are chosen as 0.1 and 0.01, respectively. After the training phase, the classification layer is removed, and for each test image, the network outputs a feature vector of 512 dimensions.

6.3.1.2 Augmentation Settings

LR face recognition involves simultaneous training on both HR and LR images. Each batch consists of 70% of HR images and 30% of LR images. The HR images are augmented with degradations to create LR versions.

Augmentation in Small-scale Dataset: In case of small scale datasets like Casia Webface, the orginal size images are available. The augmentations undergo following steps:

1. Down-sampling and Resizing: The images are down-sampled to two different resolutions of size 30×30 pixels and 20×20 pixels. After down-sampling, all the images are resized to a standard size of 112×112 pixels, which is the input size required by the CNN.

2. Alignment: The images are then aligned using five facial landmarks. However, due to down-sampling, which lowers the resolution of the images, the alignment is not perfect. Inherently, the images may also be slightly rotated and centrally cropped.

Augmentation in Large-scale Dataset: Typically, images in large-scale datasets are already aligned using five facial landmarks. The augmentation process then follows these steps:

1. Down-sampling and Resizing: The images are down-sampled to various resolutions of size $p \times p$, where $p = \{30, 25, 20, 15\}$. After down-sampling, the images are resized to a standard size of 112×112 pixels.

2. Random Cropping: The images are then randomly cropped from the center to various sizes of $q \times q$ where $q = \{100, 90, 70, 80\}$.

3. Random Rotation: Finally, the cropped images undergo a random rotation with angles varying between 0 and 4 degrees, adding slight variability to the dataset for improved generalization. Following this rotation, the images are carefully resized back to a standardized dimension of 112×112 pixels, which is the input size required by the CNN.

6.3.2 Ablation Analysis

For the ablation study, a ResNet-18 architecture was adopted as the backbone and trained on the Casia Webface dataset. The performance is evaluated on both LR and HR datasets. For HR datasets, the average 1:1 verification accuracy across LFW, CFP-FP, CPLFW, AgeDB, and CALFW datasets is calculated. For LR datasets, the Rank-1 Identification Rate (IR) on the SCface (distance d1) and Tinyface datasets is reported. Distractors were excluded in the Tinyface.

6.3.2.1 Effect of Each Component in Loss

The effect of each component in the loss function is presented in Table 6.1. Training with Curricularface serves as our baseline. Including sub-center learning significantly boosted performance on LR datasets, while showing a slight decrease on the HR datasets. This is likely because sub-center learning captures diverse representations across multiple resolutions using the power of its multiple sub-centers. Feature-constrastive and class-contrastive distillation losses slightly decrease performance on LR datasets compared to the sub-center learning, but enhance performance on HR datasets. This is because these losses aim to minimize the gap between HR and LR features, leading to better performance on HR datasets. The class-contrastive distillation loss enforces a strict constraint on both LR and HR features compared to the feature-constrastive distillation loss. It was empirically found that using the class-contrastive distillation loss alone leads to difficulties in optimizing parameters. The proposed method that utilizes the strengths of sub-center learning and contrastive distillation losses, achieves improved performance on both HR and LR datasets.

Table 6.1: Ablation Analysis of Proposed Loss

Methods	$\mathcal{L}_{cl}$	$\mathcal{L}_{sl}$	$\mathcal{L}_{f-cdl}$	$\mathcal{L}_{c-cdl}$	**LR**		**HR**
					SCface	**Tinyface**	
Baseline	✓				69.85	52.49	90.45
I		✓			**72.62**	56.44	90.07
II	✓		✓		67.08	52.87	92.22
III	✓			✓	68.15	52.68	91.46
Proposed		✓	✓	✓	72.00	**60.33**	**92.52**

6.3.2.2 Effect of Augmentations

HR and augmented LR images were used simultaneously during training. The augmentation included downs-sampling, center cropping, and rotation of images. The percentage of HR and augmented LR images in a single batch was fixed during training. The table shows the effect of different percentage of HR images on the performance. It is obvious that the selection of percentage of HR images in a batch is a trade-off between the performance of HR and LR images. The 70% of HR images in batch shows the best performance on HR as well as LR images.

TABLE 6.2: Ablation Analysis of Augmentation

Methods	**% of HR Images**	**% of LR Images**	**LR**		**HR**
			SCface	**Tinyface**	
I	60	40	**75.23**	59.95	91.34
Proposed	70	30	72.00	**60.33**	**92.52**
II	80	20	68.31	58.61	91.85
III	90	10	66.31	58.99	92.18
IV	100	0	34.77	42.27	**92.60**

6.3.3 Comparison with SOTA Methods

The proposed methodology is evaluated against SOTA techniques using two types of protocols on LR datasets. The first protocol includes fine-tuning (✓), commonly used in most published works. The second is our proposed protocol (∗), in which case only Adaface [17] is suitable for comparison due to its compatibility with our protocols. The fine-tuning achieves near-perfect results on the SCface (see Chapter 5 Section 5.2.1) and COXface datasets; therefore, test subjects were used as protocol (×) to ensure consistency with prior published works (excluding training subjects). On HR and MR datasets, only the proposed protocol is used for comparison. For each method, details such as the network size and the training dataset are explicitly mentioned in the table to enable a fair and transparent comparison, e.g, R-50 denotes the ResNet architecture with 50 layers.

6.3.3.1 LR Datasets

SCface: Table 6.3 compares identification performance with previous SOTA techniques on SCface. The performance improvement of SCface is biased towards LR images, making it challenging to achieve acceptable performance on SCface along with HR datasets using the same model. On test subjects only, our proposed method achieves an IR of 90% on distance d1, second best to DM & AGD [111] in Chapter 4. This is because this method uses JPEG degradation, which significantly enhances performance on SCface at the cost of deteriorating performance on HR datasets. In the proposed protocols with small-scale experiments, DM & AGD [111] again showed improved performance over Adaface and the proposed method. However, on distance d3, the proposed method has obtained the best results. In large-scale experiments, our proposed method outperforms Adaface by a significant margin. This validates our hypothesis that sub-center learning in the proposed scheme effectively learns diverse representations across varying resolutions of images.

TABLE 6.3: Performance Comparison on SCface. (P: Protocol)

Methods	**P**	**Arch.**	**Dataset**	**Rank-1 IR**		
				4.2m	**2.6m**	**1.0m**
T-C [68]	×	R-50	VGGface2	70.20	93.70	98.10
RIFR [69]	×	R-50	Casia	88.30	98.30	98.60
NPT Loss [18]	×	R-50	Casia	85.69	99.08	99.08
Adaface [17]	×	R-50	Casia	59.25	98.50	99.75
CCFace [78]	×	R-50	MS1M v2	74.8	94.01	99.47
CATFace [80]	×	R-101	MS1M v2	90.64	98.85	99.61
DM & AGD [111]	×	R-50	Casia	**95.25**	**99.50**	99.00
Proposed	×	R-50	Casia	90.50	98.75	**100.0**
DM & AGD [111]	∗	R-50	Casia	**92.92**	**99.54**	99.38
Adaface [17]	∗	R-50	Casia	51.38	96.92	99.54
Proposed	∗	R-50	Casia	**84.00**	**98.77**	**99.85**
Adaface [17]	∗	R-50	Webface4M	87.08	99.69	100.0
Proposed	∗	R-50	Webface4M	**94.31**	**100.0**	**100.0**

Tinyface: Table 6.4 compares identification performance to prior SOTA methods on the Tinyface dataset. The proposed scheme has enhanced performance in both small and large-scale datasets in the finetuned and proposed protocols. Adaface [17] learns discriminative embeddings using a margin-based softmax loss function with feature norm as an adaptive parameter. In contrast, our proposed scheme learns discriminative embeddings based on HR embeddings from a teacher network in a contrastive estimation framework. The results on Tinyface validate our hypothesis that our methods learn discriminative embeddings more effectively in LR images. In both small-scale and large-scale experiments, the proposed scheme outperforms Adaface with margins of 2.07% and 5.45%, respectively.

TABLE 6.4: Performance Comparison on Tinyface. (P: Protocol)

Methods	**P**	**Arch.**	**Dataset**	**Rank-1 IR**	**Rank-5 IR**
CSRI [52]	✓	-	-	44.80	-
Vivid GAN [54]	✓	-	-	47.16	56.04
DM & AGD [111]	✓	R-50	Casia	60.78	-
MIND-Net [74]	✓	-	-	66.82	-
IDEA-Net [77]	✓	-	-	68.13	-
REM [63]	✓	R-50	VGGface2	73.06	-
Proposed	✓	R-50	Webface4M	**74.10**	**79.32**
DM & AGD [111]	*	R-50	Casia	54.18	-
T-C [68]	*	R-50	VGGface2	58.60	-
Adaface [17]	*	R-50	Casia	53.48	59.46
Proposed	*	R-50	Casia	**58.93**	**64.75**
TURL [61]	*	R-100	MS1M	63.89	68.67
CCFace [78]	*	R-100	MS1M v2	65.71	69.25
CATFace [80]	*	R-101	MS1M v2	68.95	72.31
Adaface [17]	*	R-50	Webface4M	67.70	71.64
Proposed	*	R-50	Webface4M	**69.77**	**73.18**

QMUL Survface: QMUL Survface is the most challenging dataset captured from real-world surveillance cameras. The performance comparison on QMUL Survface is reported in Table 6.5. The proposed scheme demonstrates enhanced performance on both small scale and large-scale datasets in the finetuned and the proposed protocols. While performance on QMUL-Survface is biased towards HR images, our proposed

scheme effectively leverages the informative HR features from a pre-trained HR teacher network in a LR student network. Based on the results from the SCface and Tinyface datasets, Adaface has limitations in learning LR features.

TABLE 6.5: Performance Comparison on QMUL Suvface. (P: Protocol)

Methods	**P**	**TPR(%)@FAR**				**TPIR20(%)@FAR**		
		0.3	**0.1**	**0.01**	**0.001**	**0.3**	**0.2**	**0.1**
CSRI [52]	✓	78.60	53.10	18.09	12.04	-	-	-
FAN [53] (R-50, MS1M)	✓	71.30	44.59	12.94	2.75	-	-	-
RAN [60]	-	-	-	-	-	26.50	21.60	14.90
SST [115] (R-56, MS1M v1)	✓	87.00	68.21	35.72	22.18	12.38	9.71	6.61
DSN [75]		75.09	52.74	21.41	11.02	-	-	-
DDAT [55]	✓	90.40	75.40	40.40	16.40	-	-	-
IDEA-Net [77]	✓	-	-	-	-	26.24	21.82	15.61
REE [63] (R-50, VGGface2)	✓	90.21	**80.99**	**64.60**	**48.48**	**33.20**	**29.34**	**22.81**
Proposed (R-50, Webface4M)	✓	**90.45**	77.66	52.67	27.29	30.67	25.20	14.46
Adaface [17] (R-50, Casia)	*	47.12	20.34	3.45	0.92	3.05	1.95	1.02
Proposed (R-50, Casia)	*	**50.24**	**21.69**	**4.75**	**1.24**	**4.80**	**3.13**	**1.39**
Adaface [17] (R-50, Webface4M)	*	70.45	46.91	14.09	3.43	6.86	4.20	1.86
Proposed (R-50, Webface4M)	*	**71.58**	**50.49**	**23.63**	**7.81**	**9.12**	**5.99**	**3.08**

COXface: In Table 6.6, identification performance is compared with other SOTA methods on the COXface dataset. The images captured from three videos are compared with high-quality still images. IR is reported without fine-tuning, following the proposed protocols. Our proposed scheme outperforms previous methods across all cameras, which capture images at varying resolutions. In the context of Video V1, the proposed scheme demonstrated a significant enhancement in performance, achieving improvements of 7.30% in small-scale experiment and 1.51% in large-scale experiment. These results underscore the efficacy of integrating sub-center learning with contrastive

distillation losses, which together facilitate the learning of discriminative features across varying resolutions of images.

TABLE 6.6: Performance Comparison on COXface. (P: Protocol)

Methods	**P**	**Arch.**	**Dataset**	**Rank-1 IR**		
				V1-S	**V2-S**	**V3-S**
DM & AGD [111]	*	R-50	Casia	80.37	87.96	95.84
Adaface [17]	*	R-50	Casia	74.18	79.73	92.70
Proposed	*	R-50	Casia	**87.08**	**92.13**	**97.66**
Adaface [17]	*	R-50	Webface4M	94.40	96.69	99.25
Proposed	*	R-50	Webface4M	**95.91**	**97.86**	**99.54**

6.3.3.2 HR and MR Datasets

The results of both small-scale and large-scale experiments on HR datasets are presented in Table 6.7. Adaface [17] exhibits a marginally superior performance on HR datasets compared to our proposed scheme, with improvements of 0.4% and 0.54%, respectively. However, this trend is reversed when examining the MR datasets, as shown in Table 6.8. In this context, our proposed scheme, when trained on a small-scale dataset, significantly outperforms Adaface by margins of 63.64% and 69.47% on IJB-B and IJB-C, respectively. This highlights the proposed scheme's remarkable ability to learn discriminative features even with limited training data.

TABLE 6.7: Performance Comparison on HR Datasets (1:1 Verification Rate)

Method	**P**	**LFW**	**CFP-FP**	**CPLFW**	**AgeDB**	**CALFW**	**AVG**
Adaface (R-50, Casia)	*	**99.42**	**96.41**	89.97	**94.38**	**93.23**	**94.68**
Proposed (R-50, Casia)	*	99.26	96.21	**89.98**	93.05	92.83	94.26
Adaface (R-50, Webface4M)	*	**99.78**	**98.97**	**94.2**	**97.68**	**96.01**	**97.33**
Proposed (R-50, Webface4M)	*	99.68	98.27	93.78	96.55	95.65	96.79

TABLE 6.8: Performance Comparison on MR Datasets (TPR@FAR=1e-4)

Method	**P**	**Arch.**	**Dataset**	**IJB-B**	**IJB-C**
Adaface [17]	*	R-50	Casia	17.75	16.06
Proposed	*	R-50	Casia	**81.42**	**85.53**
Adaface [17]	*	R-50	Webface4M	**95.44**	**96.98**
Proposed	*	R-50	Webface4M	93.93	95.81

In large-scale experiments, Adaface [17] once again shows a slight edge over our proposed scheme, achieving 1.51% and 1.17% better performance on IJB-B and IJB-C, respectively. This advantage can be attributed to Adaface's training strategy, which involves batches containing a high proportion (approximately 80%) of HR data. While this approach enhances performance on HR datasets, it comes at the expense of low performance on LR datasets like SCface.

6.3.3.3 Combined Evaluation Metric (CEM)

The proposed method was evaluated against Adaface [17] on CEM across both small-scale and large-scale experiments. The findings, as summarized in Table 6.9, show that the proposed scheme outperforms Adaface [17] across both small-scale and large-scale

TABLE 6.9: Performance Comparison on Combined Evaluation Metric (CEM)

Dataset	**Casia Webface**		**Webface4M**	
	Proposed	**Adaface**	**Proposed**	**Adaface**
SCface	84.00	51.38	94.31	87.08
Tinyafce	58.93	53.48	69.77	67.70
COXface	87.08	74.18	95.91	94.40
QMUL	19.48	17.95	38.37	33.72
HR	94.26	94.68	96.79	97.34
IJB-B	81.42	17.75	93.93	95.44
IJB-C	85.53	16.06	95.81	96.98
CEM	**55.43**	29.60	**75.34**	71.74

experiments. In both small-scale and large-scale experiments, the proposed scheme achieves CEM scores of 55.43 and 75.34, respectively, compared to Adaface's scores of 29.60 and 71.74. It shows that our proposed scheme best suits to recognition in both HR and LR images. From a training perspective, Adaface [17] gains an edge due to its straightforward training process. In contrast, the proposed scheme relies on a teacher-student methodology, introducing additional complexity in training teacher networks.

6.4 Summary

In this chapter, a novel scheme for recognizing facial images regardless of their resolution (HR or LR) is proposed. The challenge in simultaneously learning discriminative features from HR and LR images was two-fold: Learning diverse representations across the varying resolution of images and preventing the overlooking of HR features during the learning process. These challenges were addressed by: 1) Employing sub-center learning to capture diverse representations across varying resolution of images. 2) Utilizing contrastive distillation loss to minimize the gap between HR and LR features. To adequately highlight the strengths and weaknesses in LR face recognition across LR and HR images, the proposed evaluation protocols having combined evaluation metric (CEM) were used. Our proposed scheme outperformed all other schemes on the majority of benchmarks. Using CEM as the standard for the best compromise between HR and LR face recognition, the proposed scheme achieved higher scores, surpassing the previous SOTA Adaface by margins of 25.83 and 3.60 in small-scale and large-scale experiments, respectively. This shows the efficacy of our proposed model for practical applications of face recognition where better performance is required across different resolutions.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This dissertation explores the challenges associated with recognizing LR images and presents potential solutions for accurate identification. In LR face recognition, probe images of varying resolutions, including both HR and LR, are compared to HR gallery images. This scenario is commonly encountered in surveillance applications. Surveillance cameras often capture a wide area, resulting in very small facial images. As these facial images are captured on the move, various degradation effects, such as blurring and noise, are introduced. These degradations obscure facial features essential for recognition, making LR face recognition a challenging task.

A comprehensive literature review on LR face recognition is presented, comparatively analyzing previous methods. The primary challenge in LR face recognition is learning discriminative features from both HR and LR images. Standard loss functions used in HR face recognition are insufficient for this task. Therefore, additional losses are necessary to bridge the domain gap between HR and LR images. Distillation losses, which explicitly minimize the distance between LR and HR features, have proven effective in bridging this domain gap. The literature review further identifies key research gaps: the need for more realistic synthetic LR images that simulate real-world degradation effects, the challenge of handling probe images with varying resolutions, the difficulty of

effectively bridging the LR-HR feature gap without compromising performance on HR images, and the lack of rigorous testing scenarios.

To address the identified research gaps, a scheme combining a degradation model and attention-guided distillation was proposed. The degradation model employed classical techniques to simulate real-world degradation effects such as motion blur, out-of-focus blur, noise, and compression artifacts in the synthetic LR images. These synthetic LR images help the model learn the variations found in real-world LR images, making the training data more representative of the test data. The attention-guided distillation technique bridges the domain gap between synthetic LR and HR images and prevents the model from deteriorating performance on HR images. This scheme is specifically designed for images captured by surveillance cameras, which include varying resolution of images. Faces captured at closer distances will appear as HR, while those farther away will exhibit lower quality. Our scheme outperformed existing methods on widely used surveillance benchmark datasets. However, its performance on HR images captured by digital cameras is less impressive when tested on HR benchmarks. This discrepancy is likely due to inherent differences between the image sources. The LR testing benchmarks may not adequately represent such scenarios.

Existing testing benchmarks were analyzed, revealing limitations in their ability to comprehensively evaluate LR face recognition methods. The fine-tuning mechanisms used in these protocols are insufficient for assessing generalization capability in learning discriminative features. Additionally, success on HR images from surveillance cameras does not guarantee optimal performance on HR images from digital cameras. Moreover, the varying nature of testing benchmarks means that superior performance on one does not necessarily translate to similar results on another. Finally, existing protocols lack a performance evaluation metric that can assess a model's ability to handle varying resolutions in probe images. To address these limitations, new protocols are developed for a thorough evaluation of LR face recognition models.

Finally, a scheme was proposed that effectively handles both HR and LR images during recognition and addresses the weaknesses of existing protocols. This scheme incorporated sub-center learning and contrastive distillation loss. Sub-center learning captured diverse representations across varying probe image resolutions using multiple sub-centers

defined for each class. This prevents the model from treating images of different resolutions similarly, which could lead to optimization problems. The contrastive distillation loss bridges the domain gap between HR and LR features, ensuring that corresponding HR and LR images of the same identity are closer than to others. This results in learning discriminative features from both HR and LR images. The proposed scheme is validated on existing and newly proposed protocols, outperforming majority of the benchmarks compared to previous SOTA methods.

7.2 Future Work

The research findings presented in this thesis can be extended in the following ways as future work.

1. The presence of extensive noise within large-scale training datasets for face recognition poses a significant challenge. In order to mitigate this and prevent a degradation in recognition accuracy, it is necessary to employ loss functions specifically designed to handle these noisy labeled images. The loss functions can be able to down-weight the influence of noisy images during training, ensuring the model prioritizes learning from clean and informative data.

2. In LR face recognition systems, the primary focus is often on achieving high recognition performance. However, beyond identifying individuals, a reconstructed normalized face image from the learned feature can help in the forensic analysis . Therefore decoder network can be trained in parallel with recognition model to reconstruct normalized face images from the learned features.

3. The existing benchmarks have limited variations and are built on outdated surveillance camera technology. In our newly proposed protocol, multiple benchmarks are used to assess the effectiveness and limitations of the LR face recognition model. Therefore, more diverse testing benchmarks can be developed using modern surveillance cameras to align face recognition models with today's applications.

4. Mostly, ResNet-50 and ResNet-100 architectures are used as backbone networks for face recognition models. These models have a massive number of parameters.

Efficient models can be developed through Neural Architecture Search (NAS) to transform this technology for real-time applications.

5. The current degradation model for generating synthetic LR images is based on a static approach. However, there is significant potential for developing a more complex and dynamic degradation model that can better align itself with real-world degradation effects. This can be achieved by leveraging Generative Adversarial Networks (GANs) and Uncertainty Quantification Methods.

6. In this research, deep CNNs are used for feature extraction. Recognizing the potential benefits of frequency domain techniques, which have the ability to capture both local and global features, these techniques can be integrated with deep CNNs to enhance performance.

7.3 Societal and Ethical Concerns

Face recognition technology, while offering substantial benefits in areas such as security, surveillance, personalized services, and identity verification, raises critical societal and ethical concerns. One of the most significant societal concerns surrounding face recognition is the potential erosion of privacy. The ability to identify individuals in public spaces, often without their knowledge or consent, has raised fears of mass surveillance. Governments, law enforcement agencies, and private corporations have the capability to track people's movements and behaviors, which could potentially inhibit free expression and personal freedoms. There is a need for stringent regulations and oversight to ensure that face recognition technology is not used in ways that infringe upon individuals' right to privacy and autonomy.

Another ethical issue involves the lack of informed consent when face recognition technology is deployed. Individuals are often unaware that their biometric data is being collected, analyzed, or stored, especially in public and semi-public spaces like airports, shopping malls, or even online platforms. Transparent policies and clear consent mechanisms are essential to respect individuals' autonomy and rights over their own biometric data.

The potential misuse of face recognition technology is another ethical concern. Authoritarian regimes may exploit it for political repression, targeting activists, dissidents, or ethnic minorities. Additionally, there is a risk of commercial exploitation, where companies use face recognition for intrusive marketing tactics or unauthorized data collection. The lack of clear regulations and oversight can lead to the misuse of face recognition technology, making it imperative to establish legal frameworks that protect citizens' rights.

Face recognition systems have been shown to exhibit biases, particularly against certain demographic groups, including people of color, women, and individuals from marginalized communities. These biases can result in higher error rates, leading to discriminatory outcomes in critical applications like law enforcement, hiring processes, and access to services. Ensuring fairness and equity in face recognition technology requires diverse and representative training datasets, along with ongoing audits to mitigate bias and prevent discriminatory practices.

The societal and ethical implications of face recognition technology highlight the need for a careful, balanced approach to both its development and deployment. While the potential benefits of enhanced security, convenience, and efficiency are significant, these must not come at the expense of fundamental human rights and ethical principles. Balancing the deployment of face recognition technology with ethical considerations is essential to maintain public trust and social harmony. Moving forward, it is imperative that policymakers, researchers, and industry leaders collaborate to establish comprehensive regulations, ethical standards, and best practices that prioritize privacy, fairness, and transparency. Only by addressing these concerns can society fully leverage the advantages of face recognition technology while safeguarding individual freedoms and promoting ethical use.

Bibliography

[1] W. W. Bledsoe and H. Chan, "A man-machine facial recognition system—some preliminary results," *Panoramic Research, Inc, Palo Alto, California., Technical Report PRI A*, vol. 19, p. 1965, 1965.

[2] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Identification of human faces," *Proceedings of the IEEE*, vol. 59, no. 5, pp. 748–760, 1971.

[3] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, vol. 4, no. 3, pp. 519–524, 1987.

[4] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition.* IEEE Computer Society, 1991, pp. 586–587.

[5] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. Von Der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on computers*, vol. 42, no. 3, pp. 300–311, 1993.

[6] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th international conference on pattern recognition*, vol. 1. IEEE, 1994, pp. 582–585.

[7] L. Wiskott, J.-M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *TR96-08, Institut für Neuroinformatik, Ruhr-Universität Bochum*, 1996.

[8] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[10] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137–154, 2004.

[11] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 947–954.

[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[14] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2021.

[15] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[16] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 225–14 234.

[17] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.

[18] S. S. Khalid, M. Awais, Z. Feng, C. H. Chan, A. Farooq, A. Akbari, and J. Kittler, "Npt-loss: Demystifying face recognition losses with nearest proxies triplet," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[19] 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.

[20] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[24] M. KELLY, "Visual identification of people by computer," *Stanford AI project memo AI-130*, 1970.

[25] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.

[26] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[27] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 711–720, 1997.

[28] P. Phillips, "Support vector machines applied to face recognition," *Advances in neural information processing systems*, vol. 11, 1998.

[29] Pentland, Moghaddam, and Starner, "View-based and modular eigenspaces for face recognition," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1994, pp. 84–91.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[31] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537–1543, 2008.

[32] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 2037–2041, 2006.

[33] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 786–791.

[34] T. Ahonen, J. Matas, C. He, and M. Pietikäinen, "Rotation invariant image description with local binary pattern histogram fourier features," in *Image Analysis: 16th Scandinavian Conference, SCIA 2009, Oslo, Norway, June 15-18, 2009. Proceedings 16*. Springer, 2009, pp. 61–70.

[35] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 533–544, 2009.

[36] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image*

Processing : a publication of the IEEE Signal Processing Society, vol. 11 4, pp. 467–76, 2002.

[37] C. Liu, "Independent component analysis of gabor features for face recognition," *IEEE transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, 2003.

[38] Chengjun, "Gabor-based kernel pca with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 572–581, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:594886

[39] C.-H. Chan, J. Kittler, and K. Messer, "Multi-scale local binary pattern histograms for face recognition," in *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007. Proceedings*. Springer, 2007, pp. 809–818.

[40] C.-H. Chan and J. Kittler, "Multispectral local binary pattern histogram for component-based color face verification," in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007, pp. 1–7. [Online]. Available: https://api.semanticscholar.org/CorpusID:5684065

[41] D. Zhao, Z. Lin, and X. Tang, "Laplacian pca and its applications," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[42] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[43] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Advances in neural information processing systems*, vol. 27, 2014.

[44] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, 2016, pp. 507–516.

[46] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[47] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[48] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 241–12 248.

[49] Y. Movshovitz-Attias, A. Toshev, T. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 360–368, 2017.

[50] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3238–3247.

[51] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 183–198.

[52] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 605–621.

[53] X. Yin, Y. Tai, Y. Huang, and X. Liu, "Fan: Feature adaptation network for surveillance face recognition and normalization," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[54] Y. Zhang, I. W. Tsang, J. Li, P. Liu, X. Lu, and X. Yu, "Face hallucination with finishing touches," *IEEE Transactions on Image Processing*, vol. 30, pp. 1728–1743, 2021.

[55] Q. Jiao, R. Li, W. Cao, J. Zhong, S. Wu, and H.-S. Wong, "Ddat: Dual domain adaptive translation for low-resolution face verification in the wild," *Pattern Recognition*, vol. 120, p. 108107, 2021.

[56] D. Zeng, H. Chen, and Q. Zhao, "Towards resolution invariant face recognition in uncontrolled scenarios," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[57] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, 2018.

[58] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1002–1014, 2018.

[59] M. Parchami, S. Bashbaghi, and E. Granger, "Video-based face recognition using ensemble of haar-like deep convolutional neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 4625–4632.

[60] H. Fang, W. Deng, Y. Zhong, and J. Hu, "Generate to adapt: Resolution adaption network for surveillance face recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 741–758.

[61] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6817–6826.

[62] M. S. E. Saadabadi, S. R. Malakshan, A. Dabouei, and N. M. Nasrabadi, "Aroface: Alignment robustness to improve low-quality face recognition," in *European Conference on Computer Vision*. Springer, 2025, pp. 308–327.

[63] J. C. L. Chai, T.-S. Ng, C.-Y. Low, J. Park, and A. B. J. Teoh, "Recognizability embedding enhancement for very low-resolution face recognition and quality estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9957–9967.

[64] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, and M. Gonzalez-Mendoza, "Lightweight low-resolution face recognition for surveillance applications," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5421–5428.

[65] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper, "Mixfacenets: Extremely efficient face recognition networks," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[66] M. Tan and Q. V. Le, "Mixconv: Mixed depthwise convolutional kernels," in *British Machine Vision Conference*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199064614

[67] M. Zhu, K. Han, C. Zhang, J. Lin, and Y. Wang, "Low-resolution visual recognition via deep feature distillation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3762–3766.

[68] F. V. Massoli, G. Amato, and F. Falchi, "Cross-resolution learning for face recognition," *Image and Vision Computing*, vol. 99, p. 103927, 2020.

[69] S. S. Khalid, M. Awais, Z.-H. Feng, C.-H. Chan, A. Farooq, A. Akbari, and J. Kittler, "Resolution invariant face recognition using a distillation approach," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 410–420, 2020.

[70] Y. Huang, P. Shen, Y. Tai, S. Li, X. Liu, J. Li, F. Huang, and R. Ji, "Improving face recognition from hard samples via distribution distillation loss," in *European Conference on Computer Vision*. Springer, 2020, pp. 138–154.

[71] J. Zha and H. Chao, "Tcn: Transferable coupled network for cross-resolution face recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3302–3306.

[72] P. Li, S. Tu, and L. Xu, "Deep rival penalized competitive learning for low-resolution face recognition," *Neural Networks*, vol. 148, pp. 183–193, 2022.

[73] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Derivenet for (very) low resolution image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6569–6577, 2021.

[74] C.-Y. Low, A. B.-J. Teoh, and J. Park, "Mind-net: A deep mutual information distillation network for realistic low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 354–358, 2021.

[75] S.-C. Lai and K.-M. Lam, "Deep siamese network for low-resolution face recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1444–1449.

[76] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[77] C.-Y. Low and A. B.-J. Teoh, "An implicit identity-extended data augmentation for low-resolution face representation learning," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3062–3076, 2022.

[78] M. S. E. Saadabadi, S. R. Malakshan, H. Kashiani, and N. M. Nasrabadi, "Cc-face: Classification consistency for low-resolution face recognition," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2023, pp. 1–10.

[79] Y. Song and F. Wang, "Qgface: Quality-guided joint training for mixed-quality face recognition," *ArXiv*, vol. abs/2312.17494, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266691005

[80] N. A. Talemi, H. Kashiani, and N. M. Nasrabadi, "Catface: Cross-attribute-guided transformer with self-attention distillation for low-quality face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[81] L. Chen, J. Chen, Z. Xu, Y. Liao, and Z. Chen, "Two-stage dual-resolution face network for cross-resolution face recognition in surveillance systems," *The Visual Computer*, vol. 40, no. 8, pp. 5545–5556, 2024.

[82] X. Ling, Y. Lu, W. Xu, W. Deng, Y. Zhang, X. Cui, H. Shi, and D. Wen, "Dive into the resolution augmentations and metrics in low resolution face recognition: A plain yet effective new baseline," *arXiv preprint arXiv:2302.05621*, 2023.

[83] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6898–6908, 2020.

[84] Y. Song, H. Tang, F. Meng, C. Wang, M. Wu, Z. Shu, and G. Tong, "A transformer-based low-resolution face recognition method via on-and-offline knowledge distillation," *Neurocomputing*, vol. 509, pp. 193–205, 2022.

[85] N. A. Talemi, H. Kashiani, and N. M. Nasrabadi, "Catface: Cross-attribute-guided transformer with self-attention distillation for low-quality face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[86] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," *ArXiv*, vol. abs/1411.7923, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:17188384

[87] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 87–102.

[88] J. Deng and J. Guo, "Insightface: 2d and 3d face analysis project," https://github.com/deepinsight/insightface, 2018.

[89] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 492–10 502.

[90] 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.

[91] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4804–4813.

[92] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

[93] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *ArXiv*, vol. abs/1804.09691, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21129655

[94] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[95] S. Sengupta, J.-C. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:6544744

[96] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017.

[97] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *ArXiv*, vol. abs/1708.08197, 2017.

[98] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, p. 7, 2018.

[99] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[100] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on cox face database," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5967–5981, 2015.

[101] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 90–98.

[102] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[103] X. Gao, Y. Sun, Y. Xiao, Y. Gu, S. Chai, and B. Chen, "Adaptive random down-sampling data augmentation and area attention pooling for low resolution face recognition," *Expert Systems with Applications*, vol. 209, p. 118275, 2022.

[104] R. Wang and D. Tao, "Recent progress in image deblurring," *ArXiv*, vol. abs/1409.6838, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:13985328

[105] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations*, 2022.

[106] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," 2021.

[107] M. Parchami, S. Bashbaghi, and E. Granger, "Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person," *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2017.

[108] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:4637184

[109] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*. Springer, 2016, pp. 499–515.

[110] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.

[111] M. Ullah, I. A. Taj, and R. H. Raza, "Degradation model and attention guided distillation approach for low resolution face recognition," *Expert Systems with Applications*, vol. 243, p. 122882, 2024.

[112] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[113] S. Deng, Y. Xiong, M. Wang, W. Xia, and S. Soatto, "Harnessing unrecognizable faces for improving face recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3424–3433.

[114] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[115] H. Du, H. Shi, Y. Liu, J. Wang, Z. Lei, D. Zeng, and T. Mei, "Semi-siamese training for shallow face learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 36–53.

Appendix A

A.1 Gradient of KL Divergence Loss

The KL divergence loss on logits can be represented by:

$$\mathcal{L}_{kl} = \sum_{i=1}^{N} \; p(z_i^t) \log \frac{p(z_i^t)}{p(z_i^s)}. \tag{7.1}$$

Rewriting the above equation as:

$$\mathcal{L}_{kl} = -\sum_{i=1}^{N} \; p(z_i^t) \log p(z_i^s) + \sum_{i=1}^{N} \; p(z_i^t) \log p(z_i^t). \tag{7.2}$$

Taking derivative with respective to z_j^s

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \frac{\partial}{\partial z_j^s} \log p(z_i^s), \qquad \text{where} \quad p(z_i^s) = \frac{\exp(z_i^s)}{\sum_{k=1}^{N} \exp(z_k^s)}, \tag{7.3}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \; . \; \frac{\partial}{\partial z_j^s} \left[\log \frac{\exp(z_i^s)}{\sum_{k=1}^{N} \exp(z_k^s)} \right], \tag{7.4}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \; . \; \frac{\partial}{\partial z_j^s} \left[z_i^s - \log \sum_{k=1}^{N} \exp(z_k^s) \right], \tag{7.5}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \left[\delta_{ij} - \frac{1}{\sum_{k=1}^{N} \exp(z_k^s)} \frac{\partial}{\partial z_j^s} (\sum_{k=1}^{N} \exp(z_k^s)) \right], \tag{7.6}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \left[\delta_{ij} - \underbrace{\frac{\exp(z_j^s)}{\sum_{k=1}^{N} \exp(z_k^s)}}_{p(z_j^s)} \right], \tag{7.7}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = -\sum_{i=1}^{N} \; p(z_i^t) \left[\delta_{ij} - p(z_j^s) \right], \tag{7.8}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = p(z_j^s) \sum_{i=1}^{N} p(z_i^t) - \sum_{i=1}^{N} p(z_i^t).\delta_{ij}, \tag{7.9}$$

$$\frac{\partial \mathcal{L}_{kl}}{\partial z_j^s} = p(z_j^s) - p(z_j^t). \tag{7.10}$$