OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

EIT FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIONSTECHNIK

EFFICIENT AND ROBUST FACE RECOGNITION IN THE WILD

Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

von **M.Sc. Aly Ahmed Aly Khalifa**
geb. am 14.11.1984 in Giza, Egypt

genehmigt durch die Fakultät für Elektrotechnik und Informationstechnik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr.-Ing. habil. Ayoub Al-Hamadi
Prof. Dr. Aly A. Farag
Prof. Dr. rer. nat. Andreas Wendemuth

eingereicht am 18.04.2024
Promotionskolloquium am 27.2.2025

Contents

Abstract V

Deutsche Kurzfassung VII

Related Publications IX

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Problem Definition . . . 2
- 1.3 Goal and Contributions . . . 5
- 1.4 Publications . . . 7
- 1.5 Outline . . . 9

2 Background and Literature Review 12
- 2.1 Evolution of Face Recognition Systems . . . 12
- 2.2 Preprocessing . . . 13
- 2.3 Face Detection and Tracking . . . 14
  - 2.3.1 Face Detection . . . 14
  - 2.3.2 Face Tracking . . . 17
  - 2.3.3 Effect on the Face Recognition Systems . . . 19
- 2.4 Face Alignment . . . 20
  - 2.4.1 Coordinate Regression-based Methods . . . 20
  - 2.4.2 Heatmap Regression-based Methods . . . 22
  - 2.4.3 Effect on the Face Recognition Systems . . . 23
- 2.5 Face Recognition . . . 23
  - 2.5.1 Network Architecture . . . 24
  - 2.5.2 Discriminative Loss Functions . . . 26
- 2.6 Evaluation Metrics and Datasets . . . 30
  - 2.6.1 Metrics . . . 30
  - 2.6.2 Datasets . . . 31
- 2.7 Challenges and Limitations in Face Recognition . . . 33
- 2.8 Summary . . . 34

3 Face Recognition and Tracking Framework for Human-Robot Interaction 35
- 3.1 Introduction . . . 35

3.2 State of the Art . . . 38
3.2.1 Face Detection Algorithms . . . 38
3.2.2 Facial Landmarks and Face Alignment Algorithms . . . 39
3.2.3 Face Recognition Algorithms . . . 40
3.2.4 Face Tracking Algorithms . . . 41
3.3 Human-Robot Interaction Study . . . 42
3.3.1 Concept . . . 43
3.3.2 Features . . . 43
3.3.3 Setup and Workflow . . . 44
3.4 Methodology and Proposed Framework . . . 46
3.4.1 Face Detection and Alignment . . . 47
3.4.2 Face Recognition . . . 49
3.4.3 Improved Face Recognition Using Face Tracking . . . 50
3.5 Experiments and Analysis . . . 51
3.5.1 Face Detection. . . . 51
3.5.2 Face Recognition . . . 53
3.5.3 Overall System . . . 53
3.5.4 Computational Efficiency Assessment . . . 57
3.6 Discussions and Limitations . . . 58
3.7 Summary . . . 60

4 Towards Adaptive Feature Learning For Face Recognition 61
4.1 Introduction . . . 62
4.2 Relationship to Previous Work . . . 63
4.2.1 Metric-based Methods . . . 63
4.2.2 Margin-based Methods . . . 64
4.3 Our Approach to Adaptive Feature Learning For Face Recognition . . . 65
4.3.1 Preliminary . . . 65
4.3.2 Method . . . 66
4.3.3 Comparison with other Loss Functions . . . 70
4.4 Experiments and Analysis . . . 73
4.4.1 Implementation and Training Details . . . 73
4.4.2 Ablation Study . . . 76
4.4.3 Comparison with State-of-the-Art . . . 77
4.5 Discussion . . . 85
4.6 Summary . . . 86

5 Towards Efficient and Robust Face Recognition Network 87
5.1 Introduction . . . 88
5.2 Relationship to Previous Work . . . 89

5.3 Preliminaries . . . 91
5.3.1 Efficient Mobile/Light Building Blocks . . . 91
5.3.2 Attention Mechanisms . . . 93
5.4 Our Approach to Efficient and Robust Face Recognition . . . 95
5.4.1 Enhanced Bottleneck . . . 96
5.4.2 RobFaceNet . . . 98
5.5 Experiments and Analysis . . . 102
5.5.1 Preprocessing . . . 102
5.5.2 Implementation Details . . . 103
5.5.3 Ablation Study . . . 103
5.5.4 Performance versus Computational Complexity . . . 105
5.5.5 Comparison with State-of-the-Art . . . 107
5.6 Discussions . . . 116
5.7 Summary . . . 117

6 Conclusions **118**
6.1 Summary of Thesis Achievements . . . 118
6.2 Future Directions . . . 121

Bibliography **123**

Abstract

Face recognition stands as the superior biometric technique for identity authentication, finding extensive applications in our daily lives, like access control, finance, entertainment, and public security. Despite the widespread integration of face biometrics, most current face recognition systems are tailored for environments where accurate control governs the process of capturing facial images.

In recent years, rapid advancements in face recognition techniques have unfolded across both academic and industrial sectors. This transformation has been driven by key factors, including the availability of substantial annotated training datasets, the rise of convolutional neural network based deep architectures, the affordability and power of computational resources, and the emergence of refined loss functions. Despite the considerable strides and achievements, persistent challenges await resolution.

This thesis makes significant contributions to in-the-wild face recognition, particularly concerning human-robot interaction, from three perspectives: model enhancement, loss function innovation, and network design. By enhancing current face recognition framework capabilities, designing novel loss functions, and carefully developing network architectures, this thesis aims to navigate the challenges of recognizing faces within dynamic and uncontrolled environments, where humans and robots interact.

Firstly, we address the complexities of human-robot interaction (HRI), highlighting the challenges of real-time face recognition. Emphasizing the need for fast processing and high accuracy, we adopt lightweight convolutional neural networks for our proposed face recognition framework. The integration of the state-of-the-art ArcFace loss function and the RetinaFace method for face detection, combined with an online real-time face tracker, empowers our system to adeptly handle challenges such as varying illumination, different head poses, and occlusions. By merging tracking data with recognized identities, we enhance the system's performance in unconstrained settings, resulting in improved recognition accuracy and processing speed. Evaluations within our HRI system, "RoSA," showcase significant advancements while also highlighting areas for further refinement.

Secondly, we explore the transformative role of margin-based softmax loss functions in face recognition. Traditional methods, which rely on a static, single margin, may not effectively address diverse real-world data. In response, we introduce the JAMsFace loss function, which offers flexible margin settings based on the class distribution. Harnessing joint adaptive margins in both angle and cosine spaces, JAMsFace refines feature discernibility and effectively addresses the challenge of class imbalance. Comprehensive evaluations across various datasets validate the efficacy of JAMsFace,

signaling a shift towards more adaptive face recognition methodologies.

Finally, we present RobFaceNet, a network specifically designed for face recognition. Balancing computational efficiency with accuracy, RobFaceNet employs a multi-feature approach and integrates the modified h-swish activation function. We further enhance RobFaceNet with an attention-based bottleneck, incorporating either a CA or SE attention module, to boost its facial feature discernment capabilities. Rigorous evaluations against state-of-the-art face recognition models confirm RobFaceNet's superior performance, underscoring the potential of lightweight models in real-world scenarios.

In conclusion, this thesis encapsulates a comprehensive journey through the complex landscape of face recognition in dynamic and uncontrolled environments, specifically within the context of human-robot interactions. Addressing fundamental challenges, innovating within the scope of loss functions, and devising efficient network designs underscores a clear roadmap toward achieving more seamless and natural interactions between humans and robots.

Deutsche Kurzfassung

Gesichtserkennung gilt als überlegene biometrische Technik zur Identitätsauthentifizierung und findet umfangreiche Anwendungen in unserem täglichen Leben, wie Zugangskontrolle, Finanzen, Unterhaltung und öffentliche Sicherheit. Trotz der weit verbreiteten Integration von Gesichtsbiometrie sind die meisten aktuellen Gesichtserkennungssysteme für Umgebungen maßgeschneidert, in denen eine genaue Steuerung den Prozess der Erfassung von Gesichtsbildern bestimmt.

In den letzten Jahren haben rasante Fortschritte in den Techniken zur Gesichtserkennung sowohl im akademischen als auch im industriellen Bereich stattgefunden. Diese Transformation wurde durch Schlüsselfaktoren vorangetrieben, darunter die Verfügbarkeit umfangreicher annotierter Trainingsdatensätze, der Aufstieg von tiefen Architekturen auf der Grundlage von Convolutional Neural Networks, die Erschwinglichkeit und Leistungsfähigkeit von Rechenressourcen und das Auftreten raffinierter Verlustfunktionen. Trotz der erheblichen Fortschritte und Erfolge warten weiterhin anhaltende Herausforderungen auf Lösungen.

Diese Dissertation trägt zur Gesichtserkennung unter realen Bedingungen bei, insbesondere im Zusammenhang mit der Interaktion zwischen Mensch und Roboter, aus drei Perspektiven: der Verbesserung von Modellen, der Innovation von Verlustfunktionen und dem Design von Netzwerken. Durch die Verbesserung der Fähigkeiten des aktuellen Gesichtserkennungsrahmens, die Entwicklung innovativer Verlustfunktionen und die sorgfältige Gestaltung von Netzwerkarchitekturen zielt diese Arbeit darauf ab, die Herausforderungen bei der Erkennung von Gesichtern in dynamischen und unkontrollierten Umgebungen zu bewältigen, in denen Menschen und Roboter interagieren.

Erstens behandeln wir die Komplexitäten der Mensch-Roboter-Interaktion (HRI) und betonen die Herausforderungen der Echtzeit-Gesichtserkennung. Mit Schwerpunkt auf schneller Verarbeitung und hoher Genauigkeit verwenden wir leichte Convolutional Neural Networks für unseren vorgeschlagenen Gesichtserkennungsrahmen. Die Integration der hochmodernen ArcFace-Verlustfunktion und der RetinaFace-Methode zur Gesichtserkennung, kombiniert mit einem online Echtzeit-Gesichts-Tracker, ermöglicht es unserem System, Herausforderungen wie unterschiedliche Beleuchtung, verschiedene Kopfpositionen und Verdeckungen geschickt zu bewältigen. Durch die Zusammenführung von Tracking-Daten mit erkannten Identitäten verbessern wir die Leistung des Systems in nicht eingeschränkten Umgebungen und erzielen eine verbesserte Erkennungsgenauigkeit und Verarbeitungsgeschwindigkeit. Bewertungen innerhalb unseres HRI-Systems, "RoSA", zeigen signifikante Fortschritte und weisen gleichzeitig Bereiche für weitere Verbesserungen auf.

Zweitens untersuchen wir die transformative Rolle von margenbasierten Softmax-Verlustfunktionen in der Gesichtserkennung. Traditionelle Methoden, die auf einem statischen, einzelnen Margin basieren, können vielfältige realweltliche Daten möglicherweise nicht effektiv bewältigen. Als Reaktion darauf führen wir die JAMsFace Verlustfunktion ein, die flexible Margin-Einstellungen basierend auf der Klassenverteilung bietet. Durch die Nutzung gemeinsamer anpassbarer Margen sowohl im Winkel- als auch im Cosinus-Raum verfeinert JAMsFace die Merkmalsunterscheidbarkeit und bewältigt effektiv die Herausforderung der Klassenungleichgewicht. Umfassende Bewertungen in verschiedenen Datensätzen bestätigen die Wirksamkeit von JAMsFace, was auf eine Verschiebung hin zu adaptiveren Methoden in der Gesichtserkennung hinweist.

Schließlich präsentieren wir RobFaceNet, ein speziell für die Gesichtserkennung entwickeltes Netzwerk. RobFaceNet balanciert Recheneffizienz und Genauigkeit aus und verwendet einen multi-feature Ansatz und integriert die modifizierte h-swish Aktivierungsfunktion. Wir verbessern RobFaceNet weiter mit einem aufmerksamkeitsbasierten Engpass, der entweder ein CA- oder SE-Aufmerksamkeitsmodul enthält, um seine Fähigkeiten zur Merkmalsunterscheidung im Gesicht zu steigern. Rigorose Bewertungen im Vergleich zu modernsten Gesichtserkennungsmodellen bestätigen die überragende Leistung von RobFaceNet, was das Potenzial von leichten Modellen in realen Szenarien unterstreicht.

Zusammenfassend fasst diese Dissertation eine umfassende Reise durch das komplexe Gebiet der Gesichtserkennung in dynamischen und unkontrollierten Umgebungen zusammen, insbesondere im Kontext der Interaktion zwischen Mensch und Roboter. Die Bewältigung grundlegender Herausforderungen, die Innovation im Rahmen von Verlustfunktionen und die Entwicklung effizienter Netzwerke unterstreichen einen klaren Weg zur Erreichung nahtloserer und natürlicherer Interaktionen zwischen Menschen und Robotern.

Related Publications

Most of the material contained in this dissertation is partly based on the following refereed papers and journals published in a variety of peer-reviewed journals and international conference proceedings.

Peer-reviewed Articles in International Journals & Conferences:

[1] A. Khalifa and A. Al-Hamadi, "Jamsface: joint adaptive margins loss for deep face recognition," *Neural Computing and Applications,* ***(IF 6)***, 35(26):19025–19037, 2023.

[2] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi, "Face recognition and tracking framework for human-robot interaction," *Applied Sciences,* ***(IF 2.9)***, 12(11):5568, 2022.

[3] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction," *Sensors,* ***(IF 3.9)***, 22(3):923, 2022.

[4] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, "Multimodal engagement prediction in multiperson human–robot interaction," *IEEE Access,* ***(IF 3.9)***, 10:61980–61991, 2022.

[5] M.-A. Fiedler, P. Werner, A. Khalifa, and A. Al-Hamadi, "Sfpd: Simultaneous face and person detection in real-time for human–robot interaction," *Sensors,* ***(IF 3.9)***, 21(17):5918, 2021.

[6] A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Towards efficient and robust face recognition through attention-integrated multi-level cnn," *Multimedia Tools and Applications,* ***(IF 3.6)***, pages 1-23, Springer, 2024.

[7] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, "Fine-grained gaze estimation based on the combination of regression and classification losses," *Applied Intelligence,* ***(IF 5.3)***, 54(21):10982-10994, Springer, 2024.

[8] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023.

[9] A. Khalifa and A. Al-Hamadi, "A survey on loss functions for deep face recognition network," In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–7. IEEE, 2021.

[10] D. Strazdas, J. Hintz, A. Khalifa, and A. Al-Hamadi, "Robot system assistant (rosa): Concept for an intuitive multi-modal and multi-device interaction system," In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE, 2021.

[11] T. Hempel, M.-A. Fiedler, A. Khalifa, A. Al-Hamadi, and L. Dinges, "Semantic aware environment perception for mobile human-robot interaction," In *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 200–203. IEEE, 2021.

[12] A. A. Abdelrahman, T. Hempel, A. Khalifa, D. Strazdas, and A. Al-Hamadi, "MobGazeNet : Robust Gaze Estimation Mobile Network Based On Progressive Attention Mechanisms," *Machine Vision and Applications,* ***(IF 3.3)***, Second Revision.

1 Introduction

As the convergence of artificial intelligence and robotics continues to redefine the boundaries of technological innovation, face recognition (FR) represents a key pillar that bridges human-computer interactions (HCI). While FR, considered a behavioral biometric, inherently possesses lower accuracy than physiological biometrics such as fingerprint and iris recognition, its contactless nature has pushed it to the forefront of biometric techniques. This advanced form of biometric recognition has evolved from a science fiction concept to a widespread reality, impacting numerous sectors, including mobile technology, surveillance systems, and, more prominently, human-robot interactions (HRI).

FR refers to the technology that compares a human face in an image with a database of facial images. This ability is typically divided into two primary tasks: face identification and face verification. The former, as depicted in Fig. 1.1a, the objective is to identify the subject of a given face image by comparing it with a gallery set. In contrast, the latter, illustrated in Fig. 1.1b, involves assessing a pair of facial images to determine if they represent the same individual.

(a) Face Identification **(b)** Face Verification

Figure 1.1: Face Recognition Tasks. (a)Identification: Assigning an identity to a provided face image. (b) Verification: Ascertain if two provided face images represent the same subject.

With the advent of robots entering our daily lives, whether as assistants, companions, or collaborators, the ability of these robots to accurately recognize and respond to humans in diverse settings becomes paramount. However, FR is not without its challenges, primarily arising from the dynamic, unpredictable real-world scenarios where

traditional methods often fall short.

This chapter delves deep into defining these challenges, outlines clear objectives to address them, and elucidates the contributions made by this research in enhancing deep FR, especially within the vibrant and complex landscape of HRIs.

1.1 Motivation

In the dawn of the digital age, as technologies have rapidly evolved, FR has emerged as a beacon of technological advancement. It is no longer the stuff of science fiction; it is now an integral aspect of our daily digital interactions.

Evolution of Face Recognition. Our day-to-day technologies highlight our increasing dependence on FR, from the simplicity of unlocking smartphones with a glance to the complexities of advanced security systems. This widespread adoption is powered by an array of state-of-the-art open-source techniques addressing FR's various components [48]. Its pivotal role in fields like security applications, criminal investigations, and searching for missing persons emphasizes its societal significance [100].

Human-Robot Interaction. Beyond identification, FR serves as a bridge for mutual understanding between humans and machines. As robots evolve into potential partners, caregivers, and colleagues, their ability to recognize and interpret human facial cues becomes critical. This pivotal role of FR signifies its potential to enable more intuitive HRIs.

The Challenge of the 'Wild'. Real-world environments are unpredictable. They present challenges not seen in controlled lab settings: dynamic lighting conditions, diverse facial expressions, occlusions, pose variation, and other complexities like hats, glasses, and varying image quality [66, 67]. As robots step out of controlled environments and into our homes, streets, and workplaces, the imperative for robust FR capabilities in HRI becomes paramount. However, realizing this objective carries its own technical challenges, especially when accounting for the resource constraints inherent to robotic platforms.

Motivated by these observations, this thesis seeks to address the challenges of real-world FR, particularly in the dynamic environment of HRI.

1.2 Problem Definition

Face Recognition and Computational Complexity. Traditionally, the path to optimal FR involves training a model of considerable capacity on extensive datasets, as MS1M [71]. This approach also leverages cutting-edge classification loss functions, including but not limited to, CosFace [204], ArcFace [37], and Adaptiveface [132]. While models trained this way perform impressively on face benchmarks like LFW

[90], CFP-FP [178], AgeDB-30 [154], IJB-C [149], and MegaFace [104], they require significant computational resources and time due to the volume of training data and the complexity of parameter tuning.

For real-world applications, many FR systems are deployed on devices with limited resources, from mobile phones to HRI platforms. Accommodating deep learning models, primarily designed for high-resource environments, on such platforms poses challenges. There is an inherent conflict between the computational and storage demands of these sophisticated models and the capabilities of HRI platforms.

In light of this, while large datasets undoubtedly boost the performance of FR systems, striking a balance between efficiency and accuracy is crucial. This has catalyzed research towards the development of lightweight networks for FR without compromising performance. **Concurrently**, innovating methods to speed up processing for each component of FR is of paramount importance.

Face Recognition and Global Representations. Various factors like pose, illumination, occlusion, resolution, and aging influence FR performance. Despite CNN-based representations achieving state-of-the-art performance, many existing models focus on global representations, treating entire faces as inputs [37, 136, 164, 217].

In HRI environments, where large pose variations and significant occlusions are common, global face appearances can vary drastically. Nevertheless, certain local patches remain consistent and play pivotal roles in FR.

Approaches for extracting discriminative local features often categorize into landmark-based and attention-based methods. Landmark-based techniques entail training distinct networks on facial components delineated by landmarks [42, 186]. However, the detection of these landmarks may falter under specific conditions, such as extreme poses or occlusions. On the other hand, attention-based methods, not reliant on facial landmarks, autonomously identify discriminative facial parts [99, 209]. However, these methods often focus on similar facial regions, overlooking other essential areas.

Therefore, capturing comprehensive local representations is vital. It ensures that even if certain facial features are obscured or are similar across subjects, other features can contribute effectively to recognition. Achieving this with minimal computational complexity is essential, especially considering the constraints of HRI platforms.

Face Recognition and Long-tail Data. The quality of feature descriptors significantly influences FR performance. Training and testing images often feature different identities. Distance metrics directly interact with these features to ascertain if they correspond to the same identity. While there has been significant advancement in FR recently, a persistent challenge is the generalization of learned features. These features often excel in the same domain as the training set but falter in unseen domains, a challenge especially pronounced in wild settings due to domain shifts.

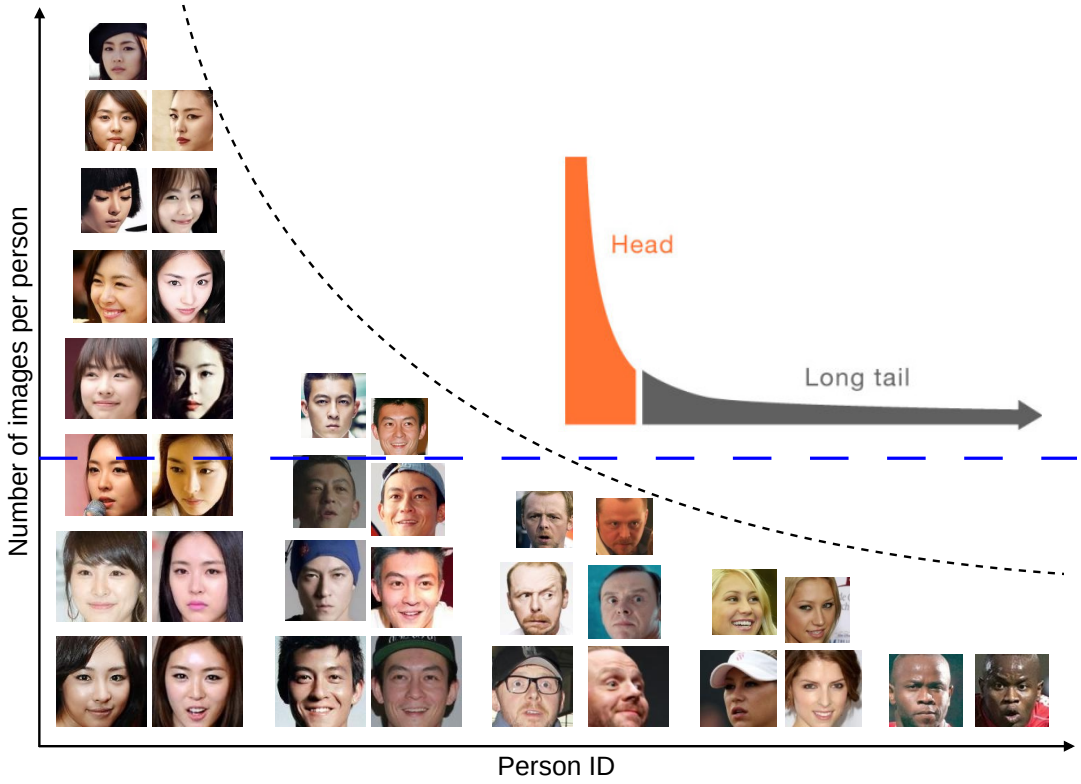


Figure 1.2: Visualization of the long-tailed distribution in the MS-Celeb-1M dataset [71].

Real-world face datasets often exhibit a long-tailed distribution, as shown in Fig. 1.2. Only a few identities are frequent (head data), with the majority being infrequent (tail data). Training on such datasets often results in well-trained head identities, but tail identities get overlooked, hindering the development of robust and generalizable features, as shown in Fig. 1.3.

Earlier studies often tackled the long-tail distribution challenge by trimming the tail, aiming for a more uniform class distribution. Nevertheless, this approach sometimes inadvertently led to the loss of numerous valuable identities. According to Zhang et al.'s research [260], preserving 40% of the tail data has been observed to significantly enhance performance. However, this selective data retention approach has its downsides. Discarding tail classes poses a risk of missing out on essential information. It is noteworthy that although tail classes are often trimmed, they can provide unique insights that are not present in the head classes. These insights can play a critical role in improving the performance of trained models.

Given this context, a pivotal challenge in this domain is determining an adaptive margin penalty tailored to each class, which would greatly aid in creating robust and generalizable features, thereby leading to superior FR models.

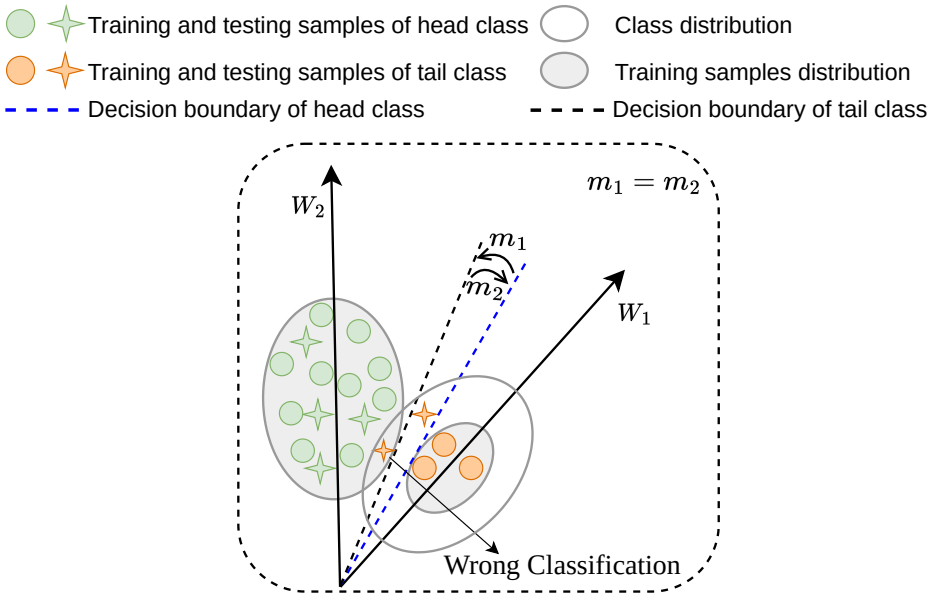


Figure 1.3: Effects of Long-tailed Distribution on Face Recognition [108]. In the presence of a static additive margin, the model is prone to errors when encountering new test samples from a tail class. Specifically, tail classes necessitate a larger margin (represented by m_1) for accurate classification, whereas head classes require a smaller margin (represented by m_2).

1.3 Goal and Contributions

This thesis presents a series of significant contributions that collectively address the challenges of deep FR within real-world scenarios. The key contributions are as follows:

1. **Enhanced FR via Integrated Face Tracking for HRI (*Model Enhancement*):** One of the cornerstone contributions of this work is the introduction of a state-of-the-art framework that cohesively incorporates FR and tracking within the environment of HRI. Motivated by the dual goals of reducing computational complexity and increasing recognition accuracy, this framework leverages the power of data obtained from tracked faces, merging it with recognized identities. Such integration enables the direct retrieval of user identity in sequential frames from the face tracker's metadata, eliminating the redundancy of reinitiating the detection and recognition processes. Consequently, this enhances both the efficiency and performance of real-time HRIs. This innovative approach directly addresses the challenge of *FR and Computational Complexity*, opening the door to more responsive and resource-efficient interactive experiences.

2. **Adaptive Feature Learning for Enhanced Face Recognition (*Loss Function Design*):** A defining contribution of this work is the formulation of a novel loss function, named JAMsFace. Rooted in the challenges presented by real-world face datasets, characterized by their long-tailed distributions, JAMsFace

is designed with precision to handle the disparities between frequently appearing identities (the head data) and the vast majority of identities that are rarely seen (the tail data). This is achieved through the incorporation of adaptive margins in both angular and cosine spaces. This adaptive approach dynamically adjusts penalty values based on the distribution of each class. The direct implication is a faster model convergence and a marked enhancement in the model's ability to discriminate between classes, even in those under-represented domains. Through this, JAMsFace directly addresses the challenges inherent in *FR and Long-tail Data*, ensuring a balanced and efficient recognition capability across all identity distributions.

3. **Efficient and Robust Face Recognition Network (*Network Design*):** A seminal contribution of this work lies in the introduction of a lightweight and robust CNN architecture designed to capture comprehensive and diverse local face representations, striking an optimal balance between accuracy and computational efficiency. Our proposed RobFaceNet network employs a multi-feature approach and attention mechanisms, leveraging local and global features extracted from input face images. This integration significantly enhances the accuracy of FR tasks in various challenging conditions like pose variations and occlusions. Additionally, we present a novel bottleneck structure with integrated attention mechanisms to enforce the model to extract diverse local features, bolstering the network's robustness and elevating its facial feature extraction capabilities. This architecture is primed to resolve the problems tied to *FR and Global Representations* while concurrently addressing the challenge of *FR and Computational Complexity*.

In essence, by integrating our key contributions, the optimized network design (RobFaceNet), the adaptive loss function (JAMsFace), and the model enhancement through face tracking, we have developed a FR system that effectively balances computational efficiency with high performance. This streamlined system directly addresses the main challenges inherent in real-world FR, especially within HRI settings. With these innovations, we aim to enable smooth, efficient, and accurate interactions between robots and humans in different real-world scenarios.

1.4 Publications

This section comprises publications that have been authored during the progression of this Ph.D. thesis.

Peer-reviewed Articles in International Journals & Conferences

[1] A. Khalifa and A. Al-Hamadi, "Jamsface: joint adaptive margins loss for deep face recognition," *Neural Computing and Applications,* ***(IF 6)***, 35(26):19025–19037, 2023.

[2] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi, "Face recognition and tracking framework for human-robot interaction," *Applied Sciences,* ***(IF 2.9)***, 12(11):5568, 2022.

[3] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction," *Sensors,* ***(IF 3.9)***, 22(3):923, 2022.

[4] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, "Multimodal engagement prediction in multiperson human–robot interaction," *IEEE Access,* ***(IF 3.9)***, 10:61980–61991, 2022.

[5] M.-A. Fiedler, P. Werner, A. Khalifa, and A. Al-Hamadi, "Sfpd: Simultaneous face and person detection in real-time for human–robot interaction," *Sensors,* ***(IF 3.9)***, 21(17):5918, 2021.

[6] A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Towards efficient and robust face recognition through attention-integrated multi-level cnn," *Multimedia Tools and Applications,* ***(IF 3.6)***, pages 1-23, Springer, 2024.

[7] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, "Fine-grained gaze estimation based on the combination of regression and classification losses," *Applied Intelligence,* ***(IF 5.3)***, 54(21):10982-10994, Springer, 2024.

[8] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023.

[9] A. Khalifa and A. Al-Hamadi, "A survey on loss functions for deep face recognition network," In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–7. IEEE, 2021.

[10] D. Strazdas, J. Hintz, A. Khalifa, and A. Al-Hamadi, "Robot system assistant (rosa): Concept for an intuitive multi-modal and multi-device interaction system," In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE, 2021.

[11] T. Hempel, M.-A. Fiedler, A. Khalifa, A. Al-Hamadi, and L. Dinges, "Semantic aware environment perception for mobile human-robot interaction," In *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 200–203. IEEE, 2021.

[12] A. A. Abdelrahman, T. Hempel, A. Khalifa, D. Strazdas, and A. Al-Hamadi, "MobGazeNet : Robust Gaze Estimation Mobile Network Based On Progressive Attention Mechanisms," *Machine Vision and Applications,* ***(IF 3.3)***, Second Revision.

1.5 Outline

Our research journey begins with the design and implementation of a comprehensive FR framework, specifically crafted for real-world HRI environments. The objective of this initial phase is to assess the effectiveness and limitations of SOTA models in real-world human-robot interaction (HRI) scenarios. Through empirical investigation, we aim to gain valuable insights and findings that will facilitate the development of novel approaches to address the identified challenges.

As the research progresses, we will explore the complexities of FR within HRI, continuously refining and expanding our framework. This exploration covers a range of advancements, from enhanced face tracking to the development of innovative loss functions and network designs. Each development phase is carefully structured to build upon the previous one, ensuring systematic and incremental advancement. The primary objective of our research is to create sophisticated solutions that significantly enhance robot capabilities in human-centric environments, focusing on everything from tracking accuracy to computational efficiency. The detailed descriptions and outcomes of these progressive enhancements, spanning the full range from improved face tracking to sophisticated network design, will be extensively covered in the following chapters of this thesis.

The thesis outline is illustrated in Fig. 1.4 as follows:

- **Chapter 2** provides a comprehensive review of existing methodologies relevant to our thesis, encompassing face detection, face alignment, face tracking, and feature learning methods and face representation networks as well. This chapter introduces representative approaches from each category, offering insights into the prevailing trends within the field.

- **Chapter 3** proposes an explanation of our FR system designed for optimizing HRI applications. Built on the solid foundations of lightweight CNNs, this framework adeptly integrates face tracking methodologies to enhance real-time recognition in HRI scenarios. Modularly packaged for easy HRI system integration, our design employs the ArcFace loss function combined with RetinaFace for detection and a uniquely developed face tracker. Preliminary tests on our HRI system, "RoSA", and the Wizard-of-Oz study dataset highlight a significant improvement in recognition robustness. However, as our studies illuminate areas of potential refinement, particularly in complex scenes, our future endeavors are directed toward refining architectural elements to further enhance intuitive HRIs.

- **Chapter 4** delves deep into our second major contribution: *Adaptive Feature Learning for Enhanced FR*. Central to this chapter is the introduction of JAMsFace, our innovative loss function designed for FR. Instead of adhering to the traditional fixed margins, JAMsFace dynamically adjusts margins based on the

distribution of each class. This approach ensures a more nuanced understanding of intra-class variations and a clearer differentiation between different classes. We rigorously test and validate JAMsFace using esteemed benchmarks such as LFW, CALFW, and MegaFace, with the results substantiating its superior efficacy and robustness. This chapter underscores the importance of utilizing adaptable and data-driven approaches in the field of FR, thereby suggesting potential avenues for future investigation.

- **Chapter 5** introduces our innovative contribution: the RobFaceNet architecture, tailored for efficient and robust FR. Leveraging the foundational principles of mobile network design, RobFaceNet ensures efficiency without compromising performance. Central to its effectiveness is the integration of an attention-driven bottleneck, which pinpoints and elevates robust facial features. Through its multi-feature approach, the architecture excels in extracting comprehensive local and global facial representations, enhancing its resilience and precision in FR tasks. Such a design allows RobFaceNet to achieve a perfect harmony between computational speed and superior accuracy. To validate the efficacy of RobFaceNet, we subjected it to intensive evaluations against esteemed benchmarks. The outcomes clearly approve its standing as a state-of-the-art FR solution optimized for real-world scenarios.

- **Chapter 6** provides a comprehensive summary of all proposed approaches discussed in the preceding chapters, highlighting their key contributions and findings. Furthermore, the chapter delves into potential avenues for future study in the field of deep FR in unconstrained environments.

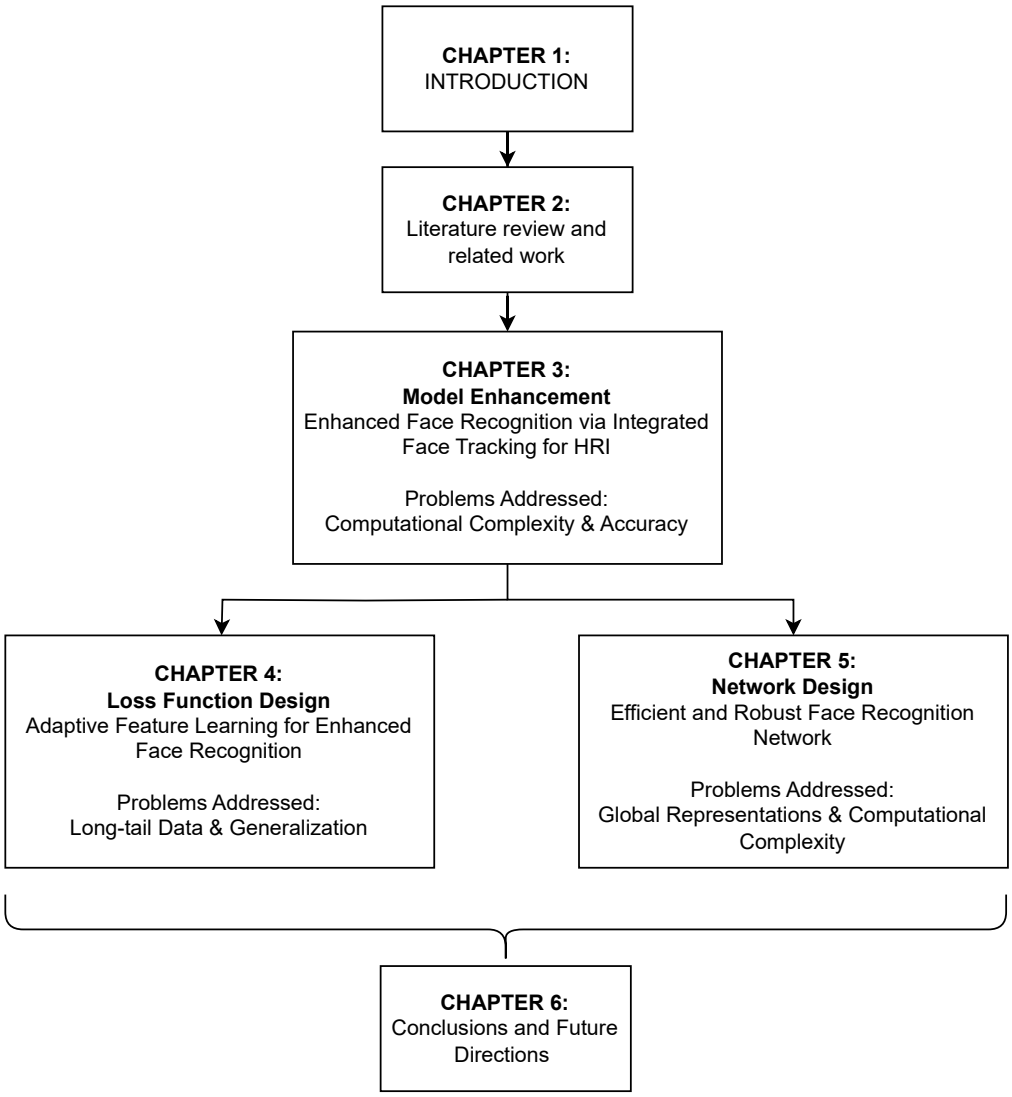


Figure 1.4: Thesis Outline.

2 Background and Literature Review

This chapter delves into the foundational principles and key scholarly works underpinning our research on deep face recognition, with a particular focus on its application within human-robot interaction (HRI). We start by exploring the development of face recognition (FR) systems, beginning with basic preprocessing stages and progressing to more sophisticated concepts like face tracking. This sequential investigation offers insights into the complexities and challenges of developing and deploying modern FR technologies. Throughout our journey, our objective is to provide the reader with a comprehensive understanding, equipping them with the essential background necessary to comprehend the innovative approaches presented in subsequent chapters.

2.1 Evolution of Face Recognition Systems

In the era of rapid technological advancement, the significance of biometric recognition i.e. biometrics, has grown exponentially in modern security. Biometrics focuses on the analysis and statistical evaluation of distinctive physical and behavioral attributes exhibited by individuals [96]. Spanning a spectrum of techniques, ranging from fingerprint analysis and retinal scanning to voice identification and DNA recognition, FR stands out as a leading contender. FR has attracted considerable attention from both industry and academia [207], due to its notable advantages, such as:

1. **Ease of Implementation:** FR stands out for its straightforward deployment, making it accessible and user-friendly.

2. **Contactless and Passive Capture:** This technology captures facial data without requiring physical contact, enhancing convenience and hygiene.

3. **Robust Tracking Capability:** FR exhibits a strong aptitude for accurately tracking and identifying individuals, bolstering its reliability.

4. **Affordable Data Acquisition Devices:** The equipment used for acquiring facial data is cost-effective. This cost-effectiveness contributes to the overall feasibility and accessibility of the approach.

At its core, FR aims to match a given human face with a stored facial profile by evaluating differences in feature embeddings [90]. This versatile technology has found

its footprints across diverse sectors, from the glitz of entertainment and precision-targeted marketing campaigns to the critical realms of health diagnostics and national security [100].

Replicating the human ability to recognize faces in machines has historically been a challenge. This effort has evolved from early algorithms based on simple computations to today's advanced models powered by artificial intelligence and deep learning [118, 213].

FR is an intricate domain that encompasses a suite of related technologies synergistically working together to build a robust system. The overarching goal of these systems is to identify or verify an individual based on their facial features. To accomplish this, the system must process and analyze facial data, ensuring that the given input is primed for optimal recognition. These technologies include image preprocessing, face detection, face alignment, face representation, and feature extraction.

2.2 Preprocessing

In uncontrolled environments, individuals might not be consciously aware of a recognition system operating in the background, automatically processing their facial data. In such scenarios, image capturing devices, like surveillance cameras, autonomously detect individuals from a distance, capture their face features, and then match these against a database for identification. The success of this recognition process largely depends on the quality of the captured image, which in turn is influenced by various factors. These factors encompass the camera's efficiency, the distance between the subject and the camera, ambient lighting conditions, and the orientation of the individual's face relative to the camera, among others [72, 73]. Given these challenges, preprocessing the captured image becomes paramount for improving recognition accuracy.

There exists a multitude of image processing techniques tailored to enhance the quality of captured images, subsequently boosting recognition rates [270]. Notable methods include image normalization, denoising, filtering, histogram equalization, image resizing and cropping. Leveraging these techniques can considerably augment the quality of images and thus enhance the overall recognition accuracy [113, 155, 167, 196].

Typical preprocessing steps include noise reduction [196], which helps in eliminating random variations or 'noise' in images, and contrast enhancement [155] that improves image visibility by adjusting the brightness and contrast levels. Image normalization [113, 167] is another crucial step that ensures that all images fed into the system are of a consistent size and scale, making it easier for subsequent algorithms to perform their tasks effectively.

With the advent of deep learning, some preprocessing steps have been incorporated directly into the network architecture, ensuring that the system learns the optimal transformations required for high-performance recognition [102].

2.3 Face Detection and Tracking

2.3.1 Face Detection

Face detection is the first step in the recognition pipeline, localizing faces within images or video frames. The algorithm isolates regions of interest (ROIs) where faces are likely present, directing subsequent procedures to focus on these regions. Face detection techniques have evolved over the years, progressing from the traditional Viola-Jones algorithm [200] to contemporary deep learning-based methods like the Single Shot MultiBox Detector (SSD) [135] and RetinaFace [36].

2.3.1.1 Traditional Methods

The groundbreaking work by Viola and Jones [200] set the stage for many face detection techniques that followed [14, 125]. Their method harnessed Haar-like features, utilizing AdaBoost to train a cascade of classifiers. The Deformable Part Models (DPM) technique, another stalwart in traditional face detection, is elaborated upon by Forsyth [56]. Its efficacy in face detection has been corroborated by Yan et al. [238].

Despite their successes, traditional methods often grapple with challenges presented by intricate real-world scenarios. Their primary shortcomings arise from a reliance on hand-crafted features and the employment of shallow classifiers. These lack the nuanced depth and adaptability essential for grappling with diverse conditions [152].

2.3.1.2 Deep Learning-Based Methods

The advent of CNNs revolutionized face detection, with numerous pioneering works significantly enhancing its efficacy [21, 36, 120, 156, 253, 258]. Broadly, these advancements can be categorized into three groups based on their seminal technical contributions: Cascade-CNN based models, Region-CNN (R-CNN) based models, and single-shot detector models.

Cascade-CNN Based Models Cascade-CNN models adopt a cascade architecture similar to Viola-Jones but substitute traditional classifiers with small CNN models. These compact models are trained in stages to distinguish between face and non-face regions, gradually removing non-face areas with increasing resolution.

Li et al. [120] introduced an early deep-learning model for face detection based on the cascade-CNN architecture. This cascade-CNN functions across various resolutions, rapidly eliminating background zones in low-resolution stages while meticulously evaluating a limited set of candidates at the final high-resolution stage. To bolster localization precision and curtail the candidate count in subsequent stages, a CNN-based calibration phase post each detection stage was incorporated. Relative

to state-of-the-art methods available at that time, their face detector was both adept and efficient, boasting a processing speed of 14 FPS on a CPU for VGA-resolution images and surging to 100 FPS with GPU support. This methodology was further refined by Zhang et al. [253], who focused on joint face detection and alignment. They pinpointed two deficiencies in Li et al.'s model: the ancillary computational overhead during calibration and the overlooked correlation between face detection and alignment. To counteract these gaps, Zhang et al. postulated a novel online hard sample mining strategy that enhances performance, eliminating the need for manual sample selection. During the training phase, they introduced a data routing mechanism that allows different layers to be trained with different types of samples. This allows the deeper layers to concentrate on more challenging samples compared to the shallower layers.

Qi et al. [168] also leveraged a cascaded architecture but used three stages of deep convolutional networks to further improve both detection performance and model generalization. Zeng et al. [249] sped up the cascade CNN face detector by generating multi-scale face proposals using a pyramid network.

To conclude, the advantage of cascade-CNN models lies in their ability to strike a good balance between speed and accuracy. The classifiers at earlier stages effectively reduce the background while retaining the faces, and they do so with relatively low computational cost. However, computational complexity can increase significantly during the inference stage when processing images containing a large number of faces.

Region-Based CNN Models The region-Based CNN (R-CNN) models involve a range of frameworks that incorporate generic R-CNN object detection algorithms.These include but are not limited to Fast R-CNN [58], R-FCN [32], Faster R-CNN [173], and Mask R-CNN [77].

Models such as Face R-CNN [203] and FDNet [250], are built upon the Faster R-CNN [173] framework, a more advanced version of the original R-CNN that combines the region proposal network (RPN) and Fast R-CNN into a single network, thus enhancing efficiency. The RPN proposes candidate object bounding boxes, while Fast R-CNN uses these proposals to classify the objects and refine their bounding boxes. By directly applying the Faster R-CNN algorithm, both models can more effectively and efficiently detect faces in images.

In addition to these models, the authors in [242] developed a different approach by creating a specialized set of CNNs with varying structures based on the Faster R-CNN [173] framework. This approach addresses the limitations of traditional face detection models, which often struggle with detecting faces of varying sizes and orientations. The different structures of the CNNs in this approach enable the detection of faces at multiple scales, thereby improving the overall detection performance.

Another innovative model, CMS-RCNN [269], integrates contextual reasoning into the detection process. Traditional object detection models often generate false posi-

tives by incorrectly classifying non-face regions as faces. CMS-RCNN minimizes these false positives by considering the surrounding context of each region proposal. For example, if a region proposal is surrounded by other regions that are highly likely to be faces, it increases the probability that the proposal is also a face.

The FA-RPN [157] model introduces an optimized anchor placement strategy to face detection. Anchors are pre-defined boxes of different sizes and aspect ratios used as starting points for detecting objects in an image. Traditional models use a dense set of anchors, which increases the computational load. The optimized anchor placement strategy of FA-RPN reduces the number of anchors required for face detection, thereby decreasing the computational load and improving detection speed while still maintaining high detection accuracy.

Face R-FCN [212], applies the R-FCN [32] algorithm directly. R-FCN is an object detection algorithm that divides an object into several parts and then aggregates the score of each part to get the final detection score. This approach allows Face R-FCN to achieve high detection accuracy while maintaining a relatively low computational load compared to other models that do not divide the object into parts. While [268], develops an Expected Max Overlapping (EMO) score to explain the capability of anchors in capturing faces. The EMO score considers the balance between recall and precision to optimize the anchor settings. This optimization leads to better detection performance by reducing the number of false positives and improving detection accuracy.

MaskFace [243] model leverages the Mask R-CNN [77] framework to detect faces and predict facial landmarks. The Mask R-CNN model is an enhanced version of the Faster R-CNN model, which has a branch for the purpose of predicting segmentation masks alongside bounding boxes. By using Mask R-CNN, MaskFace can not only detect faces but also predict the location of facial landmarks, such as the eyes, nose, and mouth. This capability is pivotal for various applications, including facial expression recognition and face alignment.

To conclude, while R-CNN based face detection models have significantly advanced in reducing false positives and predicting facial landmarks, they still face challenges in real-time applications due to high computational requirements [169, 247], and in accurately detecting faces in challenging conditions such as low light, occlusions, and extreme poses.

Single Shot Detector Models As object detection methodologies evolved, the realm of face detection witnessed a transformative shift with the advent of Single Shot Detector (SSD) and RetinaNet [127] models. Their architecture is fundamentally different from two-stage detectors. Instead of relying on subsequent R-CNN, single shot detectors directly locate faces after the RPN, resulting in a more streamlined and efficient process. Moreover, while the computational complexity of two-stage detectors can be greatly influenced by the number of faces in an image, the single-shot framework

maintains consistent computational efficiency during inference.

Utilizing the SSD architecture, the Single Shot Scale-invariant Face Detector (S3FD) [258] effectively detects small faces. This is accomplished by employing a face detection strategy that ensures fairness in scale, utilizing a scale compensation anchor matching technique to improve the accuracy of identifying small faces, and including a max-out background label to minimize false positives associated with small faces.

FaceBoxes [257] serves as another SSD-based face detection model that real-time detection, even on CPU-based systems. This model also innovates an anchor identification strategy, ensuring a balanced anchor distribution across different layers. This balance proves particularly effective in improving the detection of smaller faces.

EXTD [245] distinguishes itself in the realm of compact models by employing an iterative approach that involves the reuse of a common lightweight and shallow backbone network. This approach enables the reduction of model size while maintaining optimal performance. Meanwhile, SSH (Single Stage Headless face detection) [156] achieves both speed and compactness by removing the fully connected (FC) layers from the classification network. It also boasts scale invariance, detecting multi-scale faces in one pass and employing filters on each prediction head to broaden receptive fields and assimilate context information.

To enhance detection accuracy specifically for faces that are partially obscured, FAN [205] incorporates an anchor-level attention mechanism into the RetinaNet [127] framework. DSFD [121] introduces an innovative feature enhancement module combined with an improved anchor matching technique, aiming to boost discernibility and provide a superior starting point for the regressor. SFPD [55] utilizes a joint convolutional neural network backbone that incorporates shared feature maps to provide real-time detection of both faces and humans.

RetinaFace [36] excels in pixel-wise face localization across diverse scales. It combines face bounding box prediction, 2D facial landmark localization, and 3D vertex regression into a unified multi-level face localization task. This integration benefits from mutual feedback among these tasks during training, resulting in a significant accuracy improvement for face detection. Moreover, its lightweight backbone, enables RetinaFace to achieve real-time performance even on a single CPU core.

To conclude, Single Shot Detector models represent a significant advancement in the field of face detection. Their ability to maintain consistent computational efficiency regardless of the number of faces in an image makes them particularly promising for real-world applications. However, continuous innovation is crucial to overcome persistent challenges in detecting faces under diverse and challenging conditions.

2.3.2 Face Tracking

Face tracking, a subset of Multiple Object Tracking (MOT), has garnered substantial attention within the realm of computer vision. In domains such as surveillance or HRI,

where the primary input is often a video stream capturing one or multiple faces, it is imperative to accurately track each face to extract information critical for FR. Realizing the intrinsic connection between tracking and recognition, some methodologies have sought to unify these operations, yielding a more cohesive process [109].

To provide more detail, face tracking revolves around pinpointing a moving face or faces across a temporal span. It is initiated by face detection, which can be either human-assisted or automated. Once detected, the system traces the face's movement across succeeding video frames by evaluating its motion patterns. It is worth distinguishing this from face detection, which merely ascertains the location and dimensions of a face within a single frame. Conversely, face tracking maintains a continuous focus on the same face through numerous frames, preserving its identity [111].

At its core, face tracking is anchored by three distinct methodologies: feature-based, model-based, and learning-based. The feature-based face tracking is rooted in the principle of distinguishing the target (in this case, a face) from its background based on various discriminative features. Recognizing the inherent nature of video sequences, where deviations between consecutive frames are usually slight, this method exploits attributes like points, colors, textures, edges, and shapes. Each of these attributes consistently exhibits fluid motion dynamics across frames, ensuring a stable representation of the target face. Illustrating this, Wei et al.'s research [215] introduced an innovative technique that utilized particle filtering with the mean-shift algorithm tailored for face tracking. By leveraging motion segmentation, this method rectifies estimation errors in the particle filter, especially with non-rigid targets. A subsequent technique by Hwang et al. [93] utilized the mean shift algorithm to pursue face tracking via color distribution, thereby addressing challenges like occlusions, lighting changes, and similar background color distributions. Contreras et al. [27] and Huang et al. [88] also made significant strides in this domain, presenting novel methodologies. However, despite these advancements, the feature-based approach can sometimes grapple with intricacies in motion patterns, leading to suboptimal tracking results.

On the other hand, model-based face tracking achieves its aim via model matching. It encompasses models spanning from one-dimensional line graphs to intricate three-dimensional solid models [5, 26, 81, 126, 194]. Tewari et al. [194] introduced a deep convolutional autoencoder for model-based face tracking that accurately encodes and reconstructs 3D facial features from color images. Despite its popularity, the MeanShift model [26] struggles to adaptively adjust the tracking window size, leading to potential target loss. The CamShift model [5], an extension of MeanShift, incorporates color to refine tracking over continuous image sequences and dynamically modifies its tracking window size. Meanwhile, the Kalman filter predicts target motion based on a Gaussian distribution [126]. Shifting focus to Henriques et al.'s work [81], they formulated the Kernelized Correlation Filter (KCF) algorithm. This algorithm amalgamates correlation filters, multi-channel HOG features, and Gaussian kernel functions. By employing cyclic shift combined with the fast Fourier transform

during classifier training, KCF significantly bolsters realtime tracking performance. While KCF is proficient in tracking targets with deformations, motion blurs, complex backgrounds, and rotations, it grapples with challenges posed by swift motion and scaling changes. Despite the accuracy of model-based methods, the complexity of human faces necessitates extensive computation, impacting realtime performance.

Recently, several deep learning (DL)-based techniques have emerged, treating face tracking as a binary classification task: distinguishing faces from non-faces. Wang et al. [208] utilized a stacked auto-encoder model to learn facial feature variations and incorporated an additional classifier for tracking. Meanwhile, Doulamis et al. [47] offered a multi-layered DL model, updating its parameters dynamically to suit changing conditions. Despite these advances, challenges persist due to uncertainties in face positions, lighting, and occlusions, often resulting in reduced accuracy. In addressing these challenges, the SORT (Simple Online and Realtime Tracking) [10] algorithm presents a pragmatic approach by using a Kalman filter to predict the movement of bounding boxes between frames. Its successor, Deep SORT [220], enhances this by incorporating deep learning features to associate detection responses between frames. This integration assists in managing short-term occlusions or other challenges in the visual domain.

To conclude, achieving a balance between the performance of face tracking algorithms and the demand for low-latency responses remains a pressing concern.

2.3.3 Effect on the Face Recognition Systems

Face detection serves as the primary step in any FR system, paving the way for subsequent stages such as face alignment and representation. The accuracy of the bounding box generated during detection can profoundly influence the efficiency of the subsequent alignment phase. Two primary challenges arise in this context: the potential omission of parts of the facial region and the inclusion of excessive background context within the bounding box. These discrepancies can compromise the subsequent steps in the process. Comprehensive research, as indicated by studies such as [234] and [36], attests to the overarching influence of face detection on face alignment and recognition. Specifically, [234] demonstrates that misaligned bounding boxes can deteriorate the performance of landmark localization. Concurrently, [36] emphasizes the role of a resilient face detector in enhancing FR accuracy.

In essence, the efficacy of face detection deeply linked to the subsequent phases of face alignment and representation. As such, ensuring the precision of face detection is paramount when striving to develop a robust FR system.

2.4 Face Alignment

Upon detection, face alignment becomes the next crucial stage. It is a significant field of study within the realm of computer vision [229]. The process of face alignment plays a vital role in facilitating several advanced facial analysis tasks, including FR [148], expression recognition [123], and facial attributes analysis [264]. Face alignment is a commonly employed technique in FR systems [37, 106] to enhance their resilience against variations in face orientation, position, and scale. The face alignment procedure autonomously determines the precise location of contextual facial landmarks, such as the eyes, nose, and mouth, within a provided facial image or video frame.

The field of face alignment in 2D images boasts a rich history within computer vision. Numerous methods have been introduced to address this challenge, each achieving varying levels of success. For example, holistic approaches leverage the entirety of a facial image to predict landmarks, typically using Active Appearance Models (AAM) [28] or Active Shape Models (ASM) [29] for facial shape recognition.

In contrast, Constrained Local Models (CLM) [30] treat each landmark, or a set of landmarks, as distinct entities, predicting each one based on its localized appearance. These models also implement spatial constraints to yield coherent results. While both holistic and CLM methodologies have proven successful in facial landmark localization within constrained contexts, they face hurdles in unconstrained scenarios. These challenges arise due to various facial appearance variations, such as pose, expression, illumination, image blur, and occlusion.

To address these challenges, the Cascaded Shape Regression (CSR) technique [43, 54], equipped with handcrafted features, has gained prominence. CSR functions iteratively, progressively refining shape or landmark predictions. Each iteration aims to rectify errors from the preceding steps. While CSR-based models [18] can achieve precise facial landmark localization, they are often limited by their reliance on hand-engineered features like SIFT (Scale-Invariant Feature Transform) [142], HOG (Histogram of Oriented Gradients) [33], and LBP (Local Binary Patterns) [2].

With deep learning's advent, CNNs and other advanced architectures, such as auto-encoder [218, 252] and recurrent neural networks (RNN) [231], have been employed in face alignment, demonstrating enhanced efficacy by extracting hierarchical features for robust landmark prediction. From an overall perspective, these deep learning approaches to face alignment can be divided into coordinate regression-based and landmark heatmap-based.

2.4.1 Coordinate Regression-based Methods

Coordinate regression-based models aim to directly predict the coordinates of facial landmarks from the input face image. Leveraging deep architectures allows these methods to abstract rich hierarchical features, which leads to robust landmark predic-

tions even under challenging scenarios. These models, in essence, view face alignment as a regression problem where the objective is to map the input facial features to a set of continuous coordinate values representing landmark positions [59].

To ensure precision, many methodologies follow a coarse-to-fine strategy. Prominently, cascaded regression techniques [144, 184, 252] and RNN [195, 231] are utilized to incrementally refine landmark coordinate predictions. Furthermore, multi-task learning is often harnessed, synergizing landmark localization with associated facial tasks. This includes but is not limited to, face detection [36, 170, 237, 253] and facial attribute recognition [170, 235].

Coordinate Regression-based face alignment techniques can be divided into two primary categories: local-based and global-based models.

Local-based approaches can be visualized as an advanced iteration of shape-index features built on deep learning paradigms. An illustrative example is the mnemonic descent method presented by [195]. This approach efficiently utilizes information across all cascade levels by integrating a memory unit that shares information among them. Consequently, the network can extract features tailored for predicting facial landmarks. However, these methods have a drawback, as the local shape-index features anchored in deep learning are still vulnerable to inaccuracies arising from initial landmark estimates or the predefined mean shape

On the flip side, global-based models take a more encompassing approach. An entire facial region, captured as an image patch, is presented to the regression CNN, which outputs the 2D coordinates of facial landmarks. One of its benefits is that it prevents the need for pre-defined landmark initialization. For instance, Sun et al. [184] laid out a meticulously structured three-level network. The overarching principle here is to process the entire image as an input and yield facial key points as the output. By doing so, they managed to harness both the textural context and inherent geometric constraints in the image, making it conducive for pinpointing each key point. The obtained model remains resilient against challenges such as occlusions, significant face orientations, and extreme lighting variations. In addition to these models, regression methodologies tend towards employing loss functions such as L2 [46, 144]. Although they're known for their precision, they are sensitive to outliers.

In response to these challenges, Rashid et al. [172] used the smooth L1 loss function instead of L2. Feng et al. [53] introduced the "Wing loss" function, specifically designed to enhance the influence of samples with minor to moderate errors. While many of the mentioned techniques focus on facial landmark localization in still images, there is an evident need to harness temporal information across frames when dealing with video-based face landmarks. Addressing this, the Two-Stream Transformer Networks (TSTN) [131] crafted a dual-stream architecture. This design not only identifies the landmark in an individual frame but also maintains temporal consistency across frames for refined results. Additionally, Dong et al. [46] presented a novel unsupervised strategy termed Supervision-by-Registration (SBR). This method capitalizes on the

optical flow consistency of identified landmarks during video data training.

2.4.2 Heatmap Regression-based Methods

Rather than directly predicting landmark coordinates as in coordinate regression, the heatmap regression-based methods predict heatmaps for each landmark instead of the direct x,y coordinates on the input image.

Newell et al. [159] introduced a stacked hourglass (HG) network for human pose estimation, which achieved significant success. Recognizing the similarities between human pose estimation and facial landmark localization, several studies [15, 39, 91, 211, 240] have leveraged the stacked HG network for facial landmark localization, resulting in notable performance improvements. Bulat et al. [15] replaced the basic block with a multi-scale version, enabling the network to capture more information and further enhance its performance.

Various effective architectures have been developed for heatmap regression, as demonstrated in studies such as [34, 45, 151, 206]. For example, DeCaFA [34] integrates stacked U-nets with landmark-wise attention maps to preserve spatial resolution while extracting local information. Another noteworthy approach is the High-Resolution Network (HR-Net) [206], which is specifically tailored to maintain high-resolution representations, offering advantages for tasks involving landmarks.

Wang et al. [211] proposed the adaptive wing loss to address the limitations of wing loss [53], namely the issue of pixel distribution discrepancy between the foreground and background in heatmap regression. This variant adjusts to the ground-truth heatmap pixels, penalizing foreground pixels more than the background. Additionally, PropNet [91] introduces a variation termed "focal wing loss". It adapts the penalty for incorrect predictions and adjusts the loss weight for each sample in every batch during training, thereby addressing data imbalance concerns.

Multiple techniques have been devised to tackle the difficulties presented by specific facial landmarks caused by their unclear definitions or occlusions [22, 115, 139, 226, 272]. For instance, Wu et al. [226] introduced the facial boundary heatmap, which offers a clear representation of facial geometric structure, helping to reduce semantic ambiguities. In their work [227], they advocate for the use of boundary lines as geometric structures for human faces, effectively mitigating the inherent ambiguities associated with facial landmark definitions. Another approach to tackle semantic ambiguities, treated as noisy annotations, is proposed by Liu et al. [139]. They present a probabilistic model to estimate the actual landmark location.

Recent advancements in facial landmark detection have focused on predicting detection reliability [22, 115]. For instance, LUVLi [115] introduces a framework that not only determines the landmark position but also estimates uncertainty and visibility simultaneously. This comprehensive approach utilizes additional information from uncertainty and visibility factors, resulting in improved accuracy across various

datasets.

2.4.3 Effect on the Face Recognition Systems

Face alignment is pivotal in optimizing FR systems. The primary purpose of alignment is to adjust facial images to predefined spatial coordinates using predicted landmarks, ensuring the face representation model learns from an organized layout. However, the alignment process can be compromised if these landmarks are inaccurately predicted, causing the facial image to shift from its optimal layout. Studies conducted by Guo et al. [68] and Deng et al. [36] underscore this concern, noting that poor landmark localization can lead to shift variations. Conversely, recognition accuracy is markedly improved when alignment is robust across different facial poses.

Additionally, the strategy for face alignment, encompassing elements like the count of facial landmarks, the dimensions to which an image is cropped, and the extent of vertical adjustments, plays a significant role in determining FR performance. [236] suggests that a well-conceived alignment approach can enhance recognition across various contexts. It is also worth noting that the right degree of spatial transformation is crucial during alignment. Insufficient and excessive adjustments can introduce potential disturbances, as findings in [214] indicate.

2.5 Face Recognition

FR, a fundamental task in pattern recognition and machine learning, aims to identify individuals based on their unique facial features. This topic encompasses two primary tasks: face verification and face identification. Face verification is the process of determining if two face images belong to the same person. Face identification, on the other hand, goes farther by recognizing a specific face (known as the probe) within a set of known faces (known as the gallery). Additionally, the process of open-set face identification introduces complexity by initially determining if the face is a member of the gallery.

Despite its significance, FR faces numerous challenges, from low-resolution images and varying poses to complex lighting conditions and motion blur. These complications can significantly degrade recognition accuracy.

The journey of FR, depicted in Figure 2.1, began earnestly in the early '90s with the introduction of the Eigenface method [197]. Early efforts took a holistic approach, attempting low-dimensional representations through techniques like linear sub-space [9, 153] and sparse representation [224, 254]. However, these struggled with unpredictable facial changes.

By the turn of the millennium, there was a noticeable shift to local feature-based recognition. Methods utilizing Gabor [129] and LBP [2] gained prominence, offering

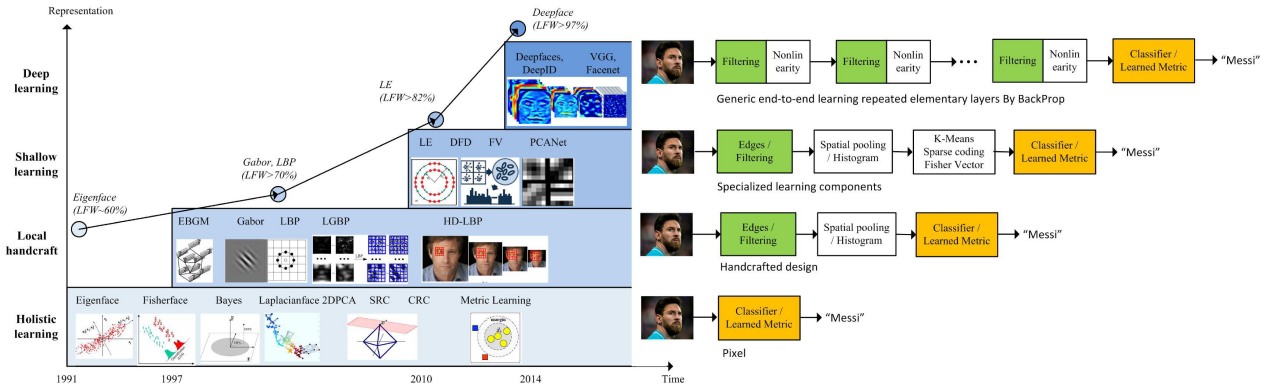


Figure 2.1: Evolution of face recognition techniques [207]. Holistic approaches were prevalent in the 1990s, giving way to handcrafted local descriptors in the early 2000s and local feature learning in the late 2000s. A significant paradigm shift occurred in 2014 with the introduction of DeepFace [191] and DeepID [186].

greater robustness against variances. However, the limitations of these handcrafted features became evident in their precision.

The 2010s marked another paradigm shift, this time towards learning-based local descriptors [19, 20, 119]. The focus shifted from solely feature extraction to refining local filters and building more effective encoding codebooks for better compactness. Despite these advances, challenges persisted, especially with complex facial variations.

In summary, while traditional methods have incrementally improved accuracy, they have often proven inadequate when confronting the wide range of facial changes encountered in real-world scenarios.

In 2012, the advent of AlexNet's victory in the ImageNet competition marked a pivotal moment in computer vision, highlighting the capability of deep learning [114]. By 2014, the surge of deep convolutional neural networks changed the dynamics of the FR domain. DeepFace [191], utilizing a 9-layer CNN on a vast dataset of 4 million facial images, achieved state-of-the-art (SOTA) performance on the LFW benchmark [90] and, for the first time, rivaled human-level recognition in uncontrolled conditions. This breakthrough catalyzed a comprehensive shift towards deep-learning research in FR [177, 185–187, 217].

In the following sections, we provide a comprehensive review of deep face representation learning methods, focusing on two primary dimensions: network architectures and discriminative loss functions.

2.5.1 Network Architecture

The evolution of network architectures in deep FR has largely paralleled the strides made in deep object classification. From the inception of AlexNet to the rise of SENet, the transformation has been rapid and remarkable. Figure 2.2 chronologically illustrates the seminal architectures that have influenced both deep object classification and FR, emphasizing their symbiotic progression.

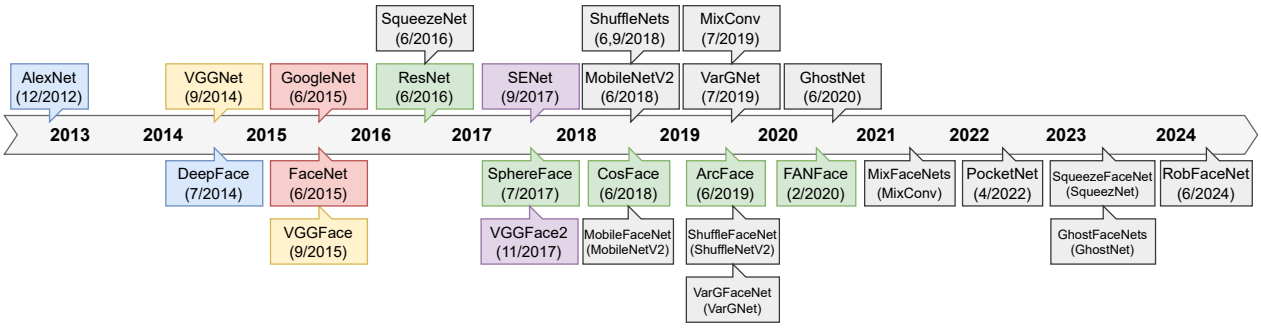


Figure 2.2: Evolution of network architectures in object classification and face recognition (FR). The top row displays common network architectures used in object classification, while the bottom row shows FR models that apply these architectures. FR models utilizing the same architecture are consistently represented by rectangles of the same color.

Initially, deep FR architectures were relatively straightforward with fewer convolutional layers, as seen in VGGFace [164]. The landscape shifted with the advent of GoogleNet [190], a more sophisticated 22-layer architecture. Its inception structure amalgamated multiple feature maps, a concept later adopted by FaceNet [177] for FR.

The Residual Network (ResNet) [79] marked another significant evolution. With its novel residual connections, ResNet enabled the training of much deeper networks, often ranging from 18 to an astounding 152 layers. ResNet soon established itself as an essential tool in many visual tasks, FR being a primary application. The introduction of the Squeeze and Excitation network (SENet) [86] was another leap forward. It seamlessly integrated the SE block, thereby facilitating the automatic weighting of convolution channels. Notably, the SE block amplifies model efficacy without introducing considerable complexity. However, the quest for deeper architectures brings its own challenges, primarily the exorbitant computational demands and increasing memory requirements. Consequently, deploying SOTA deep CNN models in real-time contexts, especially on resource-limited platforms like autonomous vehicles, robots, healthcare devices, and mobile devices, remains a formidable challenge.

In response, research pivoted towards creating efficient deep networks without sacrificing accuracy. This resulted in the emergence of more lightweight architectures, including MobileNets [83, 84, 174], SqueezeNet [94], ShuffleNets [145, 259], CondenseNet [89], EfficientNet [192], VarGNet [256], Ghostnet [74] and MobileOne [199]. These networks reduced memory and computational demands via strategies like convolution factorization, bottleneck convolution introduction, and parameter adjustments. Simultaneously, there is an increasing focus on improving existing networks by employing methods such as compressing pre-trained networks, training of smaller networks, or knowledge distillation using techniques like Huffman coding, quantization, and pruning [24, 50, 60, 75, 225, 255].

The recent wave in the deep learning arena underscores a momentum towards crafting efficient neural networks expressly for FR, and these networks have demonstrated

impressive accuracy [3, 6, 23, 146, 228, 239]. Without compromising the number of parameters in the model, several researchers are breaking new ground with compact embeddings from large face datasets [239]. Others have adapted proven lightweight mobile architectures, infusing them with significant alterations to enhance both their discriminative power and generalization capabilities in FR tasks [3, 6, 23, 146]. These cutting-edge FR models strike a balance between compactness and computational efficiency, addressing the shortcomings of traditional mobile networks by elevating accuracy in FR tasks.

Presently, the research objective emphasizes the development of innovative blocks to elevate network representation by capitalizing on feature maps. Simultaneously, there is an overarching trend towards engineering lightweight designs that harmonize efficiency and performance, ensuring adaptability even on resource-limited platforms.

2.5.2 Discriminative Loss Functions

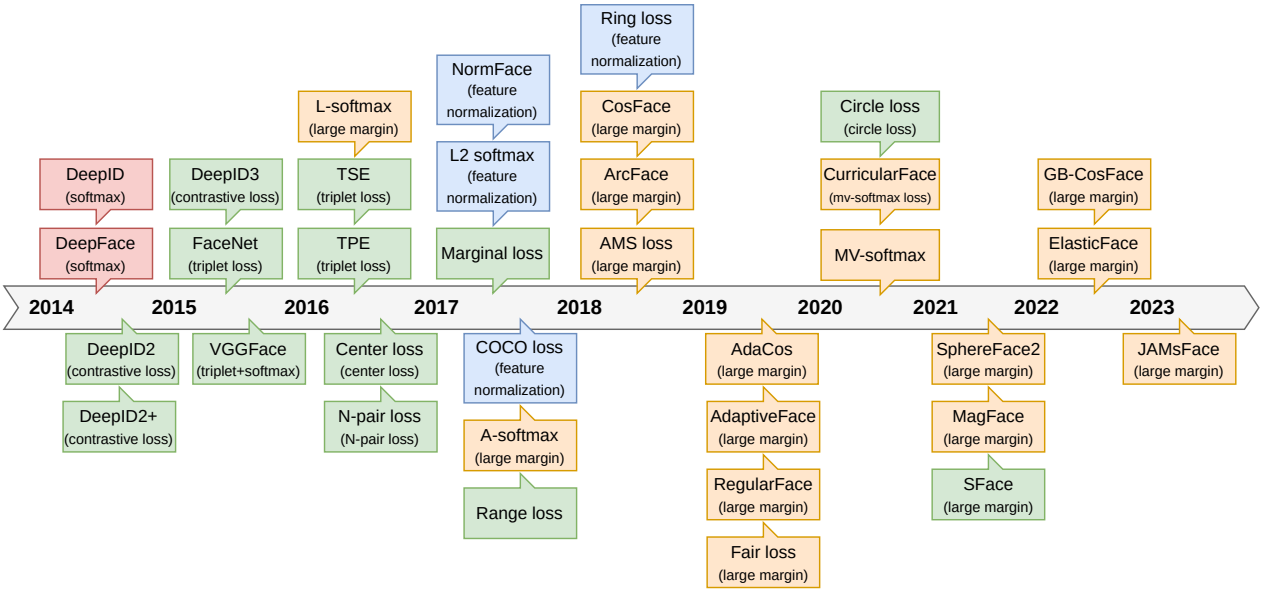


Figure 2.3: Timeline of Loss Function Evolution in Deep Face Recognition. Beginning in 2014 with the introduction of DeepFace [191] and DeepID [186], we observed the dominant role of Euclidean-distance-based losses, including contractive loss, triplet loss, and center loss. 2016 and 2017 marked the rise of large-margin feature learning, prominently with L-softmax [137] and A-softmax [136]. Subsequently, 2017 also saw a surge in performance using feature and weight normalization, leading to innovations in softmax variations. The figure's color-coded rectangles categorize these advancements: deep methods employing softmax (red), Euclidean-distance-based losses (green), angular/cosine-margin-based losses (yellow), and modified softmax approaches (blue).

Despite the improvements in accuracy achieved through CNNs, one of the critical components determining their effectiveness remains the choice of the loss function. Particularly in FR, loss functions serve a dual purpose: they guide the optimization

of network weights and crucially shape the discriminative capability of the extracted features. This ensures that facial features from the same identity are clustered closely while distinctly separating those from different identities. Recent advancements in this domain have been marked by introducing specialized loss functions tailored for FR, as illustrated in Fig. 2.3.

2.5.2.1 Euclidean-distance-based Loss

In deep metric learning, the contrastive loss [182, 183, 187, 188, 244] and the triplet loss [41, 134, 164, 175, 177] stand out as pivotal Euclidean-distance-based loss functions.

The contrastive loss operates on pairs of face images. Its primary objective is to decrease the distance between images of the same identity (positive pairs) and increase the distance between images of different identities (negative pairs). The formulation for the contrastive loss is as follows:

$$\mathcal{L}_{contra} = y_{ij}\, max\big(0, \|x_i - x_j\|_2 - \varepsilon^+\big) + (1 - y_{ij})\, max\big(0, \varepsilon^- - \|x_i - x_j\|_2\big), \quad (2.1)$$

where $y_{ij} = 1$ means x_i and x_j are matching samples and $y_{ij} = 0$ means non-matching samples. x is the feature embedding, ε^+ and ε^- are margin parameters of positive pairs and negative pairs, respectively. DeepID2 [182] adeptly merges the softmax loss (employed for face identification) with the contrastive loss (utilized for face verification), creating a more discriminative face embedding suitable for FR tasks. This approach was later enhanced in DeepID2+ [187] and DeepID3 [183]. Nevertheless, a primary challenge in using contrastive loss remains: selecting appropriate margin parameters can be intricate and non-intuitive.

On the other hand, the triplet loss goes a step further by considering triplets of images: an anchor, a positive sample (same identity as the anchor), and a negative sample (different identity from the anchor). The objective is to ensure that the anchor is closer to the positive sample than the negative one in the embedding space. The triplet loss is formulated as:

$$\mathcal{L}_{triplet} = \|x_i^a - x_i^p\|_2^2 - \|x_i^a - x_i^n\|_2^2 + \varepsilon, \quad (2.2)$$

where x_i^a, x_i^p and x_i^n denote the anchor, positive and negative samples, respectively. x is the feature embedding and ε is the margin.

Google introduced the triplet loss in FaceNet [177] to train feature embeddings within a Euclidean space. They later refined this approach with various triplet selection strategies to enhance performance. Building on FaceNet's foundation, [175] and [176] developed a linear projection for constructing the triplet loss. Additionally, [41, 134, 267] and [31] have combined the strengths of triplet loss and softmax loss. Typically, this involves initially training networks with softmax and then fine-tuning them using triplet loss to achieve improved convergence.

The contrastive and triplet losses can occasionally face high training complexity and instability issues, primarily arising from the selection of effective training samples. In response to these issues, researchers have sought simpler and more direct alternatives. One notable approach is the Center loss [217]. Its primary purpose is to identify a central point for each identity and subsequently impose penalties for any deviations of the features from their corresponding class centers. It can be expressed as:

$$\mathcal{L}_{center} = \|x_i^a - x_i^p\|_2^2 - \|x_i^a - x_i^n\|_2^2 + \varepsilon, \tag{2.3}$$

where c_{y_i} represents the center of deep features for the y_i-th class, while x_i refers to the i-th deep feature that belongs to the y_i-th class. Building on the concept introduced by the center loss, various similar auxiliary loss functions [40, 230, 260] have been developed to address specific challenges. For instance, the Range loss [260] focuses on the harmonic mean of samples with the most extensive intra-class range, addressing issues related to long-tailed data. To stabilize the training of centers, the center-invariant loss [230] was introduced, penalizing discrepancies between centers. Deng et al. [40] devised a margin loss targeting the most distant intra-class and the closest inter-class samples. However, a common challenge with the center loss and its variants is the introduction of numerous additional parameters. This can lead to a substantial increase in GPU memory usage, particularly as the count of identity labels expands during the training process.

2.5.2.2 Margin-based Loss

FR can be viewed as a multi-class classification problem in the training phase. Each class represents an identity, with multiple face samples corresponding to the same identity. A prevalent approach for classification tasks is the categorical cross-entropy loss, commonly coupled with a softmax activation, known as the Softmax loss. This loss is mathematically expressed as:

$$\mathcal{L}_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log P_i = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}. \tag{2.4}$$

Here, the primary goal of the training process is to optimize the probability variable P_i, which represents the projected probability that the embedded feature x_i corresponds to the true classification. Within the training dataset, the aggregate count of classes is denoted as n, the dimension of the embedding feature is represented as d, and the batch size is denoted as N. The i-th training sample contains an embedded feature denoted as $x_i \in \mathbb{R}^d$, which corresponds to the class y_i. The weight $W \in \mathbb{R}^{d \times n}$ has its j-th column represented by $W_j \in \mathbb{R}^d$. The bias is denoted as b_j.

Building on the Softmax (Eq.2.4), NormFace [202] and COCO loss [138] emphasized the importance of normalization, employing L_2 normalization constraints on both

feature vectors and weights, while omitting the bias term b_j. The insertion of a scaling factor s effectively re-scales the cosine similarity, transforming Eq.2.4 to:

$$\mathcal{L}_{norm} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\,cos(\theta_{y_i})}}{e^{s\,cos(\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{n}e^{s\,cos\,\theta_j}}, \tag{2.5}$$

where θ_j signifies the angle between weight W_j and feature x_i.

To enhance both intra-class compactness and inter-class distinction, L-softmax [137] introduced an angular multiplier $m > 1$ to the cosine term, expanding the decision boundary. Additionally, it utilizes a piece-wise function to address the non-monotonic nature of the cosine function. On the other hand, A-softmax [136] normalized weight vectors $|W| = 1$ by L2 norm, ensuring they are situated on a hypersphere, thus addressing boundary imbalances between classes. The A-softmax loss (SphereFace) can be written as:

$$L_{sphereface} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\,cos\,(m\theta_{y_i})}}{e^{s\,cos(m\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{n}e^{s\,cos\,\theta_j}}. \tag{2.6}$$

While L-softmax [137] and Sphereface [136] pioneered the integration of margin into the softmax loss, they employed an integer-based multiplicative angular margin. This approach made the target logit curve steeper, posing challenges for model convergence. Recognizing this challenge, researchers sought alternative methods to stabilize the training process. CosFace [204] built upon the foundations laid by SphereFace [136], emphasizing the use of cosine similarity over angular losses by introducing an additive cosine margin. Meanwhile, ArcFace [37] proposed a novel approach by incorporating an additive angular margin, targeting the dual objectives of maximizing intra-class similarity while also ensuring inter-class diversity.

Pushing the boundaries further, AdaptiveFace [132] and Fair loss [128] introduced class-wise adaptive margins during training to handle the long-tailed data distribution. Innovations continued with the emergence of ring loss [265], AdaCos [261], MagFace [150], Anchorface [133], and KappaFace [162].

Contrasting with deep metric learning methodologies such as triplet loss [177], margin-based softmax techniques globally compare class centers, but at the expense of increased memory consumption. The comparison of samples to classes proves to be more efficient and stable because the number of classes is significantly less than the sample count, and each class can be represented by a dynamically updated center vector. Consequently, margin-based softmax methods have garnered more attention than their metric-learning counterparts in addressing large-scale real-world FR challenges.

2.6 Evaluation Metrics and Datasets

2.6.1 Metrics

FR is often benchmarked across two primary tasks: verification and identification. Both tasks employ distinct metrics and rely on gallery and probe samples. The gallery encompasses registered faces with known identities, whereas the probe set includes faces awaiting recognition.

Before delving into the specific metrics, we need to comprehend a few basic terms. The recognition system determines the similarity between a probe face and a gallery face, matching them if their similarity surpasses a predefined threshold. TA, TR, FA, and FR emerge as fundamental concepts based on this threshold [62–64].

- **True Acceptance (TA):** When a probe and a gallery face of the same identity have a similarity surpassing the threshold.
- **False Rejection (FR):** When a probe and a gallery face of the same identity do not meet the threshold similarity.
- **True Rejection (TR):** When a probe and a gallery face of different identities have a similarity below the threshold.
- **False Acceptance (FA):** When a probe and a gallery face of different identities surpass the threshold similarity.

Verification Task: In most biometric access control systems, face verification is a common application where a user presents their face and claims a specific identity. The system's task is to confirm or deny this claim, making the verification process a one-to-one face-matching task. In this scenario, assessing the system's performance relies on metrics like the False Accept Rate (FAR) and True Accept Rate (TAR). By adjusting the similarity threshold, different TAR and FAR values can be obtained, leading to the creation of a Receiver Operating Characteristic (ROC) curve. This curve depicts TAR against FAR at various thresholds, and its Area Under the Curve (AUC) serves as a comprehensive performance indicator. Generally, a higher AUC suggests better verification performance.

- **False Accept Rate (FAR):** This metric evaluates the frequency with which an impostor, someone not enrolled in the system, gets incorrectly verified. Specifically, FAR is computed as $\frac{FA}{FA+TR}$.
- **True Accept Rate (TAR):** Conversely, TAR measures how often genuine users, those enrolled, get rightly accepted. It is mathematically expressed as $\frac{TA}{FA+TR}$.

Identification Task: While verification focuses on a one-to-one match, identification involves a one-to-N matching scenario. In identification, a probe face is compared against every face in the gallery to determine a match or determine its absence. This one-to-N face matching process encompasses two main sub-tasks: open-set and closed-set identification.

In the **open-set identification**, the probe face might not align with any gallery identity. Two primary metrics emerge for open-set identification: TPIR and FPIR. We obtain an ROC curve derived from TPIR vs. FPIR points by adjusting the threshold with a fixed rank. This curve evaluates open-set identification efficacy.

- **True Positive Identification Rate (TPIR):** Within this task, when a probe face does have a corresponding gallery identity (mate probe), the TPIR quantifies the frequency with which these mates get correctly matched.

- **False Positive Identification Rate (FPIR):** On the flip side, when a probe does not have a corresponding gallery identity (non-mate probe), FPIR measures how often such probes get incorrectly matched to any gallery face.

Conversely, in **closed-set identification**, every probe face has a specific identity in the gallery, making it more constrained than open-set identification. The **Cumulative Matching Characteristic (CMC)** curve emerges as the evaluation metric. This curve visualizes the ratio of accurately identified probe faces (identification rate) against rank. Essentially, the CMC curve conveys the percentage of successful matches within a specified rank, with the rank-one identification rate frequently cited as the primary performance indicator. Notably, CMC becomes a TPIR subset when thresholds are disregarded.

2.6.2 Datasets

Datasets play a pivotal role in FR research, determining the robustness and reliability of any developed algorithm. A comprehensive overview of major training and testing datasets widely adopted in the FR field is presented in this section.

Training Datasets: Modern deep FR mandates the availability of an expansive and precise training dataset. For instance, DeepID models [182, 185–187] were trained using CelebFace [141], encompassing 0.2M images from 10K individuals. Initial methods for deep FR often relied on proprietary large-scale datasets. Facebook's DeepFace [191] was trained using 4M images of 4K individuals, while Google's FaceNet [177] utilized a vast dataset of 200M images from 3M people. Though both tech giants achieved revolutionary outcomes, replicating such models using publicly accessible datasets remains challenging.

Recognizing this limitation, CASIA [244] released a public dataset sourced from the web, comprising 0.5M images of 10K celebrities. Its relatively moderate size, paired with a balanced distribution, has made it a popular choice for fair academic comparisons. Concurrently, UMDFace [7] gathered 367,888 face images from 8,277 subjects paired with human-annotated bounding boxes. This dataset also includes estimated poses, twenty-one keypoint locations, and gender details, all discerned by a pre-trained neural network. In comparison, VGGFace2 [17] offers a more expansive dataset with a similar subject count but more images for each subject, pushing the model to tackle diverse intra-class discrepancies such as varying lighting, aging, and postures.

However, datasets like CASIA, UMDFace, and VGGFace2 primarily feature around 10K identities, failing to showcase the potential of cutting-edge deep learning techniques. Recently, databases like MS1M [71], Celeb-500K [16], and MegaFace2 [158] have emerged, championing the significance of extensive training datasets. Despite their expansive scope, these datasets often grapple with high noise ratios and a long tail distribution. Notably, the noise rates for MS1M and Celeb-500K hover around 50% [201], while MegaFace2 surpasses 30%.

Highlighting the significant impact of noisy labels on accuracy, Wang et al. [201] introduced the IMDB-Face dataset, a curated collection of 1.7M images from 59K celebrities. Nevertheless, refining this dataset required a concerted effort from 50 annotators over a month, underscoring the challenges of curating vast, clean datasets. To mitigate noise, Deng et al. [37] utilized an automated cleaning approach, releasing an improved MS1M dataset and a novel Asian celebrity dataset.

Testing Datasets: Beyond training datasets, testing datasets serve as instrumental tools in measuring the performance and efficacy of FR models. Over the years, several benchmarks have gained prominence in the research community. LFW [90], for instance, is revered for evaluating uncontrolled face verification capabilities. In contrast, datasets like CFP-FP [178] and CPLFW [262] are tailored for pose-invariant face verification. AgeDB-30 [154], and CALFW [263] for age-invariance face verification.

In addition to specialized datasets, the research landscape has seen the emergence of large-scale testing datasets, facilitating evaluations on an unprecedented scale, often involving multi-million face pair comparisons. For example, the MegaFace dataset [104] comprises 1 million images, with the gallery set representing 690,000 unique individuals, and 100,000 photos from 530 unique individuals sourced from the FaceScrub dataset [160] serving as the probe set. The IJB-B dataset [219] includes 1,845 subjects with 21,800 still images and 55,000 frames extracted from 7,011 videos. Expanding on IJB-B, the IJB-C dataset [149] incorporates 1,661 new subjects, totaling 3,531 individuals, with 31,300 still images and 117,500 frames sourced from 11,779 videos.

The Trillion-Pairs dataset [38] has significantly broadened the evaluation spectrum by providing a vast array of image pairings for thorough testing. It includes 1.58

million images sourced from Flickr as the gallery set and 274,000 images from 5,700 LFW [90] identities as the probe set. To highlight its scale, the Trillion-Pairs dataset facilitates a total of 0.4 trillion pair evaluations.

In addition to image-based FR, the YouTube Faces (YTF) dataset [221] stands out for video-based recognition. It comprises 3,425 videos sourced from YouTube, featuring 1,595 subjects, thereby offering a unique dimension for evaluations.

2.7 Challenges and Limitations in Face Recognition

FR, while making significant strides over the years, still encounters challenges that limit its efficacy across diverse real-world scenarios. Understanding these challenges is pivotal in the quest for advancements in the field.

- **Variability in Lighting Conditions:** Dramatic changes in illumination can severely affect the performance of FR systems. Shadows can distort facial features, and excessive lighting can lead to overexposure. Though some modern systems have robustness against lighting variations, they remain a challenge in uncontrolled environments.
- **Facial Occlusions:** Objects such as glasses, hats, scarves, or even hands can occlude parts of the face. Additionally, facial hair, like beards or mustaches, can introduce variability. These occlusions can impede accurate face detection and recognition.
- **Pose Variations:** Faces can be presented in multiple orientations as frontal, profile, or tilted. Recognizing faces in extreme profiles or at acute angles remains a challenging task for many systems.
- **Expression Variability:** Facial expressions can significantly alter the appearance of facial features. A broad smile or a scowl can look dramatically different from a neutral face, adding complexity to the recognition process.
- **Aging Effects:** The human face undergoes changes over time due to aging. These changes can affect the shape, texture, and overall appearance, posing challenges for systems trained on younger versions of faces.
- **Low Resolution and Quality:** Often, surveillance cameras or certain sensors capture low-resolution images, making it difficult to discern finer facial details. Image blur, noise, or compression artifacts further compound this issue.
- **Cross-Dataset and Domain Adaptability:** Models trained on one dataset might struggle when deployed on data from a different source. This discrepancy underscores the need for models that can generalize well across various datasets and real-world scenarios.

- **Ethical and Bias Concerns:** Recent discussions in the community highlight potential biases in FR models, where performance might vary across different ethnicities, genders, or age groups. Addressing these biases is crucial for the fair and ethical application of the technology.

2.8 Summary

While FR has achieved remarkable accuracy in controlled settings, these challenges highlight the complexities of real-world applications. Addressing these challenges paves the way for more robust and universally applicable systems, a focus of ongoing research and the subsequent chapters of this thesis.

3 Face Recognition and Tracking Framework for Human-Robot Interaction

In the dynamic field of human-robot Interaction (HRI), where robotics and artificial intelligence converge, there is a critical need to precisely assess the adaptability and effectiveness of current state-of-the-art (SOTA) face recognition (FR) models in the real-world HRI settings. Our research is driven by the objective of ensuring robots can accurately and promptly recognize and interpret human facial features, an essential component of seamless human-robot coexistence.

Our research begins with the design and implementation of a comprehensive FR framework tailored specifically for real-world HRI environments. This initial phase serves as a platform to evaluate the performance and limitations of existing SOTA models in practical HRI scenarios. The insights and findings from this empirical analysis will pave the way for developing innovative approaches to address the identified challenges.

The goal of this research is to enhance human-robot interaction by improving face recognition in dynamic, real-world settings. To achieve this, we present a face recognition (FR) framework [106] that integrates state-of-the-art techniques across all critical stages of face recognition: detection, alignment, and feature extraction. To ensure the effectiveness and robustness of these techniques, particularly when integrated within our framework, we conduct an empirical study within a real-world human-robot interaction context. This study not only evaluates the performance of current state-of-the-art approaches but also uncovers novel insights that advance the application of FR and tracking in HRI. Specifically, we demonstrate how integrating continuous face tracking into real-time FR systems can enhance accuracy under challenging conditions.

3.1 Introduction

The intersection between practical HRI and the theoretical objectives of achieving efficient FR and tracking opens a complex and exploratory quest. Our inspiration evolves within varied and complex real-world contexts, including homes, schools, hospitals, and workplaces, where robots are increasingly integrated. Such integration surfaces new challenges in areas such as security, automation, and recognition [246].

Humans possess the remarkable ability to remember and recognize individuals based on facial and speech features, enabling them to interact and cooperate smoothly and safely. To integrate robots into this collaboration and foster a seamless HRI envi-

ronment, we need to equip them with more sophisticated systems, such as face and speech recognition. Fortunately, recent advancements in face detection and recognition (FR) powered by deep neural networks have made it possible for robots to rapidly approach human-level performance and handle various challenging conditions. These conditions include large pose variations, occlusions, challenging lighting settings, and poor-quality images with significant motion blur [36, 37].

However, challenges persist in enabling real-world applications to function effectively in unconstrained environments, such as limitations in computing power and the scarcity of training data for user-specific face identification. Recently, the levels of interaction between humans and robotics have become increasingly complex. In order to obtain a more comprehensive understanding of the key determinants that impact behavior in HRIs, a study was done using the Wizard-of-Oz framework [180]. The objective of this study was to examine the prevalent communication intuitions among individuals who engage in new interpersonal interactions, as seen in Fig. 3.1. Fig. 3.1 illustrates the real-world study setup, capturing the interactions between the subjects and the industrial robot. This visual aids in understanding the complexities of human-robot interaction, including how varying participant positions and lighting conditions affect face recognition performance.

Figure 3.1: Wizard-of-Oz field study [180]. A video summary can be found here: `https://youtu.be/JL409R7YQa0` (accessed on 02 February 2025).

The empirical insights gained from this research facilitated the development of the comprehensive multi-modal robotic system, RoSA (Robot System Assistant) [181]. RoSA was designed to overcome the complexities associated with intuitive and user-centered HRI by integrating multiple interaction modalities, including speech, gesture, object, body, and FR. Utilizing this advanced system, a real-world study on HRI was conducted to assess the effectiveness and robustness of state-of-the-art approaches

within an integrated FR framework. While the theoretical framework and RoSA design provide insights into the operational challenges of human-robot interaction, the experimental findings highlighted a key challenge in face recognition during actual interactions with RoSA. This challenge stemmed from face orientation during interactions with RoSA. Participants frequently deviated from the expected positioning for FR due to the necessity of looking away from the camera to perform tasks effectively, as elaborated in Section 3.3. This undesirable head orientation, specifically, the varying face pitch angle, prevents the overhead camera's ability to capture optimal face poses for the FR module. Additionally, the accuracy of FR was found to be influenced by the viewpoint, with variations across different axes (pitch, yaw, and roll) impacting recognition performance. Notably, pitch rotations were observed to notably compromise recognition accuracy [51], as illustrated in Fig. 3.2.

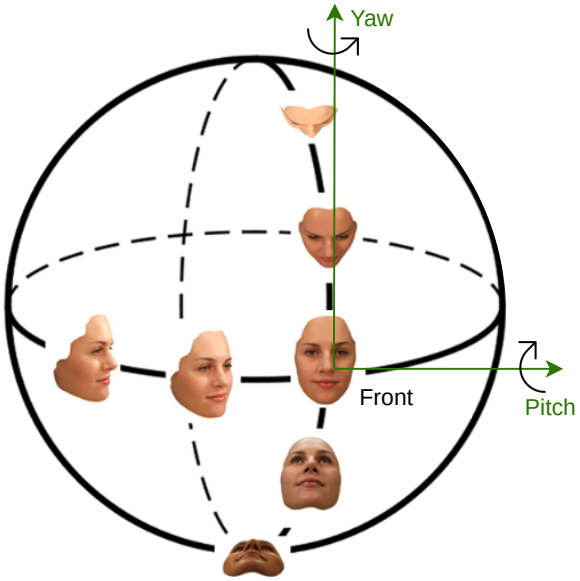


Figure 3.2: The effects of changes in yaw and pitch angles on head pose [106]. Alterations in pitch angle significantly influence facial features, with deviations from the frontal pose leading to reduced distinctiveness in these features.

To improve interaction smoothness and recognition accuracy, we propose an enhanced FR system with tracking capabilities. This enhanced system handles continual variations in subject appearance and lighting conditions. Moreover, it equips the robot with the capacity to acquire new facial characteristics and authenticate them instantaneously, facilitating a more inclusive form of social cognition.

The remainder of this chapter is structured as follows: Section 3.2 surveys the state-of-the-art on face detection, alignment, recognition, and tracking algorithms. The RoSA study is detailed in Section 3.3. Our proposed methodology and framework are detailed in Section 3.4. The empirical aspects, encompassing experiments and their subsequent evaluations, are explored in Section 3.5. In a bid to provide an overall view, potential limitations associated with our approach are discussed in Section 3.6.

Finally, Section 3.7 concludes this chapter, paving the way for subsequent explorations and discussions.

3.2 State of the Art

Most current deep FR systems typically comprise three primary stages: Firstly, face detection, where faces are localized in an input image. This is followed by face alignment, where the detected faces are transformed into a canonical 2D or 3D representation. The final stage is the FR, where the aligned faces are classified into different identities.

Each of these components has been the focal point of various research, resulting in performance metrics that closely mirror human capabilities across a myriad of benchmark datasets [4, 36, 37]. Subsequently, we present a brief overview of the latest advancements associated with each of these stages, emphasizing the merits of the chosen methodologies and their integration into our proposed framework.

3.2.1 Face Detection Algorithms

Face detection algorithms play a critical role in identifying and locating the primary facial region within input images or video sequences, enabling robots to distinguish humans from other entities within a scene.

Historically, before the emergence of deep learning, cascade-based techniques and deformable part models (DPM) led the field of face detection. However, these methods often struggled with unconstrained facial images due to various challenges such as diverse resolutions, lighting conditions, expressions, skin tones, postures, and occlusions [152].

Deep learning techniques, specifically convolutional neural networks, have demonstrated exceptional achievements in the fields of computer vision and pattern recognition in recent times. As a result, numerous face identification methods based on CNN have been suggested in order to address the constraints associated with conventional approaches [36, 55, 58, 135, 156, 173]. The conventional approach generally consists of two distinct phases: firstly, the extraction of features using a CNN-backbone network, and secondly, the prediction of bounding box positions [98]. The majority of detection algorithms comprise either multiple stages or a single stage.

Primarily inspired by Faster R-CNN [173], two-stage algorithms adopt a two-phase approach. During the initial phase, a sliding window is employed to suggest candidate bounding boxes at a specific scale. This is followed by a phase dedicated to eliminating false positives and refining the candidate bounding boxes [157, 250, 253]. While these models offer high precision, they come with increased computational complexity.

On the other hand, single-stage algorithms, largely influenced by the single-shot multi-box detector (SSD) [135], treat object detection as a simple regression problem.

They derive classifications and bounding box coordinates directly from feature maps in a singular stage, without additional proposal stages [36, 156]. While these models offer expedited processing speeds, they may slightly compromise on accuracy.

Among the various single-stage variants, RetinaFace [36] stands out as a top-performing model, setting benchmarks in face detection performance. RetinaFace features a multi-tasking design, predicting bounding boxes, facial landmarks, and 3D face poses concurrently. Leveraging advanced techniques such as deformable convolution, dense regression, and a Pyramid Feature Attention Module, it effectively addresses faces of varied scales. Additionally, the MobileNet-based version of RetinaFace offers lightweight efficiency, making it suitable for real-time performance in FR frameworks. Hence, we integrated this lightweight RetinaFace variant into our system to enhance detection speed.

3.2.2 Facial Landmarks and Face Alignment Algorithms

Face alignment plays a pivotal role in various computer vision tasks, enhancing the robustness of FR against in-plane rotations and pose variations [90]. The core of alignment involves identifying the geometric structure of detected faces and normalizing them to a canonical pose by determining the position and shape of facial elements like the eyes, nose, mouth, and eyebrows. Facial landmarks, central to many alignment algorithms, aid in a similarity transformation to achieve the optimal facial shape, making their localization essential for alignment.

Methods for face alignment can be broadly categorized into model-based and regression-based techniques [229]. Regression-based methods, known for their superior precision, efficiency, and robustness, have outperformed model-based approaches [59]. Model-based approaches often struggle to capture the complex appearance of individual landmarks.

Recent advancements in regression-based strategies are noteworthy. Trigeorgis et al. [195] introduced a unified convolutional recurrent neural network architecture that combines training across stages by incorporating a shared memory unit. Addressing the sensitivity of initialization in face alignment, Valle et al. [198] proposed the Deeply-initialized Coarse-to-Fine Ensemble (DCFE) methodology, refining a CNN-driven initialization phase using an Ensemble of Regression Trees (ERT) to predict landmark location probability maps. In their study, Feng et al. [52] tackled alignment accuracy against diverse facial poses by leveraging a cascade of experts in their Random Cascaded Regression Copse (R-CR-C), employing three parallel cascaded regressions. Further refining this approach, Zhu et al. [271] applied a probabilistic technique to implement coarse-to-fine shape searching.

Deep learning has seen notable advancements in face alignment. Kumar and Chellapa [116] introduced the Pose Conditioned Dendritic Convolution Neural Network (PCD-CNN), fusing a classification network with a subsequent modular one to pin-

point landmark positions with enhanced accuracy. Expanding on this, Wu et al. [116] unveiled a boundary-sensitive face alignment algorithm that interprets facial geometric structure through boundary lines, optimizing landmark positioning.

In recent works, Guo et al. proposed a more efficient compact model called the Practical Facial Landmark Detector (PFLD) [70]. It employs a network branch to estimate geometric information for each face sample, enhancing model robustness. Impressively, the PFLD's compactness ensures a size of 2.1 Mb, delivering over 140 fps per face on mobile devices. Its adeptness in handling complex facial nuances—including variable poses, expressions, lighting, and occlusions—makes it especially suitable for HRI applications. Leveraging these advantages, we have seamlessly integrated PFLD into our proposed framework.

3.2.3 Face Recognition Algorithms

FR systems play a pivotal role in identifying or verifying individuals from images or video frames. Recent advances in machine learning have led to the dominance of deep learning-based FR systems, particularly those leveraging convolutional neural networks [107]. These systems have achieved remarkable accuracy, revolutionizing the field with several innovative models [37, 122, 177, 191, 204, 266].

Central to these models is the concept of face embeddings, unique mathematical representations of facial features. These embeddings are extracted and compared against a database to verify a person's identity. Each embedding is distinct, encoding the features of each human face into a discernible signature.

One seminal work in this domain is DeepFace, proposed by Taigman et al. [191]. This multi-stage methodology, influenced by the AlexNet architecture [114], begins with aligning faces to a generic 3D model, followed by feature extraction using a nine-layer neural network. Utilizing a Siamese network trained via standard cross-entropy loss, DeepFace focuses on face verification tasks.

Building on DeepFace foundation, Sun et al. [187] introduced DeepID2+, an enhanced deep convolutional neural network for FR. It achieved superior results by integrating additional supervisory signals into the earlier layers and expanding the feature representation space.

In a different innovative approach, Schroff et al. presented FaceNet [177], drawing on the GoogleNet architecture [190]. FaceNet uniquely optimizes the embedding space itself, employing a triplet loss function to refine the features extracted by the deep convolutional network, ensuring greater accuracy in capturing facial variances.

The need for optimal feature discrimination has prompted the development of several specialized loss functions [37, 40, 136, 204, 266]. For instance, SphereFace [136] introduced angular margin optimization but faced challenges with training stability. CosFace [204] improved upon this by enforcing a decision margin in cosine space, enhancing performance stability and ease of implementation. ArcFace [37], also known

as Additive Angular Margin Loss, improves discriminative feature learning by emphasizing the angle between facial feature vectors, leading to separated embeddings in the feature space. This approach not only ensures robustness and higher accuracy but also guarantees a stable training process, outperforming many state-of-the-art methods [57, 85, 107]. Given these advantages, ArcFace becomes a natural choice for integration into our proposed framework. Its robust and discriminative feature learning ensures that our system can effectively recognize and differentiate faces even in challenging and dynamic environments typical of HRIs. Moreover, its stable training process guarantees that our model remains reliable and consistent throughout its operational life. The heightened accuracy provided by ArcFace is crucial for real-world applications in HRI, where precision is critical, and errors can have significant consequences.

Recognizing practical computational constraints, especially in real-time applications like HRI, the research community has focused on developing lightweight network architectures [48]. Models like SqueezeNet [94], MobileNets [84], MobileNetV2 [174], and others have emerged as leading lightweight architectures for FR. Among these, MobileFaceNets [23] stand out for their robust performance and efficiency. Inspired by the MobileNets framework, MobileFaceNets employ depthwise separable convolutions to reduce computational cost, making them well-suited for resource-constrained environments. With a compact design and significant accuracy, MobileFaceNets ensure quality FR capabilities even with limited computational resources. Given the advantages offered by MobileFaceNets [23], we have chosen it as the feature extraction backbone for our proposed framework. Its architecture aligns with our goals of ensuring efficient and accurate FR, making it an ideal choice for our system.

3.2.4 Face Tracking Algorithms

Visual object tracking, especially face tracking, has been a focal point in computer vision research due to its numerous applications and inherent challenges. Fundamentally, face tracking involves determining the position of a human face within digital video frames, typically based on the initial detected face location. This task is challenging because faces can vary in pose, view, illumination, occlusion, and posture changes over time, making it a complex problem.

Despite these complexities, face tracking offers various advantages. It enables the enumeration of human faces in video feeds, tracking specific face movements, and offers computational savings by reducing the processing time associated with face detection and recognition.

Numerous visual object tracking techniques have been introduced, including traditional methods like the Kalman filter [87] and template matching [130]. However, a standout approach in this domain is the SORT (Simple Online and Real-time Tracking) algorithm by Bewley et al. [10]. SORT efficiently associates multiple objects

in real-time scenarios using the Kalman filter coupled with the Hungarian method, achieving excellent performance at an impressive frame rate of 260 Hz. Deep-SORT, an enhancement by Wojke et al. [220], integrates appearance information through a trained CNN, further optimizing tracking accuracy and reducing the number of identity switches, even during occlusions.

While deep learning methods in face tracking show promise, the simplicity, efficiency, and robustness of SORT [10] make it an ideal foundation for further innovation. Recognizing this potential, we have built upon the foundation established by SORT to develop our tracking algorithm. However, our innovation doesn't stop there; we have seamlessly integrated this improved tracking algorithm into our FR framework. This integration promises not only enhanced tracking but also optimized FR, bringing together the best of both worlds for superior performance and real-world applicability.

3.3 Human-Robot Interaction Study

While the conceptual framework provides a foundational understanding of the interaction dynamics, the real-world application of these principles is equally crucial. In the following section, we examine the RoSA system [181], a multi-modal system dedicated to facilitating contactless human-machine interactions. Harnessing speech, gesture, and facial recognition, RoSA encapsulates our vision for intuitive and seamless collaborations between humans and machines. This practical implementation highlights the challenges and innovations encountered in deploying our FR framework in a live environment.

At the heart of RoSA effectiveness lies its FR module. The variability of human-machine interaction settings presents inherent challenges, particularly in accurately recognizing faces amidst variable lighting conditions and changing subject orientations and appearances. Viewing these challenges as opportunities for innovation, we enhanced RoSA's FR module by integrating advanced tracking capabilities. This integration is crafted to adapt to continuous changes in a subject's appearance and ambient illumination. Thus, RoSA stands resilient against these variable conditions, ensuring unparalleled recognition accuracy.

To validate RoSA's superiority, we conducted a comprehensive evaluation of our system. Our study involved multiple subjects, assessing the system's user experience and interaction efficiency. Our empirical findings not only affirm RoSA's capabilities but also underscore the efficacy of its FR module, highlighting its pivotal role in the system's overall performance.

In this section, we provide a brief overview of RoSA and our associated study, with particular emphasis on the innovations of the FR module.

3.3.1 Concept

Our vision was to merge disparate information streams, creating a cohesive and synergistic network. The overall structure of this conceptual framework is visually represented in Fig. 3.3. The system architecture comprises seven distinct modules: face, attention, speech, gesture, robot, scene, and cube. These modules, each specializing in processing specific information types, exchange data through middle-ware to a central unit: the interaction module. This pivotal module coordinates the system's logic and drives its actions. Adopting a modular design enhances flexibility, allowing independent development and evaluation of each module.

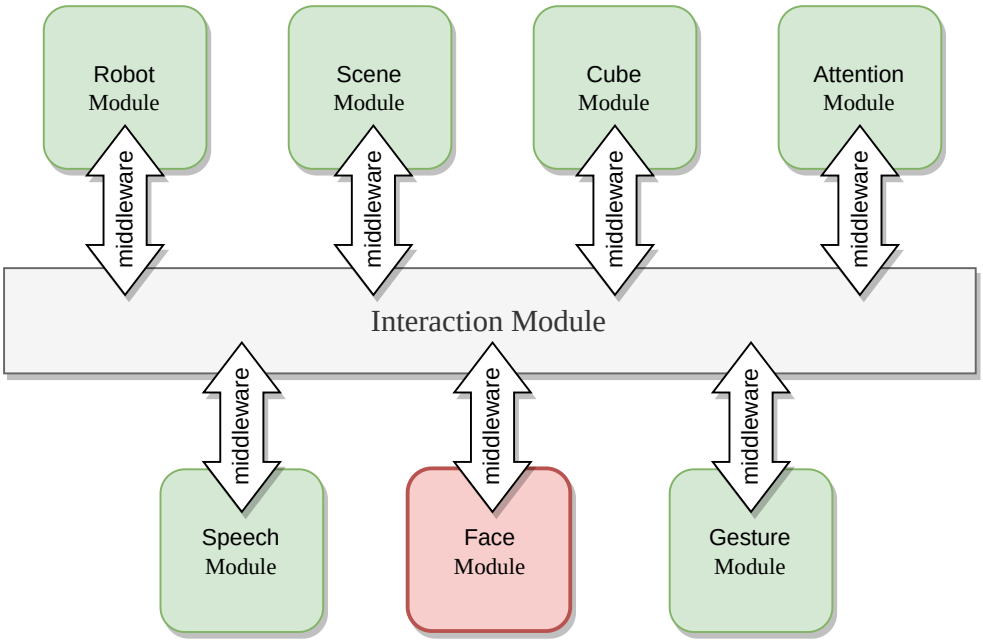


Figure 3.3: Schematic representation of the RoSA conceptual framework. Seven distinct modules intercommunicate through middle-ware to the central interaction module. This modular architecture facilitates synchronized and synergistic human-machine interactions. The interaction module, serving as the system's heart, coordinates the logic and actions based on inputs from the surrounding modules.

3.3.2 Features

In the previous study by Strazdas et al. [180], RoSA functioned as a system that relied on the expertise of the "wizard", an operator controlling the robot, to recognize various human cues, including speech, gestures, facial expressions, body language, and attention. While this model showcased the potential for dynamic human-machine interactions, its operation was primarily dependent on manual control, thus limiting scalability and real-time responsiveness.

In our endeavor to overcome these constraints and propel RoSA into the realm of automation, we integrated the cobot (collaborative robot) with artificial intelligence.

This fusion endowed RoSA with a new array of features aimed at automating its functions, rendering it more self-reliant and adaptable to real-time interactions. Table 3.1 delineates the comprehensive list of these newly incorporated features, while Fig. 3.4 illustrates an exemplary user's interaction with the system, showcasing the detected features in action.

Table 3.1: Extracted feature stream [181].

Stream	Feature	Description	Methods
Face	Face embedding	512 features $\in [0, 1]$	ArcFace [37]
	Facial expression	7 features $\in [0, 1]$	Residual Masking Network [143]
	Face box	4 features for each $\in (x, y)$ (in pixels)	RetinaFace [36]
	Face center	1 features for each $\in (x, y)$ (in pixels)	post processed
	Facial landmark	5 features for each $\in (x, y)$ (in pixels)	RetinaFace [36]
	No. detected faces	1 feature $\in \mathbb{Z}_{>0}$	post processed
	Face Id	1 feature $\in \mathbb{Z}_{>0}$	post processed (cosine similarity)
Head	Head angles	3 features [yaw, pitch, roll] (in degrees)	Img2pose [4]
Gaze	Gaze direction	2 features [yaw, pitch] (in degrees)	Gaze360 [103]
	Attention visual	1 feature $\in 0, 1$	post processed
Speech	Wakeword	1 feature $\in 0, 1$	Piccovoice [166]
	VAD	1 feature $\in 0, 1$	Deepspeech [76]
	Speech-to-text	n features $\in$"spoken text"	WebRCT [61]
	NLU	2 features $\in$ [intent, entity]	RASA [171]
Distance	3D head position	3 features [x, y, z]	post processing using kinect
	Face distance	1 feature (in meter)	post processed
Gesture	Hand Pose	4 Features (Open, Closed, Finger, None)	Kinect SDK
Body	Body Joints	26 Features [x,y,z]	Kinect SDK
Object	Cube Location	4 Features (Letter, Color, Bounding Box, Angle)	CubeDetector [80]

3.3.3 Setup and Workflow

The RoSA framework, depicted in Fig. 3.5, is a network of interconnected components and workstations, each playing a unique role in facilitating seamless HRI.

System Infrastructure. The system comprises two main workstations: Workstation 1 (WS1) and Workstation 2 (WS2). These are integrated with seven dedicated modules, each responsible for different aspects of interaction: face, speech, gesture, attention, robot, cube, and scene. Communications within this complex network are coordinated using the Robot Operating System (ROS) infrastructure, specifically utilizing ROS network and ROS messages for intra- and inter-workstation communication.

WS1: Interaction & Collaboration Hub. WS1 serves as the core of the human-robot interactive experience. At its center is the industrial robot UR5e, equipped with an RG6 gripper, mounted on a robust metal table, ready to execute collaborative

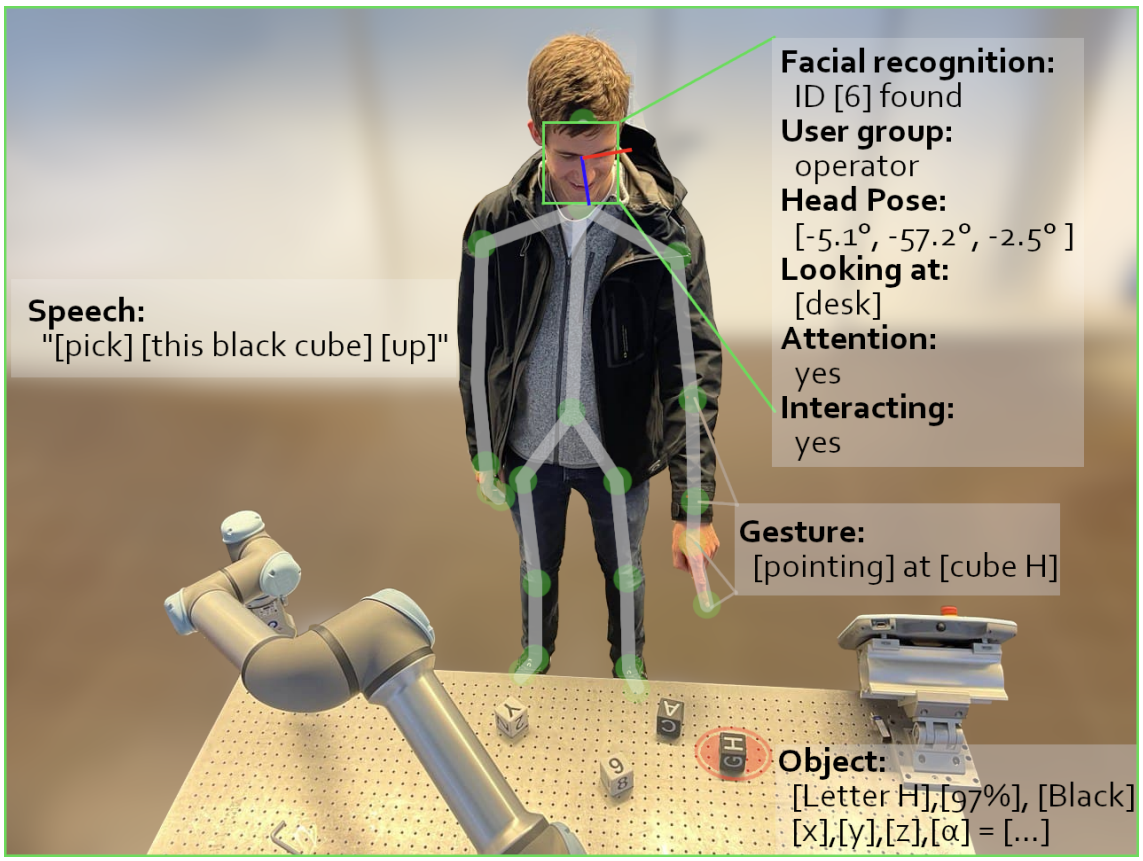


Figure 3.4: User pointing at a letter Cube. Multiple detected features are displayed for clarification [181].

tasks. Monitoring the environment from above is a time-of-flight (ToF) Kinect V2 camera, providing a real-time feed of ongoing activities. Additionally, there is a set of black and white lettered cubes, serving as mission objects and elements that can be manipulated by the robot gripper. For enhanced visualization and feedback, a projector illuminates the cubes and the workspace.

WS2: Registration & Feedback Console. WS2 serves dual purposes: subject registration and feedback acquisition. A smart touchscreen interface, complemented by built-in speakers, aids the registration process. Users input data via a graphic user interface, while the system captures their facial features from both frontal and profile views. This collected information and face embeddings are stored in a dedicated database. After participating in the human-robot collaborative tasks at WS1, users return to WS2, where they provide feedback on their experience via structured questionnaires. RoSA then aids in the collection of additional data appropriate for module assessment and benchmarking.

Face Recognition. The RoSA framework heavily relies on its FR module. Maintaining an active session crucially depends on the face module's capability to consistently

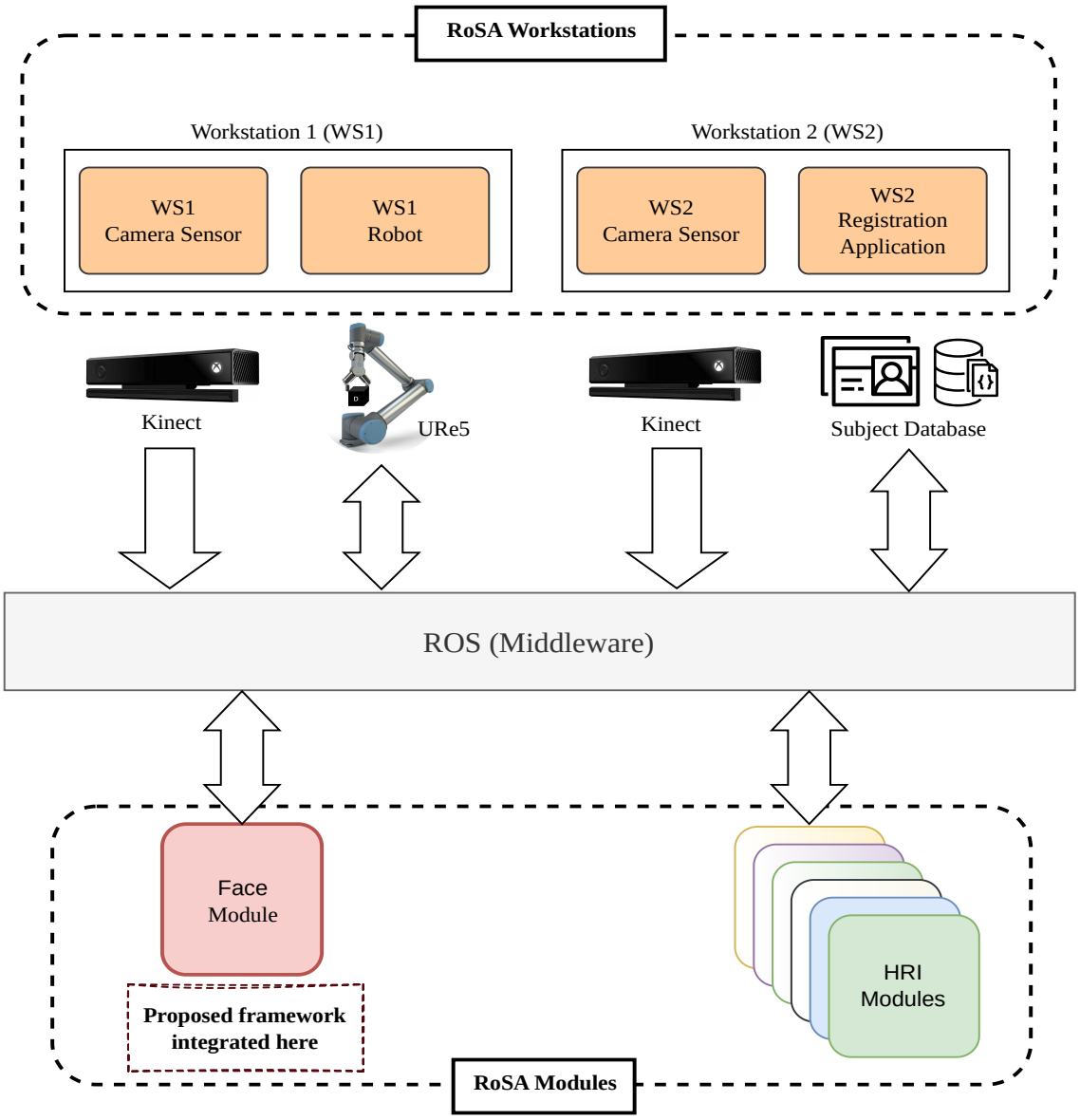


Figure 3.5: Robot System Assistant (RoSA) setup, showing the proposed framework is integrated as the face module.

recognize and track the subject's identity during the experiment. The operational environment presents numerous challenges, from unpredictable lighting shifts to diverse facial orientations and potential occlusions, making consistent FR a significant challenge. Our proposed enhancements aim to improve the FR system, equipping RoSA to adeptly navigate these complexities and deliver unparalleled interaction accuracy.

3.4 Methodology and Proposed Framework

In our proposed framework, we aim to redefine FR by seamlessly integrating advanced tracking algorithms, ensuring real-time adaptability and heightened accuracy. The

system's core begins by processing the current frame through the face detection module, responsible for localizing faces within each video frame. Upon successful detection, a dedicated face tracker is initiated for each identified face, ensuring continuous and smooth tracking across the entire video sequence.

Simultaneously, each detected face undergoes an accurate alignment process. Leveraging detected facial landmarks, these faces are aligned to a canonical face structure, ensuring standardization and enhancing the accuracy of subsequent processes.

Once aligned, the faces proceed to the centerpiece of our framework: the FR module. This module delves deep into each face's unique features, extracting identity embeddings and matching them against a robust database. Upon successful recognition, the identified face's identity is associated with its dedicated face tracker. This associative step is crucial, ensuring not only the face's recognition but also seamless mapping of its movements and interactions within the frame. Concluding the process, these identified identities are seamlessly published to RoSA's other modules, enabling cohesive and unified interaction within the overall system.

A visual representation of this comprehensive approach can be viewed in Fig. 3.6. The framework is structured around three pivotal modules:

1. Face Detection and Alignment
2. Face Recognition
3. Face Tracking

Each of these modules represents a specific aspect of our recognition process, and their collaborative work is central to the system's effectiveness. In the subsequent sections, we will delve into the details of each module.

3.4.1 Face Detection and Alignment

Face Detection with RetinaFace. For the face detection task, we utilize RetinaFace [36], a deep CNN-based face detector that employs a single-shot, multi-level face localization method. RetinaFace stands out for its ability to integrate three diverse face localization tasks: predicting the face box, pinpointing 2D facial landmarks, and regressing 3D vertices. Notably, all regression tasks are executed directly on the image plane.

At the core of RetinaFace's success is its multi-task loss function, formulated in Equation 3.1. This function effectively merges these tasks into a unified objective, with balanced parameters ensuring equal importance for each task. This approach guarantees comprehensive and proficient face detection.

$$\mathcal{L} = \mathcal{L}_{cls}(p_i, p_i^*) + \lambda_1 p_i^* \mathcal{L}_{box}(t_i, t_i^*) + \lambda_2 p_i^* \mathcal{L}_{pts}(l_i, l_i^*) + \lambda_3 p_i^* \mathcal{L}_{mesh}(v_i, v_i^*), \tag{3.1}$$

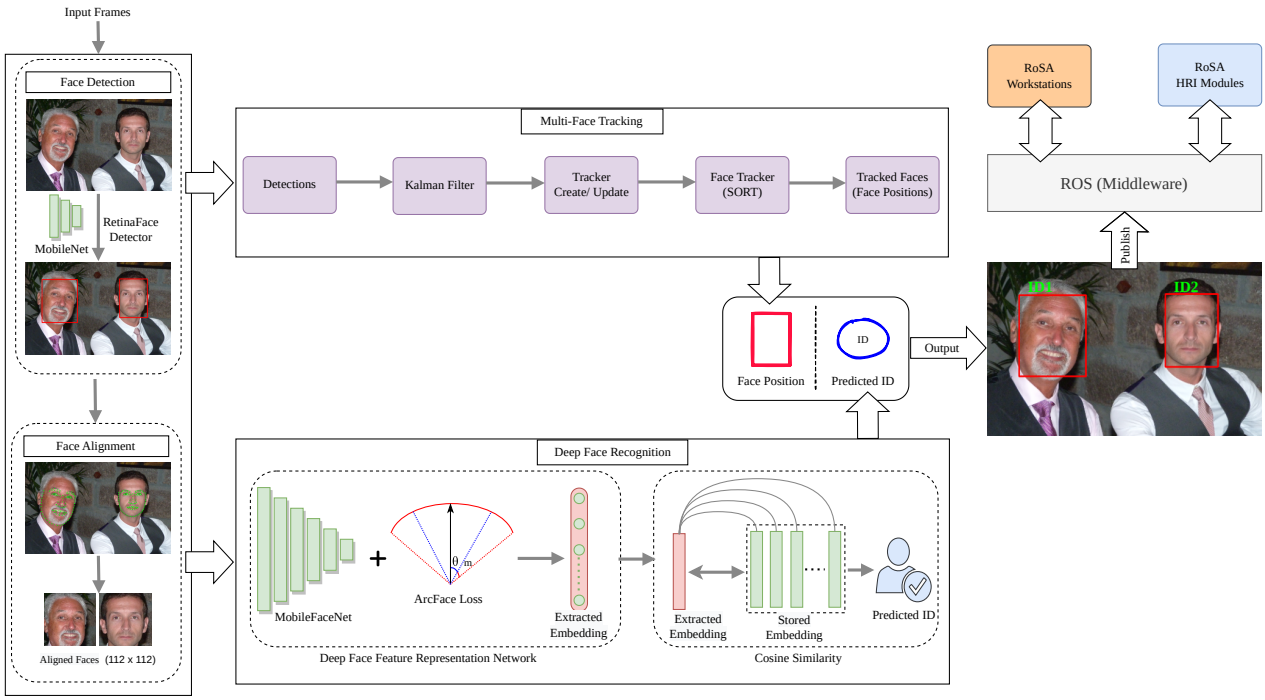


Figure 3.6: The proposed face recognition and tracking framework [106]. The face bounding box and identity predictions are disseminated to the ROS network for the purpose of broadcasting to RoSA workstations and modules.

where t_i, l_i, v_i represent the predictions for box coordinates, five landmarks, and 1,000 vertices, respectively. t_i^*, l_i^*, v_i^* denote the corresponding ground-truth values. p_i is the predicted probability of anchor i being a face, while p_i^* equals 1 for positive anchors and 0 for negative anchors. The classification loss $\mathcal{L}_{cls}$ refers to the softmax loss for binary classes (face/not face). Furthermore, the loss-balancing parameters are set as $\lambda_1 = 0.25$ and $\lambda_2 = 0.1$.

Face Landmarks and Alignment with PFLD. After face detection, the next crucial step is alignment, where we employ the Practical Facial Landmark Detector (PFLD) proposed by Guo et al. [70]. PFLD utilizes a dedicated network branch to estimate geometric information for each detected face, ensuring precise and regularized landmark localization. A notable feature of PFLD is its multi-scale fully connected (MS-FC) layer, which expands the receptive field to capture the global essence of facial structure more effectively.

In order to achieve a harmonious equilibrium between precision and efficiency, PFLD integrates the MobileNet network [84] as its backbone. This integration results in efficient processing speeds exceeding 140 frames per second (fps) for individual face processing on mobile devices, with a compact model size of just 2.1 Mb. Remarkably, despite its efficiency, the system maintains excellent accuracy, handling a wide range of facial complexities, including challenging poses, varied expressions, lighting conditions, and potential occlusions. This integration underscores the framework's ability

to address real-time applications without sacrificing precision.

Operational Flow. In our face detection and alignment module, each frame first undergoes RetinaFace detection. Detected faces are then paired with bounding boxes, five crucial facial landmarks (representing the eyes, nose, and mouth), and a confidence score. To ensure real-time performance, we utilize MobileNet-0.25 [84] as the backbone. This choice strikes a balance between processing speed and performance efficacy, with the compact model achieving real-time speeds of 40 fps on a GPU for 4K images without compromising performance.

Following detection, only faces with high confidence scores proceed to the alignment phase. Here, the lightweight PFLD model aligns the detected faces to a canonical view and crops them to 112×112 pixels, preparing them for subsequent face feature extraction in the FR phase.

3.4.2 Face Recognition

As previously highlighted, our FR approach is rooted in the state-of-the-art additive angular margin loss model, also known as ArcFace, introduced by Deng et al. [37]. ArcFace improves both intra-class compactness and inter-class discrepancy by introducing an additive angular margin penalty m between the face feature x_i and the target weight W_{y_i}. It is defined as follows:

$$L_{arc} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(cos\,(\theta_{y_i}+m))}}{e^{s(cos\,\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\,cos\,\theta_j}}, \tag{3.2}$$

where, n denotes the number of classes in the training dataset, while N denotes the batch size. $x_i \in \mathbb{R}^d$ denotes the deep feature 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 $W \in \mathbb{R}^{d\times n}$. The embedding feature dimension d, the feature scale s, and the angular margin m are set to 512, 64, and 0.5, respectively, as in [37].

The process of extracting facial features involves using aligned and normalized face images with the ArcFace model. The employed backbone extracts the face features, producing the feature vector x_i that encapsulates the unique facial characteristics of the subject. This feature vector is normalized so that it resides on a hypersphere. Recognizing the processing and size challenges posed by the deep backbone network, especially during testing, we effectively employed the lightweight MobileFaceNet network [23] for the feature extraction backbone.

The embedding phase then computes the logit $cos\,\theta_j$ for each class as $W_j^T\,x_i$. Then, the angle between the extracted feature and its corresponding ground truth weight W_{y_i} is calculated as $arccos\,\theta_{y_i}$. An additive angular margin penalty m is introduced to this angle, aiming to decrease angles for the correct class and widen them for incorrect classes. This ensures features are more compact within the same class while being

distinct between different classes. After that, we calculate $cos(\theta_{y_i} + m)$ and multiply all logits by the feature scale s. These logits or the scaled angular embeddings are then transformed via a modified softmax function, turning them into class probabilities that inherently emphasize correct class prediction due to the earlier introduced angular margin.

For real-time recognition tasks, as in our online experiments, the FR module outputs 512-dimensional feature vectors (embeddings). These embeddings are compared with stored features in the database by measuring the cosine similarity [161]. The identity corresponding to the most similar feature is assigned to the recognized face. Finally, this recognized identity is published, and the database is updated with newer or superior-quality face embeddings.

It is essential to highlight the empirical success of ArcFace. A comprehensive ablation study conducted by Deng et al. [37] showcased ArcFace's superiority over 11 other established loss functions, including powerful functions such as Softmax, Center Loss, SphereFace, and CosFace. The benchmarking results on datasets like LFW, CALFW, and CPLFW show ArcFace's effectiveness, with impressive accuracy scores of 99.82%, 95.45%, and 92.08%, respectively. This empirical success underscored our decision to incorporate ArcFace into our FR module.

3.4.3 Improved Face Recognition Using Face Tracking

In HRI applications, the primary challenge of FR is to find the right balance between speed and accuracy. This necessitates the invention of novel solutions. Traditional approaches to face detection and recognition often prove computationally intensive, especially when constantly detecting and recognizing faces in a dynamic environment. To tackle this, our proposal adeptly integrates face tracking using the principles from the Simple Online and Real-Time Tracking (SORT) methodology [10].

SORT primarily employs the Kalman filter to estimate the face current location based on its position in the preceding frame. For instance, once a face is detected in a given frame, let is term it as frame i, the Kalman filter predicts its location in the next frame $(i + 1)$. However, this estimation is an approximation and requires fine-tuning. To optimize this, we utilize the Hungarian algorithm, which excels at both pinpointing the face accurate location and associating it across frames.

One notable shortcoming in FR is the degradation of feature quality when a face deviates from a direct frontal pose. Recognizing this, our methodology pivots from continuous detection to efficient tracking. After the initial detection, each face is assigned a unique tracker. This means that instead of repeated detections in each frame, the position of the face is now tracked. This approach not only saves computational resources but also enhances recognition accuracy, especially for faces in non-frontal poses.

For every new tracker, the system inferred the face embedding and matched it

against existing embeddings through cosine similarity. The outcome of this is the determination of a user's identity (ID) that is subsequently stored in the tracker's metadata. In the frames that follow, the system can instantly fetch the user ID from the tracker, negating the need to redo the recognition procedure, thereby ensuring rapid and effective recognition.

Our advanced face tracking approach, described in Algorithm 1, and visually illustrated in Fig. 3.6, integrates face detection and recognition into a seamless procedure. For every input frame, face detection is performed using RetinaFace as described in Section 3.4. Subsequently, a new tracker is established for each detected face box using the SORT method [10]. The SORT model utilizes both the Kalman filter and the Hungarian algorithm to predict face locations in real-time, leveraging information from the current frame and its predecessor.

After the detection and initialization of the trackers, identification is crucial. Our methodology utilizes the ArcFace model described in Section 3.4 for FR. Once a face is recognized, we assign a unique user ID, linking it directly to the ongoing tracker. This user-centric approach is central to expediting recognition in subsequent frames, ensuring rapid and accurate identification.

But, the process does not end with identification. Continuous verification is essential. With each frame progression, we update the tracker to validate the alignment of each face within its designated tracker boundary to improve the tracking quality. If discrepancies arise and a face is not where it is expected to be, that tracker is terminated. This dual strategy not only optimizes operational efficiency by preventing unbounded growth in the number of trackers but also guarantees that each active tracker, embedded with a unique user ID, remains a key player in rapid FR.

By combining tracking and FR, our method achieves heightened computational efficiency and unmatched recognition accuracy, even for faces at oblique angles.

3.5 Experiments and Analysis

The crucial components of the FR and tracking framework are the face detection and FR models. To thoroughly assess the effectiveness of the proposed tracking approach, we trained and evaluated these two models separately.

3.5.1 Face Detection.

For face detection, aiming to achieve the optimal balance between computational efficiency and accuracy, we trained the RetinaFace model from scratch on the comprehensive WIDER FACE dataset [241]. WIDER FACE dataset comprises 32,203 images and 393,703 face bounding boxes, reflecting significant variability in aspects such as scale, pose, expression, occlusion, and illumination. This dataset is partitioned into training (40%), validation (10%), and testing (50%) subsets by randomly sampling

Algorithm 1: The Proposed Face Tracking Algorithm

Inputs : Video, Detections, KalmanFilter, HeadJoints, SubjectIDs
Output: Recognized Tracked Faces
Initialize $KalmanFilterTracker$;
foreach *frame* $f_i \in Video$ **do**
  $Trackers \leftarrow Predict()$;
  $Trackers \leftarrow Assign(Detections, Trackers)$;
  $TrackersID \leftarrow Attach(Trackers, SubjectIDs)$;
  $TrackersID \leftarrow Assign(TrackersID, HeadJoints)$;
  Update $KalmanFilterTracker$;
  foreach *tracker* $t_i \in TrackersID$ **do**
    $ROS \leftarrow Publish(t_i)$;
  end
end

from 61 scene categories. For validation and testing, three levels of difficulty are defined: Easy, Medium, and Hard, where each successive level incrementally introduces hard samples.

We carefully evaluated the performance of RetinaFace using four diverse neural network architectures: ResNet50 [79], MobileNets [84], MobileNetV3 [83], and GhostNet [74]. The results of these evaluations are detailed in Table 3.2. Notably, the reduced MobileNet backbone with a multiplier of 0.25 stood out, yielding a total average precision of 83% and resulting in a model size of just 1.7 Mb. These findings underscore that the MobileNet0.25 configuration offers the most favorable trade-off between speed and accuracy for real-time face detection in our application.

Table 3.2: Performance comparison of RetinaFace trained on WIDER FACE dataset using different neural network architectures. The evaluation was conducted on images with VGA resolution at 640 × 480. All metrics and processing speeds reported are based on a CPU backend. MobileNet0.25 model achieves the best balance between computational efficiency and accuracy. The models are ordered based on the run time.

Backbone	**Run Time**	**Size**	**Average Precision (%)**		
	(ms)	(MB)	**Easy**	**Medium**	**Hard**
ResNet50	1571	106	95.48	94.04	84.43
MobileNetV3	576	8	92.95	90.73	78.39
GhostNet	403	12	93.35	90.84	76.11
MobileNet0.25	187	1.7	90.70	88.16	73.82

3.5.2 Face Recognition

For training our FR models, we utilize the MS1M-RetinaFace dataset [37]. This dataset is a semi-automatically refined version of the MS1M dataset [71], comprising 5.1M images from 93K identities. Each face image in this dataset has been detected and aligned using five facial landmarks as predicted by RetinaFace [37]. Post alignment, images are resized to dimensions of 112 x 112. These images undergo normalization into the range [1, 1] by subtracting a mean pixel value of 127.5 and are subsequently divided by 128.

To ensure a robust and fair performance assessment, our benchmark models include ResNet50, VarGFaceNet [239], AirFace [124], and ShuffleFaceNet [146]. In addition to these, our selected MobileFaceNet model. All these models are trained from scratch using the MS1M-RetinaFace dataset [37] and the ArcFace loss function [37]. Such a uniform training process promises a fair performance comparison across all the models.

Throughout the training phase, parameters remain consistent across models. The weight decay parameter is set at 4e-5. Leveraging the SGD optimizer with a momentum of 0.9, we set our batch size at 512. Our learning rate starts at 0.1, and we reduce it by a factor of 10 at specified iteration milestones: 36K, 52K, and 58K. The complete training ends at 60K iterations.

To assess the effectiveness of our lightweight face models, we conducted tests using various benchmarks that exemplify the principal characteristics of FR scenarios. These benchmarks include LFW [90], cross-pose datasets (CFP-FP [178] and CPLFW [262]), and cross-age datasets (AgeDB-30 [154] and CALFW [263]). Our findings, as summarized in Table 3.3, indicate that the MobileFaceNets model strikes the optimal balance between computational efficiency and recognition accuracy.

Table 3.3: Performance comparison of ArcFace trained on MS1MV2 dataset using different neural network architectures. The models are ordered based on the number of FLOPs. Results are in % and higher values are better. MobileFaceNets model achieves the best balance between computational efficiency and accuracy.

					Cross-Age		Cross-Pose	
Model	**#FLOPs** (M)	**#Params.** (M)	**Size** (MB)	**LFW** (%)	**CA-LFW** (%)	**AgeDB-30** (%)	**CP-LFW** (%)	**CFP-FP** (%)
ResNet50 [37]	24211	65.2	261.22	99.82	95.45	98.15	92.08	98.40
VarGFaceNet [239]	1022	5	20.0	99.85	95.15	98.15	88.55	98.50
AirFace [124]	1000	4.23	-	99.27	-	93.25	-	94.11
ShuffleFaceNet [146]	577.5	2.60	10.5	99.67	95.05	97.32	88.5	97.26
MobileFaceNets [23]	439.8	0.99	8.2	99.55	95.20	96.07	89.22	96.90

3.5.3 Overall System

The metrics utilized to assess the overall system performance include precision, recall, F-score, and recognition rate. Predictions are categorized into True Positives (TP),

False Positives (FP), False Negatives (FN), and True Negatives (TN). A True Positive in recognition occurs when the model accurately predicts the subject class (i.e., subject ID), matching the ground truth, while a False Positive prediction indicates otherwise.

Similarly, a True Negative is recorded in recognition when the model correctly abstains from predicting a subject not present in the database, while a False Negative occurs when a valid subject is not correctly recognized.

Precision represents the probability of correctly identifying a subject relative to the ground truth identity, calculated as follows:

$$Precision = \frac{TP}{TP + FP}. \tag{3.3}$$

Recall measures the likelihood of correctly recognizing subjects among the ground truth, calculated as follows:

$$Recall = \frac{TP}{TP + FN}. \tag{3.4}$$

The F-score, a harmonic mean of precision and recall, provides insight into the overall performance of the model, calculated as follows:

$$F\text{-}score = \frac{Precision * Recall}{Precision + Recall} * 2. \tag{3.5}$$

The FR rate FR_R indicates the ratio of correctly recognized faces to the total detected or tracked faces, calculated as follows:

$$FR_R = \frac{TP}{Total faces} * 100. \tag{3.6}$$

To evaluate the proposed framework, we conducted two types of evaluations: dataset-based evaluations and online evaluations.

3.5.3.1 Dataset Evaluation

We evaluate the proposed framework using the ChokePoint dataset [222], a comprehensive video dataset specifically designed for experiments on person identification and verification in real-world surveillance scenarios. This dataset comprises videos featuring 25 subjects, including six females and 19 males. It encompasses 48 video sequences and a total of 64,204 face images, capturing variations in illumination, pose, sharpness, and misalignment resulting from automatic face localization and detection.

The experimental analysis assesses the tracking performance across all 25 subjects in the ChokePoint dataset. To demonstrate the refinement in recognition, we compare the proposed FR framework with and without tracker assistance. Table 3.4 presents the average results obtained from these experiments. Additionally, we visualize the Receiver Operating Characteristic (ROC) curve in Fig. 3.7, illustrating how the tracking approach enhances the recognition rate, particularly at high false positive rates,

while simultaneously reducing the false classification rate.

Table 3.4: The average results of precision, recall, and F-score on ChokePoint dataset.

Tracking	Precision	Recall	F-Score
No	0.83	0.79	0.81
Yes	0.96	0.93	0.94

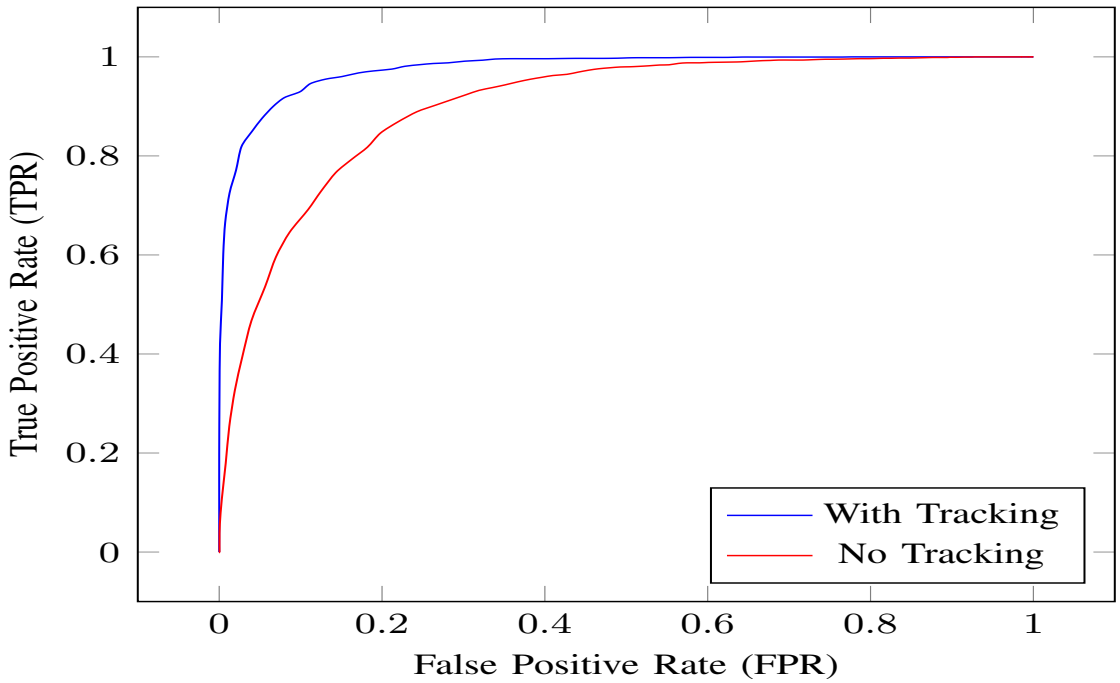


Figure 3.7: ROC Curve of ChokePoint Dataset for the Proposed Framework.

3.5.3.2 Online Evaluation

We apply the proposed framework in a real-time HRI study [181] to further assess its performance and robustness in practical HRI scenarios. During the study, data for evaluation were collected from 11 subjects, including two females and nine males, aged between 20 and 34 years.

The experimental analysis showcases the tracking performance and recognition rate for the 11 subjects during interactions with RoSA [181]. To illustrate the refinement in recognition, we compare the proposed FR framework with and without tracker assistance. With tracking, the proposed framework achieves a FR rate of 94%, whereas without tracking, it achieves 76%. Fig. 3.8 illustrates the impact of tracking on the precision of the proposed framework, while Fig. 3.9 depicts its impact on recall. Additionally, Fig. 3.10 presents the F-score results of the proposed framework with and without tracker assistance.

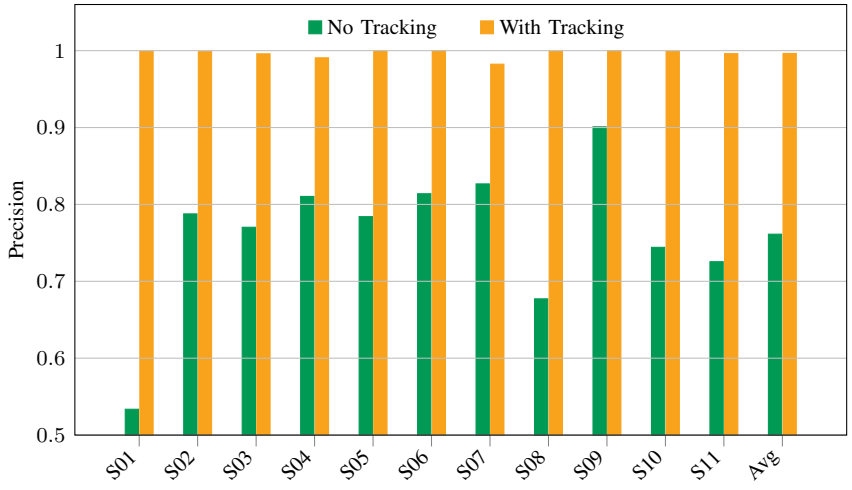


Figure 3.8: Impact of Tracking on Precision of Face Recognition.

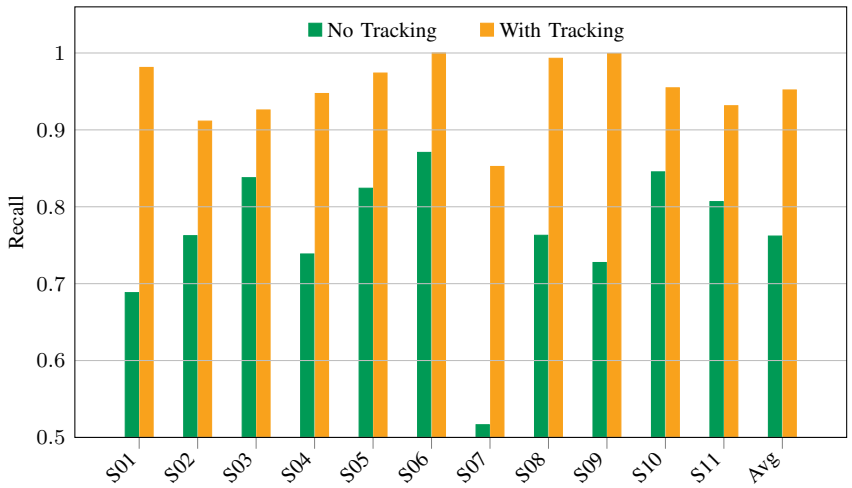


Figure 3.9: Impact of Tracking on Recall of Face Recognition.

The proposed framework exhibits faster processing times compared to the standard FR framework, achieving frame rates ranging from 25 to 40 frames per second (fps). Fig. 3.11 illustrates some results obtained from the proposed framework during real HRI experiments conducted with our RoSA system [181].

To validate the obtained results, we conducted additional experiments using recorded videos from the Wizard-of-Oz study [180], yielding consistent outcomes. This dataset comprises videos featuring 36 subjects performing tasks similar to those in the RoSA study, recorded on various days under different lighting conditions.

For each subject (video), we selected three exemplary face images with distinct poses and added the extracted embeddings to the database for comparison with the faces in the videos. The precision and recall results for 37 subjects, delineated by the top ten outcomes, are presented in Table 3.5.

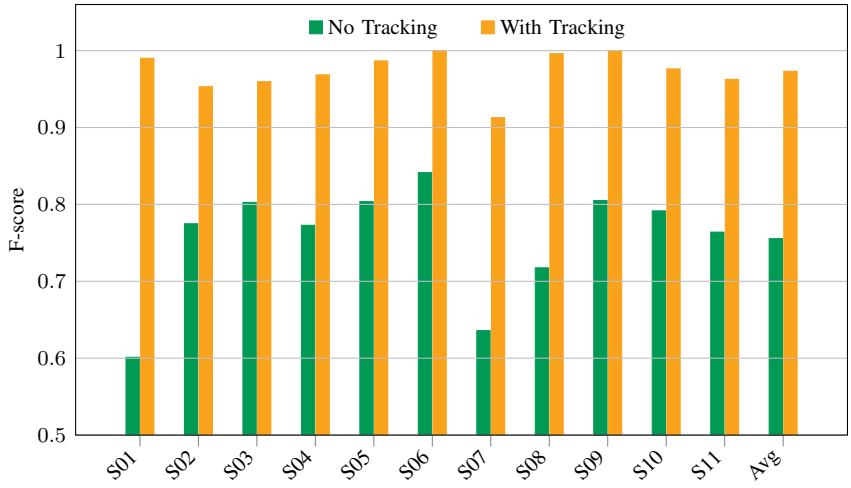


Figure 3.10: F-score results of the proposed framework with tracker-assisted and without tracking.

3.5.4 Computational Efficiency Assessment

Lightweight face networks demonstrate promising results in FR, often performing comparably to state-of-the-art deep face models across various scenarios. For instance, ResNet100-ArcFace, proposed by Deng et al. [37], ranks among the top-performing models in different evaluation scenarios, however demanding significant computational resources. Notably, in the challenging DeepGlint-Image dataset (one of the most challenging databases), ResNet100-ArcFace exhibits an 8% accuracy difference compared to MobileFaceNet, our chosen network. However, in other datasets, this margin is less than 3%. Despite its superior accuracy, ResNet100-ArcFace requires significantly more resources, requiring 19 times more storage space and involving 26 times more FLOPs and 32 times more parameters than MobileFaceNet.

Incorporating face tracking into our framework offers the advantage of avoiding face detection and recognition for every input frame. To enhance the accuracy of our framework and minimize tracking errors, we conduct the entire recognition process once every fifth frame.

To evaluate the computational efficiency of the proposed framework, we tested it on videos collected during both the RoSA study [181] and the Wizard-of-Oz study [180] (total of 47 videos). The average processing time for each FR module was recorded using a hardware setup featuring a NVIDIA GeForce GTX 1080 Ti Desktop GPU (12 GB GDDR5, 3584 CUDA cores). Table 3.6 presents the average execution time of individual methods utilized in the proposed framework. In summary, the average execution time per frame for the entire process is approximately 6.7 milliseconds, with an average frame rate of around 35 frames per second.

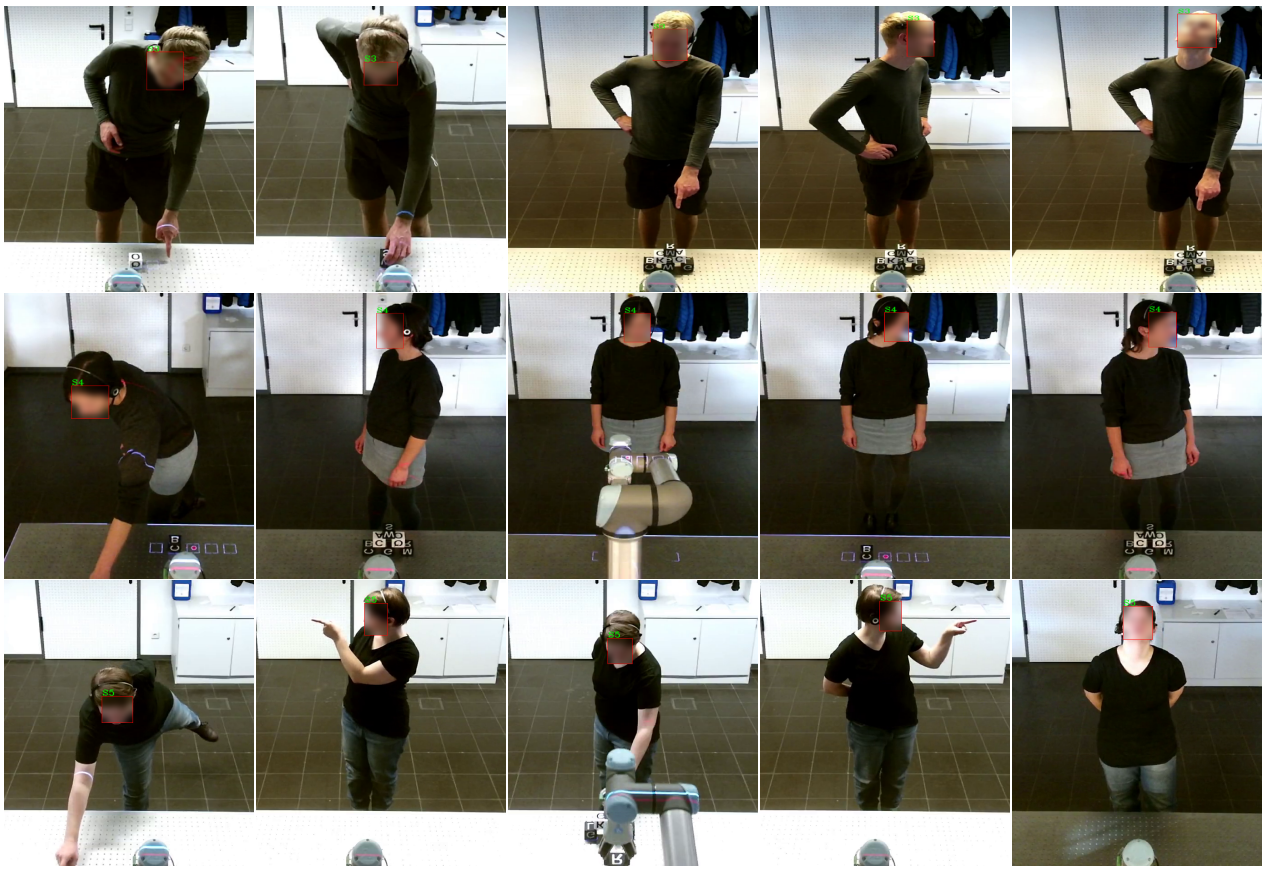


Figure 3.11: Online experimental results, showing the framework's stability in different head pose and lighting conditions.

3.6 Discussions and Limitations

The study conducted in this research entails a complex setting involving two synchronized workstations (WS1 and WS2) interconnected through the robot operating system. Despite encountering challenges such as varying illumination conditions, extreme head pose angles, and occlusions, the framework demonstrated effectiveness in extracting face features and recognizing subjects' identities within a multi-person environment.

Although the framework exhibited promising results, instances of misidentification were observed. These were primarily attributed to incomplete registration processes and suboptimal face feature embeddings lacking sufficient discriminability. Addressing this challenge necessitated certain subjects to undergo re-registration to enhance model recognition accuracy.

One notable benefit of our approach is its utilization of lightweight CNNs for all stages of FR, including face detection, alignment, and feature extraction. This ensures compliance with real-time requirements in HRI systems. Furthermore, the framework effectively identified collaborating participants in various positions, facial expressions, lighting conditions, and even when participants were wearing face masks during the

Table 3.5: Result of precision and recall for the proposed framework.

No	ID	Precision		Recall	
		Tracking	No Tracking	Tracking	No Tracking
1	4	0.97	0.76	0.81	0.70
2	7	1	0.84	0.92	0.59
3	11	1	0.68	1	0.73
4	16	1	1	1	0.66
5	18	0.96	0.88	1	0.81
6	24	0.98	0.65	0.95	0.77
7	25	0.89	0.53	0.92	0.63
8	29	0.98	0.83	0.98	0.79
9	32	0.99	0.68	0.98	0.60
10	36	0.95	0.76	0.97	0.75

Table 3.6: Average execution time of individual methods used in the proposed framework.

Method	Average Time (ms)
Detection	3.2
Alignment	1.4
Tracking	0.8
Recognition (Embedding Inference)	1.3
Identification (Similarity)	0.08
Visualization & Delays	7.5

experiment.

Nevertheless, it is crucial to recognize that while our framework demonstrates strong performance across various scenarios, it may not be entirely optimal. As the number of faces within a scene increases, there is a rise in computational demands, sometimes resulting in system delays that can hinder the fluidity and intuitiveness of HRI.

Moreover, in certain scenarios, particularly with challenging poses, the facial embeddings extracted may lack the robustness required for consistent differentiation between subjects. Overcoming these limitations necessitates progress in two crucial areas. Firstly, the introduction of a robust loss function capable of generating more discriminative features becomes essential. This would ensure that individual features remain distinct even under challenging conditions, thereby facilitating accurate identification. Secondly, refining the current CNN architecture into a more streamlined and efficient design is paramount. This enhancement would enable real-time FR capabilities and mitigate computational strain as the number of subjects in a scene increases. These

two challenges will be addressed in Chapters 4 and 5, respectively.

Through these enhancements, our goal is to further strengthen the framework's capabilities, rendering it more adaptable and efficient for real-world scenarios, particularly in intuitive HRIs.

3.7 Summary

In our pursuit of optimizing HRI dynamics, we introduced a FR system enhanced by advanced face tracking techniques, leveraging deep convolutional neural networks. The inherent challenges posed by real-time HRI, such as the need for fast processing, led us to base our framework on lightweight CNNs. This ensured proficiency across all integral stages of FR, from detection to alignment, tracking, and feature extraction.

Our design approach was further streamlined for HRI system compatibility. By encapsulating our methodology within a modular ROS package, we facilitated its seamless integration into varied HRI systems. Preliminary evaluations suggest that incorporating face tracking along with FR significantly boosts the recognition rate.

At the core of our FR framework lies the SOTA ArcFace loss function, coupled with the RetinaFace detection method and an intuitive real-time face tracker. This comprehensive system adeptly tackles challenges such as varying illumination, diverse head poses, and occlusions.

We further augmented this framework with a tailored face tracker, designed to integrate tracking data with recognized identities. This fusion of tracking and recognition markedly enhances processing speed and recognition accuracy, particularly for faces in wild environments.

To assess the effectiveness of our framework, we conducted real-time tests within our HRI system RoSA, involving 11 participants interacting with the robot to accomplish various tasks. To ensure the reliability of our findings, we extended our evaluations to recorded videos from the Wizard-of-Oz study. This dataset comprises videos of 36 subjects engaging with RoSA, mirroring the tasks and outcomes of our real-time tests. Our results underscored the framework's significant enhancement of FR robustness. On average, the system exhibited a 25% improvement in real-time recognition, achieving precision, recall, and F-score values of 99%, 95%, and 97%, respectively.

However, amidst these achievements, our studies revealed limitations, particularly in complex scenes. As number of faces increased, computational demands escalated, leading to occasional system lags detrimental to HRI fluidity. Additionally, under challenging conditions, facial feature extraction occasionally lacked the robustness required for consistent differentiation. Addressing these challenges is pivotal for the next phase of our research. A focused effort on refining the loss function and CNN architecture aims to optimize the framework for seamless, intuitive HRIs.

4 Towards Adaptive Feature Learning For Face Recognition

The field of large-scale face recognition (FR) has witnessed a significant transformation with the advent of deep feature learning, where margin-based loss functions have played a pivotal role. These functions expand the feature margin between disparate classes, aiming to improve the discriminative capability of the softmax loss. A fundamental assumption in many existing methodologies is the balance among classes, assuming a one static margin that uniformly reduces intra-class variations, thereby treating all classes equally.

However, a thorough examination of numerous real-face datasets reveals a different reality. These datasets often follow a long-tail distribution, where a small subset of identities (the 'head') is over-represented, while a larger number of identities (the 'tail') are under-represented. In such scenarios, a static margin becomes less effective, potentially constraining the discriminative and generalization capabilities of FR models. This imbalance poses unique challenges, especially in learning from the less frequent tail identities.

Conventional margin-based strategies, while concentrating on improving differentiation within cosine or angular spaces, may overlook the nuances inherent in diverse class representations. This underscores the significance of our innovative approach, the Joint Adaptive Margins Loss Function (JAMsFace). Departing from uniform margin assumptions, JAMsFace dynamically calibrates adaptive margins tailored to the distinct characteristics of each class. This adaptive approach effectively addresses the critical challenges posed by long-tail data distributions in FR.

Building upon the framework established in Chapter 3, this chapter presents a detailed geometric analysis of JAMsFace, demonstrating its adaptability and effectiveness through extensive empirical evaluations across several FR benchmarks. The results are compelling: JAMsFace not only holds its ground but often surpasses existing FR losses. Its performance is validated through testing on benchmarks like CPLFW, LFW, CFP-FP, and even in the more challenging environments of IJB-C and IJB-B datasets.

This chapter aims to offer a comprehensive exploration of the transformative potential of adaptive margin methodologies in FR. It specifically addresses the inherent problems posed by long-tail data distributions and underscores the necessity for adaptive, class-specific approaches within the FR landscape.

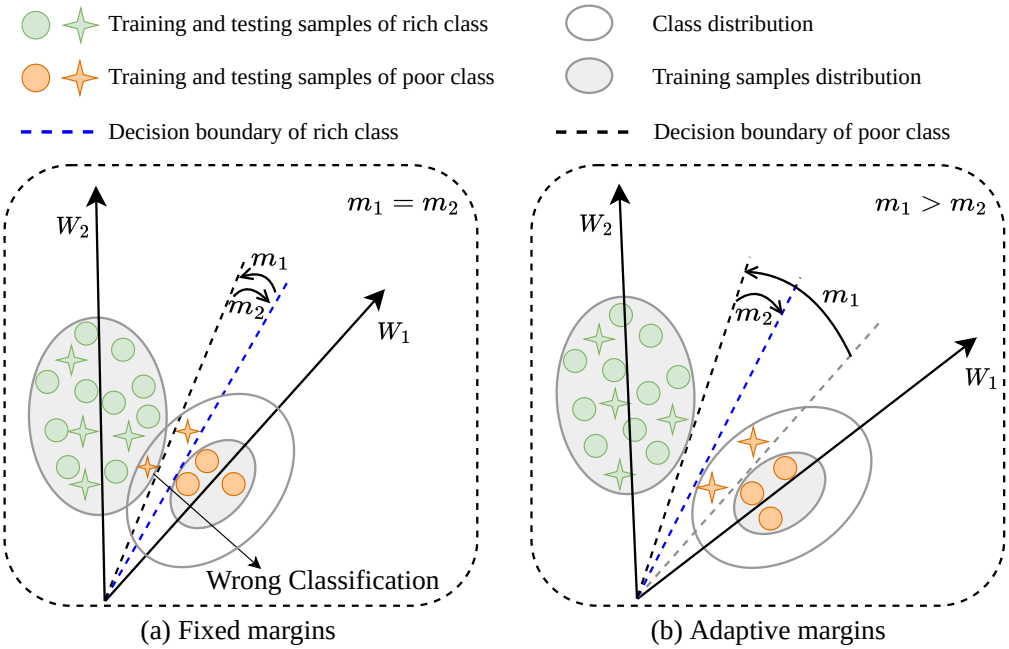


Figure 4.1: Impact of class imbalance [108]. (a) The model exhibits errors on new test samples from underrepresented classes (poor classes) with a fixed additive margin. (b) Adaptive margins offer a more suitable solution to these errors, wherein an underrepresented class requires a relatively larger margin, while a well-represented class (rich class) requires a relatively smaller margin.

4.1 Introduction

The advancement of deep learning, particularly through deep convolutional neural networks (DCNNs), has led to remarkable progress across various domains, including speech recognition, natural language processing, and notably, computer vision. Among these, FR models have gained significant benefits, as the auto-encoding capability of these techniques allows for the extraction of feature vectors with highly discriminative power [177]. This progress in FR techniques has resulted in their incorporation into various applications, spanning from security protocols and surveillance systems to human-robot interactions and mobile devices [95, 101, 106].

The importance of loss functions in the optimization of DCNNs cannot be overemphasized. While the softmax loss is commonly employed, it fails to fully meet the demands of FR tasks, as it struggles to produce feature vectors with highly discriminative power [37, 136, 204]. In response to this limitation, several enhanced loss functions have been developed, with the goal of balancing between inter-class separability and intra-class compactness. Ultimately, these advancements aim to enhance generalization and accuracy of the trained FR models [12, 37, 107, 132, 136, 204, 217].

The current peak of FR largely relies on margin-based softmax loss techniques. Approaches like SphereFace [136], CosFace [204], and ArcFace [35, 37] embed margins into identity features to enhance class separation. However, these methods implicitly

assume class balance, which is often not the case in public facial datasets. As illustrated in Fig. 4.1, while some classes may be well represented, others are sparse. This imbalance highlights the necessity for adaptive margins, allowing for flexibility in margin values based on class-specific sampling distributions [132]. Moreover, the exclusive focus of these methods on either cosine or angle area may overlook the other dimension, potentially impairing performance. To address these limitations, this chapter proposes the Joint Adaptive Margins Loss (JAMsFace), a loss function designed for dynamic margin adjustments that maximize discrimination across cosine and angular dimensions.

This chapter illustrates our contributions in FR feature learning, which include:

- The crafting of the Joint Adaptive Margins Loss (JAMsFace), formulated to recalibrate decision boundaries and yield more refined facial feature representations.
- The innovative design of JAMsFace enables superior feature differentiation by judiciously operating in both cosine and angular dimensions, elevating the overall FR performance.
- Comprehensive experimentations across several FR benchmarks, revealing that JAMsFace sets new standards in FR across a majority of the mainstream benchmarks.

4.2 Relationship to Previous Work

Our approach is linked to margin-based loss functions tailored for the FR task. In the context of FR, loss functions guide models in discerning unique features and enhancing overall performance. Selecting an appropriate loss function is crucial and can significantly influence the efficacy of FR models. The loss functions commonly employed in FR models may be branched into two main types: metric-based approaches and margin-based methods [207]. Each of these methods plays a distinct role in enhancing the accuracy of the models. The margin-based techniques maximize the separation between two feature vectors, while metric-based techniques focus on quantifying the distance between them.

4.2.1 Metric-based Methods

During the initial stages of FR research, metric-based methods were predominant. These strategies leverage deep metric learning networks to learn a similarity metric between images [65], aiming to cluster images with similar visual characteristics in an embedding manifold while concurrently distancing dissimilar ones. The most prevalent metric-based methods losses are triplet and contrastive losses.

Contrastive loss, introduced by Chopra et al. [25], focuses on decreasing the separation between positive pairs and boosting the separation between negative ones, hence its alternative name, distance metric learning. The utilization of this methodology played a pivotal role in the initial advancements of FR, hence greatly enhancing our comprehension of inter-sample associations. However, it faces limitations such as slow convergence, as it compares only one negative class per update. Triplet loss, another metric-based method, was first employed in FaceNet by Parkhi et al. [164]. Unlike contrastive loss, triplet loss uses a tripartite structure comprising negative, positive, and anchor samples. Its objective is to reduce the distance between positive and anchor samples while simultaneously increasing the distance between negative and anchor samples. This approach marked a significant advancement in metric-based methods, demonstrating a more sophisticated means of learning facial features.

However, the adaptability and scalability of these techniques have been hindered by the inherent challenges of optimization against a single negative class, resulting in slow convergence and model instability [40, 179, 217]. To overcome these challenges, innovations like the $(N + 1)$-tuplet loss were introduced [179]. While achieving improved convergence, the $(N + 1)$-tuplet loss significantly expands the samples batch in quadratic behavior. Center loss [217], aimed to navigate around these challenges by learning the center of features for each class and penalizing deviations. However, when tested against the open set protocol, center loss exhibits limited discrimination [40].

4.2.2 Margin-based Methods

Recently, there have been several innovations in adjusting the decision boundaries of softmax loss to enhance discriminative feature learning, as demonstrated by studies such as [12, 37, 92, 97, 132, 136, 137, 162, 204].

The L-softmax loss function [137], improves upon the conventional softmax loss function by incorporating a large margin and utilizing a piecewise function. These improvements help maintain the monotonicity of the cosine function while enhancing both the separability between classes and the compactness within each class.

Their subsequent work, SphereFace [136], expands on L-softmax approach by introducing a multiplicative angular margin that allows for quantitative control. However, this enhancement also complicates the training process.

To address this challenge, CosFace [204] advocates for an additive cosine margin, striking a balance to ensure both inter-class separability and intra-class compactness. Similarly, ArcFace by Deng et al. [37] boosts recognition accuracy through the introduction of an additive angular margin, emphasizing its geometric interpretation.

Further innovations like AdaptiveFace [132], and Dyn-arcFace [97] addressed data imbalance by incorporating adaptive margins, demonstrating the versatility and adaptability of margin-based methods in varying data landscapes. Similarly, methods like

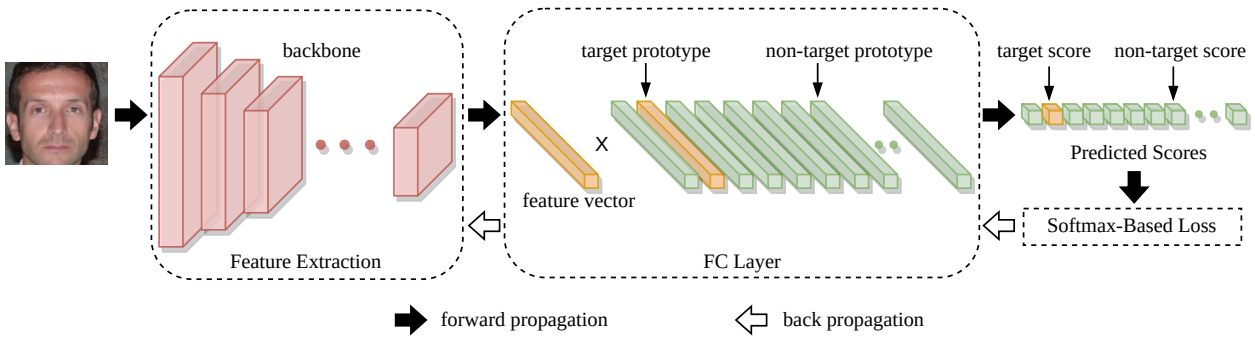


Figure 4.2: General face recognition training framework using softmax-based loss.

CurricularFace [92] and KappaFace [162] have enriched the domain by adaptively modulating margins and recalibrating the importance of samples, respectively. Incorporating image quality considerations, AdaFace [110] introduced quality adaptive margins, indicating a trend toward holistic approaches considering varied aspects of data for improved model performance.

In summation, while metric-based methods laid the foundational stone, margin-based techniques have advanced the SOTA FR performance. Building upon this comprehensive review, the following sections introduce a novel loss function, JAMsFace, and explore its efficacy across contemporary FR datasets [90, 149, 154, 178, 219, 262, 263].

4.3 Our Approach to Adaptive Feature Learning For Face Recognition

4.3.1 Preliminary

FR, as a critical domain of computer vision, has witnessed rapid advancements in training mechanisms and feature representation. The general FR training framework using softmax-based loss, depicted in Fig. 4.2, can be deconstructed into three pivotal modules.

Firstly, datasets play a foundational role in training, validating, and testing FR models. These datasets consist of facial instances captured under various conditions, serving as the basis upon which facial feature extraction algorithms are trained and tested. The training dataset facilitates model learning by iteratively adjusting its parameters. The validation dataset provides an interim evaluation mechanism, crucial for hyperparameter tuning. Lastly, the testing dataset offers an objective measure of the model's performance under generalized conditions.

Following the datasets is the backbone, which consists of a CNN network. Its primary function is to transform raw facial images into a compact, representative feature space. These feature vectors, inherently high-dimensional, encapsulate distinguishing facial attributes crucial for recognition tasks. With the continuous progress of deep

CNN architectures, these networks have improved in capturing subtle facial nuances, thereby enhancing recognition accuracy.

The last component comprises the loss function and its associated prototypes. It assesses the variance between the model's predictions and the ground truth. In FR, this function gauges the likeness or disparity among sample feature vectors. Prototypes are representative exemplars for each unique class in the training dataset. During forward propagation, the final layer computes a correlation measure between an input feature vector and the prototypes. Subsequently, these measures are refined during backpropagation using gradients derived from the loss function, thereby enhancing both the backbone and the classifier.

4.3.2 Method

The major goal of our loss function is to improve the discrimination power across both cosine and angular domains, while also tackling the common issue of long-tail data. In pursuit of this goal, we present an innovative approach that encompasses dynamic additive penalties for cosine and angular margins. Instead of proposing an entirely novel loss, our JAMsFace loss adopts the foundations of extant cosine and angular margin-based loss methods, seamlessly integrating adaptive margin penalties. This integration not only enhances the model's sensitivity to nuances in sample distribution but also strengthens its inherent discriminative power. Through the integration of these flexible penalties, our methodology successfully achieves a harmonious equilibrium between the compactness within each class and the separability across different classes, leading to enhanced performance in FR.

4.3.2.1 Revisiting the Softmax Loss

In classification problems, the softmax loss function is commonly employed as the principal loss function. The fundamental nature of this concept resides in its capacity to maximize the probability of accurately categorizing a given sample into the right class. The mathematical expression can be expressed as follows:

$$\mathcal{L}_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log P_i = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{4.1}$$

where x_i is the embedded feature of i-th training sample, and its probability of being correctly classified as class y_i is represented by P_i. $W_j \in \mathbb{R}^d$ is the j-th column of the weight $W \in \mathbb{R}^{d\times n}$. The bias is denoted as b_j, and the batch size is represented as N. The training dataset consists of n classes, and the dimensionality of the embedded feature is d.

In practical applications, it is common to assign a bias value of $b_j = 0$, as demonstrated in [37]. Subsequently, the weight vector $W_j^T x_i + b_j$ is transformed into $W_j^T x_i$,

which is then expressed as $W_j^T x_i = \|W_j\|\|x_i\| cos\,\theta_j$. Here, θ_j represents the angle created between the weight vector W_j and the feature vector x_i. In order to enhance the process of feature learning, the weight assigned to each individual is adjusted to $\|W_j\| = 1$ using l_2 normalization techniques [37, 136, 202, 204]. To enhance the efficiency of the classification outcome, the deep feature $\|x_i\|$ is additionally normalized by l_2 and re-scaled to a constant value s. The original softmax can be altered, as shown in Eq. 4.2. This modified version is referred to as the Normalized Softmax loss (NSL).

$$\mathcal{L}_{nsl} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(cos\,\theta_{y_i})}}{e^{s(cos\,\theta_{y_i})} + \sum_{j=1, j\neq y_i}^{n} e^{s\,cos\,\theta_j}}. \tag{4.2}$$

Nevertheless, the normalized softmax loss exhibits a restricted capacity to distinguish features optimally for real-world FR applications. In order to address this constraint, several margin-based variations have been suggested [37, 136, 137, 204]. The aforementioned variant methods incorporate a margin in order to establish the boundary between target scores and their non-target entities. The systematic incorporation of this margin not only improves the level of detail in distinguishing features, but also strengthens the overall effectiveness of FR systems. The margin-based variations can be generalized, with each incorporating the margin function $g(m,\, \theta_{y_i})$ that reflects the introduced margin. The general form of these variations can be expressed as follows:

$$\mathcal{L}_{general} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\,g(m,\theta_{y_i})}}{e^{s\,g(m,\theta_{y_i})} + \sum_{j=1, j\neq y_i}^{n} e^{s\,cos\,\theta_j}}. \tag{4.3}$$

The function $g(m_1,\, \theta_y) = cos(m_1\theta_y)$ is introduced by SphereFace [136]. In this function, m_1 represents a multiplicative angular margin, $m_1 \geq 1$, and is an integer. The idea of CosFace, as described by Wang et al. [204], is given by the equation $g(m_2,\, \theta_y) = cos(\theta_y) - m_2$, where m_2 represents an additive cosine margin, and $m_2 \geq 0$. In the study conducted by Deng et al. [37] on ArcFace, the equation $g(m_3,\, \theta_y) = cos(\theta_y + m_3)$ is presented, where $m_3 \geq 0$ represents an additive angular margin. Therefore, the incorporation of margin penalties in the softmax loss resulted in improved distinguishing characteristics compared to the initial softmax loss. Ultimately, the margin-based variations can be combined into a unified expression, denoted as $g(m,\, \theta_y) = cos(m_1\theta_y + m_3) - m_2$.

However, implementing a consistent margin across all classes, as exemplified by the aforementioned margin techniques, poses difficulties, particularly when confronted with imbalanced datasets. Universally applying the same margin may fail to consider the intrinsic variations in class distribution within the training data, potentially resulting in mediocre performance [128, 132]. Moreover, these margin techniques exhibit a tendency to prioritize either angular or cosine spaces, thereby overlooking the possible advantages associated with the other.

In order to tackle these challenges, we propose a novel methodology that adapts to

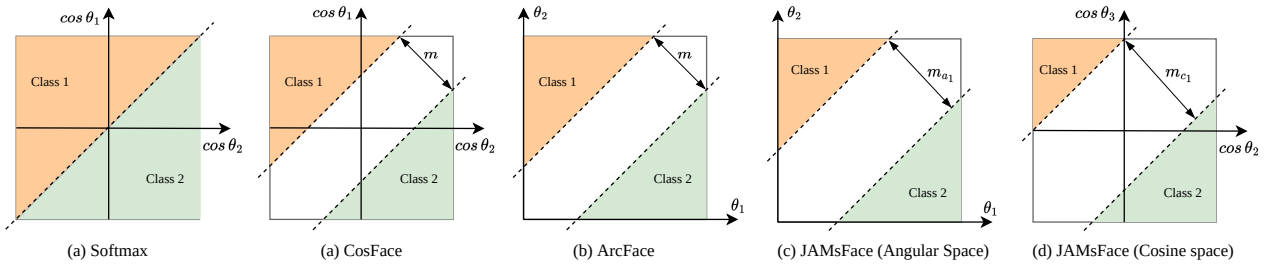


Figure 4.3: Decision margins for various loss functions are illustrated in the context of binary classification. The decision boundary is represented by the dashed line, while the decision margins are denoted by the white areas. Class 1 is associated with underrepresented samples (poor class), while Class 2 is associated with well-represented samples (rich class). JAMsFace assigns a greater joint margin to the poor class to achieve more compactness, thereby implicitly optimizing the underlying feature space. It is important to note that m_{a_1} represents the angular margin for class 1, m_{c_1} represents the cosine margin for class 1, and $\theta_3 = \theta_1 + m_{a_1}$.

the diverse distribution of classes in the training data. At the core of our approach is the integration of a dynamic penalty specifically designed to account for both cosine and angular margins. By incorporating both cosine and angular dimensions, this technique not only ensures comprehensive differentiation across the feature space but also significantly enhances the effectiveness of the FR process.

4.3.2.2 Joint Adaptive Margins Softmax Loss

Building upon existing methodologies, rather than creating an entirely novel loss function, our research proposes an intuitive methodology aimed at concurrently incorporating both cosine and angular dynamic margins. For a clearer explanation of the foundational mechanics of our approach, we present a binary-class illustrative example.

In a binary classification scenario (with two classes C_1 and C_2), the angle between the learned feature vector x and the ground truth weight vector W_i associated with class C_i $(i = 1, 2)$ is represented by θ_i. A conventional normalized softmax loss requires that $cos(\theta_1) > cos(\theta_2)$ for correct classification of x as C_1, and similarly for C_2 requires $cos(\theta_2) > cos(\theta_1)$ for correct classification of x as C_2. However, this decision boundary, as illustrated in Fig. 4.3(a), has limited discrimination power for practical FR tasks.

To navigate this challenge, CosFace [204] advocates for a classifier with an augmented large margin. The correct classification criterion for x as C_1 becomes $cos(\theta_1) - m > cos(\theta_2)$, and reciprocally for correct classification of x as C_2, it requires $cos(\theta_2) - m > cos(\theta_1)$. In a divergent approach, ArcFace [37] proposes to enhance the discriminative power by adding an additive angular margin, refining the target logit to $cos(\theta_i + m)$. The resultant classification boundaries for classes C_1 and C_2 are defined as $s(cos(\theta_1 + m) - cos\,\theta_2) = 0$, and $s(cos(\theta_2 + m) - cos\,\theta_1) = 0$, respectively. The

angular margin limitations enhance separability and classification performance by reducing the gap between feature vectors belonging to the same class and increasing the distance between feature vectors belonging to separate classes.

The aforementioned margin methodologies predominantly focus their efforts on amplifying discrimination within either the angular or cosine dimensions, prioritizing one domain to the potential expense of the other. For example, methodologies such as CosFace and ArcFace employ a pre-established margin m that remains constant throughout the entire training procedure. These static single-margin approaches inadvertently hinder the ability of the techniques to accurately capture the intrinsic complexity of facial data. Furthermore, it is important to note that these methods may not fully leverage the ability of the feature space to differentiate between different classes, particularly when considering the diverse variances within each class.

In light of these challenges, our methodology endorses the incorporation of adaptive margins that flexibly adjust across both the cosine and angular domains. Instead of relying on a static margin, our approach allows the margins to evolve in accordance with the data distribution, resulting in superior discriminative capabilities and enhanced FR performance. In formal terms, the Joint Adaptive Margins Softmax Loss (JAMsFace) is defined as:

$$\mathcal{L}_{JAMs} = -\frac{1}{N}\sum_{i=1}^{N}\log\Big(\frac{e^{s\,(cos(\theta_{y_i}+m_{a_{y_i}})-m_{c_{y_i}})}}{e^{s(cos(\theta_{y_i}+m_{a_{y_i}})-m_{c_{y_i}})}+\sum_{j=1,j\neq y_i}^{n}e^{s\,cos\,\theta_j}}\Big), \tag{4.4}$$

where $m_{a_{y_i}}$ is the angular margin corresponding to the target class y_i, which denotes the extent of the angle increase, while $m_{c_{y_i}}$ is the cosine margin corresponding to the target class y_i and denotes the degree of increment of the cosine.

Illuminating the essence of these joint adaptive margins, the previously discussed binary classification paradigm serves as an optimal illustration. The fundamental condition for accurate classification of a sample x as C_1 defines that $cos(\theta_1) > cos(\theta_2)$. Nonetheless, in the ambit of joint adaptive margins, the required condition evolves to $(cos(\theta_1 + m_a) - m_c) > cos(\theta_2)$ with the constraints $m_a, m_c > 0$. This results in a more stringent decision, since both $cos(\theta + m_a)$ and $cos(\theta) - m_c$ are lesser than $cos(\theta)$. In essence, this approach refines the decision boundary between classes, blessing it with specificity and flexibility, thereby amplifying the model's discriminative powers.

To provide a graphical intuition, Fig. 4.4 visually contrasts our proposed JAMsFace with other prevailing margin-based softmax losses in a geometric context. This representation is grounded in a geometric interpretation: methods like CosFace, ArcFace, and JAMsFace can be interpreted as projections of face features onto a hyperspherical surface. In this high-dimensional spatial interpretation, each facial feature finds representation as a distinct point. The goal of FR is to establish a decision boundary that effectively separates these points into their respective classes. This decision boundary, akin to a hyperplane in this scenario, divides the hypersphere into distinct zones, each corresponding to a specific facial class. While each method aims to enhance FR by

Table 4.1: Decision boundaries for class 1 under binary classification. Note that, θ_i, $i = 1, 2$ is the angle between W_i and x. s is the scale factor, and m is the constant margin. m_{c_1} and m_{a_1} are the cosine and angular margins of class 1, respectively.

Loss Functions	Decision Boundaries
Norm-Softmax	$\parallel x \parallel (cos\,\theta_1 - cos\,\theta_2) = 0$
SphereFace [136]	$\parallel x \parallel (cos\,m\theta_1 - cos\,\theta_2) = 0$
CosFace [204]	$s(cos\,\theta_1 - m - cos\,\theta_2) = 0$
ArcFace [37]	$s(cos(\theta_1 + m) - cos\,\theta_2) = 0$
JAMsFace (Ours)	$s(cos(\theta_1 + m_{a_1}) - m_{c_1} - cos\,\theta_2) = 0$

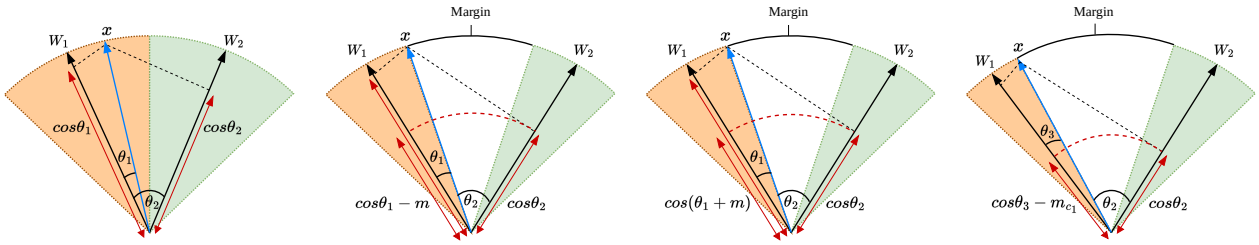
determining an optimal hyperplane that emphasizes inter-class disparity and reduces intra-class variance, JAMsFace stands out. It assigns a larger joint margin that further compresses the underrepresented class (poor class in orange arc), resulting in implicit optimization within the dimensional space.

4.3.3 Comparison with other Loss Functions

In this section, we delve deeper into the comparative nuances between the proposed method, JAMsFace, and other prevalent margin-based softmax loss functions, highlighting the main distinctions in the respective decision boundaries formed by these methods. Fundamentally, the differentiation among various margin-based softmax loss functions revolves around the configuration and adaptability of margins. This can be analyzed into two key facets:

1. *Placement of Margin:* This aspect concerns the spatial arrangement of the margin within the feature space. The margin acts as a buffer or protective barrier around decision boundaries. Its primary role is to prevent potential misclassification, particularly for features situated close to these boundaries. By positioning this margin adeptly, certain loss functions ensure that the decision boundary exhibits enhanced discriminative capacity.

2. *Uniformity vs. Adaptiveness:* Another crucial aspect of distinction lies in whether the margin remains static and uniform across all classes or dynamically adapts to the specific characteristics of each class. Uniform margins treat each class equivalently, imposing a consistent buffer irrespective of the class's distribution or inherent complexities. Conversely, adaptive margins demonstrate flexibility by adjusting themselves based on the properties of each class. This adaptiveness can prove particularly advantageous for classes with long-tail distributions, ensuring that the decision boundary for underrepresented classes is as robust as that for well-represented ones.



(a) Modified Softmax (b) CosFace (c) ArcFace (d) JAMsFace

Figure 4.4: Geometric analysis of various loss functions. The feature space of the poor Class 1 is represented by the orange area, while the feature space of the rich Class 2 is depicted by the green area. (a) Modified softmax loss. (b, c) CosFace and ArcFace assign an equal margin m to both classes, resulting in low compactness for the poor class. (d) JAMsFace assigns a greater joint margin to further compress the poor class, optimizing the underlying feature space. Note that, the angular margin for Class 1 is denoted as m_{a_1}, the cosine margin for Class 1 is denoted as m_{c_1}, and θ_3 is defined as $\theta_1 + m_{a_1}$.

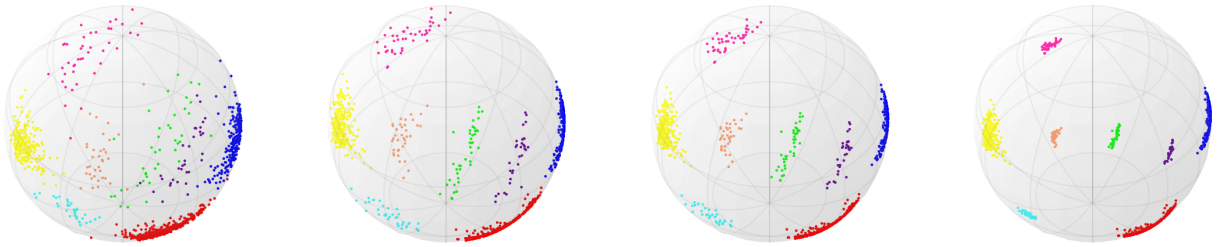
4.3.3.1 Decision Boundaries

Understanding the intricacies of decision boundaries formed by margin-based softmax loss functions requires illustrative comparisons. Referencing Table 4.1 and the visual representation in Fig. 4.3, we can observe how different loss functions shape the decision boundaries.

The original softmax loss simply divides the feature space, resulting in a basic dividing line as shown in Fig. 4.3(a). However, this simplistic boundary might falter when confronted with closely situated samples, leading to potential misclassifications due to its limited discriminatory power.

In contrast, methods like CosFace and ArcFace adopt a reformative approach by introducing uniform margins to enhance the decision boundary's discriminative capability, creating a margin between classes as depicted in Fig. 4.3(b) and 4.3(c). Nevertheless, the inherent limitation of a uniform approach is its disregard for the sample distribution of each class, which can result in poor generalization, especially in the presence of class imbalances in real-world datasets. Thus, a uniform strategy is not always optimal.

Recognizing this limitation, JAMsFace introduces an innovative approach: class-specific adaptive margins in both cosine and angular spaces. Unlike static margins, these adaptive margins dynamically adjust to the distributions of each class. For underrepresented classes, such as class 1, a significantly larger margin is introduced. This not only leads to more compact feature extraction but also pushes the boundary of class 1 further from class 2. The result is a model with enhanced discriminatory power tailored to the specific needs of each class.



(a) Softmax (b) CosFace (c) ArcFace (d) JAMsFace

Figure 4.5: Feature distribution visualization of several loss functions.

4.3.3.2 Toy Example

In order to better illustrate the effectiveness and adaptability of the proposed JAMsFace in contrast to other loss functions, a toy experiment was designed, focusing on the feature distributions produced by each loss function.

Experimental Design: For this experiment, a toy dataset was synthesized using face images from eight distinct identities derived from the MS1MV2 dataset [37]. Subsequently, several 10-layer ResNet models were trained to produce 3-dimensional feature vectors. This dataset's composition is designed to mimic a common real-world distribution, with some classes rich in samples while others significantly less so. Specifically:

- *Class 0* (represented in red): Boasts the richest distribution with over 500 samples.
- *Classes 1 & 2* (illustrated in yellow and blue respectively): Possess a substantial number of samples, with a count close to 250 for each.
- *Classes 3 to 7:* These are the poorly represented classes with approximately 60 samples each.

For effective visualization, the extracted 3-dimensional features from the model were normalized and subsequently plotted on a spherical coordinate system.

Analysis and Interpretations: Fig. 4.5 illustrates the distinct behaviors exhibited by various loss functions.

Softmax Loss Function: One of the primary observations is the distinct bias of the softmax loss function towards richer classes, namely classes 0, 1, and 2. This preference towards richer classes is displayed in the form of ambiguous decision boundaries. Consequently, this could compromise the model's discriminative capacity, potentially leading to misclassification, especially for instances located near these boundaries.

CosFace and ArcFace: Both these loss functions exhibit a principled approach aiming for an equitable feature space allocation. Through a reduction of intra-class variances, CosFace and ArcFace endeavor to provide a symmetrical distribution of feature space to every class. This is irrespective of the underlying sample distribution to each class. However, while this might seem reasonable at first glance, practical implications emerge. Notably, for classes like those delineated in pink and yellow in Fig. 4.5(b) and 4.5(c), despite having different sample sizes, they are accorded almost indistinguishable feature spaces. This could potentially hinder the generalization capacity of models trained with such a loss function, especially when faced with real-world imbalanced datasets.

JAMsFace Function: The JAMsFace function showcases a behavior that is both nuanced and distinct from its counterparts. Its core premise revolves around optimizing the feature space allocations across all classes, irrespective of their respective sample counts. When one examines Fig. 4.5(d), it is apparent that the feature space dedicated to the extensively rich class 0 (marked in red) remains relatively unaltered across CosFace, ArcFace, and JAMsFace. However, the transformative capability of JAMsFace comes to the fore with sparser classes. Classes represented by colors of pink, green, orange, and purple exhibit enhanced compactness and separation, underscoring the adaptability and efficacy of the JAMsFace function. This sharp calibration ensures optimal representation and discernibility across classes, addressing the main challenge of imbalanced datasets.

4.4 Experiments and Analysis

In this section, experimental results of our proposed Joint Adaptive Margins Softmax Loss are presented to provide a comprehensive understanding of its efficacy and robustness. Initially, we delineate the implementation and training specifics. Subsequently, we undertake an ablation study, illustrating the impact of joint adaptive margins within our loss function. Ultimately, we benchmark our JAMsFace model against contemporary state-of-the-art FR models to underscore its comparative performance.

4.4.1 Implementation and Training Details

Data Preprocessing. Central to the effectiveness of any deep learning model is the quality and preparation of the input data. We follow the standard practices in FR as in recent works [12, 37, 266]. Each face image undergoes a systematic preprocessing pipeline: First, it is cropped to a dimension of 112×112. This cropping leverages a similarity transformation based on five pivotal facial landmarks (two eyes, the nose, and the two extremities of the mouth). The detection of these landmarks

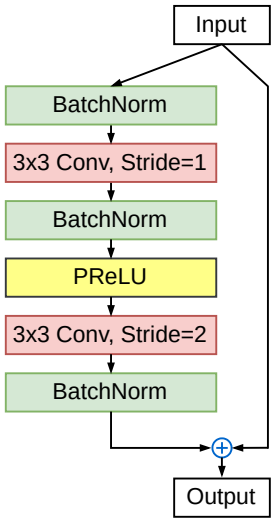


Figure 4.6: Illustration of improved residual unit in ArcFace [37]: BN-Conv-BN-PReLU-Conv-BN.

is facilitated by the MTCNN framework [253]. Subsequent to this, the RGB pixel values are normalized from their conventional range $[0, 255]$ to $[-1, 1]$, priming them for neural processing.

Learning Strategy. To ensure consistency and reproducibility in our experiments, we anchored our model implementation on the publicly accessible code provided by [37]. For a fair comparison, we utilized the same ResNet architecture as outlined in ArcFace [37]. Notably, this differs from the conventional ResNet [79] block, but rather an improved version, visualized in Fig. 4.6, offers unique modifications for the FR domain. These modifications, as showcased in [37], have empirically demonstrated superior results in FR benchmarks compared to the conventional ResNet [79] variant.

Our training operations harnessed the computational capabilities of an NVIDIA Quadro RTX 8000. Each training cycle processed batches of 512 images. The optimization strategy adopted was the Stochastic Gradient Descent (SGD) algorithm [105], configured with a momentum of 0.9 and a weight decay parameter set to 5e−4. For the CASIA-WebFace [244] and VGGFace2 [17] datasets, the learning rate was initialized at 0.001. This rate was subsequently scaled down by a factor of ten at distinct epoch milestones: the 20th, 28th, and 32nd epochs. The training was concluded after 34 epochs. For datasets of a more expansive scale, the learning rate was adjusted at the 10th, 18th, and 22nd epochs, concluding the training process after the 24th epoch.

In the context of our memory buffer settings, we designated the momentum coefficient, α, with a value of 0.3. Depending on the dataset scale - smaller or larger - the γ parameter was calibrated to 0.5 and 0.7, respectively. This granularity ensured that our model was finely tuned to the peculiarities of each dataset. Our experimental implementations were realized using the PyTorch platform [165], chosen for its flexibility, efficiency, and broad support for deep learning operations.

Table 4.2: Face datasets for training and testing

Dataset	#subjects	#imgs/videos	Task	Metrics	Key features
CASIA [244]	10K	0.5M/-	train	-	Unconstrained images
VGGFace2 [17]	9.1K	3.3M/-	train	-	Unconstrained images
MS1MV2 [37]	85K	5.8M/-	train	-	Unconstrained images
LFW [90]	5,749	13,233/-	1:1	Acc	Unconstrained images
CFP-FP [178]	500	2,000/-	1:1	Acc	Cross-pose
CPLFW [262]	3,968	11,652/-	1:1	Acc	Cross-pose
CALFW [263]	4,025	12,174/-	1:1	Acc	Cross-age
AgeDB-30 [154]	568	16,488/-	1:1	Acc	Cross-age
MegaFace [104]	690,572	1M/-	1:1	VR@FAR	Large-scale,
			1:N	Rank-1	Full pose variation
IJB-B [219]	1,845	21.8K/7,011	1:1	TAR@FAR	Large-scale,
			1:N	TPIR@FPIR	Full pose variation
IJB-C [149]	3,531	31.3K/11,779	1:1	TAR@FAR	Large-scale,
			1:N	TPIR@FPIR	Full pose variation

Training data. We utilized several publicly available datasets for our experiments, each with unique characteristics. These datasets have been widely used in FR research, making them suitable for a comprehensive and fair evaluation against other state-of-the-art methods. Table 4.2 provides an overview of these datasets.

MS1MV2 [37] was central to most of our experiments. Originating from the MS-Celeb-1M dataset [71], a semi-automatically refinement process led to the creation of MS1MV2. This dataset is among the largest for FR in unconstrained environments, encompassing 98,685 celebrities and a total of 10 million images. After refinement, about 5.8 million images spanning 85k identities remain. An essential dataset in our training is VGGFace2 [17]. As an enhancement over its predecessor, VGGFace [164], VGGFace2 is a diverse collection with 3.31M images representing 9,131 subjects. This dataset captures a broad spectrum of facial attributes and conditions, such as different poses, age groups, lighting conditions, ethnicities, and even professions. For our ablation studies, VGGFace2 was primarily used. Further evaluations with models trained on this dataset were conducted on benchmarks like LFW [90], AgeDB-30 [154], CALFW [263], and CPLFW [262]. The CASIA-Webface dataset [244] is another dataset in our training process. It comprises 0.49M face images from a diverse group of 10,575 subjects. While it might not be as vast as MS1MV2, CASIA-Webface's diversity serves as a crucial asset for training FR models.

Using multiple datasets in our research was a strategic choice. The wide-ranging facial images and attributes from these datasets aim to provide our model with robustness, adaptability, and superior generalization abilities for various real-world scenarios.

Test Settings. During the testing phase, we process each face image, ensuring it is cropped and aligned to a size of 112 × 112 pixels. Once preprocessed, the image is fed into our trained model, which produces a 512-dimensional feature vector. This vector encapsulates the unique characteristics and attributes of the face image. For consistency and to enhance the robustness of subsequent comparisons, we normalize this feature vector to unit length. Once we have the normalized feature vector, we proceed to classification. Here, the cosine similarity metric plays a pivotal role. This metric measures the cosine of the angle between two vectors, making it particularly suitable for high-dimensional feature comparisons in FR. For every image in the test dataset, we compute its cosine similarity with feature vectors from our reference set (or gallery). This similarity score aids in determining the category or identity of the face.

Benchmarks. FR benchmarks play an instrumental role in evaluating the performance of various FR models. In the initial testing phase, we turn to efficient face verification datasets such as LFW [90], CFP-FP [178], and AgeDB-30 [154]. These datasets were chosen for their capacity to facilitate rapid evaluations, enabling us to assess the effectiveness of different training settings and hyperparameters quickly. After satisfactory preliminary evaluations, we advance to more challenging datasets. In addition to LFW [90] and AgeDB-30 [154], we evaluate our model's performance on datasets specifically designed to challenge FR systems with pronounced pose variations and age differences, such as CPLFW [262] and CALFW [263]. To determine the scalability and robustness of our model, we test it on extensive image datasets like IJB-B [219] and IJB-C [149]. These benchmarks pose considerable challenges due to their wide variations in pose, illumination, expression, and occlusion. Employing these benchmarking strategies is essential to gauge how effectively our model performs in large-scale, real-world scenarios encompassing diverse challenges.

4.4.2 Ablation Study

In our ablation study, we present results from LFW, AgeDB-30, CALFW, CPLFW, and the merged dataset from [216], which combines the four constituent validation datasets.

We demonstrate the effectiveness of the joint adaptive margins in our proposed loss function, JAMsFace, through a comparative analysis with competing approaches that utilize distinct static/adaptive margin configurations. Equation 4.3 outlines three categories of margins: multiplicative angular margin (MA), additive angular margin (AA), and additive cosine margin (AC). Our experimental analysis evaluates the impact of these joint adaptive margins.

Table 4.3 provides an initial assessment of the performance of static joint alternatives

Table 4.3: The impacts of varying joint margins. A 64-CNN architecture was employed to train all models on VGGFace2. The abbreviations MA, AA, and AC represent multiplicative angular margin, additive angular margin, and additive cosine margin, respectively. Static or adaptive settings are denoted by F and A.

	MA	AA	AC	LFW	AgeDB-30	CALFW	CPLFW	Combined
Static	✓	✗	✓	99.317	88.367	90.650	90.433	91.949
	✓	✓	✗	99.250	87.700	90.450	89.783	91.527
	✗	✓	✓	99.300	86.217	89.133	88.833	90.690
Mixed	✗	F	A	99.567	93.300	93.150	92.233	94.324
	A	F	✗	99.483	92.250	92.283	91.800	93.798
	A	✗	F	99.400	90.350	91.633	89.317	90.525
	✗	A	F	99.633	93.733	93.250	92.183	94.468
Adaptive	✓	✗	✓	99.383	91.100	92.250	89.450	90.460
	✓	✓	✗	99.283	90.883	90.183	89.133	90.325
	✗	✓	✓	**99.667**	**94.383**	**93.683**	**92.300**	**94.883**

to examine the effects of employing joint static margins. We then incorporate adaptive margins in conjunction with static margins. Finally, static margins are replaced with adaptive margins to dynamically adjust to the data distribution.

The examination of Table 4.3 reveals that employing a hybrid methodology, which integrates adaptive and static margins, leads to enhanced performance across all datasets in comparison to the static margin alternatives. It is worth mentioning that in comparison to hybrid alternatives with an adaptive multiplicative margin, mixed alternatives with adaptive additive angular or cosine margins demonstrate greater performance. To further boost performance, the integration of adaptive algorithms for both cosine and angular margins surpasses all other alternative approaches.

These findings provide compelling evidence that the adaptive versions outperform the static versions, highlighting the effectiveness of adaptive margins in enhancing the distinguishing capability of our methodology.

In summary, our methodology demonstrates superior verification performance across various datasets, surpassing alternative techniques that utilize either joint static margins or adaptive margins alone. Our results indicate that our combined adaptive margin technique successfully balances inter-class separability and intra-class compactness, resulting in exceptional performance in FR tasks.

4.4.3 Comparison with State-of-the-Art

FR has witnessed remarkable advancements over the years, leading to the introduction of numerous algorithms and approaches. Evaluating these algorithms against established benchmarks provides valuable insights into their strengths and potential areas

Table 4.4: Verification comparison with state-of-the-art methods on LFW benchmark reported in terms of accuracy (%). JAMsFace consistently extend state-of-the-art performances.

Method	Training Data	**LFW**
NormFace [202]	0.5M	98.28
Center Loss [217]	0.7M	98.75
SphereFace [136]	0.5M	99.55
CosFace [204]	5M	99.73
ArcFace [37]	5.8M	99.82
AdaptiveFace [132]	5M	99.62
CurricularFace [92]	5.8M	99.80
Dyn-arcFace [97]	5.8M	99.80
SFace [266]	5.8M	99.82
ElasticFace [12]	5.8M	99.80
AdaFace [110]	5.8M	99.82
JAMsFace (Ours)	5.8M	**99.86**

of improvement. In this section, we compare our approach with several state-of-the-art methods using widely recognized FR benchmarks, including IJB-B [219], IJB-C [149], MegaFace [104], LFW [90], CFP-FP [178], CPLFW [262], CALFW [263], and AgeDB-30 [154].

For every experiment, we ensured fair comparison standards, referencing performance metrics of prior methods directly from their original publications. It is noteworthy that our approach consistently demonstrated competitive, if not superior, performance across these benchmarks, underscoring its robustness and adaptability in diverse scenarios.

4.4.3.1 LFW Dataset

We conducted comparisons with other methodologies using the LFW dataset [90], which comprises 13,233 web-collected images across 5,749 individuals. Among the most common nuisance transformations are illumination, pose, color jittering, and aging. With the evolution of methodologies and the standardization of protocols, the LFW dataset, while foundational, has been noted to exhibit minor differentials in performance. This observation, particularly discernible in the last few percent of accuracy, is attributed to the dataset's limited volume. As such, while it remains an invaluable reference, drawing categorical conclusions about the generalizability of models solely from their LFW performance can be problematic.

Nevertheless, for the purpose of completeness and adherence to traditional benchmarks, we undertook a comparative evaluation using the LFW dataset. Adhering to

Table 4.5: Verification comparison with state-of-the-art methods on cross-age benchmarks reported in terms of accuracy (%). JAMsFace scores comparable results to the state-of-the-art.

Method	**CALFW**	**AgeDB-30**
NormFace [202]	85.61	88.63
Center Loss [217]	85.48	-
SphereFace [136]	92.55	92.88
CosFace [204]	95.76	98.11
ArcFace [37]	95.45	98.28
AdaptiveFace [132]	95.05	97.68
MagFace [150]	96.15	98.17
Dyn-arcFace [97]	-	97.76
SFace [266]	96.07	-
ElasticFace [12]	96.17	**98.35**
AdaFace [110]	96.08	98.05
JAMsFace (Ours)	**96.18**	98.26

the unrestricted protocol augmented by external labeled data, as delineated in [90], we present the performance evaluations of our model on 6,000 face pairs from the LFW dataset in Table 4.4. The results show that our JAMsFace model outperforms the state-of-the-art on the LFW dataset. This underscores that the joint adaptive margins penalty significantly enhances the discriminative capabilities of deeply learned features, thereby underlining the robustness and capability of JAMsFace in FR tasks.

4.4.3.2 Cross-Age Datasets: CALFW and AgeDB-30

Age-invariant FR remains a topic of significant interest due to its profound implications in real-world scenarios, such as identifying missing individuals over the years. The challenges posed by the natural aging process, which brings about considerable morphological changes to facial features, often complicate the recognition task. Besides the LFW dataset [90], we also present a comparative analysis of the proposed JAMsFace against the state-of-the-art models, focusing primarily on age-related datasets CALFW [263] and AgeDB-30 [154]. The Cross-Age LFW (CALFW) dataset [263] has been collected to include images of LFW [90] spanning different ages with the same identities, making it a robust testing ground for age-invariant models. AgeDB-30 [154] further intensifies the challenge by introducing a more concentrated age variation set over a 30-year span.

In the realm of open-source FR models, our JAMsFace model stands out as the leading model on the CALFW benchmark [263], as demonstrated in Table 4.5. When evaluated on the AgeDB-30 dataset [154], JAMsFace obtains accuracy of 98.26% closely

Table 4.6: Verification comparison with state-of-the-art methods on cross-pose benchmarks reported in terms of accuracy (%). JAMsFace consistently extend state-of-the-art performances.

Method	**CPLFW**	**CFP-FP**
NormFace [202]	78.71	90.21
Center Loss [217]	77.48	-
SphereFace [136]	81.40	-
CosFace [204]	92.28	98.12
ArcFace [37]	92.08	98.27
AdaptiveFace [132]	88.80	94.96
CurricularFace [92]	93.13	98.37
Dyn-arcFace [97]	-	94.25
SFace [266]	92.88	-
ElasticFace [12]	93.27	98.67
AdaFace [110]	93.30	98.49
JAMsFace (Ours)	**93.40**	**98.73**

aligned with the state-of-the-art. However, it falls marginally short of the leading 98.35% accuracy achieved by the ElasticFace model [12]. Notwithstanding this minor discrepancy, the overall results underscore the robustness of the JAMsFace model and its intrinsic ability to adeptly handle the complexities presented by cross-age facial variations.

4.4.3.3 Cross-Pose Datasets: CPLFW and CFP-FP

In line with our comprehensive evaluation strategy, we extended our focus beyond cross-age datasets to delve into evaluations centered around a model's resilience to pose variations, specifically targeting the CFP-FP [178] and CPLFW [262] datasets. This emphasis is crucial since, in real-world scenarios, faces are rarely perfectly aligned or frontal. The Celebrities Frontal-Profile (CFP) dataset [178] serves as a vital and rigorous benchmark in this domain. The unique aspect of this dataset is its design, where algorithms are subjected to stringent evaluations based on matches between frontal and profile faces. The dataset contains roughly 7,000 matches, evenly split between genuine and impostor matches. Collectively, it encompasses around 500 unique subjects. To elaborate, the Frontal-Profile (FP) face verification experiment within the CFP dataset [178] includes 350 same-person pairs and 350 different-person pairs, iterated over ten splits. Notably, each probe image in the dataset is almost entirely in profile, amplifying the challenge by introducing extreme pose, lighting, and expression variations. Furthermore, our JAMsFace model was precisely evaluated on the CPLFW dataset [262]. CPLFW, synonymous with pose variations, complements our evaluation

spectrum by further magnifying the challenges associated with pose discrepancies.

The results from these evaluations, as detailed in Table 4.6, underscore that JAMsFace significantly outperforms all other approaches by a clear margin, highlighting the effectiveness of JAMsFace. Several factors contribute to the superior performance of JAMsFace. Primarily, its design emphasizes learning intricate facial details across a range of poses. This ensures that the extracted features are both representative and discriminative, even in the face of extreme pose variations. Moreover, the adaptive margins penalty ensures that the deeply learned features are effectively separated in the embedding space, making recognition tasks more accurate.

4.4.3.4 IJB-B and IJB-C Datasets

In order to thoroughly assess the effectiveness of our loss function, JAMsFace, we employ two highly demanding FR benchmarks, specifically IJB-B and IJB-C. In order to guarantee an even comparison, we use additional SOTA techniques, such as SphereFace, CosFace, ArcFace, and Circle loss. In order to maintain uniformity, we train all the implemented models on the extensively utilized VGGFace2 dataset using the identical CNN architectures as described in [136, 204, 216]. Therefore, VGGFace2 consists of 3.1 million photos derived from 8.6 thousand distinct identities. Through the utilization of this testing methodology, we can objectively assess the effectiveness of JAMsFace in comparison to other contemporary alternatives.

The dataset IJB-B [219] involved a total of 1,845 participants and included 21.8K still photos (11.8K faces and 10k non-faces) as well as 55K frames from 7K videos. The experiments employ the conventional 1:1 verification and 1:N identification techniques. There are 12,115 templates and a list of 8,010,270 comparisons defined by the protocol. More precisely, the 1:1 verification process generates 10,270 authentic matches and 8 million imposter matches, while the 1:N identification approach generates 10,270 probes and 1,875 galleries. IJB-C, as described by Maze et al. [149], is an expansion of IJB-B that employs comparable assessment procedures. The IJB-B dataset is expanded by 1,661 subjects, encompassing a combined total of 31.3K still photos (21.3K faces and 10k non-faces) and 117.5K frames extracted from 11.8K videos.

The evaluations are displayed in Table 4.7 and Table 4.8, measuring the true accept rates (TAR) at different false accept rates (FAR) for verification and the true positive identification rates (TPIR) at different false positive identification rates (FPIR) for identification. In comparison to the baseline methods SphereFace, CosFace, ArcFace, and Circle loss, the proposed JAMsFace loss demonstrates superior performance in both identification and verification tasks. The IJB-B benchmark (Table 4.7) revealed that myLoss achieved a TAR at FAR 1e–4 of 89.09%, surpassing the performance of CosFace [204] and ArcFace [37] with TARs of 86.75% and 88.79%, respectively. Similarly, the IJB-C benchmark (Table 4.8) demonstrated that JAMsFace achieved

Table 4.7: Results on IJB-B. We cite the results from the original papers for [17, 232, 233]. For the reimplemented methods, we use the hyperparameters that lead to the best results on the validation set. Results are in % and higher values are better.

	1:1 Veri. TAR@FAR			**1:N Iden. TPIR@FPIR**		
Method	1e-5	1e-4	1e-3	rank-1	1e-2	1e-1
VGGFace2 (SENet) [17]	67.10	80.00	88.80	90.10	70.60	93.90
MN-vc [233]	-	83.10	90.90	-	-	-
Comparator Nets [232]	-	84.90	93.70	-	-	-
SphereFace	79.10	88.45	93.72	92.83	71.63	86.40
CosFace	76.22	86.75	93.35	92.39	69.01	84.73
ArcFace	80.35	88.79	94.26	93.21	74.39	87.32
Circle Loss	75.15	86.69	93.18	91.71	69.14	84.36
JAMsFace (Ours)	**81.66**	**89.09**	**94.46**	**93.49**	**74.72**	**87.82**

Table 4.8: Results on IJB-C. We cite the results from the original papers for [17, 232, 233]. For the reimplemented methods, we use the hyperparameters that lead to the best results on the validation set. Results are in % and higher values are better.

	1:1 Veri. TAR@FAR			**1:N Iden. TPIR@FPIR**		
Method	1e-5	1e-4	1e-3	rank-1	1e-2	1e-1
VGGFace2 (SENet) [17]	74.70	84.00	91.00	91.20	74.60	84.20
MN-vc [233]	-	86.20	92.70	-	-	-
Comparator Nets [232]	-	88.50	94.70	-	-	-
SphereFace	84.53	90.96	95.29	94.13	80.46	87.92
CosFace	82.08	89.55	94.93	93.66	76.84	85.89
ArcFace	86.02	91.47	95.45	94.51	82.22	88.93
Circle Loss	81.40	89.19	94.67	92.94	76.40	85.33
JAMsFace (Ours)	**86.79**	**91.81**	**95.91**	**94.63**	**83.19**	**89.17**

a Target Accuracy Rate (TAR) at FAR1e–4 of 91.81%. This performance surpassed that of CosFace and ArcFace, which reached TARs of 89.55% and 91.47% respectively.

Visually, the prowess of JAMsFace is evident from the Receiver Operating Characteristic (ROC) curves depicted in Fig. 4.7 for IJB-B and Fig. 4.8 for IJB-C. Notably, JAMsFace continues to demonstrate exceptional performance even at the challenging FAR=1e-5 setting, underscoring its robustness.

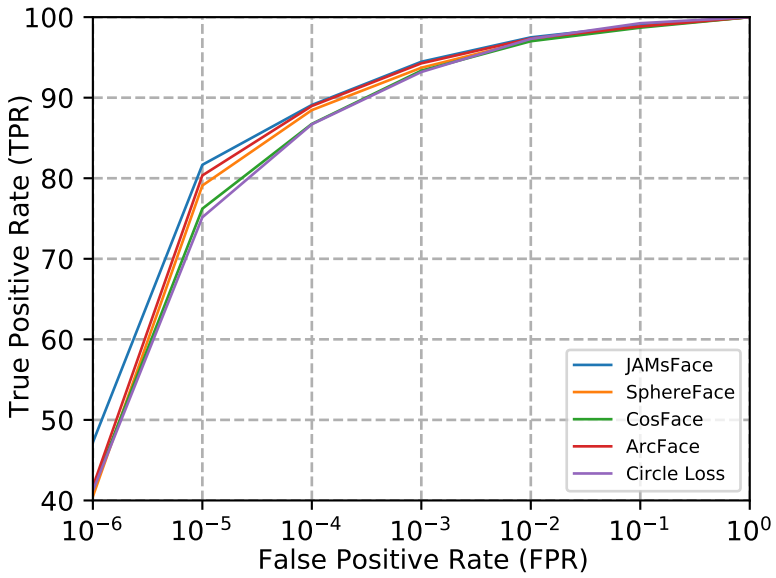


Figure 4.7: The ROC curves of JAMsFace and other start-of-art methods on IJB-B dataset.

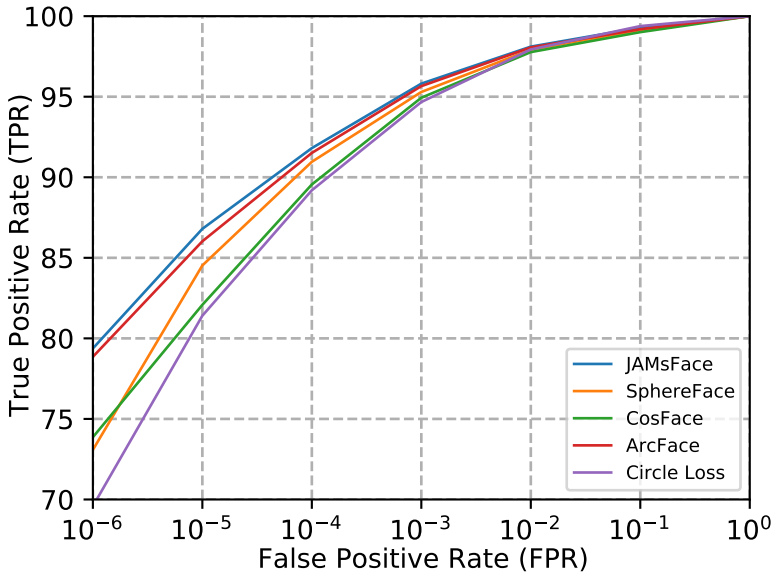


Figure 4.8: The ROC curves of JAMsFace and other start-of-art methods on IJB-C dataset.

4.4.3.5 MegaFace Dataset

Finally, we turn our attention to the MegaFace Challenge to demonstrate the efficacy of the proposed JAMsFace loss function. MegaFace [104] is a particularly demanding benchmark that serves as a stress test for evaluating FR performance at million-level distractors. The MegaFace gallery set consists of 1 million images spanning 690,000 unique subjects, and the probe set, named FaceScrub [160], includes 100,000 photos from 530 distinct individuals. To maintain the quality and reliability of the evaluation, we adopted the refinement strategy from [37], where mislabeled face images are cleaned, ensuring a more refined dataset for performance benchmarking.

The comprehensive results of our evaluation, as well as those of other leading meth-

Table 4.9: Face identification and verification results on MegaFace Challenge 1 using FaceScrub as the probe set. Identification refers to rank-1 face identification accuracy with 1M distractor and Verification refers to face verification TAR (True Acceptance Rate) at 10^{-6} FAR (False Acceptance Rate).

Method	**Identification (%)**	**Verification (%)**
NormFace [202]	89.24	90.76
SphereFace [136]	97.91	97.91
CosFace [37, 204]	97.91	97.91
ArcFace [37]	98.35	98.48
AdaptiveFace [132]	95.02	95.61
MV-Softmax [210]	97.76	97.80
CurricularFace [92]	98.71	98.64
Circle Loss [189]	98.50	98.73
GroupFace [112]	**98.74**	98.79
JAMsFace (Ours)	**98.71**	**98.95**

Table 4.10: Verification performance results reported in terms of accuracy (%).

Method	**LFW**	**AgeDB-30**	**CALFW**	**CPLFW**	**CFP-FP**	**VGG2-FP**
CosFace	99.533	93.533	93.100	91.983	96.374	93.540
ArcFace	99.583	92.417	92.367	91.733	96.569	93.476
AdaptiveFace	99.333	90.983	91.900	89.850	94.929	93.400
ArcPlusCos	99.300	86.217	89.133	88.833	94.411	90.656
JAMsFace (Ours)	**99.667**	**94.383**	**93.683**	**92.300**	**97.286**	**94.700**

ods, are detailed in Table 4.9. Two major tasks are considered: identification and verification. For the identification task, the prowess of JAMsFace shines through, as it narrowly trails the benchmark's best performance. Notably, it falls short by a mini margin of 0.03% when compared with the top performance, GroupFace [112]. This demonstrates that JAMsFace is adept at recognizing individuals across various images and is virtually on par with the best models available. However, it is the verification task where JAMsFace truly showcases its superior discriminative capability. It surpasses all other SOTA methods with a clear margin. This underscores the method's capacity to consistently verify the authenticity of facial identities, even in an extensive dataset like MegaFace.

4.5 Discussion

This chapter presents a novel loss function that provides substantial versatility in determining margins according to the distribution of classes. In contrast to previous approaches like CosFace [204], ArcFace [37], and AdaptiveFace [132], the JAMsFace employs joint adaptive margins in both the angle and cosine spaces to decrease intra-class variation and enhance inter-class variance. This methodology promotes the acquisition of more distinguishing characteristics by the model, hence enhancing the accuracy of FR. Moreover, the incorporation of joint adaptive margins not only increases the depiction of faces but also optimizes overall performance.

In order to confirm the efficacy of the suggested loss function, we introduced a modified version of our JAMsFace denoted as ArcPlusCos loss, which has a static margin. Additionally, we re-implemented three other cutting-edge loss functions: CosFace, ArcFace, and AdaptiveFace. In order to achieve equitable comparisons, the losses that were implemented and the JAMsFace function were trained on the VGGFace2 model [17] using a 64-CNN architecture derived from the works in [136, 204]. The cosine and angular margins were established using the optimal values documented in prior studies [37, 204]. The verification performance results, measured in terms of accuracy score, on many widely used benchmark datasets are presented in Table 4.10. The findings indicate that our suggested approach exhibits superior performance compared to both the static margin variations and the single adaptive margin variant across all evaluation datasets. This underscores the notable enhancement attained with the incorporation of joint adaptive margins.

In addition to its performance advantages, the JAMsFace provides practical benefits by efficiently resolving the issue of class imbalance in FR. Furthermore, it addresses the issues provided by unconstrained environments by using both cosine and angular margins. The proposed approach offers resilience against class imbalance as well as the long-tail problem, unrestricted contexts, the ability to generalize to unfamiliar classes, and the potential for integration into pre-existing FR frameworks. Nevertheless, it is crucial to thoroughly investigate the computational complexity and take into account the specific requirements of the application when evaluating the feasibility of utilizing the JAMsFace.

However, it is important to acknowledge that our proposed methodology still necessitates a significant quantity of training data in order to attain optimal performance. The computing complexity of the training process is increased by the combined penalty imposed by the cosine and angular margins in the JAMsFace function. In order to mitigate these constraints, future investigations may delve into methodologies aimed at enhancing the computational efficacy of margin-based softmax losses, particularly in practical contexts encompassing mobile devices or cloud-based systems. Furthermore, although JAMsFace has demonstrated superior performance on the cross-pose CPLFW dataset, there remains room for enhancement in this domain. Additional

investigation can be conducted to examine different loss functions that possess improved capabilities in managing cross-pose FR and augmenting the model's capacity to accommodate pose fluctuations.

4.6 Summary

Building upon the foundational framework established in Chapter 3, this chapter delves into the design and implementation of a novel loss function tailored to enhance FR accuracy.

Our motivation is based on the observation that a fixed margin, applied uniformly across different classes, may not comprehensively capture the variations and distinctions inherent within and between classes, particularly in the context of diverse and heterogeneous real-world data. Real-face datasets often exhibit a long-tail distribution, which can hinder model learning when using fixed margin loss functions.

Additionally, traditional margin-based methods, while effective in many contexts, tend to prioritize improvements in either angle or cosine space discrimination, potentially neglecting the other dimension. This singular focus limits the model's comprehensive discriminative capabilities. Addressing these gaps, we introduce the Joint Adaptive Margins Loss Function (JAMsFace), which employs an adaptive cosine margin along with angular margin penalties to avoid relying on a single constant penalty margin. JAMsFace not only enhances discrimination in both spaces but also dynamically adjusts the margin for each class. This dual advantage ensures that the model is sensitive to intra-class nuances while maintaining robustness in inter-class discrimination.

Empirical evaluations and extensive experimentation across several benchmarks validate the effectiveness and robustness of JAMsFace. Significantly, our proposed method demonstrates advancements on benchmarks such as LFW, CALFW, CPLFW, and CFP-FP datasets, as well as competitive results on the AgeDB-30 benchmark. The adaptability and effectiveness of JAMsFace are further demonstrated on challenging datasets, achieving state-of-the-art outcomes on IJB-B, IJB-C, and MegaFace datasets.

Beyond its empirical successes, the introduction of an adaptive, class-sensitive margin underscores the importance of nuanced, data-responsive methodologies in FR. It emphasizes the need to move away from rigid structures and towards more flexible, data-driven designs.

Future directions for this research could explore the extension and adaptation of JAMsFace or similar adaptive loss functions across various tasks within computer vision. The results presented here advocate strongly for the broader adoption of such adaptive methodologies not only in FR but also in other domains within computer vision and beyond.

5 Towards Efficient and Robust Face Recognition Network

Traditional face recognition (FR) models often struggle to balance performance and efficiency, especially when deployed in resource-constrained environments like human-robot interaction (HRI) systems. These models, typically comprising millions of parameters, struggle with complex computational demands. Furthermore, their reliance on global facial representations limit their effectiveness in various conditions, such as significant pose variations or occlusions, which are common in HRI scenarios. To address these dual challenges of high computational requirements and dependence on global face representations, this chapter introduces our proposed network, RobFaceNet.

Drawing inspiration from the insights gained in previous chapters, RobFaceNet is designed to efficiently extract both local and global facial features, thereby enhancing robustness and accuracy in varied and unpredictable settings. Its architecture, featuring an innovative attention-enhanced bottleneck, adeptly identifies and prioritizes crucial facial features at different levels. This strategic design significantly reduces computational complexity and the number of model parameters, without sacrificing the accuracy and robustness crucial for effective FR.

Furthermore, we have integrated the JAMsFace loss function, introduced in Chapter 4, into our network architecture. This integration forms a comprehensive FR framework, combining the strengths of an adaptive loss function with an efficient and robust network. JAMsFace, with its unique class-specific adaptive margins, optimizes the feature space for enhanced class separability and discriminability, thus amplifying RobFaceNet's capability to process the intricate details of diverse facial features.

Collectively, RobFaceNet and JAMsFace create a complete FR framework ideally suited for HRI environments. This system not only tackles the challenges of high computational demands and the necessity for comprehensive feature extraction but also offers a balanced, practical, and highly effective solution for real-world applications. Our extensive empirical evaluations demonstrate RobFaceNet's exceptional efficiency and its competitive performance against state-of-the-art accuracy across various FR datasets.

5.1 Introduction

Convolutional Neural Networks (CNNs) have distinctly delineated their impact in the area of computer vision, attaining notable achievements in various visual recognition tasks by adeptly discerning and analyzing relevant visual attributes from images [1, 44]. Particularly in the domain of FR, they have facilitated a paradigm shift, substantively enhancing the fidelity with which facial features are recognized and interpreted [42, 177].

Concurrently, FR technology has opened new horizons in various application realms, including, but not limited to, security [117], video surveillance [8], and most pivotally, HRI [106, 181]. In the realm of HRI, there is a crucial demand for the deployment of robust and efficient FR mechanisms, aiming to enable seamless and personalized interactions between robots and humans. Nevertheless, this necessitates addressing the challenge posed by the low computational constraints inherent to robotic systems, particularly those relevant to power and processing capabilities.

Contemporary FR methodologies, despite their demonstrated performance, predominantly hinge on the computational extensiveness of deep CNNs [37, 48, 177], thus posing significant challenges in resource-limited, real-time contexts. This computational complexity becomes particularly prominent in dynamic HRI environments, where real-time processing and interpretation of facial data are pivotal for facilitating meaningful and interactive engagements.

The imperative to develop lightweight neural networks without compromising accuracy is evident. A spectrum of strategies, including network pruning, knowledge distillation, and depth-wise convolutions, has emerged to navigate the challenge of reducing computational demands without significantly attenuating performance [50, 75, 83, 84, 140, 174, 255]. Despite their exploration in image classification and object detection contexts, a substantial research gap persists regarding FR tasks, highlighting a tangible need for further scholarly investigation.

A limited number of studies have presented accurate and lightweight architectures tailored specifically for FR purposes [49, 124, 146, 228, 239]. Moreover, these lightweight architectures frequently ignore essential low-level, local details, as they predominantly capitalize on the semantic-rich high-level features derived from the last convolution layer. This consequently suppresses the model's capacity to leverage vital low-level information, thereby necessitating a balanced approach that concurrently harnesses both local and high-level information to enhance FR capabilities, particularly within the context of HRI.

In light of these considerations, this chapter introduces RobFaceNet, an innovative CNN architecture uniquely crafted for adept FR, yet carefully conscious of computational and resource limitations. RobFaceNet seeks to harmonize high-level semantic and low-level feature extraction, ensuring a comprehensive extracting of critical facial information whilst aligning with the computational confines of resource-restricted plat-

forms such as robotic systems. Generally, RobFaceNet embeds the bottleneck residual block from MobileNetV2 [174], augmented with attention mechanisms, thereby enhancing its capacity to discern and prioritize salient features within varied face regions.

This chapter illustrates our contributions in FR network design, which can be summarized as follows:

- Multi-feature approach: We devise an efficient and accurate lightweight FR architecture that adopts a multi-feature approach. This approach facilitates the extraction of comprehensive feature information, enhancing the network's FR capabilities while ensuring the feasibility of real-time processing.
- Enhanced bottleneck with attention: Our approach involves the integration of various attention blocks (Channel Attention and Squeeze-and-Excitation) within the bottleneck, tailored to specific layers. This augmentation significantly improves the discriminative power of RobFaceNet.
- Nonlinearity activation function: In RobFaceNet, we employ the h-swish function as the nonlinearity activation function, replacing PReLU. The implementation of this substitution greatly improves the performance of the model while also reducing the computational burden.
- Comprehensive experimental evaluation: We evaluate our method on a series of popular FR benchmarks and demonstrate that our proposed model consistently outperforms other SOTA counterparts, even when compared to other models with similar or larger parameter sizes.

To further illustrate the significance of our approach, the rest of this chapter is organized into the following sections: first, we review related work in Section 5.2, then we introduce our proposed architecture in Section 5.4, followed by detailed experiments and performance analysis in 5.5. Finally, we conclude our work in Section 5.7.

5.2 Relationship to Previous Work

FR has gained immense popularity, especially in mobile devices and robotic systems. This surge has amplified the need to develop computationally efficient models without compromising accuracy, particularly in real-world, dynamic environments where real-time processing is crucial. Although numerous advancements have sought to develop FR systems that are not only accurate and robust across various environments and lighting conditions but also lightweight and computationally efficient, the challenge persists.

Various lightweight network architectures have been proposed for common visual tasks, including SqueezeNet [94], MobileNets [84], MobileNetV2 [174], ShuffleNet [259],

MobileNetV3 [83], and MobileOne [199]. Moreover, efficient lightweight network architectures (Table 5.1) have adapted these networks for FR by designing compact convolution building blocks, such as SqueezerFaceNet [6], MobileFaceNets [23], AirFace [124], VarGFaceNet [239], ShuffleFaceNet [146], Mixfacenets [11], and PocketNet [13].

These models draw inspiration from the advancements in deep image classification models and the evolution of depthwise separable convolutions [174, 193, 256, 259]. These models address the specific challenge of the high number of parameters in fully connected (FC) layers. To overcome this issue, these efficient FR models replace FC layers with global depthwise convolutions (GDC). The GDC layer weights different units of the feature map differently, providing a more effective architecture for FR tasks. The GDC layer has a computational cost of only $W \times H \times C$, where W, H, and C represent the width, height, and channels of the input feature map. Moreover, this approach effectively reduces the parameter count while maintaining or even improving performance.

For instance, the MobileFaceNets architecture [23] is built upon the residual bottlenecks introduced by MobileNetV2 [174], incorporating approximately 1M parameters with 439M FLOPs. The authors finetuned the MobileNetV2 architecture by incorporating a GDC layer instead of a global average pooling (GAP) layer. Additionally, they opted to use the PReLU [78] function as the nonlinearity in all convolutional layers. These design choices have yielded improved performance in facial recognition tasks. While MobileFaceNets demonstrate improved performance across various datasets, they encounter limitations when applied to the MegaFace dataset, where accuracy experiences a slight decrease.

ShuffleFaceNet [146] proposed a compact FR model by finetuning ShuffleNetV2 [145]. This approach replaced the last GAP layer with a GCD layer and the Rectified Linear Unit (ReLU) activation function with PReLU. ShuffleFaceNet is slightly larger than MobileFaceNet; however, it offers better accuracy.

Following a similar pattern as in [23] and [146], the VarGFaceNet [256] and Mixfacenets [11] model architectures adopted the VarGNet [256] and MixNets [193], respectively. In [256], the authors introduced modifications to the VarGNet block, including incorporating a squeeze and excitation block (SE), replacing ReLU with PReLU, and introducing variable group convolutions before the FC layer. These modifications significantly reduced the number of parameters to 5M and the computational cost to 1G FLOPs. Recursive knowledge distillation was also employed to enhance the model's generalization capability. VarGFaceNet achieved an impressive accuracy of 99.85% on the LFW dataset. However, the computational cost of VarGFaceNet is still higher compared to ShuffelFaceNet and MobileFaceNet. Similarly, in [11], the authors introduced a family of efficient FR models (MixFaceNets) by incorporating the MixConv block [193] with a channel shuffle operation, which enhances the discriminative ability of the model. With 3.95M parameters and 626M FLOPs, the MixFaceNets

Table 5.1: Proposed lightweight models in the literature for face recognition

Network	Vector Size	Base Architecture
MobileFaceNets [23]	256	MobileNetv2
AirFace [124]	512	MobileFaceNets
VarGFaceNet [239]	512	VarGNet
ShuffleFaceNet [146]	128	ShuffleNet
Mixfacenets [11]	512	MixNets
PocketNet [13]	128-256	PocketNet
RobFaceNet(Ours)	512	MobileNets

model achieved an accuracy of 99.68% on the LFW dataset.

Inspired by the successes of MobileFaceNets [23], Li et al. [124] developed AirFace, a lightweight FR model based on the deeper MobileFaceNet(y2) [37] architecture. In their approach, they increased the network width and depth to further improve the model's performance. Additionally, they incorporated the Convolutional Block Attention Module (CBAM) [223] into every bottleneck within the network. However, the model still expensive in terms of computational complexity.

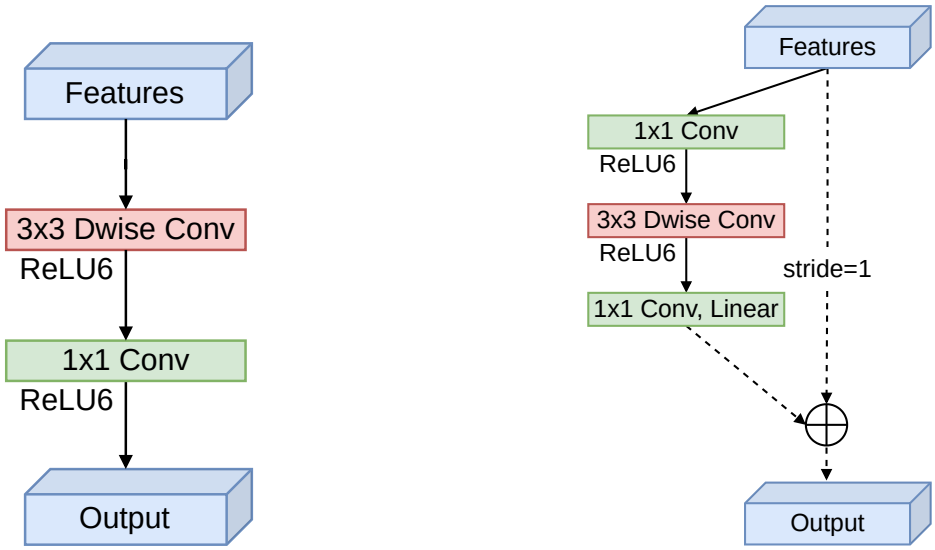
In [13], Boutros et al. introduced a family of lightweight FR models called PocketNets. They utilized Neural Architecture Search (NAS) techniques to automatically discover an FR-specific lightweight architecture, optimizing it for performance and computational efficiency. In addition to the architectural design, Boutros et al. proposed a novel KD paradigm to address the challenges arising from the significant performance gap between the teacher and student models.

Among the previously mentioned works, MobileFaceNets [23] and MixFaceNets [11] have particularly shined in achieving impressive accuracy with minimal computational costs. However, our proposed RobFaceNet architecture surpasses these by delivering superior results with even fewer computational demands. It stands as a potential solution of optimizing both accuracy and efficiency in FR tasks.

5.3 Preliminaries

5.3.1 Efficient Mobile/Light Building Blocks

Depthwise Separable Convolutions have emerged as pivotal building blocks within numerous efficient neural network architectures, underscored by works such as [83, 84, 174, 259]. Our work also leverages depthwise separable convolutions, mainly motivated by their computational and parameter efficiency. Depthwise separable convolutions strategically substitute a full convolutional operator with a factorized version, effectively splitting the convolution into two discrete layers, as illustrated in Fig. 5.1(a). The first layer involves a depthwise convolution, which executes lightweight filtering



(a) MobileNet V1 **(b)** MobileNet V2

Figure 5.1: Comparative visualization of convolutional blocks across different architectures. (a) MobileNetV1 [84]. (b) MobileNetV2 [174], where each block features narrow input and output (bottleneck) dimensions, devoid of nonlinearity, subsequently expanding into a higher-dimensional space before projecting to the output. Notably, the residual connection links the bottleneck stages, bypassing the expanded dimension and only applied when the stride is equal to 1.

by applying a singular convolutional filter per input channel. Subsequent to this, the second layer, termed a 1×1 convolution or pointwise convolution, focuses on crafting new features through the computation of linear combinations of the input channels.

In a standard convolution, an input tensor T_{in} of dimensions $h \times w \times c_{in}$ is subjected to a convolutional kernel $K \in R^{k \times k \times c_{in} \times c_{out}}$, producing an output tensor T_{out} with dimensions $h \times w \times c_{out}$. The computational cost of standard convolutional layers can be formulated as:

$$h \cdot w \cdot k \cdot k \cdot c_{in} \cdot c_{out} \tag{5.1}$$

Conversely, depthwise separable convolutions, a key efficiency gain in RobFaceNet, offer a significantly lower computational cost than traditional convolutions. The computational cost for depthwise separable convolutions is:

$$h \cdot w \cdot c_{in}(k^2 + c_{out}), \tag{5.2}$$

which reduces the overall model complexity and accelerates processing by representing the cumulative cost of the depthwise and 1×1 pointwise convolutions. This mechanism allows depthwise separable convolution to notably reduce computational demand in comparison to traditional layers, diminishing it by nearly a factor of k^2. MobileNetV1 [84], employing $k = 3$, manages to reduce computational costs by a factor of 8 to 9, compared to standard convolutions, while only marginally sacrificing accuracy [84].

MobileNetV2 [174] introduced the linear bottleneck and inverted residual structure,

leveraging the low-rank nature of neural networks to foster even more efficient layer structures. As illustrated in Fig. 5.1(b), this structure is characterized by a 1×1 expansion convolution, succeeded by depthwise convolutions and a 1×1 projection layer. A residual connection connects the input and output if and only if they have an identical number of channels. This methodology preserves a compact representation at both the input and output while internally expanding to a higher-dimensional feature space to augment the expressiveness of nonlinear per-channel transformations.

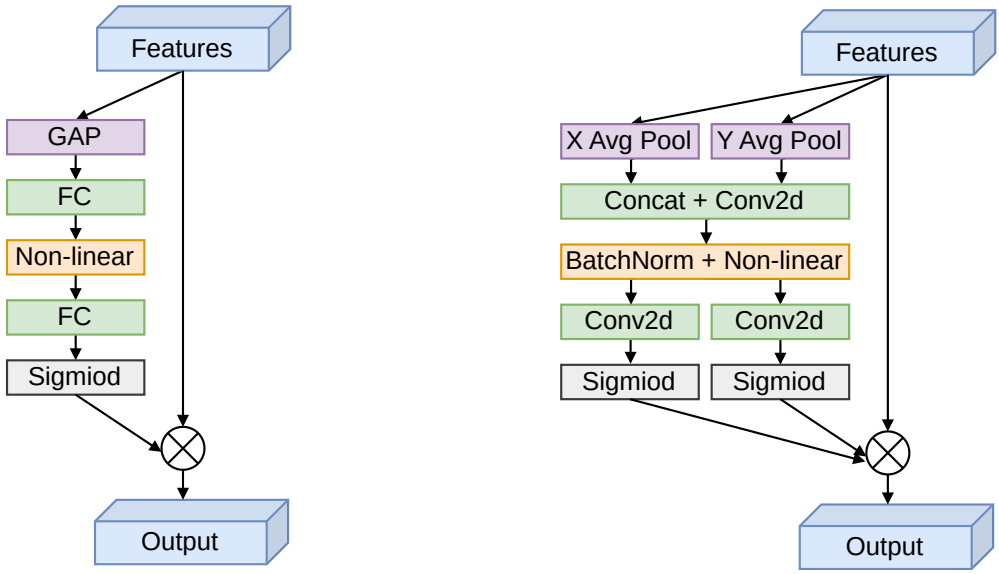
Building upon the foundation of MobileNetV2 [174], MobileNetV3 [83] integrates lightweight attention modules into the bottleneck structure, based explicitly on the squeeze and excitation approach. It's imperative to note that the integration point of the squeeze and excitation module deviates from the location used in ResNet-based modules as proposed in [86]. In MobileNetV3, the module is situated after the depthwise filters within the expansion, facilitating the application of attention to the most voluminous representation.

In developing our approach, we harness a combination of these layers as foundational building blocks to construct models of optimum efficacy. Layers are further enhanced with modified hswish nonlinearities, a nuance introduced in MobileNetV3 [83]. Beyond merely employing attention modules based on squeeze and excitation, our strategy also embraces the utilization of a coordinate attention mechanism, as elaborated in Section 5.4.

5.3.2 Attention Mechanisms

Inspired by humans' intrinsic capability to adeptly identify salient regions within complex scenes, we introduce attention mechanisms into our proposed model, aiming to emulate this particular facet of the human visual system. The attention mechanism can be understood as a variable process of adjusting weights, which is dependent on the characteristics of the input facial image. In this endeavor, we employ two distinct categories of attention mechanisms: channel attention and a combined form of channel and spatial attention. Specifically, the Squeeze-and-Excitation (SE) [86] mechanism is utilized to facilitate channel attention, while Coordinate Attention (CA) [82] is employed to manage both channel and spatial attention.

Squeeze-and-Excitation Attention (SE). The Squeeze-and-Excitation (SE) block, introduced by Hu et al. [86], improves the network's performance by enabling it to capture channel-wise dependencies and enhance the representational capacity. This mechanism is essential for optimizing feature extraction and classification. As illustrated in Fig. 5.2(a), SE block is divided into two main segments: a squeeze module and an excitation module. The compression module captures spatial information on a global spatial by utilizing global average pooling. Conversely, the excitation module employs fully-connected layers and non-linear layers (ReLU and sigmoid) to capture



(a) SE Block **(b)** CA Block

Figure 5.2: Attention mechanisms utilized in our approach.

channel-wise dependencies and generate an attention vector as output. Subsequently, the input feature's different channels are adjusted by multiplying them by the relevant element in the attention vector. SE blocks enhance important channels while simultaneously reducing noise. We incorporate the SE block into our work due to its minimal computational resource demands.

Coordinate Attention (CA). While an SE block aggregates global spatial information via global pooling and subsequently models cross-channel relationships, it overlooks the criticality of positional information. Although Convolutional Block Attention Module (CBAM) [223] and Bottleneck Attention Module (BAM) [163] utilize convolutions to capture local associations, they fall short in modeling long-range dependencies. To address these limitations, Hou et al. [82] proposed the Coordinate Attention (CA), an innovative attention mechanism that incorporates positional information into channel attention, enabling the network to concentrate on substantial, important regions with minimal computational cost.

In Fig. 5.2(b), we illustrate the coordinate attention mechanism, which unfolds through two phases: first, embedding coordinate information, followed by coordinate attention generation. This approach allows the network to focus on important spatial regions in facial images. Initially, two spatial extents of pooling kernels encode each channel horizontally and vertically. The process of the coordinate attention mechanism consists of two sequential phases: coordinate information embedding and coordinate attention generation. The initial step involves encoding each channel horizontally and vertically using two spatial extents of pooling kernels. Following this, a convolutional transformation function with dimensions of $1times1$ is applied to the combined outputs

of the two pooling layers. The network, utilizing coordinate attention, can precisely ascertain the position of a targeted object. This method expands the receptive field compared to BAM and CBAM, and, similar to the SE block, it models cross-channel relationships to enhance the expressive power of the features. Owing to its lightweight design and versatility, it is integrated into the building blocks of our approach.

5.4 Our Approach to Efficient and Robust Face Recognition

FR technologies have been incorporated into many applications, spanning from secure authentication to HRI. The design imperatives for constructing robust and efficient deep learning models have significantly evolved. These models require a refined understanding that model architectures must be both computationally optimized and capable of nuanced feature discernment. Consequently, we present our approach to an efficient and robust FR network. This network architecture not only draws inspiration from prevailing mobile network design strategies but also infuses a series of solid enhancements to augment its proficiency in FR tasks.

Detailed exploration of our architectural design is pivotal to understanding the functionality and efficacy of our approach. Anchoring our discussion in Section 5.3, it is crucial to note that the elementary building block of our architecture is rooted in the bottleneck depth-separable convolution. Initiating the architectural cascade of RobFaceNet, an introductory fully convolutional layer with 64 filters takes the stage, which is sequentially followed by 16 enhanced residual bottlenecks and three branches of depthwise convolution layers, as described in Table 5.2 and visually represented in Fig. 5.3. We have opted for the modified h-swish activation function as the nonlinearity, primarily attributable to its demonstrated robustness in scenarios employing low-precision computation [83]. A kernel size of 3×3 is employed, aligning with contemporary network design norms. Lastly, we establish a hyperparameter s to control the stride within the model's head and stages, offering the option to apply a rapid downsampling strategy at the network's beginning or not. If $s = 2$, then the fast downsampling is activated.

The layers within RobFaceNet incorporate innovative connections, capitalizing on the deep features encoded in both low- and high-level representations of face regions. This integration facilitates the extraction of a broader spectrum of informative features, substantially enhancing the network's ability to capture and represent facial characteristics. While the proposed network adopts the depthwise separable convolutions utilized by numerous efficient neural network architectures [83, 84, 174, 259] to ensure model efficiency, it distinguishes itself by embodying an enhanced version. This version integrates strategic refinements, precisely optimized to simultaneously boost computational efficiency and promote robust performance across a variety of FR applications and scenarios. Subsequent sections illuminate the enhancements and architecture integral to our approach.

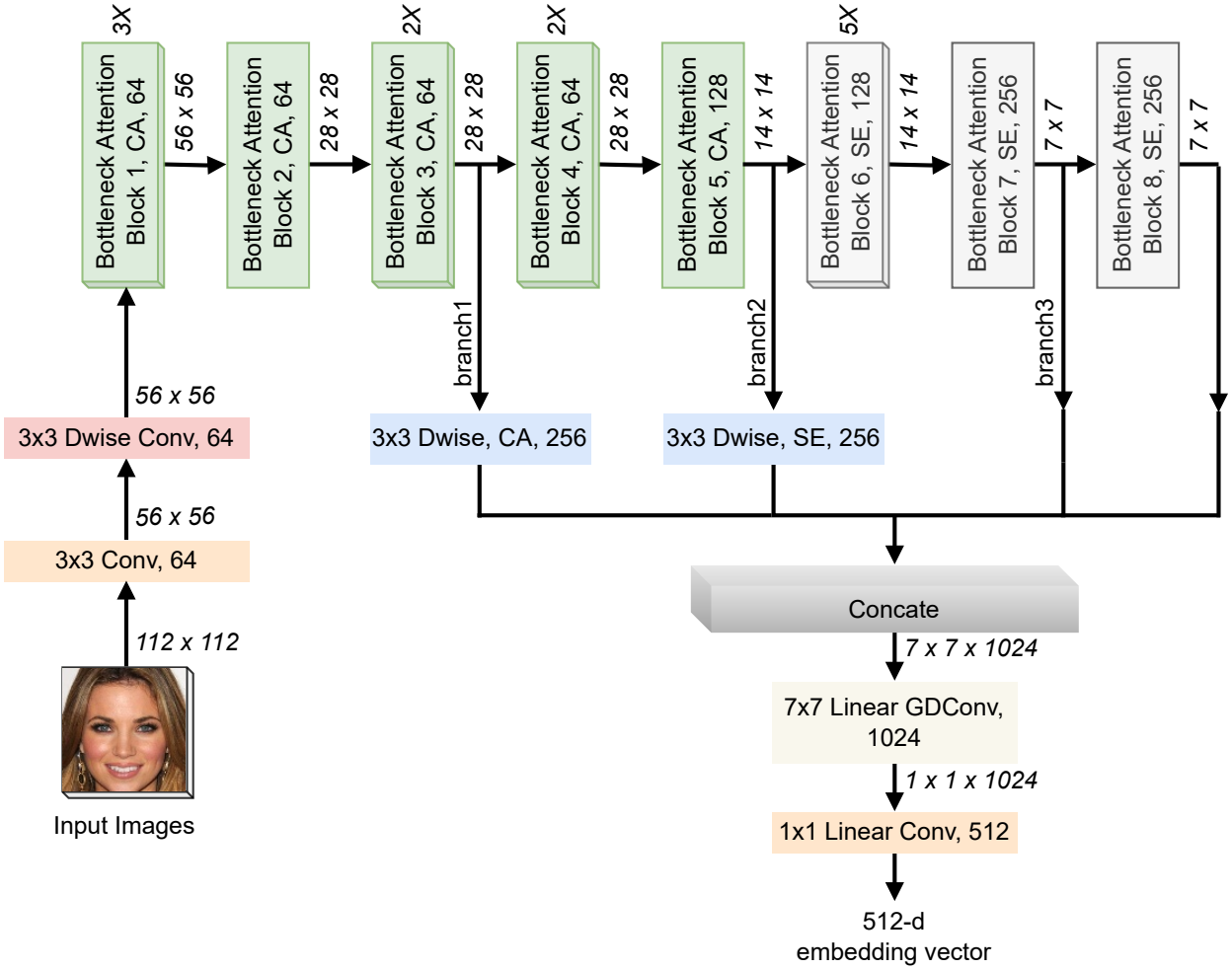


Figure 5.3: Architecture of the proposed network. The RobFaceNet architecture incorporates multi-feature networks that consider both low-level and high-level features in the embedding process. This information is extracted from the middle blocks of the network.

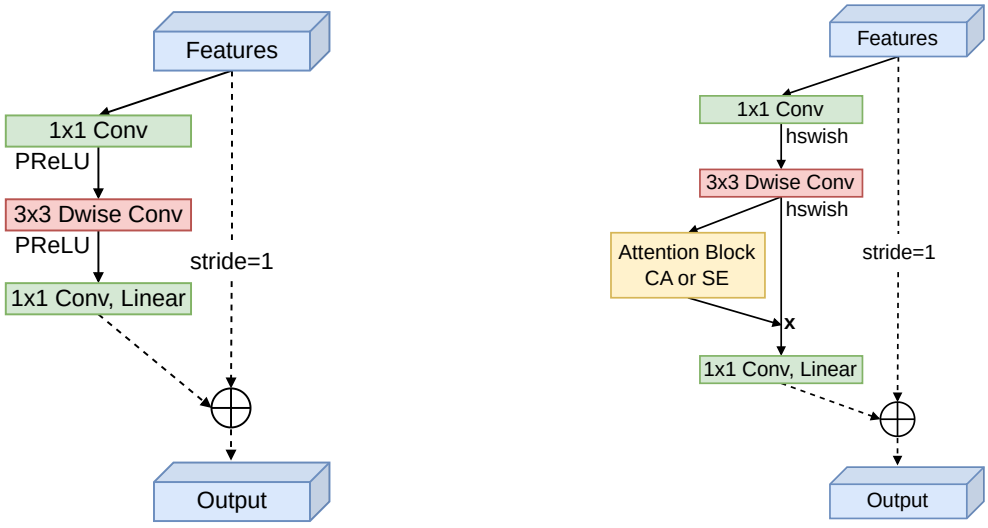
5.4.1 Enhanced Bottleneck

In the context of using CNNs for facial feature extraction, it is crucial to assign more weight to the most recognizable face regions. Similarly, channel features that convey pivotal distinguishing information should be assigned more weight [69]. To achieve superior performance, we adeptly synthesize these principles by introducing an attention-based enhanced bottleneck into our approach, as illustrated in Fig. 5.4. We comprehensively explore the impact of this attention mechanism approach in Section 5.5.3.

The enhanced attention-based bottleneck is characterized as an inverted bottleneck that smoothly integrates either a Coordinate Attention (CA) or Squeeze-and-Excitation (SE) attention module subsequent to the depthwise convolution layer. This combination of attention modules augments the network's proficiency in decoding both channel and spatial features, enhancing its capability to differentiate between various

Table 5.2: The proposed network architecture. Each line describes a sequence of operators, repeated n times with stride s. All layers in the same sequence have the same number c of output channels

Input	Operator	n	s	c	Attention
$112^2 \times 3$	conv3 × 3	1	2	64	No
$56^2 \times 64$	depthwise conv3 × 3	1	1	64	No
$56^2 \times 64$	bottleneck	3	1	64	CA
$56^2 \times 64$	bottleneck	1	2	64	CA
$28^2 \times 64$	bottleneck	2	1	64	CA
$28^2 \times 64$	depthwise branch1	1	4	256	CA
$28^2 \times 64$	bottleneck	2	1	64	CA
$28^2 \times 64$	bottleneck	1	2	128	CA
$14^2 \times 128$	depthwise branch2	1	2	256	SE
$14^2 \times 128$	bottleneck	5	1	128	SE
$14^2 \times 128$	bottleneck	1	2	256	SE
$7^2 \times 256$	bottleneck	1	1	256	SE
$7^2 \times 1024$	linear GDConv7 × 7	1	1	1024	No
$1^2 \times 1024$	linear conv1 × 1	1	1	512	No



(a) MobileFaceNet **(b)** RobFaceNet

Figure 5.4: Comparison of the bottlenecks used in (a) MobileFaceNet and (b) RobFaceNet. MobileFaceNet incorporates MobileNetV2 [174] bottlenecks, replacing ReLU with PReLU. RobFaceNet utilizes MobileNetV2 bottlenecks along with attention mechanisms, such as the Squeeze-and-Excite block or coordinate attention. Unlike [23, 174], we apply the squeeze and excite block or coordinate attention in the residual layer and use hswish as the nonlinearity. The dashed blocks are only applied when the stride is equal to 1.

facial characteristics. A significant aspect of our approach is the discerning application of attention modules, the selection of which is dictated by the specific layer, as delineated in Table 5.2.

The CA module, deployed in the initial layers, captures dependencies and correlations among different positions within a feature map. This boosts the network's ability to differentiate a wide array of facial attributes and structures adeptly. The inclusion of the CA block in our network architecture ensures that the model effectively identifies and prioritizes pertinent spatial and channel information throughout the network, thereby facilitating a more nuanced understanding of spatial hierarchies and channel-wise dependencies within the facial features being analyzed. This detailed attention to spatial and channel aspects guarantees that pivotal facial features are optimally captured during the feature extraction process.

Conversely, the SE module, employed in subsequent layers, focuses on channel-wise interactions, deftly selecting the most fitting representation through a process of channel weight recalibration. The SE blocks assert their importance by emphasizing channels with high-variance features while simultaneously mitigating the influence of channels encompassing redundant or non-informative features, thus ensuring a cleaner, more focused forward propagation of useful information through the network.

Precisely deploying these attention modules across diverse network layers boosts the model's overall performance and allows the network to concentrate on pivotal features at varying stages, thereby enhancing its proficiency for FR tasks. Furthermore, the incorporation of these attention mechanisms enables RobFaceNet to extract facial features with heightened proficiency, achieving superior recognition accuracy without compromising computational efficiency.

5.4.2 RobFaceNet

Multi-feature CNN. Existing methodologies commonly exhibit a singular dependency on high-level features, primarily extracted from the last convolutional layer, while occasionally neglecting the potential comprehensiveness of the features extracted by individual layers [251]. Such an approach, although frequently effective, may overlook the valuable insights offered by low- and mid-level features. In pursuit of a more encompassing, informed feature extraction methodology, we introduce a nuanced, multi-feature strategy within our proposed model.

Our approach integrates features from various network layers to extract a rich and informative feature pool. Consequently, RobFaceNet adeptly extracts, processes, and leverages a broad, insightful feature scope that is especially relevant for FR tasks, where recognizing and understanding fine-grained features and subtle variations are imperative for achieving high recognition precision.

Specifically, as illustrated in Fig. 5.5, we merge the outputs from the *block*3, *block*5, and *block*7 layers in RobFaceNet through separate network branches. To align the

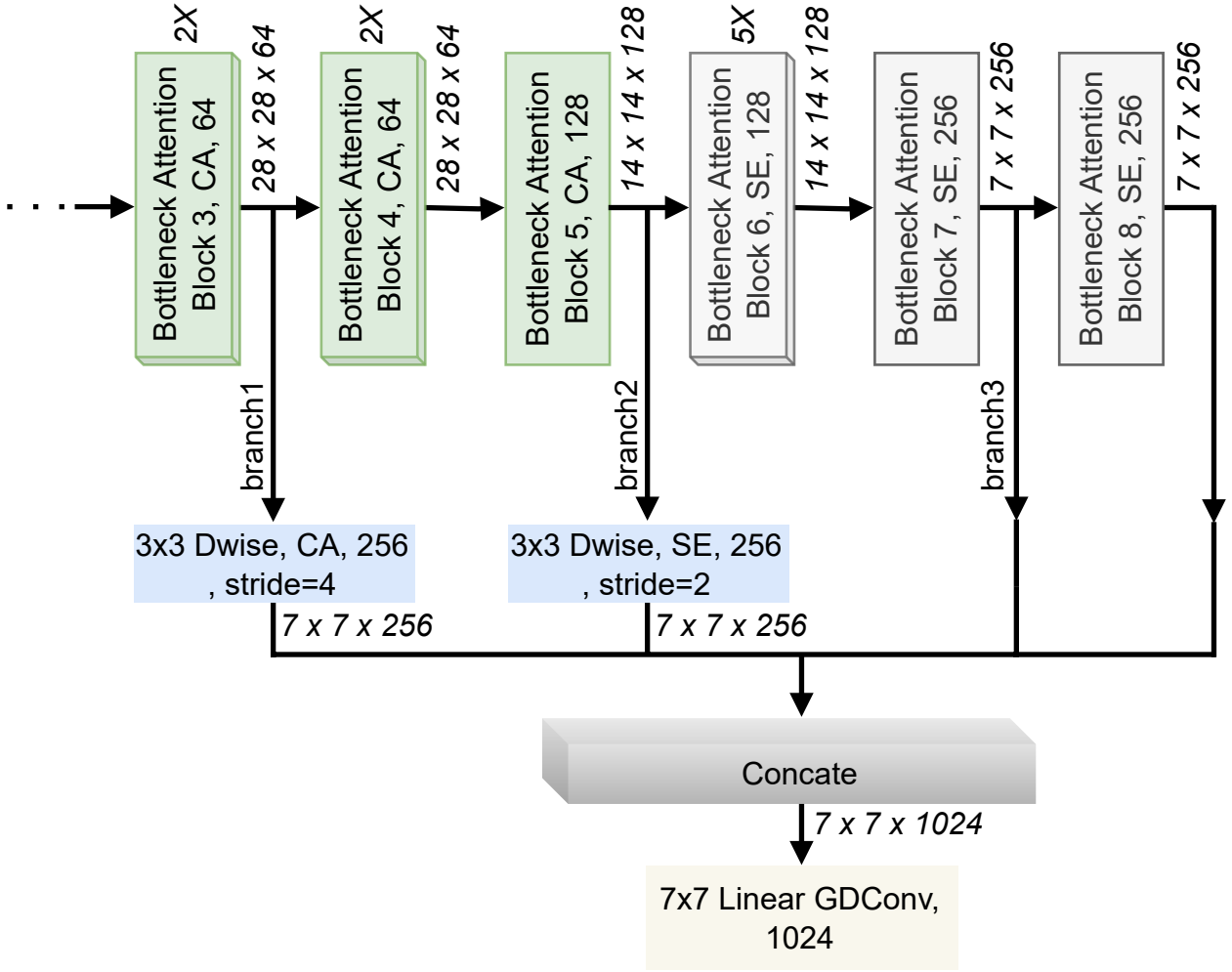


Figure 5.5: Schematic Representation of the Multi-Feature Extraction and Concatenation Process. The diagram illustrates the intricate process of extracting features from various network layers and concatenating them to form a comprehensive and rich feature map, adept for subsequent processing and analysis in our proposed model.

dimensions of the feature maps, we introduce $branch1$ and $branch2$ depthwise convolutional layers. Subsequently, the outputs of these layers are concatenated with the output of $block8$, resulting in feature maps of dimensions $7 \times 7 \times 1024$ that are fed into the following global depthwise convolution layer, $GDCConv7x7$.

Our strategy hinges on insightful feature integration. By amalgamating information from distinct network layers, our model not only encapsulates high-level abstractions generally derived from deeper layers but also retains and utilizes subtler, detailed feature information evident in the network's earlier and mid-layers. This ensures that the extracted features, now representative of a wider, more insightful feature spectrum, serve as a robust and detailed foundation for recognizing and differentiating faces, even in nuanced and challenging recognition scenarios.

By integrating both low-level and mid-level features besides the high-level features,

our model becomes proficiently equipped for accurate and effective facial characteristic recognition. Moreover, our multi-feature approach adeptly balances between the richness of information and computational efficiency, ensuring the model's practical applicability across a spectrum of FR scenarios, thereby enhancing its versatility and overall effectiveness as a promising solution for various applications.

Nonlinearities. Traditional mobile and lightweight networks, ranging from SqueezeNet [94] to the iterations of MobileNets and ShuffleNet [84, 174, 259], have primarily anchored their architectures around the ReLU activation function [248] as a nonlinearity. This widespread adoption can be attributed to ReLU's simplistic implementation and its capacity to mitigate vanishing gradient issues. Nonetheless, the ReLU function is not without its limitations, notably its restriction of activations to non-negative values, which could potentially inhibit the network when navigating through the complex and intricate landscape of facial feature representations.

In light of these challenges, mainly when focusing on lightweight FR networks [6, 11, 13, 23, 124, 146, 239], PReLU [78] emerges as a potent alternative. PReLU, allowing for the propagation of negative activation values, augments the network's ability to learn more intricate feature representations by introducing a degree of asymmetry into the activation function. Empirically, this asymmetry has demonstrated a propensity to elevate performance metrics across various FR tasks and scenarios, making it a valuable addition to the model's architecture.

Navigating a different trajectory from established works, we opt for the modified h-swish activation function, unveiled in MobileNetV3 [83], as the nonlinearity function of choice for RobFaceNet. This pivotal decision is fueled by two critical factors: firstly, the h-swish activation function has demonstrated a remarkable ability to attenuate computational expenses while maintaining a competitive performance in our model for diverse FR tasks, ensuring that the network remains both precise and effective. Secondly, the efficient piece-wise implementation of h-swish minimizes memory access and, consequently, reduces latency costs [83], thereby strengthening its alignment with the computational and resource sensitivities inherent in lightweight, mobile-oriented applications.

By adopting h-swish, our ambitions extend towards refining both the computational efficiency and effectiveness of our network, making it ideally suited for FR tasks, especially within resource-limited environments and embedded systems. The modified h-swish function is defined as:

$$H\text{-}Swish(x) = x\,\frac{ReLU6(x+3)}{6}, \tag{5.3}$$

where, $ReLU6(x) = min(max(x, 0), 6)$.

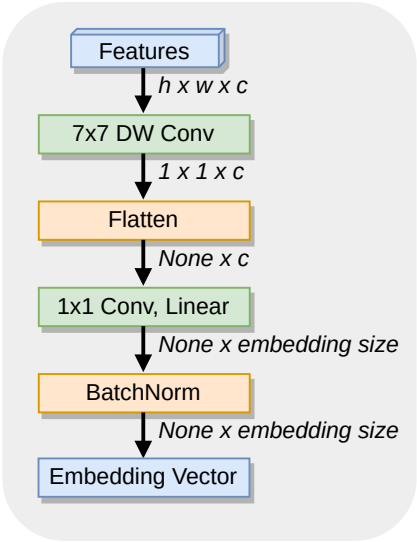


Figure 5.6: Structure of the Global Depthwise Convolution (GDC) layer utilized in the proposed approach, with the embedding size set to 512.

Embedding Setting. In traditional lightweight networks, with MobileNetV2 as an outstanding exemplar, the global average pooling layer is commonly employed to derive an embedding vector. Although prevalent, the GAP layer reveals certain inefficiencies when applied to FR tasks [23, 37, 146, 228], predominantly arising from its equal treatment of every unit within the output feature map, thereby neglecting the varied discriminative power inherent to different units during face feature extraction.

As we delve deeper into the intricacies of FR, it becomes evident that the different units mapped across the feature map correlate to distinct facial features. These units, in turn, render disparate contributions toward the network's discriminative capability. Therefore, creating a feature vector necessitates an approach that precisely weights these units based on their specific contribution to discrimination. However, a straightforward replacement of the GAP layer with a fully connected layer, while facilitating the learning of specific weights for each unit, introduces significant computational overhead and enlarges the model size, which is not suitable for lightweight and mobile architectures.

To mitigate these challenges, our proposed approach adopts an alternative inspired by MobileFaceNet [23], wherein we replace the GAP layer with a Global Depth-wise Convolution (GDC) layer. The GDC layer in our model harmonizes the need to weight feature map units differently with the imperative to maintain a streamlined model. Unlike GAP, which applies a single scalar coefficient across all units, GDC employs a depth-wise convolution to generate a weighted sum of the feature map units, thereby constructing the embedding. The weights learned during the training process allow the network to accentuate or attenuate the contribution of specific feature map units in the embedding, depending on their relevance to the task. This approach promotes a model that is both discerning in its feature extraction and computationally efficient.

Table 5.3: Face datasets for training and testing

Input	#Identity	#Image/Videos	Task	Key features
MS1MV2 [37]	85K	5.8M/-	train	Unconstrained images
MS1MV3 [38]	91K	5.1M/-	train	Unconstrained images
LFW [90]	5,749	13,233/-	1:1	Unconstrained images
CFP-FP [178]	500	2,000/-	1:1	Cross-pose
AgeDB-30 [154]	568	16,488/-	1:1	Cross-age
CPLFW [262]	3,968	11,652/-	1:1	Cross-pose
CALFW [263]	4,025	12,174/-	1:1	Cross-age
IJB-B [219]	1,845	21.8K/7,011	1:1 1:N	Large-scale, Full pose variation
IJB-C [149]	3,531	31.3K/11,779	1:1 1:N	Large-scale, Full pose variation

The comprehensive structure of the GDC layer is illustrated in Fig. 5.6.

The incorporation of the GDC layer, together with strategic refinements across our model architecture, formulates a model proficiently balanced between computational efficiency and perceptive feature extraction, yielding a network finely tuned for FR tasks, especially within resource-constrained environments.

5.5 Experiments and Analysis

In this section, experimental results of our proposed Efficient and Robust FR Network (RobFaceNet) are presented. We first present preprocessing methods and then the implementation details. After that, we conduct an ablation study to analyze the impacts of various settings in our model, focusing on how they influence accuracy and computational efficiency. Next, we visually demonstrate the efficiency of our approach. Finally, the proposed model is compared with state-of-the-art methods for FR

5.5.1 Preprocessing

For data preprocessing, we adhere to the widely employed methodology as described in previous studies [12, 37, 266]. This involves cropping each face image to dimensions of 112×112, employing a similarity transformation that relies on the five face landmarks identified by MTCNN [253]. Ultimately, the resolution of the RGB pixel values is normalized from the range of $[0, 255]$ to $[-1, 1]$. The RobFaceNet model utilizes aligned and cropped facial images with dimensions of $112 \times 112 \times 3$ to generate feature embeddings with a dimensionality of 512.

5.5.2 Implementation Details

Training Datasets. Our RobFaceNet model was trained using the MS1MV2 dataset [37], and for the ablation study, we utilized the VGGFace2 dataset [17]. The MS1MV2 dataset is an improved version of MS-Celeb-1M [71], with around 5.8M images belonging to approximately 85k identities. On the other hand, VGGFace2 comprises 3.14M face images that cover a wide range of poses, ages, and ethnicities.

Validation and Test Datasets. We used the LFW [90], CFP-FP [178], and AgeDB-30 [154] datasets for validation purposes to assess the improvements achieved with different settings. Additionally, we used different benchmarks to evaluate the effectiveness of our proposed lightweight face model in various FR tasks, highlighting their main characteristics in Table 5.3. In addition to efficient face verification datasets like LFW, we also evaluated the performance of our lightweight networks on larger-scale image datasets, such as IJB-B [219] and IJB-C [149]. Furthermore, we extensively tested our models on cross-pose datasets, including CFP-FP [178] and CPLFW [262], as well as cross-age datasets, such as AgeDB-30 [154] and CALFW [263]. These evaluations demonstrate the robustness and effectiveness of our lightweight FR RobFaceNet in various challenging scenarios.

Training Setup. The models introduced in this work are implemented using PyTorch. For a fair performance comparison with other SOTA models, all models are trained using the ArcFace loss [37] with an angular margin of $m = 0.5$ and a feature scale of $s = 64$. During training, we set the batch size to 512 and utilized an NVIDIA Quadro RTX 8000 GPU. The optimization is performed using the Stochastic Gradient Descent (SGD) optimizer [105] with an initial learning rate of 1e-1, momentum of 0.9, and weight decay parameter of 5e-4. The learning rate is reduced by a factor of ten at 80k, 140k, 210k, and 280k training iterations. To monitor the model's performance during training, we evaluate it on LFW, CFP-FP, and AgeDB datasets after every 5000 training iterations. The training process is stopped after 300k iterations. For verification, we use the cosine distance between feature vectors in all experiments.

5.5.3 Ablation Study

5.5.3.1 Impact of Attention Mechanisms

The infusion of attention modules into deep learning architectures, specifically into an inverted bottleneck configuration in our context, is driven by an overarching objective: to confer the network with the adaptive capability to prioritize the learning and extraction of pivotal features, while simultaneously attenuating the influence of less informative elements.

In this experiment, we emphasize our focus on two distinguished attention mechanisms: Channel Attention and Squeeze-and-Excitation, as discussed in section 5.3. These mechanisms are examined both individually and in pairs to deduce their respective and collective impacts on FR capabilities, as detailed in Table 5.4.

Phase 1: Individual Application of Attention Mechanisms. Initially, a singular attention mechanism, either CA or SE, is infused throughout all layers of the network. This approach is undertaken to derive insights into the inherent capabilities and potential limitations of each attention mechanism when applied uniformly across the architecture.

Phase 2: Hybrid Application of Attention Mechanisms. Subsequent investigations involve a hybrid strategy, incorporating both CA and SE attention modules. In this phase, we intersperse the attention mechanisms, embedding CA in the initial layers and SE in the latter and reciprocally, aiming to find an optimal configuration that harmonizes the benefits offered by both mechanisms.

Table 5.4 encapsulates the empirical outcomes, revealing a compelling result: the architecture that incorporates CA modules within the initial layers and SE modules in the succeeding ones displays a superior recognition performance across all investigated datasets. This configuration not only enhances performance in a generic setting but also sustains this elevated performance across datasets that introduce additional complexities, such as age and pose variations.

This substantive performance enhancement is not merely an empirical observation but a testament to the effective learning and representative capabilities bestowed upon our network by the amalgamated attention-based bottleneck. Consequently, the results emphasize the strategic significance of employing a precisely configured attention mechanism, one that is not arbitrarily incorporated, but is architecturally harmonious with the learning nuances demanded by diverse FR scenarios and datasets.

These outcomes not only validate but also illuminate the pivotal role of our enhanced attention-based bottleneck in amplifying FR performance, offering a versatile solution adept at navigating through a myriad of scenarios and conditions prevalent in real-world applications.

5.5.3.2 Impact of Nonlinearities

The role of nonlinear activation functions in deep neural networks is pivotal, acting as a determinant in the learning capabilities and computational demands of the network. Within the context of our experimentation, we build a systematic exploration of various nonlinear functions, namely the Rectified Linear Unit (ReLU), Parametric Rectified Linear Unit (PReLU), and the modified Hard Swish (h-swish), highlighting their implications on the FR performance of our proposed model. Our experimentation begins with a pragmatic evaluation of h-swish nonlinearities, an activation function renowned for its balanced trade-off between computational efficiency and model capa-

Table 5.4: Effects of different attention modules. F and L denote the application of attention modules to the first and last layers of the network, respectively. All models are trained on VGGFace2 [17]. The last row indicates the RobFaceNet settings

SE	CA	MFLOPs	LFW	AgeDB-30	CALFW	CPLFW	CFP-FP
✗	✗	333.7	99.56	93.98	93.25	91.28	96.94
✓	✗	335.4	99.50	94.11	93.50	91.65	97.42
✗	✓	339.0	99.61	94.08	93.40	92.00	97.51
F	L	336.6	99.55	93.92	93.40	92.13	97.35
L	**F**	**337.3**	**99.65**	**94.53**	**93.66**	**92.33**	**97.79**

bility. Refer to Table 5.5 for a quantitative delineation of our findings.

The initial phase of our experimentation adopts a conventional approach, substituting ReLU with PReLU, attributed to the latter's capability to learn negative activation values, an attribute which has been correlated with improved performance in FR tasks [78]. Following this, we delve into the domain of h-swish nonlinearities, replacing ReLU and contrasting the impacts. This analytical comparison, encompassing single-type nonlinearity adjustments, paves the way for our subsequent exploration into the realm of mixed nonlinearities.

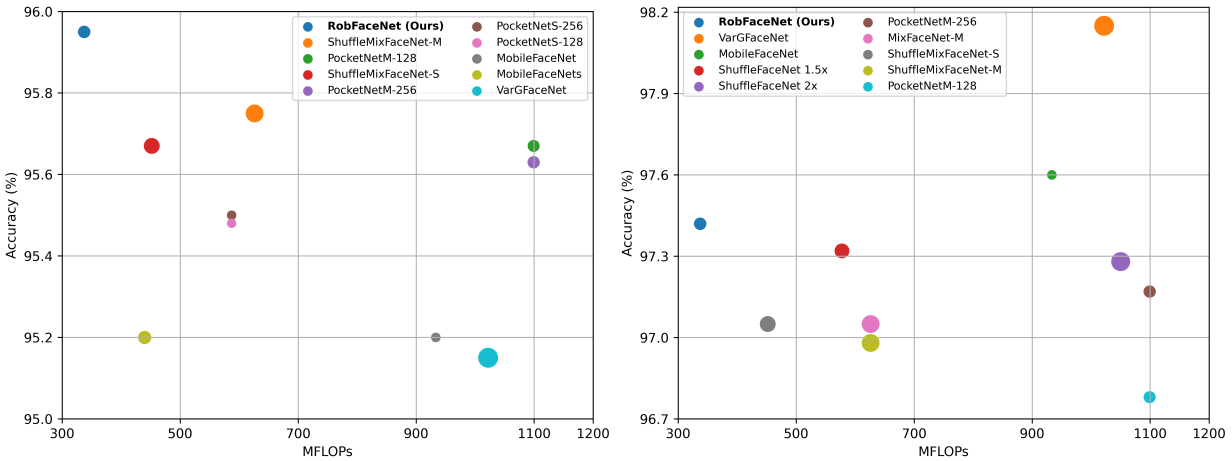
In the pursuit of refined performance and computational balance, we explore a hybrid nonlinearity configuration, incorporating both h-swish and PReLU or ReLU in different network layers. This approach derives from the hypothesis that different network depths may benefit variably from distinct nonlinearities, thereby enriching the feature extraction and representation capabilities of RobFaceNet. Our experimentations with various configurations of these mixed nonlinearities clarify their direct impact on the network's FR prowess.

A precise examination of the results (Table 5.5) reveals significant insights into the interplay between nonlinear activation functions and FR performance. A discernable improvement is noted with the consistent deployment of h-swish across all layers of RobFaceNet, signifying not only an enhancement in recognition accuracy but also a commendable reduction in computational demand when compared with other configurations.

Our findings emphasize the salient benefits of harnessing the h-swish nonlinearity, shedding light on its efficacy in not only preserving but enhancing both the computational efficiency and effectiveness of our network across diverse FR tasks.

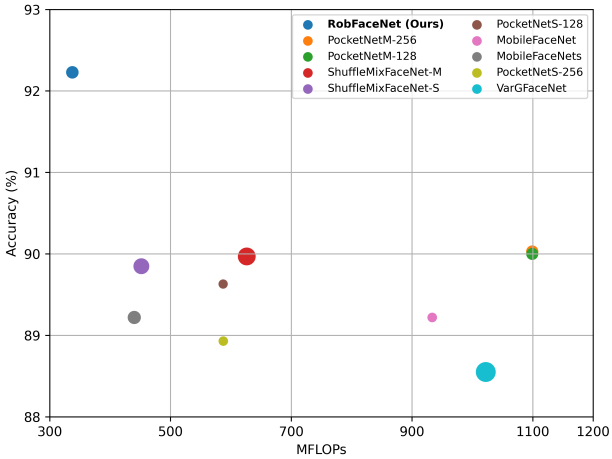
5.5.4 Performance versus Computational Complexity

Navigating the intricate intersections between model performance and computational complexity constitutes a captivating challenge, particularly in domains such as FR,

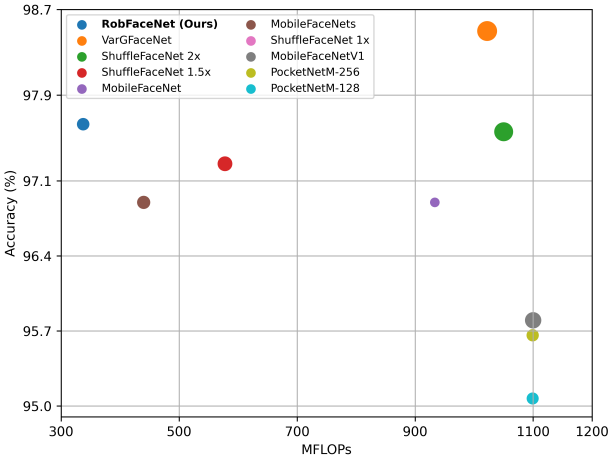


(a) CALFW

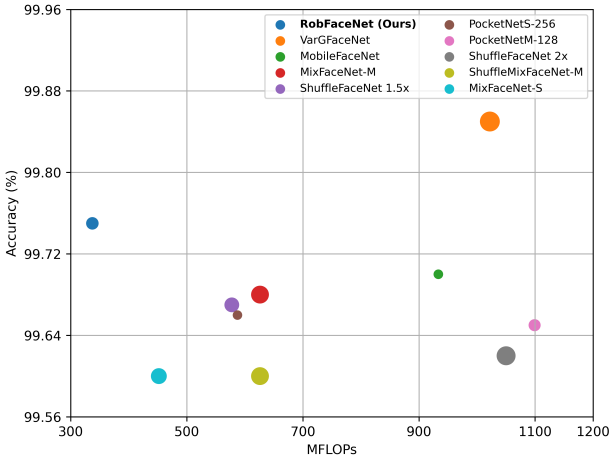
(b) AgeDB-30



(c) CPLFW



(d) CFP-FP



(e) LFW

Figure 5.7: Illustrating the Balance: Computational Complexity, Performance, and Model Size Across Benchmarks. Model size is proportionally represented by marker area. RobFaceNet, highlighted with a blue circle, exemplifies a leading trade-off between FR performance, FLOPs, and compactness, consistently occupying the top-left corner and compared with the top-performing compact models from recent literature on each benchmark.

Table 5.5: Effects of different nonlinearities. All models are trained on VGGFace2 [17]. The last row indicates RobFaceNet settings

ReLU	PReLU	HSwish	MFLOPs	LFW	AgeDB-30	CALFW	CPLFW	CFP-FP
✓	✗	✗	341.4	99.60	94.21	93.40	92.20	97.27
✗	✓	✗	341.4	99.60	94.35	93.61	92.11	97.65
✓	✗	✓	340	99.53	94.13	93.26	92.17	97.20
✗	✓	✓	340	99.50	94.40	93.60	92.10	97.51
✘	**✘**	**✔**	**337.3**	**99.65**	**94.53**	**93.66**	**92.33**	**97.79**

where precision and efficiency are paramount. In our quest to uncover an optimal balance, Fig. 5.7 offers a visual exposition, clarifying the computational efficiency of our model compared to its verification performance, with quantitative results summarized in Tables 5.6, 5.7, and 5.8.

To carve a holistic representation and further deepen the insights derived from our analysis, we introduce model size, an essential determinant of deployment feasibility in resource-constrained environments, as an additional evaluative metric. In Fig. 5.7, the size of each marker is directly proportional to the model size, seamlessly weaving this dimension into our comparative visualization and providing a multi-faceted view of the models under consideration.

Our analysis integrates a diverse array of models, particularly emphasizing the top ten compact models that have demonstrated exceptional performance in recent literature across each benchmark. Each model is encapsulated by a distinct colored circle within the visual representation, its precise position reflecting its respective computational complexity and performance. Within this context, the upper-left corner of the graph represents the coveted 'sweet spot,' symbolizing a model of high performance coupled with minimal computational cost.

As delineated by Fig. 5.7, our model RobFaceNet consistently anchors itself in this aspirational position, illustrating a compelling narrative of optimal symbiosis between reduced model complexity and elevated FR performance. This not only underscores the robust accuracy achieved by our model but also shines a spotlight on its minimal computational demands.

In summary, the compelling efficacy of our approach, evidenced through its delicate balance of precision and computational cost, establishes it as a highly pragmatic and potent solution for FR tasks. This is especially relevant in scenarios constrained by resources, such as embedded systems and robotic deployments, where the ability to perform with a minimal computational footprint is pivotal.

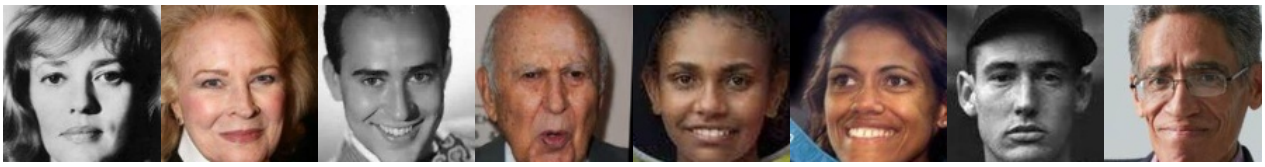
5.5.5 Comparison with State-of-the-Art

To guarantee fairness and impartiality in our comparative analysis, we've implemented stringent measures to align the experimental conditions for RobFaceNet and the state-

(a) LFW

(b) CFP-FP

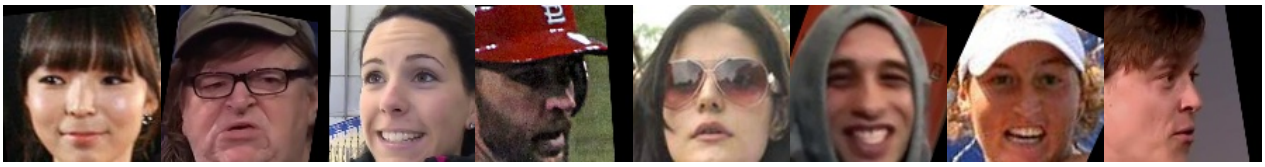
(c) CPLFW



(d) CALFW

(e) AgeDB-30

(f) IJB-B



(g) IJB-C

Figure 5.8: Example face images derived from various benchmarks utilized in the evaluations, illustrating the distinct challenges associated with (a) LFW, (b, c) Cross-Pose, (d, e) Cross-Age, (f) IJB-B, and (g) IJB-C datasets. LFW predominately explores variations in lighting and expressions, Cross-Pose introduces challenges in head pose variations, Cross-Age provides a spectrum of age variations, while IJB-B and IJB-C introduce a diverse and complex array of real-world conditions.

of-the-art lightweight models under comparison. Specifically, RobFaceNet and all comparably examined models are trained using the MS1MV2 dataset [37] and utilize the ArcFace [37] loss function, maintaining consistent training settings to assure an unbiased comparison.

In our comparisons, we categorize models based on computational complexity (MFLOPs), with results reported as shown in related works. For context regarding the current state-of-the-art (SOTA) performance in large-scale deep FR, we initially report the results for the prevailing SOTA ReNet100 model [37]. The table is subsequently organized into three parts: the first section showcases results for models with complexities above 1000M FLOPs, while the second and third sections present results for models with less than 1000M and 500M FLOPs, respectively. This structured approach facilitates a comprehensive comparison between RobFaceNet and models with varying complexities, accentuating RobFaceNet's efficiency and performance within the FR domain. This organizational strategy is applied across all comparative tables. **Notably**, among all models with computational complexity less than 500M FLOPs, our RobFaceNet surpasses all the listed models, including MobileFaceNets [23]. This underscores the efficiency and effectiveness of our proposed RobFaceNet architecture for FR tasks.

Subsequent sections will unveil insights and observations from this comprehensive comparison, highlighting the arenas where our model excels and where it encounters stiff competition from other models. Through this in-depth examination, we aim not only to underscore the strengths and distinctive advantages of our approach but also to spotlight areas that warrant further research and optimization in the multifaceted and perpetually evolving domain of FR.

5.5.5.1 Evaluation on LFW Dataset

LFW [90] comprises 13,233 web-collected images, capturing a wide spectrum of 5749 unique identities, featuring a substantial range of poses, expressions, and lighting conditions, thereby reflecting a multitude of real-world scenarios. LFW's evaluation methodology hinges upon 6000 face pairs, systematically partitioned into ten subsets, each containing 300 positive and 300 negative pairs.

Adhering to the standard protocol, specifically the unrestricted with labeled outside data approach, our evaluations pivot around the verification accuracy across these 6000 face pairs. Detailed results and insights into the network and model sizes are delineated in Table 5.6, illuminating the performance and efficiency of the models under comparison.

VarGFaceNet [239], achieving a top verification accuracy of 99.85%, occupies the prime position in the LFW benchmark. However, this superior accuracy comes at the cost of substantial computational demand, registering at 1022M FLOPs.

Conversely, RobFaceNet stands out not just as a competitor but as a benchmark in

Table 5.6: Comparative Analysis of Face Recognition Models on the LFW Benchmark. The models are ordered based on the number of FLOPs. Results and the number of decimal points are reported as in the respective works. Our model is highlighted to underscore its proficient balance between high accuracy and computational efficiency in the challenging face verification task amidst real-world conditions encapsulated in the LFW benchmark. Results represent the accuracy (%), with higher values indicating better performance. The best performance in each category on each benchmark is emphasized in bold.

Model	**#FLOPs** (M)	**#Params.** (M)	**Size** (MB)	**LFW** (%)
ArcFace (ResNet100) [37]	24211	65.2	261.22	99.82
MobileFaceNetV1 [147]	1100	3.40	13.1	99.40
PocketNetM-256 [13]	1099.15	1.75	7.0	99.58
PocketNetM-128 [13]	1099.02	1.68	6.7	99.65
ShuffleFaceNet 2x [146]	1050	4.5	18.0	99.62
VarGFaceNet [239]	1022	5	20.0	**99.85**
AirFace [124]	1000	4.23	-	99.27
MobileFaceNet [147]	933.30	2	4.0	**99.70**
ProxylessFaceNAS [147]	900	3.20	12.5	99.20
MixFaceNet-M [11]	626.1	3.95	15.8	99.68
ShuffleMixFaceNet-M [11]	626.10	3.95	15.8	99.60
PocketNetS-256 [13]	587.22	0.99	3.9	99.66
PocketNetS-128 [13]	587.11	0.92	3.7	99.58
ShuffleFaceNet 1.5x [146]	577.5	2.60	10.5	99.67
MixFaceNet-S [11]	451.7	3.07	12.28	99.60
ShuffleMixFaceNet-S [11]	451.7	3.07	12.28	99.58
MobileFaceNets [23]	439.8	0.99	8.2	99.55
ShuffleFaceNet 1x [146]	275.8	1.40	5.6	99.45
MixFaceNet-XS [11]	161.9	1.04	4.2	99.60
ShuffleMixFaceNet-XS [11]	161.9	1.04	4.2	99.53
Ours	337.3	1.90	7.27	**99.75**

computational efficiency. With a verification accuracy of 99.75% on LFW, it closely matches the top-performing VarGFaceNet while achieving an impressive 67% reduction in computational complexity, with just 337M FLOPs.

This comparison highlights RobFaceNet as a model that effectively balances high-precision recognition with computational efficiency, demonstrating its potential as a practical solution for face recognition, particularly in resource-constrained environments.

5.5.5.2 Evaluation on Cross-Age Datasets

Amid the complexities of face appearance alterations that grow over time, cross-age variations stand as a particularly hard challenge within the FR community due to the myriad of subtle and nonlinear changes introduced by aging. To analyze these complexities, focused evaluations on cross-age FR are executed, utilizing the Cross-Age LFW (CALFW) [263] and AgeDB-30 [154] datasets, specifically chosen for their robust testing capabilities in navigating the nuanced challenges posed by age variance in FR tasks. Fig. 5.8 presents selected face images from these datasets, offering visual insights into the varied age-related face changes and the challenges they pose to recognition models, thereby emphasizing the need for efficient cross-age FR systems.

Like the original LFW dataset, CALFW defines an evaluation protocol divided into ten distinct subsets of image pairs, each subset with 300 positive and 300 negative pairs. Meanwhile, AgeDB-30, recognized as the most challenging group of the AgeDB [154], enforces a minimum age gap of 30 years. It formulates an age-invariant face verification protocol, divided into ten splits of face images, each incorporating 300 positive and 300 negative examples. We adhere to the 'Accuracy' metric for face verification evaluations across both datasets, aligning with the same benchmark established by LFW.

Table 5.7 presents a detailed comparison of the verification accuracy of lightweight networks against state-of-the-art results reported in the literature. A discerning exploration reveals our model clearly dominates the CALFW dataset, not only surpassing the deep learning model ArcFace-ReNet100 [37] and all lightweight models but also achieving this while preserving a reduced computational complexity. When examined against the challenging benchmark of the AgeDB-30 dataset, our model not only gained a peak of performance amongst lightweight models with computational complexity under 900M FLOPs but also firmly held its ground, showcasing strongly competitive results against both deep models and those exceeding 930M FLOPs in computational complexity. Specifically, our model achieved verification accuracies of 95.95% and 97.42% on the CALFW and AgeDB-30 datasets, respectively, with a computational cost of 337M FLOPs. In contrast, the very deep ArcFace-ReNet100 model recorded verification accuracies of 95.45% and 98.15% on the CALFW and AgeDB-30 datasets, respectively, but at a computational cost 71 times that of our model. Concerning the best results from lightweight models, PocketNetM-128 [13] achieved 95.67% on CALFW, while VarGFaceNet [239] recorded 98.15% on the AgeDB-30 dataset, each with computational complexities triple that of our model.

5.5.5.3 Evaluation on Cross-Pose Datasets

Evaluations on the cross-pose datasets underscore the significant effectiveness and resilience of our proposed RobFaceNet model, especially in addressing the intrinsic challenges associated with varying face poses. In this context, two datasets, specif-

Table 5.7: Comparative Analysis of Face Recognition Models on the CALFW and AgeDB-30 Datasets. The models are ordered based on the number of FLOPs. Results and the number of decimal points are reported as in the respective works. Our model notably outperforms competitors on CALFW and maintains commendable results on AgeDB-30, all while ensuring reduced computational complexity. Results represent the accuracy (%), with higher values indicating better performance. The best performance in each category on each benchmark is highlighted in bold.

Model	**#FLOPs** (M)	**#Params.** (M)	**Size** (MB)	**CALFW** (%)	**AgeDB-30** (%)
ArcFace (ResNet100) [37]	24211	65.2	261.22	95.45	98.15
MobileFaceNetV1 [147]	1100	3.40	13.1	94.47	96.40
PocketNetM-256 [13]	1099.15	1.75	7.0	95.63	97.17
PocketNetM-128 [13]	1099.02	1.68	6.7	**95.67**	96.78
ShuffleFaceNet 2x [146]	1050	4.5	18.0	-	97.28
VarGFaceNet [239]	1022	5	20.0	95.15	**98.15**
AirFace [124]	1000	4.23	-	-	93.25
MobileFaceNet [147]	933.30	2	4.0	95.20	**97.60**
ProxylessFaceNAS [147]	900	3.20	12.5	92.55	94.40
MixFaceNet-M [11]	626.1	3.95	15.8	-	97.05
ShuffleMixFaceNet-M [11]	626.10	3.95	15.8	**95.75**	96.98
PocketNetS-256 [13]	587.22	0.99	3.9	95.50	96.35
PocketNetS-128 [13]	587.11	0.92	3.7	95.48	96.10
ShuffleFaceNet 1.5x [146]	577.5	2.60	10.5	95.05	97.32
MixFaceNet-S [11]	451.7	3.07	12.28	-	96.63
ShuffleMixFaceNet-S [11]	451.7	3.07	12.28	95.67	97.05
MobileFaceNets [23]	439.8	0.99	8.2	95.20	96.07
ShuffleFaceNet 1x [146]	275.8	1.40	5.6	-	96.33
MixFaceNet-XS [11]	161.9	1.04	4.2	-	95.85
ShuffleMixFaceNet-XS [11]	161.9	1.04	4.2	94.93	95.61
Ours	337.3	1.90	7.27	**95.95**	**97.42**

ically CFP-FP [178] and Cross-Pose LFW (CPLFW) [262], have been employed for assessment. Fig. 5.8 for a visual exhibit of some face images extracted from these datasets, showcasing a spectrum of poses and angles intended to rigorously evaluate the robustness of FR models against cross-pose challenges. We also adhere to the 'Accuracy' metric for face verification evaluations across both datasets, aligning with the same benchmark established by LFW.

CFP-FP, originating from the Frontal-Profile (FP) face verification experiment of the CFP dataset [178], encompasses 3,500 face pairs, divided into ten subsets, and showcases a stark contrast between frontal and profile views, providing a challenging environment for assessing model robustness against pose variations. Within this

Table 5.8: Comparative Analysis of Face Recognition Models on the CPLFW and CFP-FP Datasets. The models are ordered based on the number of FLOPs. Results and the number of decimal points are reported as in the respective works. Our model notably outperforms competitors on CPLFW and maintains commendable results on CFP-FP, all while ensuring reduced computational complexity. Results represent the accuracy (%), with higher values indicating better performance. The best performance in each category on each benchmark is highlighted in bold.

Model	**#FLOPs** (M)	**#Params.** (M)	**Size** (MB)	**CPLFW** (%)	**CFP-FP** (%)
ArcFace (ResNet100) [37]	24211	65.2	261.22	94.20	95.60
MobileFaceNetV1 [147]	1100	3.40	13.1	87.17	95.80
PocketNetM-256 [13]	1099.15	1.75	7.0	**90.03**	95.66
PocketNetM-128 [13]	1099.02	1.68	6.7	90.00	95.07
ShuffleFaceNet 2x [146]	1050	4.5	18.0	-	97.56
VarGFaceNet [239]	1022	5	20.0	88.55	**98.50**
AirFace [124]	1000	4.23	-	-	94.11
MobileFaceNet [147]	933.30	2	4.0	89.22	96.90
ProxylessFaceNAS [147]	900	3.20	12.5	84.17	94.70
MixFaceNet-M [11]	626.1	3.95	15.8	-	-
ShuffleMixFaceNet-M [11]	626.10	3.95	15.8	**89.97**	94.96
PocketNetS-256 [13]	587.22	0.99	3.9	88.93	93.34
PocketNetS-128 [13]	587.11	0.92	3.7	89.63	94.21
ShuffleFaceNet 1.5x [146]	577.5	2.60	10.5	88.5	**97.26**
ShuffleMixFaceNet-S [11]	451.7	3.07	12.28	89.85	94.10
MobileFaceNets [23]	439.8	0.99	8.2	89.22	96.90
ShuffleFaceNet 1x [146]	275.8	1.40	5.6	-	96.04
ShuffleMixFaceNet-XS [11]	161.9	1.04	4.2	86.93	91.25
Ours	337.3	1.90	7.27	**92.23**	**97.63**

complex scenario, RobFaceNet achieved a remarkable verification accuracy of 97.63%, confirming its adept cross-pose recognition capabilities while maintaining a computational efficiency of 337M FLOPs.

Conversely, CPLFW, adhering to a structured evaluation protocol that mirrors the original LFW dataset, seeks to challenge face verification models with its disparate pose variations. Nevertheless, RobFaceNet unambiguously showcased its prowess by securing a stellar 92.23% verification accuracy, highlighting its robustness amidst diverse face orientations.

The quantitative comparison underscores the competitive advantage encapsulated by RobFaceNet. Specifically, when contrasted with the very deep ArcFace-ReNet100 model, which secured 94.20% and 95.60% accuracies on CPLFW and CFP-FP respec-

tively, RobFaceNet not only demonstrates compelling performance parity but also achieves this with a significantly reduced computational demand of 337M FLOPs, approximately 71 times less than its deep-model counterpart, thereby enhancing its practicality, especially in resource-limited deployments.

In comparison with lightweight models, RobFaceNet outperforms PocketNetM-256 [13], which secured a comparable accuracy of 90.03% on CPLFW. In addition, our model offers commendable competition to VarGFaceNet [239], which achieved a 98.50% accuracy on CFP-FP, all while maintaining a computational complexity of just one-third of these models.

In essence, RobFaceNet not only establishes a distinguished position as a proficient, resource-efficient FR model amidst pose variations but also sets a new benchmark. It melds robust recognition capabilities with resource-efficient deployment, thereby holding the potential to redefine the metrics for evaluating and deploying lightweight, highly accurate FR models in real-world scenarios. This highlights not just the theoretical capabilities of RobFaceNet, but also its practical utility across various applications and deployment contexts.

5.5.5.4 Evaluation on IJB-B and IJB-C Datasets

Navigating through the complex environments of the IARPA Janus Benchmark-B (IJB-B) and IJB-C datasets [149, 219], RobFaceNet demonstrates robust face verification capabilities amidst a wide variety of still images and video frames. The IJB-B dataset encapsulates data from 1,845 subjects, spread across 21,798 images and 55,026 video frames, while IJB-C expands this further, encompassing 3,531 subjects, 31,334 images, and a notable 117,542 video frames. Both benchmarks encapsulate myriad operational FR scenarios, establishing a complex evaluation benchmark to depict RobFaceNet's prowess clearly.

Subjected to the stringent 1:1 verification protocol within these comprehensive datasets, our model faced challenges against IJB-B's 12,115 templates (10,270 genuine, 8M impostor matches) and IJB-C's 23,124 templates (19,557 genuine, 15,639 impostor matches). The pivotal metric was the True Acceptance Rate (TAR) at varying False Acceptance Rate (FAR), providing a rigorous evaluation framework.

The evaluation results detailed in Table 5.9 indicate that our model, securing a commendable 93.86% TAR at 1e-4 FAR on the IJB-C dataset, with a mere 1.9M parameters and 337.3M FLOPs, substantiates its ability to efficiently and effectively verify faces. This capability is pivotal in real-world applications where computational resources may be limited, yet a high degree of accuracy in face verification remains paramount.

In a clear contrast, while MobileFaceNet [147] and VarGFaceNet [239] marginally outperform with a 94.7% TAR, they do so at the cost of significantly higher computational and parameter demands, 933.3M/1022M FLOPs and 2M/5M parameters,

Table 5.9: Comparative Analysis of Face Recognition Models on IJB-B and IJB-C Datasets. Models are arranged based on FLOP count, with results and decimal points reported as per original works. Our model, highlighted for emphasis, exemplifies an optimal balance between high performance and computational efficiency in the demanding task of face verification amidst the real-world scenarios encapsulated by the IJB-B and IJB-C benchmarks. Results represent the TAR at FAR 1e–4 (%), with higher values indicating better performance. The best performance in each category on each benchmark is emphasized in bold.

Model	**#FLOPs** (M)	**#Params.** (M)	**Size** (MB)	**IJB-B** (%)	**IJB-C** (%)
ArcFace (ResNet100) [37]	24211	65.2	261.22	89.90	92.10
MobileFaceNetV1 [147]	1100	3.40	13.1	92.00	93.90
PocketNetM-256 [13]	1099.15	1.75	7.0	90.74	92.70
PocketNetM-128 [13]	1099.02	1.68	6.7	90.63	92.63
ShuffleFaceNet 2x [146]	1050	4.5	18.0	-	-
VarGFaceNet [239]	1022	5	20.0	**92.90**	**94.70**
AirFace [124]	1000	4.23	-	-	-
MobileFaceNet [147]	933.30	2	4.0	**92.80**	**94.70**
ProxylessFaceNAS [147]	900	3.20	12.5	87.10	89.70
MixFaceNet-M [11]	626.1	3.95	15.8	91.55	93.42
ShuffleMixFaceNet-M [11]	626.10	3.95	15.8	91.47	93.5
PocketNetS-256 [13]	587.22	0.99	3.9	89.31	91.33
PocketNetS-128 [13]	587.11	0.92	3.7	89.44	91.62
ShuffleFaceNet 1.5x [146]	577.5	2.60	10.5	92.30	94.30
MixFaceNet-S [11]	451.7	3.07	12.28	90.17	92.30
ShuffleMixFaceNet-S [11]	451.7	3.07	12.28	90.94	93.08
MobileFaceNets [23]	439.8	0.99	8.2	-	-
ShuffleFaceNet 1x [146]	275.8	1.40	5.6	-	-
MixFaceNet-XS [11]	161.9	1.04	4.2	88.48	90.73
ShuffleMixFaceNet-XS [11]	161.9	1.04	4.2	87.86	90.43
Ours	337.3	1.90	7.27	**92.08**	**93.86**

respectively. This comparison illuminates the computational efficiency aspect of RobFaceNet without a significant compromise on the performance, marking it as a viable option in resource-constrained environments.

When RobFaceNet is contrasted against other lightweight models, its performance becomes particularly notable, demonstrating not only the capability to deliver comparable results but also its capability to outperform its counterparts. It not only holds its ground in comparison with contemporaries but, crucially, does so with a finely

Table 5.10: Comparison of very deep SOTA FR models, SOTA FR models with computation complexity under 500M FLOPs, and our proposed RobFaceNet. The best performance in each category for each benchmark is highlighted in bold

Model	#FLOPs (M)	#Params. (M)	Size (MB)	LFW (%)	Cross-Age CA-LFW (%)	Cross-Age AgeDB-30 (%)	Cross-Pose CP-LFW (%)	Cross-Pose CFP-FP (%)	IJB IJB-B (%)	IJB IJB-C (%)	MegaFace Rank-1 (%)	MegaFace Ver. (%)
FaceNet [37]	451.7	3.07	-	99.63	-	-	-	-	-	-	70.49	86.47
SphereFace [136]	24211	65.2	261.22	99.42	90.30	92.88	81.40	-	-	-	72.73	85.56
CosFace [204]	24211	65.2	261.22	99.73	95.76	98.11	92.19	98.12	**94.80**	**96.36**	80.56	96.56
ArcFace [37]	24211	65.2	261.22	**99.82**	95.45	**98.15**	92.08	**98.40**	94.20	95.60	**81.03**	**96.98**
RobFaceNet (Ours)	337.3	1.90	7.27	99.75	**95.95**	97.42	**92.23**	97.63	92.08	93.86	80.20	96.38
MixFaceNet-S [11]	451.7	3.07	12.28	99.60	-	96.63	-	-	90.17	92.30	76.49	92.23
ShuffleMixFaceNet-S [11]	451.7	3.07	12.28	99.58	95.67	97.05	89.85	94.10	90.94	93.08	77.41	93.60
MobileFaceNets [23]	439.8	0.99	8.2	99.55	95.20	96.07	89.22	96.90	-	-	-	90.16
MixFaceNet-XS [11]	161.9	1.04	4.2	99.60	-	95.85	-	-	88.48	90.73	74.18	89.40
ShuffleMixFaceNet-XS [11]	161.9	1.04	4.2	99.53	94.93	95.61	86.93	91.25	87.86	90.43	73.85	89.24
GhostFaceNetV2-2 [3]	76.51	6.84	13.66	99.71	95.70	96.55	89.58	93.07	91.76	93.03	79.31	95.21
RobFaceNet (Ours)	337.3	1.90	7.27	**99.75**	**95.95**	**97.42**	**92.23**	**97.63**	**92.08**	**93.86**	**80.20**	**96.38**

balanced trade-off between computational efficiency and verification accuracy.

This evaluation underscores RobFaceNet's adeptness in balancing efficiency and accuracy, confirming its prominence as a powerful model in real-world FR tasks and highlighting its promising utility in future resource-conscious applications.

5.6 Discussions

In this chapter, we introduced a novel lightweight, efficient, and robust neural network explicitly tailored for FR tasks. Unlike existing lightweight FR networks such as MobileFaceNets [23], VarGFaceNet [239], and PocketNet [13], our proposed model, RobFaceNet, leverages both low-level and high-level features to enable the extraction of diverse and comprehensive feature information. This approach results in more robust and accurate FR performance across various conditions, including differences in lighting, poses, and occlusions.

Furthermore, RobFaceNet has been designed with a focus on efficiency without compromising performance. This is achieved through the careful selection of architectural components and attention mechanisms, allowing RobFaceNet to achieve competitive performance while maintaining low computational complexity.

To further demonstrate the effectiveness of the proposed network, we recorded the verification performance results, measured in terms of accuracy, on several popular benchmark datasets. Table 5.10 compares the performance of RobFaceNet against the SOTA deep FR baseline models and lightweight FR models ($<$ 500M FLOPs). The results indicate that our proposed model outperforms the lightweight FR models in all evaluation datasets, highlighting the significant improvement achieved by incorporating multi-feature and attention mechanisms.

Moreover, our proposed model outperforms the very deep baseline models on two datasets, namely CA-LFW and CP-LFW, and achieves comparable results on the remaining benchmarks. This is accomplished with a more lightweight model complexity.

For instance, in terms of computation cost, RobFaceNet has 337M FLOPs compared to ArcFace's 24211M, with only 3% of the parameters.

In addition to its performance benefits, RobFaceNet offers practical advantages by effectively addressing the limitations of the traditional lightweight networks and allowing for more reliable and accurate recognition.

A limitation of this work is that the study mainly focused on standard face images and did not extend to low-resolution facial images. The lack of information in low-resolution images poses a challenge for effective recognition compared to standard face images. This limitation underscores the need for further research in developing network architectures that can effectively bridge the representation gap between low-resolution images and their high-resolution counterparts. Furthermore, although RobFaceNet has demonstrated robust performance on the cross-pose CP-LFW dataset, there remains scope for improvement in this domain. Future research opportunities could explore alternative network architectures capable of extracting more relevant features for cross-pose FR, thereby enhancing the model's ability to handle variations in pose.

5.7 Summary

In the pursuit of efficient and effective FR solutions for HRI applications, this chapter introduces RobFaceNet, a novel lightweight, efficient, and robust network specifically designed for FR. Despite its simplicity, RobFaceNet excels in both accuracy and computational efficiency. For example, RobFaceNet achieves 95.95% and 92.23% accuracy on the CA-LFW and CP-LFW datasets, respectively, compared to 95.45% and 92.08% for the much deeper ArcFace model. Simultaneously, RobFaceNet maintains a lighter model complexity, with only 337M FLOPs, a 67% reduction compared to ArcFace's 24,211M FLOPs, and just 3% of the parameters.

When compared to other lightweight FR models (less than 500M FLOPs), RobFaceNet outperforms state-of-the-art models across all evaluation datasets. It achieves 99.75%, 97.42%, and 97.63% accuracy on the LFW, AgeDB-30, and CFP-FP datasets, respectively, while MobileFaceNets [23] achieves 99.55%, 96.07%, and 96.9% accuracy at a higher computational cost of 439.8M FLOPs.

The architecture of RobFaceNet incorporates a novel multi-feature approach, which leverages features from various network levels, and utilizes a modified h-swish activation function to balance computational cost with performance. Additionally, the attention-enhanced bottleneck improves the network's ability to prioritize crucial features, enhancing its FR capabilities. Through extensive experimentation on comprehensive public face verification benchmarks, we demonstrate RobFaceNet's competitive performance against deeper FR networks. These results underscore RobFaceNet as a robust and efficient solution for FR tasks, particularly in dynamic HRI environments where real-time processing and interpretation of facial data are crucial for meaningful interactions.

6 Conclusions

6.1 Summary of Thesis Achievements

In this thesis, we undertook a transformative journey, revisiting the intersections of artificial intelligence, robotics, and face recognition (FR). At the core of our exploration was the realm of human-robot interaction (HRI), a rapidly advancing field that binds humans and robots in complex patterns of collaboration, companionship, and connection. Our pursuit ventured beyond the mere identification of faces; it sought to harness the power of FR to enhance HRI, especially in real-world settings characterized by unpredictable variables. From understanding the importance of FR in daily life, such as unlocking smartphones and supporting security systems, to dealing with the challenges of 'in-the-wild' scenarios, full of variable lighting, occlusions, and diverse facial expressions, **our ambition was clear: develop a robust and efficient FR system tailored for HRI in-the-wild scenarios**. As we reflect on the research undertaken, let us recap the milestones achieved in each chapter, understanding the significance of our contributions and the path forward.

In **Chapter 3**, the main objective revolves around enhancing the capabilities of HRI. Recognizing the challenges posed by FR in real-time HRI, most notably, the need for fast and efficient processing with high precision. We leveraged the power of convolutional neural networks (CNNs), particularly lightweight variants to design and develop our proposed FR framework. The driving force behind this choice was to ensure rapid processing while maintaining accuracy, two attributes crucial for seamless interactions between humans and robots. Central to our FR system is the integration of the state-of-the-art ArcFace loss function, which enhances recognition capabilities. This, combined with the RetinaFace method for face detection and a specially designed online real-time face tracker, equips our system to handle common challenges like varying illumination conditions, various head poses, and occlusions. However, our framework offers more than just recognition capabilities. By integrating a face tracker designed to merge tracking data with recognized identities, we have ensured that the system is adept at recognizing faces in unconstrained settings. This merger promises a boost in recognition accuracy and processing speed. To ascertain the real-world efficacy of our framework, we introduced it to our HRI system, aptly named "RoSA". A total of 11 participants took part in real-time interactions with the robot, and their experiences offered crucial insights into the system's performance. Additionally, by evaluating our system using videos from the Wizard-of-Oz study, we further bolstered the reliability of our findings, yielding an impressive 25% improvement in real-time

recognition. While the research highlighted impressive results, it also identified challenges, such as system lags in dense face scenes, which form the cornerstone of our future improvements. The conclusion of this chapter sets the stage for future research, where we aim to further fine-tune the architecture and loss function to achieve seamless and natural HRI.

In **Chapter 4**, we highlight the profound impact of margin-based softmax loss functions on current achievements. Historically, many methodologies in this domain have relied on the presumption of a static, constant margin. However, such an approach might not be the most effective, especially when handling heterogeneous real-world data. To address this gap, we proposed the JAMsFace loss function. Contrary to existing methodologies such as CosFace [204], ArcFace [37], and AdaptiveFace [132], JAMsFace introduces greater flexibility in margins based on class distribution, setting it apart from existing models. By concurrently harnessing joint adaptive margins in both angle and cosine spaces, JAMsFace ambitiously seeks to reduce intra-class variance while enhancing inter-class variance. This duality ensures the model's ability to extract significantly discriminative features, thereby enhancing FR accuracy. The effectiveness of JAMsFace does not terminate at feature discernibility. This loss function also addresses the longstanding challenge of class imbalance in FR, an issue that many contemporary techniques have faced. The utilization of joint adaptive margins champions not just improved face representation but an overall improvement in performance, especially in the scenarios of unconstrained settings. This flexibility is useful in providing robustness against challenges like class imbalance, unconstrained environments, and the task of generalizing to unseen classes. We undertook an extensive empirical evaluation to ascertain the capabilities of JAMsFace. The results were profound, with our method redefining benchmarks across datasets like LFW [90], CFP-FP [178], CPLFW [262], CALFW [263], and AgeDB-30 [154]. Moreover, JAMsFace showcased its robustness on more challenging datasets like IJB-B [219], IJB-C [149], and MegaFace [104]. Beyond empirical success, this chapter underscores a paradigm shift in how we perceive loss function design for FR. It advocates a move away from rigid structures, emphasizing the necessity for more adaptive, responsive methodologies.

In **Chapter 5**, we delved into network design, presenting RobFaceNet, a novel lightweight, efficient, and robust network designed specifically for FR. It demonstrates remarkable performance in both accuracy and computational efficiency. A keystone in RobFaceNet's design is its multi-feature modality. Our approach precisely intertwines features from various network layers, navigating beyond mere fusion to extract a rich and informative feature pool. Consequently, RobFaceNet adeptly extracts, processes, and leverages a broad, insightful feature scope that is especially relevant for FR tasks, where recognizing and understanding fine-grained features and subtle variations are imperative for achieving high recognition precision. However, the innovation does not end here. We have incorporated the modified h-swish activation function, a strategic move that decreases computational demands while preserving high-performance stan-

dards. Furthermore, we enhanced RobFaceNet by combining it with an innovative attention-based bottleneck, aimed to highlight the most distinct facial regions and feature channels. This enhancement employs an advanced inverted bottleneck, integrating either a CA [82] or SE [86] attention module, after the depthwise convolution layer. Such an integration improves the network's ability to interpret both channel and spatial features, thereby boosting its ability to discriminate between different facial characteristics. The positioning of attention modules throughout the network enables RobFaceNet to capture complex facial attributes early on with the CA module and refine feature representation in later stages using the SE module. This nuanced approach not only elevates the network's FR capabilities but also retains computational efficiency. To assess the effectiveness of RobFaceNet, we conducted a comprehensive experimental evaluation comparing it to contemporary lightweight FR models, using various face verification benchmarks. The findings were promising, as RobFaceNet not only met but often surpassed the performance criteria of deeper networks. These results provide compelling evidence that lightweight models like RobFaceNet can effectively address the complexities of real-world applications despite their compact nature.

As illustrated in Figure 6.1, our proposed face recognition system integrates RobFaceNet as the feature extraction network with the JAMsFace loss function to optimize facial feature representation. Coupled with the recognition-tracking approach, this system achieves a robust and efficient solution for face recognition in unconstrained, real-world environments, such as human-robot interaction (HRI) scenarios.

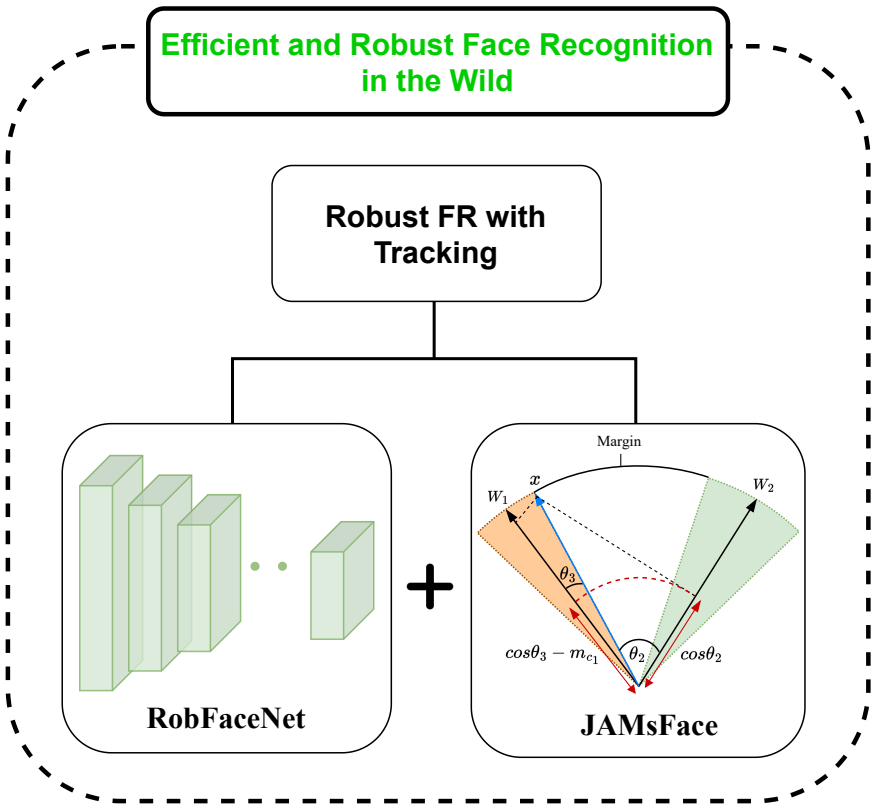


Figure 6.1: Proposed Complete Face Recognition System for Human-Robot Interaction.

6.2 Future Directions

There are plenty of possible directions to explore to extend the research in this thesis.

Pose-invariant face recognition. While there has been significant advancement in the field of FR, with many solutions now integrated into everyday applications, a glaring limitation remains. These solutions mostly excel at recognizing near-frontal facial orientations. As highlighted in [178], a shift from frontal-frontal to frontal-profile verification sees many existing methods experience a performance drop of over 10%. This underscores the challenge of achieving truly pose-invariant FR, which is still an unresolved problem and provides opportunities for further research. Our future endeavors will delve into the theoretical aspects of facial orientations, aiming to uncover the inherent connections between images of an individual from diverse angles. Based on the theoretical support, our goal is to conceptualize and train a novel network adept at extracting pose-invariant features.

Data-free face recognition. Deep FR's success relies heavily on the vast annotated data it employs. However, the online publishing of these datasets presents a spot of privacy concerns. This has led to the withdrawal of pivotal datasets, such as MegaFace, from the web. Similarly, while many social media giants possess in-house datasets that enhance performance, they often withhold them, cautious of assorted concerns. A promising solution is "deep inversion", which involves retrieving the original training image from a model's features. Though this approach has found some traction in general object classification, its applicability to FR remains largely unexplored, mainly due to the optimization complexities involved. Should we overcome this problem, it would enable the extraction of datasets directly from pre-trained models, facilitating their use in new model training. Thus, exploring face inversion could pave the way for a partnership between academia and the tech industry in advancing FR.

Low-resolution facial images face recognition. Real-world scenarios often result in facial images captured at distances that yield a much lower resolution compared to those obtained under controlled conditions. This holds true even with high-definition 1080P cameras, where facial segments might measure as small as 20×20 pixels. These small image patches sparsely contain very little information. Compared with standard face images, the information insufficiency in low-resolution hinders effective classification. Conventional algorithms falter when tasked with comparing these low-resolution images to their high-resolution counterparts, given the absence of shared feature representation. As such, the quest for a novel network architecture that adeptly bridges this representation gap in low-resolution facial recognition datasets is a compelling avenue for exploration.

Bibliography

[1] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006.

[3] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 2023.

[4] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021.

[5] J. G. Allen, R. Y. Xu, J. S. Jin, et al. Object tracking using camshift algorithm and multiple quantized feature spaces. In *ACM international conference proceeding series*, volume 100, pages 3–7. Citeseer, 2004.

[6] F. Alonso-Fernandez, K. Hernandez-Diaz, J. M. Buades Rubio, and J. Bigun. Squeezerfacenet: Reducing a small face recognition cnn even more via filter pruning. In *VIII International Workshop on Artificial Intelligence and Pattern Recognition, IWAIPR*, 2023.

[7] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[8] S. Bashbaghi, E. Granger, R. Sabourin, and M. Parchami. Deep learning architectures for face recognition in video surveillance. *Deep Learning in Object Detection and Recognition*, pages 133–154, 2019.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[11] 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)*, pages 1–8. IEEE, 2021.

[12] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2022.

[13] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[14] S. C. Brubaker, J. Wu, J. Sun, M. D. Mullin, and J. M. Rehg. On the design of cascades of boosted ensembles for face detection. *International journal of computer vision*, 77:65–86, 2008.

[15] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

[16] J. Cao, Y. Li, and Z. Zhang. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410. IEEE, 2018.

[17] 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)*, pages 67–74. IEEE, 2018.

[18] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International journal of computer vision*, 107:177–190, 2014.

[19] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2707–2714. IEEE, 2010.

[20] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE transactions on image processing*, 24(12):5017–5032, 2015.

[21] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 122–138. Springer, 2016.

[22] L. Chen, H. Su, and Q. Ji. Face alignment with kernel density deep neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate

real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.

[24] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. In *International conference on machine learning*, pages 2285–2294. PMLR, 2015.

[25] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[26] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[27] V. Contreras-Gonzalez, V. H. Diaz-Ramirez, and R. Juarez-Salazar. Facial landmark detection and tracking with dynamically adaptive matched filters. *Journal of Electronic Imaging*, 29(3):033004–033004, 2020.

[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 484–498. Springer, 1998.

[29] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[30] D. Cristinacce, T. F. Cootes, et al. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1(2), page 3. Citeseer, 2006.

[31] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *Image and Vision Computing*, 79:35–48, 2018.

[32] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

[33] 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)*, volume 1, pages 886–893. Ieee, 2005.

[34] A. Dapogny, K. Bailly, and M. Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[35] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[36] 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 (CVPR)*, June 2020.

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[38] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[39] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019.

[40] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 60–68, 2017.

[41] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE transactions on Multimedia*, 17(11):2049–2058, 2015.

[42] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1002–1014, 2017.

[43] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE, 2010.

[44] S. Dong, P. Wang, and K. Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[45] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[46] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018.

[47] N. Doulamis and A. Voulodimos. Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 318–323. IEEE, 2016.

[48] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022.

[49] C. N. Duong, K. G. Quach, I. Jalata, N. Le, and K. Luu. Mobiface: A lightweight

deep learning face recognition on mobile devices. In *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–6. IEEE, 2019.

[50] A. P. Fard and M. H. Mahoor. Facial landmark points detection using knowledge distillation-based neural networks. *Computer Vision and Image Understanding*, 215:103316, 2022.

[51] S. Favelle and S. Palmisano. View specific generalisation effects in face recognition: Front and yaw comparison views are better than pitch. *PloS one*, 13(12):e0209927, 2018.

[52] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu. Random cascaded-regression copse for robust facial landmark detection. *IEEE Signal Processing Letters*, 22(1):76–80, 2014.

[53] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018.

[54] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2490, 2017.

[55] M.-A. Fiedler, P. Werner, A. Khalifa, and A. Al-Hamadi. Sfpd: Simultaneous face and person detection in real-time for human–robot interaction. *Sensors*, 21(17):5918, 2021.

[56] D. Forsyth. Object detection with discriminatively trained part-based models. *Computer*, 47(02):6–7, 2014.

[57] M. T. H. Fuad, A. A. Fime, D. Sikder, M. A. R. Iftee, J. Rabbi, M. S. Al-Rakhami, A. Gumaei, O. Sen, M. Fuad, and M. N. Islam. Recent advances in deep learning techniques for face recognition. *IEEE Access*, 9:99112–99142, 2021.

[58] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[59] I. Gogić, J. Ahlberg, and I. S. Pandžić. Regression-based methods for face alignment: A survey. *Signal Processing*, 178:107755, 2021.

[60] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4852–4861, 2019.

[61] Google. Webrtc. `https://webrtc.org`, 2020.

[62] P. Grother, P. Grother, M. Ngan, and K. Hanaoka. Face recognition vendor test

(frvt) part 2: Identification, 2019.

[63] P. Grother, R. J. Micheals, and P. J. Phillips. Face recognition vendor test 2002 performance metrics. In *International conference on audio-and video-based biometric person authentication*, pages 937–945. Springer, 2003.

[64] P. J. Grother, P. J. Grother, M. Ngan, and K. Hanaoka. *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology, 2014.

[65] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE 12th international conference on computer vision*, pages 498–505. IEEE, 2009.

[66] M. Günther, P. Hu, C. Herrmann, C.-H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, et al. Unconstrained face detection and open-set face recognition challenge. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 697–706. IEEE, 2017.

[67] G. Guo and N. Zhang. A survey on deep learning based face recognition. *Computer vision and image understanding*, 189:102805, 2019.

[68] J. Guo, J. Deng, N. Xue, and S. Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. *arXiv preprint arXiv:1812.01936*, 2018.

[69] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

[70] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019.

[71] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[72] Z. Haiyang. Image preprocessing methods in face recognition. In *2011 Symposium on Photonics and Optoelectronics (SOPO)*, 2011.

[73] H. Han, S. Shan, X. Chen, and W. Gao. A comparative study on illumination preprocessing in face recognition. *Pattern Recognition*, 46(6):1691–1699, 2013.

[74] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[75] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[76] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[77] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[78] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[79] 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*, pages 770–778, 2016.

[80] T. Hempel. RoSA: Cube Detector.

[81] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014.

[82] Q. Hou, D. Zhou, and J. Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[83] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[84] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[85] G.-S. J. Hsu, H.-Y. Wu, and M. H. Yap. A comprehensive study on loss functions for cross-factor face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 826–827, 2020.

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

[87] W.-C. Hu, C.-H. Chen, T.-Y. Chen, D.-Y. Huang, and Z.-C. Wu. Moving object detection and tracking from video captured by moving camera. *Journal of Visual Communication and Image Representation*, 30:164–180, 2015.

[88] D.-Y. Huang, C.-H. Chen, T.-Y. Chen, W.-C. Hu, Z.-B. Guo, and C.-K. Wen. High-efficiency face detection and tracking method for numerous pedestrians through face candidate generation. *Multimedia Tools and Applications*, 80:1247–1272, 2021.

[89] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2752–2761, 2018.

[90] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[91] X. Huang, W. Deng, H. Shen, X. Zhang, and J. Ye. Propagationnet: Propagate points to curve to learn structure information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7265–7274, 2020.

[92] 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*, pages 5901–5910, 2020.

[93] Y. Hwang, M.-H. Jeong, S.-R. Oh, and C. Yoon. Adaptive mean shift based face tracking by coupled support map. *International Journal of Fuzzy Logic and Intelligent Systems*, 17(2):114–120, 2017.

[94] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[95] N. S. Irjanto and N. Surantha. Home security system with face recognition based on convolutional neural network. *International Journal of Advanced Computer Science and Applications*, 11(11), 2020.

[96] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

[97] J. Jiao, W. Liu, Y. Mo, J. Jiao, Z. Deng, and X. Chen. Dyn-arcface: dynamic additive angular margin loss for deep face recognition. *Multimedia Tools and Applications*, 80(17):25741–25756, 2021.

[98] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu. A survey of deep learning-based object detection. *IEEE access*, 7:128837–128868, 2019.

[99] B.-N. Kang, Y. Kim, B. Jun, and D. Kim. Attentional feature-pair relation networks for accurate face recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5472–5481, 2019.

[100] M. Karpagam, R. B. Jeyavathana, S. K. Chinnappan, K. Kanimozhi, and M. Sambath. A novel face recognition model for fighting against human trafficking in surveillance videos and rescuing victims. *Soft Computing*, pages 1–16, 2022.

[101] H. Kavalionak, C. Gennaro, G. Amato, C. Vairo, C. Perciante, C. Meghini, and F. Falchi. Distributed video surveillance using smart cameras. *Journal of Grid Computing*, 17:59–77, 2019.

[102] J. D. Kelleher. *Deep learning*. MIT press, 2019.

[103] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360:

Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019.

[104] 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*, pages 4873–4882, 2016.

[105] N. Ketkar and N. Ketkar. Stochastic gradient descent. *Deep learning with Python: A hands-on introduction*, pages 113–132, 2017.

[106] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi. Face recognition and tracking framework for human-robot interaction. *Applied Sciences*, 12(11):5568, 2022.

[107] A. Khalifa and A. Al-Hamadi. A survey on loss functions for deep face recognition network. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–7. IEEE, 2021.

[108] A. Khalifa and A. Al-Hamadi. Jamsface: joint adaptive margins loss for deep face recognition. *Neural Computing and Applications*, 35(26):19025–19037, 2023.

[109] K. Kim, Z. Yang, I. Masi, R. Nevatia, and G. Medioni. Face and body association for video-based face recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 39–48. IEEE, 2018.

[110] 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*, pages 18750–18759, 2022.

[111] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[112] Y. Kim, W. Park, M.-C. Roh, and J. Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020.

[113] S. Koley, H. Roy, S. Dhar, and D. Bhattacharjee. Illumination invariant face recognition using fused cross lattice pattern of phase congruency (fclppc). *Information Sciences*, 584:633–648, 2022.

[114] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[115] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[116] A. Kumar and R. Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 430–439, 2018.

[117] P. M. Kumar, U. Gandhi, R. Varatharajan, G. Manogaran, and T. Vadivel. Intelligent face recognition and navigation system using neural learning for smart security in internet of things. *Cluster Computing*, 22:7733–7744, 2019.

[118] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[119] Z. Lei, M. Pietikäinen, and S. Z. Li. Learning discriminant face descriptor. *IEEE Transactions on pattern analysis and machine intelligence*, 36(2):289–302, 2013.

[120] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[121] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[122] L. Li, X. Mu, S. Li, and H. Peng. A review of face recognition technology. *IEEE Access*, 8:139110–139120, 2020.

[123] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[124] X. Li, F. Wang, Q. Hu, and C. Leng. Airface: Lightweight and efficient model for face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[125] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):211–223, 2015.

[126] G. Ligorio and A. M. Sabatini. A novel kalman filter for human motion tracking with an inertial-based dynamic inclinometer. *IEEE Transactions on Biomedical Engineering*, 62(8):2033–2043, 2015.

[127] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[128] B. Liu, W. Deng, Y. Zhong, M. Wang, J. Hu, X. Tao, and Y. Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10061, 2019.

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

Image processing, 11(4):467–476, 2002.

[130] F. Liu, C. Gong, X. Huang, T. Zhou, J. Yang, and D. Tao. Robust visual tracking revisited: From correlation filter to template matching. *IEEE Transactions on Image Processing*, 27(6):2777–2790, 2018.

[131] H. Liu, J. Lu, J. Feng, and J. Zhou. Two-stream transformer networks for video-based face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2546–2554, 2017.

[132] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019.

[133] J. Liu, H. Qin, Y. Wu, and D. Liang. Anchorface: Boosting tar@ far for practical face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(2), pages 1711–1719, 2022.

[134] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[135] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[136] 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*, pages 212–220, 2017.

[137] 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*, pages 507–516, 2016.

[138] Y. Liu, H. Li, and X. Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv preprint arXiv:1710.00870*, 2017.

[139] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[140] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[141] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[142] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[143] P. Luan, V. Huynh, and T. Tuan Anh. Facial expression recognition using residual masking network. In *IEEE 25th International Conference on Pattern*

Recognition, pages 4513–4519, 2020.

[144] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017.

[145] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[146] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[147] Y. Martinez-Diaz, M. Nicolas-Diaz, H. Mendez-Vazquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, pages 1–44, 2021.

[148] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018.

[149] 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)*, pages 158–165. IEEE, 2018.

[150] 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*, pages 14225–14234, 2021.

[151] D. Merget, M. Rock, and G. Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[152] S. Minaee, P. Luo, Z. Lin, and K. Bowyer. Going deeper into face detection: A survey. *arXiv preprint arXiv:2103.14983*, 2021.

[153] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. In *Proceedings third IEEE international conference on automatic face and gesture recognition*, pages 30–35. IEEE, 1998.

[154] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*,

pages 51–59, 2017.

[155] P. Musa, F. Al Rafi, and M. Lamsani. A review: Contrast-limited adaptive histogram equalization (clahe) methods to help the application of face recognition. In *2018 third international conference on informatics and computing (ICIC)*, pages 1–6. IEEE, 2018.

[156] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.

[157] M. Najibi, B. Singh, and L. S. Davis. Fa-rpn: Floating region proposals for face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2019.

[158] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2017.

[159] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[160] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

[161] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[162] C. Oinar, B. M. Le, and S. S. Woo. Kappaface: Adaptive additive angular margin loss for deep face recognition. *arXiv preprint arXiv:2201.07394*, 2022.

[163] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[164] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[165] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[166] Picovoice. Porcupine. `https://github.com/Picovoice/porcupine`, 2020.

[167] Y. Qian, W. Deng, and J. Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019.

[168] H. Qin, J. Yan, X. Li, and X. Hu. Joint training of cascaded cnn for face

detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3456–3465, 2016.

[169] L. Ramos and B. Morales. Swiftface: real-time face detection. *arXiv preprint arXiv:2009.13743*, 2020.

[170] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017.

[171] RASA. Rasa. `https://rasa.com/open-source/`, 2020.

[172] M. Rashid, X. Gu, and Y. Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6894–6903, 2017.

[173] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[174] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[175] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2016.

[176] S. Sankaranarayanan, A. Alavi, and R. Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[177] 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*, pages 815–823, 2015.

[178] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[179] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[180] D. Strazdas, J. Hintz, A.-M. Felßberg, and A. Al-Hamadi. Robots and wizards: An investigation into natural human–robot interaction. *IEEE Access*, 8:207635–207642, 2020.

[181] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi. Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction. *Sensors*, 22(3):923, 2022.

[182] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation

by joint identification-verification. *Advances in neural information processing systems*, 27, 2014.

[183] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[184] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[185] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Proceedings of the IEEE international conference on computer vision*, pages 1489–1496, 2013.

[186] 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*, pages 1891–1898, 2014.

[187] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[188] Y. Sun, X. Wang, and X. Tang. Sparsifying neural network connections for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4856–4864, 2016.

[189] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

[190] 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*, pages 1–9, 2015.

[191] 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*, pages 1701–1708, 2014.

[192] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[193] M. Tan and Q. V. Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[194] A. Tewari, M. Zollhoefer, F. Bernard, P. Garrido, H. Kim, P. Perez, and C. Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):357–370, 2018.

[195] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[196] N. M. Tun, A. I. Gavrilov, and N. L. Tun. Facial image denoising using convolutional autoencoder network. In *2020 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pages 1–5. IEEE, 2020.

[197] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[198] R. Valle, J. M. Buenaposada, A. Valdes, and L. Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[199] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan. Mobileone: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7907–7917, 2023.

[200] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[201] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[202] 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*, pages 1041–1049, 2017.

[203] H. Wang, Z. Li, X. Ji, and Y. Wang. Face r-cnn. *arXiv preprint arXiv:1706.01061*, 2017.

[204] 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*, pages 5265–5274, 2018.

[205] J. Wang, Y. Yuan, and G. Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017.

[206] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[207] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[208] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. *Advances in neural information processing systems*, 26, 2013.

[209] Q. Wang and G. Guo. Ls-cnn: Characterizing local patches at multiple scales

for face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1640–1653, 2019.

[210] 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*, volume 34(7), pages 12241–12248, 2020.

[211] X. Wang, L. Bo, and L. Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[212] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*, 2017.

[213] D. Wanyonyi and T. Celik. Open-source face recognition frameworks: A review of the landscape. *IEEE Access*, 10:50601–50623, 2022.

[214] H. Wei, P. Lu, and Y. Wei. Balanced alignment for face recognition: A joint learning approach. *arXiv preprint arXiv:2003.10168*, 2020.

[215] S. Wei and R. Jianxin. Real-time tracking of non-rigid objects. In *Proceedings of the 2016 International Conference on Communication and Information Systems*, pages 11–15, 2016.

[216] Y. Wen, W. Liu, A. Weller, B. Raj, and R. Singh. Sphereface2: Binary classification is all you need for deep face recognition. In *International Conference on Learning Representations*, 2022.

[217] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[218] R. Weng, J. Lu, Y.-P. Tan, and J. Zhou. Learning cascaded deep auto-encoder networks for face alignment. *IEEE Transactions on Multimedia*, 18(10):2066–2078, 2016.

[219] 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*, pages 90–98, 2017.

[220] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[221] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.

[222] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 81–88. IEEE, June 2011.

[223] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[224] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008.

[225] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4820–4828, 2016.

[226] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[227] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[228] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[229] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[230] Y. Wu, H. Liu, J. Li, and Y. Fu. Deep face recognition with center invariant loss. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 408–414, 2017.

[231] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 57–72. Springer, 2016.

[232] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 782–797, 2018.

[233] W. Xie and A. Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018.

[234] Y. Xiong, Z. Zhou, Y. Dou, and Z. Su. Gaussian vector: An efficient solution for facial landmark detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[235] X. Xu and I. A. Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 642–649. IEEE, 2017.

[236] X. Xu, Q. Meng, Y. Qin, J. Guo, C. Zhao, F. Zhou, and Z. Lei. Searching

for alignment in face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(4), pages 3065–3073, 2021.

[237] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He. Centerface: joint face detection and alignment using face as point. *Scientific Programming*, 2020:1–8, 2020.

[238] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2504, 2014.

[239] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[240] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 79–87, 2017.

[241] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[242] S. Yang, Y. Xiong, C. C. Loy, and X. Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017.

[243] D. Yashunin, T. Baydasov, and R. Vlasov. Maskface: multi-task face and landmark detector. *arXiv preprint arXiv:2005.09412*, 2020.

[244] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[245] Y. Yoo, D. Han, and S. Yun. Extd: Extremely tiny face detector via iterative filter reuse. *arXiv preprint arXiv:1906.06579*, 2019.

[246] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas. Safety bounds in human robot interaction: A survey. *Safety science*, 127:104667, 2020.

[247] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, 126:103514, 2022.

[248] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al. On rectified linear units for speech processing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3517–3521. IEEE, 2013.

[249] D. Zeng, F. Zhao, S. Ge, and W. Shen. Fast cascade face detection with pyramid network. *Pattern Recognition Letters*, 119:180–186, 2019.

[250] C. Zhang, X. Xu, and D. Tu. Face detection using improved faster rcnn. *arXiv preprint arXiv:1802.02142*, 2018.

[251] H. Zhang and M. Xu. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Transactions on Multimedia*, 23:2033–2044, 2020.

[252] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 1–16. Springer, 2014.

[253] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[254] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *2011 International conference on computer vision*, pages 471–478. IEEE, 2011.

[255] L. Zhang, C. Bao, and K. Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021.

[256] Q. Zhang, J. Li, M. Yao, L. Song, H. Zhou, Z. Li, W. Meng, X. Zhang, and G. Wang. Vargnet: Variable group convolutional neural network for efficient embedded computing. *arXiv preprint arXiv:1907.05653*, 2019.

[257] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.

[258] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.

[259] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[260] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE international conference on computer vision*, pages 5409–5418, 2017.

[261] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.

[262] 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*, 5:7, 2018.

[263] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for study-

ing cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[264] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He. A survey of deep facial attribute analysis. *International Journal of Computer Vision*, 128:2002–2034, 2020.

[265] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018.

[266] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 2021.

[267] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.

[268] C. Zhu, R. Tao, K. Luu, and M. Savvides. Seeing small faces from robust anchor's perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5127–5136, 2018.

[269] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep learning for biometrics*, pages 57–79. Springer, 2017.

[270] J.-Y. Zhu, W.-S. Zheng, F. Lu, and J.-H. Lai. Illumination invariant single face image recognition under heterogeneous lighting condition. *Pattern Recognition*, 66:313–327, 2017.

[271] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015.

[272] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

A Written Declaration of Honor

"I hereby declare that I prepared this thesis without the impermissible help of third parties and that none other than the aids indicated have been used; all sources of information are clearly marked, including my own publications. In particular I have not consciously:

- Fabricated data or rejected undesirable results.
- Misused statistical methods with the aim of drawing other conclusions than those. warranted by the available data.
- Plagiarized external data or publications.
- Presented the results of other researchers in a distorted way.

I am aware that violations of copyright may lead to injunction and damage claims by the author and also to prosecution by the law enforcement authorities. I hereby agree that the thesis may be electronically reviewed with the aim of identifying plagiarism. This work has not yet been submitted as a doctoral thesis in the same or a similar form in Germany, nor in any other country. It has not yet been published as a whole."

Magdeburg, den 11.03.2025
Aly Khalifa