PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
Mohamed Kheider University – BISKRA
Faculty of Science and Technology
Department of Electrical Engineering

Order No: ……../../2025

Doctoral Thesis

For the purpose of obtaining a Doctor of Science degree in Electrical Engineering

Facial Recognition in Low-Quality Images

Presented and publicly defended by: **Sana BELLILI**

Publicly defended on : 01/11/2026

Before the jury composed of :

Athmane ZITOUNI	Professor	University of Biskra	Chairperson
Abdelmalik OUAMANE	Professor	University of Biskra	Thesis Supervisor
Ammar CHOUCHANE	MCA	University Center of Barika	Examiner
Abderrazak DEBILOU	Professor	University of Biskra	Examiner

Academic Year 2025-2026

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider – BISKRA

Faculté des Sciences et de la Technologie

Département de Génie Électrique

Arrêté N° : ……../../2025

Thèse de Doctorat

En vue de l'obtention du diplôme de Doctorat ès Sciences en Génie Électrique

Reconnaissance Faciale dans des Images de Faible Qualité

Présentée et soutenue publiquement par : **Sana BELLILI**

Soutenue le : 01/11/2026

Devant le jury composé de :

Athmane ZITOUNI	Professeur	Université de Biskra	Président
Abdelmalik OUAMANE	Professeur	Université de Biskra	Rapporteur
Ammar CHOUCHANE	MCA	Centre Universitaire de Barika	Examinateur
Abderrazak DEBILOU	Professeur	Université de Biskra	Examinateur

Année universitaire 2025-2026

Dedication

First and foremost, I dedicate this work to myself, for the patience, perseverance, and strength that accompanied me throughout this journey.

To my dear parents, the pillars of my life, who have always been a source of safety and endless support.

To my kind and loving mother, a symbol of tenderness and sacrifice, who has always been by my side to console and encourage me with her gentle words. Her big heart has been a warm haven in all times.

To my beloved father, the man who never withheld his guidance and advice, the beacon who illuminated my path with his wisdom and sacrifices. A father who has selflessly given everything for his children.

Today, from the depths of my heart, I say: **Thank you, my dear parents**, you are the fountain of courage and support.

To my beloved sisters, you have been my greatest supporters and sources of unwavering strength. Your love, encouragement, and belief in me were reflected in every achievement. I am profoundly grateful to share this success with you.

To my dear brother Mohamed Amine, your steadfast presence and reassuring words have been a constant source of comfort. You have always reminded me that I am never alone, and your support has been my driving force during the hardest times.

To my dear friends and classmates, thank you for every kind word, every moment of support, and every shared memory.

To everyone who knows me, loves me, and wishes me well, to everyone whose kind words left a lasting impact on me, you are all cherished in my heart without exception.

With all my love, appreciation, and deepest gratitude.

Acknowledgements

All praise is due to Allah, the Glorious, the Sovereign, the Holy, the Source of Peace, the Guardian, the Almighty, the Compeller, the Supreme. Glory be to Him and praise be to Him; there is no god but Allah, and Allah is the Greatest. He is the Possessor of Majesty and Honor, Owner of all that is in the heavens and the earth, and He is capable of all things. He is the First and the Last, the Evident and the Hidden. No praise can truly encompass His greatness; He is the Rich, the Praiseworthy, who created mankind and taught them speech, fashioned man in the best form, and provided sustenance from sources unseen. May all thanks and glory be to Him alone.

First and foremost, all praise and thanks are due to Allah, the Almighty and the Most Merciful, who granted me strength, patience, and perseverance to complete this work. Without His guidance and blessings, none of this would have been possible.

I extend my deepest gratitude to my supervisor, Professor **Ouamane Abdelmalik**, whose profound knowledge, insightful guidance, and unwavering support were instrumental throughout this research journey. His mentorship has greatly enriched both my academic and personal growth. He was, in every sense, a teacher of both great knowledge and high moral character. May Allah reward him abundantly.

I am sincerely grateful to Professor **Chouchane Ammar** for his unwavering support, insightful guidance, and continuous encouragement throughout the course of this work. I am truly grateful for his patience, generosity, he was a great teacher both in knowledge and in character. May Allah reward him abundantly.

I would also like to sincerely thank all the professors who have contributed to my academic and personal growth. Your teachings have shaped my path and inspired me to pursue knowledge with passion.

Special thanks to the lab director, **Debilou Abdelrezzak**, and all committee members for their time and effort in reviewing this thesis.

I would like to thank all the professors who contributed to facilitating my path to success.

Finally, I would like to thank my colleagues and friends who shared this journey with me, offering support, motivation, and camaraderie through both challenges and successes.

With heartfelt appreciation and gratitude to all who have accompanied me on this path.

List of Publications

Journal Publications

- Bellili, S., Ouamane, A., Chouchane, A., Himeur, Y., Atalla, S., Mansoor, W., ... Bensaali, F. (2025). Tensor-driven face recognition: Integrating super-resolution and multilinear subspace learning for low-resolution images. Information Fusion, 120, 103075.

- Sana Bellili, Abdelmalik Ouamane, Ammar Chouchane, Yassine Himeur, Shadi Atalla, Wathiq Mansoor, Salah Bourennane, Enhancing Face Verification for Low-Resolution Images with Super-Resolution and Vision Transformers, Expert Systems with Applications, 2026, 131976, ISSN 0957-4174.

International Conferences

- Bellili, S., Ouamane, A., Chouchane, A., Himeur, Y., Atalla, S., Mansoor, W., Al Ahmad, H. (2024, July). Very Low-Resolution Face Recognition Based On Multilinear Side-Information Based Discriminant Analysis. In 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService) (pp. 104-108). IEEE.

- Gharsa, O., Touba, M. M., Boumehraz, M., Abderrahman, N., Bellili, S., Titaouine, A. (2024, April). Autonomous landing system for A Quadrotor using a vision-based approach. In 2024 8th International Conference on Image and Signal Processing and their Applications (ISPA) (pp. 1-5). IEEE.

- Abderrahmani, N., Boumehraz, M., Touba, M. M., Gharsa, O., Bellili, S., Titaouine, A. (2024, May). Diagnosing Faults in Electrical Machines Using VMD: The Effect of the Sampling Frequancy. In 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC) (pp. 1-5). IEEE.

- Ghiloubi, I. B., Gharsa, O., Bellili, S., Abderrahmani, N., Boutiba, S. (2025). Trajectory tracking control of a quadrotor: A comparative study of PSO-and FPA-optimized PID controllers.

National Conferences

- Bellili, S., Ouamane, A., Chouchane, A., Bourennane, S. Advancing Face Recognition for Low-Resolution with Multi-linear Side Information-Based Discriminant Analysis. National Conference on Artificial Intelligence in Electrical Engineering, NCAIEE2024.

- Bellili, S., Ouamane, A., Chouchane, A. Super-Resolution and Vision Transformer-Based Approach for Low-Resolution Face Verification. The 1st National Conference on Electrical Engineering, Renewable Energy, and Artificial Intelligence (NCEERA2026).
- Bellili, S., Ouamane, A., Chouchane, A. Multilinear Subspace Learning for Improved Face Recognition in Low-Resolution Scenarios. The 1st National Conference on Electrical Engineering, Renewable Energy, and Artificial Intelligence (NCEERA2026).

Abstract

Low-resolution face verification remains a major challenge in computer vision and biometrics, particularly in real-world surveillance scenarios where images suffer from poor resolution, noise, blur, and illumination variations. These degradations significantly reduce identity-discriminative features and limit the robustness of conventional recognition systems. This work proposes a comprehensive framework that combines multilinear subspace learning and Vision Transformer models enhanced by super-resolution techniques to improve verification performance under extreme low-resolution conditions.

First, a tensor-based verification framework is introduced using multilinear representations and discriminant tensor learning. Advanced metric learning methods including TXQDA, MSIDA, SILD and XQDA are integrated within a multi-scenario verification scheme, where TXQDA and MSIDA belong to multilinear subspace learning, while XQDA and SILD are linear subspace learning methods used for comparison and complementary evaluation. TXQDA preserves the intrinsic multi-dimensional structure of facial data without vectorization, leading to stronger discriminative power and improved robustness. Experiments are conducted on extremely low-resolution scales (16×16, 32×32, and 48×48). Super-resolution enhancement using SRResNet, SRGAN, and Real-ESRGAN with a ×4 upscaling factor further boosts recognition accuracy. The proposed approach achieves verification rates of 99.17% on LFW, 90.90% on CelebA, and 76.90% on QMUL-SurvFace.

Second, the study investigates Vits for low-resolution face verification by evaluating multiple architectures including SimpleViT, DistillableViT, DeepViT, and CaiT combined with super-resolution preprocessing using SRResNet and SRGAN. Evaluations on QMUL-TinyFace and QMUL-SurvFace show that super-resolution significantly improves transformer-based verification, with SimpleViT and DistillableViT achieving near-perfect accuracy. The proposed SR-ViT framework surpasses state-of-the-art methods, reaching 52.92% TAR@FAR=1% and 92.65% AUC on QMUL-TinyFace, with strong AUC gains on QMUL-SurvFace.

Keywords: Face recognition, Low-resolution face verification, Super-resolution, Multilinear subspace learning, Tensor Cross-View Quadratic Discrimi nant Analysis, Vision Transformers, feature enhancement, Optimized model SR-Vit, Metric learning, Biometric recognition.

Résumé

La vérification faciale en basse résolution demeure un défi majeur en vision par ordinateur et en biométrie, en particulier dans les scénarios réels de surveillance où les images sont affectées par une faible résolution, le bruit, le flou et les variations d'illumination. Ces dégradations réduisent fortement les caractéristiques discriminantes de l'identité et limitent la robustesse des systèmes de reconnaissance classiques. Ce travail propose un cadre complet combinant l'apprentissage de sous-espaces multilineaires et les modèles Vision Transformer, renforcés par des techniques de super-résolution, afin d'améliorer les performances de vérification dans des conditions de très basse résolution.

Premièrement, un cadre de vérification basé sur les tenseurs est introduit en utilisant des représentations multilineaires et un apprentissage discriminant tensoriel. Des méthodes avancées d'apprentissage de métriques, incluant TXQDA, MSIDA, SILD et XQDA, sont intégrées dans un schéma de vérification multi-scénarios, où TXQDA et MSIDA relèvent de l'apprentissage de sous-espaces multilineaires, tandis que XQDA et SILD appartiennent aux méthodes de sous-espaces linéaires, utilisées à des fins de comparaison et d'évaluation complémentaire. TXQDA préserve la structure multidimensionnelle intrinsèque des données faciales sans vectorisation, offrant ainsi un pouvoir discriminant plus élevé et une meilleure robustesse. Des expériences sont menées sur des résolutions extrêmement faibles (16×16, 32×32 et 48×48). L'amélioration par super-résolution à l'aide de SRResNet, SRGAN et Real-ESRGAN avec un facteur d'agrandissement ×4 augmente davantage la précision de vérification. L'approche proposée atteint des taux de vérification de 99.17% sur LFW, 90.90% sur CelebA et 76.90% sur QMUL-SurvFace.

Deuxièmement, l'étude analyse les Vision Transformers pour la vérification faciale en basse résolution en évaluant plusieurs architectures, notamment SimpleViT, DistillableViT, DeepViT et CaiT, combinées avec un prétraitement par super-résolution utilisant SRResNet et SRGAN. Les évaluations sur QMUL-TinyFace et QMUL-SurvFace montrent que la super-résolution améliore significativement la vérification basée sur les transformers, SimpleViT et DistillableViT atteignant des performances quasi parfaites. Le cadre proposé SR-ViT dépasse les méthodes de l'état de l'art, atteignant 52.92% TAR@FAR=1% et 92.65% d'AUC sur QMUL-TinyFace, avec des gains importants d'AUC sur QMUL-SurvFace.

Mots-clés : Reconnaissance faciale, Vérification faciale à basse résolution, Super-résolution, Apprentissage de sous-espaces multilineaires, Tensor Cross-View Quadratic Discriminant Analysis (TXQDA), Vision Transformers, Amélioration des caractéristiques, Modèle optimisé SR-ViT, Apprentissage de métriques, Reconnaissance biométrique.

الملخص

لا تزال عملية التحقق من الهوية بالوجه في الصور منخفضة الدقة تمثل تحديًا كبيرًا في مجال الرؤية الحاسوبية والقياسات الحيوية، خاصة في سيناريوهات المراقبة الواقعية حيث تعاني الصور من ضعف الدقة، والضجيج، والطمس، وتغيرات الإضاءة. تؤدي هذه العوامل إلى تقليل الخصائص المميّزة للهوية وتحد من متانة أنظمة التعرف التقليدية. في هذا العمل نقترح إطارًا شاملاً يجمع بين تعلم الفضاءات الفرعية متعددة الخطية ونماذج المحولات البصرية المدعّمة بتقنيات رفع الدقة، بهدف تحسين أداء التحقق في ظروف الدقة المنخفضة جدًا.

أولًا، يتم تقديم إطار تحقق قائم على التنسورات باستخدام تمثيلات متعددة الخطية وتعلم تمييزي تنسوري. تم دمج طرق متقدمة في تعلم المقاييس تشمل TXQDA و MSIDA و SILD و XQDA ضمن مخطط تحقق متعدد السيناريوهات، حيث إن TXQDA و MSIDA تنتميان إلى تعلم الفضاءات الفرعية متعددة الخطية، بينما XQDA و SILD تنتميان إلى طرق الفضاءات الفرعية الخطية وتستعملان للمقارنة والتقييم التكميلي. تحافظ طريقة TXQDA على البنية متعددة الأبعاد الأصلية لبيانات الوجوه دون تحويلها إلى متجهات، مما يمنح قدرة تمييزية أعلى ومتانة أفضل. أُجريت التجارب على دقات منخفضة جدًا (16×16 و 32×32 و 48×48). كما ساهم تحسين الصور باستخدام SRResNet و SRGAN و Real-ESRGAN بمعامل تكبير ×4 في رفع دقة التحقق بشكل إضافي. حققت الطريقة المقترحة معدلات تحقق بلغت 99.17 على LFW و 90.90 على CelebA و 76.90 على QMUL-SurvFace.

ثانيًا، تدرس هذه الدراسة نماذج Transformers Vision للتحقق الوجهي في الدقة المنخفضة من خلال تقييم عدة معماريات تشمل SimpleViT و DistillableViT و DeepViT و CaiT مع دمج المعالجة المسبقة برفع الدقة باستخدام SRResNet و SRGAN. أظهرت النتائج على مجموعتي QMUL-TinyFace و QMUL-SurvFace أن رفع الدقة يحسّن بشكل ملحوظ أداء التحقق المعتمد على المحولات، حيث حقق SimpleViT و DistillableViT أداءً قريبًا من الكمال. الإطار المقترح SR-ViT تفوّق على أحدث الطرق المرجعية، محققًا TAR@FAR=1 52.92 و AUC 92.65 على QMUL-TinyFace، مع تحسنات قوية في AUC على QMUL-SurvFace.

الكلمات المفتاحية: التعرف على الوجوه، التحقق الوجهي منخفض الدقة، رفع الدقة ، تعلم الفضاءات الفرعية متعددة الخطية، تحليل تمييزي تربيعي تنسوري متعدد المشاهد، محولات الرؤية ، تحسين السمات، النموذج المحسّن SR-ViT، تعلم المقاييس، التعرف البيومتري.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Face verification from low-resolution images remains a challenging problem in modern biometric and computer vision systems. When facial images are captured or stored at very small sizes, such as 16×16, 32×32, or 48×48 pixels, important visual details are often missing, which negatively affects the reliability of recognition and verification methods. To improve system robustness under these conditions, it is necessary to combine image enhancement techniques with powerful representation and learning models.

This chapter presents a structured framework for low-resolution face verification based on super-resolution and advanced feature learning approaches. The general idea is to first enhance small facial images using super-resolution methods in order to recover useful visual information, and then apply different representation models to extract discriminative features for verification.

Several complementary approaches are considered in this chapter. Convolutional neural network models are used for deep feature extraction, followed by subspace learning techniques for discrimination and matching. Both linear and multilinear (tensor-based) subspace methods are discussed to highlight the benefit of preserving structural information in image data. In parallel, transformer-based vision models are also explored as an alternative representation strategy for face verification after super-resolution enhancement.

The chapter describes the overall processing pipelines, from low-resolution input images to enhanced images, feature extraction, subspace projection, and final verification decisions. Different configurations are organized and compared under a unified evaluation protocol. The goal is to provide a clear and coherent view of how super-resolution, subspace learning, and transformer-based models can be combined to improve the robustness of face verification in low-resolution scenarios.

1.2 Problem Formulation

Face Verification is the task of determining whether two facial images belong to the same person by extracting discriminative facial features and comparing their representations using machine learning or deep learning models. In a Low-Resolution scenario, the input face images have very small spatial dimensions or limited pixel information, such as those captured by surveillance cameras or distant sensors, which makes feature extraction significantly more difficult. This scenario introduces several major challenges, including loss of texture, where fine facial details disappear; blur, caused by low quality or motion; and feature instability, where the extracted features become sensitive to noise, pose, and illumination variations. To address these issues, facial images are mathematically represented as matrices or higher-order tensors, which preserve spatial and channel-wise structure, enabling advanced models to learn multidimensional relationships and produce more robust representations even under low-resolution conditions.

1.2.1 Face Verification

Most face verification approaches operate by independently extracting descriptive features from each of the two face images under comparison, and then measuring their similarity in a learned feature space. Over the years, many types of low-level facial features have been employed, including both hand-crafted and learned representations. Traditional hand-crafted descriptors such as Local Binary Patterns (LBP) and its variants, SIFT, and Gabor features have been widely used due to their ability to capture local texture and structural information. In addition, learned feature representations derived from data-driven models have also been introduced to improve discrimination capability [6].

Face verification itself is the task of determining whether two facial images belong

to the same person or to two different individuals. Unlike face identification, which assigns an identity from a known gallery, face verification focuses on pairwise comparison and decision making based on similarity scores. A typical verification pipeline includes feature extraction, feature representation or projection, similarity measurement, and a final decision step based on a threshold. The effectiveness of the system largely depends on the quality of the extracted features and their robustness to variations such as resolution, illumination, pose, and noise.

1.2.2 Challenges in Low-Resolution Face Verification

Low-resolution face verification faces several critical challenges that directly affect recognition robustness. One major issue is loss of texture, where fine facial details such as skin patterns, wrinkles, and micro-structures disappear due to the limited number of pixels, reducing the discriminative power of the image. Another challenge is blur, which may result from low-quality sensors, motion, or compression, and leads to the smoothing of edges and facial contours that are important for feature extraction. In addition, feature instability becomes a significant problem, as the extracted representations tend to vary strongly with small changes in illumination, pose, noise, or scale, making the matching process less reliable. Together, these factors degrade the quality of learned embeddings and make accurate verification more difficult in low-resolution scenarios.

1.2.2.1 loss of texture

One of the primary challenges in low-resolution face images is the loss of texture, where fine-grained facial details are significantly reduced or completely removed due to the small number of pixels representing the face region. Texture information such as skin micro-patterns, subtle edges, and local intensity variations plays a crucial role in distinguishing between individuals. When resolution decreases, these discriminative cues become blurred or merged, leading to less informative feature maps and weaker identity representations. As a result, feature extraction models struggle to capture stable and distinctive descriptors, which negatively affects verification robustness and matching accuracy [7].

1.2.2.2 Blur

Blur refers to the loss of sharpness or fine detail in an image, often caused by motion, defocus, or low-quality acquisition systems. It reduces the clarity of edges and textures, making it more difficult to perceive distinct features. Blur is also a fundamental attribute of human spatial vision, and sensitivity to it has been widely studied in experimental and theoretical research. Previous studies often proposed specialized mechanisms to explain blur perception, such as intrinsic blur, multiple spatial frequency channels, or dedicated blur estimation units. However, recent findings suggest that such specialized mechanisms are not strictly necessary. Instead, blur discrimination can be effectively explained using models based on visible contrast energy, where two spatial patterns are distinguished when the integrated difference in their local contrast energy exceeds a certain threshold [8]. In low-resolution face verification, blur poses a significant challenge as it smooths important facial features, complicating the extraction of stable and discriminative representations.

1.2.2.3 Feature instability

Feature instability refers to the variability or sensitivity of extracted features under small changes in the input image, such as illumination, pose, noise, or resolution. An unstable feature is one whose representation can change significantly even when the underlying identity remains the same, making it unreliable for verification. This concept is analogous to the idea of replaceability in linguistic features, where a feature can vary while still representing the same underlying content. In face verification, unstable features reduce the robustness of the model because small perturbations in low-resolution images may lead to large changes in the feature representation, thus degrading matching accuracy. Measuring and controlling feature instability is therefore crucial for designing systems that are resilient to low-resolution and degraded facial images [9].

1.3 Super-Resolution for Low-Resolution Faces

Super-Resolution (SR) for low-resolution faces is a critical technique aimed at enhancing the spatial resolution of facial images while preserving identity-related features. In scenarios such as surveillance, remote sensing, or old datasets, face images often have very small dimensions, which severely degrade recognition performance. SR methods recon-

struct high-resolution images from low-resolution inputs, recovering lost textures, sharpening edges, and improving the clarity of facial features. Modern approaches leverage deep learning techniques, including CNNs and generative adversarial networks (GANs), to learn mappings from low-resolution to high-resolution images. By enhancing image quality, Super-Resolution not only improves the visual appearance but also increases the robustness and accuracy of face verification systems, enabling more reliable feature extraction and matching even under challenging low-resolution conditions.

1.3.1 Super-Resolution Models

In our work, we employed several state-of-the-art Super-Resolution (SR) models to enhance low-resolution face images before verification. The first model, SRResNet, is a deep residual network designed to reconstruct high-resolution images from low-resolution inputs by learning residual mappings, effectively recovering fine facial details. The second model, SRGAN [5], extends SRResNet [5] by incorporating a generative adversarial network (GAN) framework, which improves perceptual quality and produces sharper, more realistic faces by training a generator to fool a discriminator. Finally, we used Real-ESRGAN [10], a more recent model that combines GAN-based learning with practical degradation modeling, making it highly effective for real-world low-resolution and degraded images. These SR models were applied to our datasets to improve texture, reduce blur, and stabilize facial features, ultimately enhancing the performance of the face verification system.

1.3.2 Resolution Scenarios

In our study, we considered multiple resolution scenarios to evaluate the robustness of face verification models under varying image qualities. Specifically, we generated low-resolution images at 16×16, 32×32, and 48×48 pixels, representing extreme, moderate, and relatively high low-resolution conditions, respectively. These scenarios reflect real-world situations such as surveillance or distant capture, where facial images can be extremely small. Additionally, previous studies included in our review utilized even smaller resolutions, highlighting the challenges posed by severely limited pixel information. By systematically testing these resolution levels, we were able to assess how different Super-Resolution models and feature extraction techniques perform under progressively

degraded image conditions, providing insights into the limits of current low-resolution face verification methods.

1.4 CNN Feature Extraction using VGG-19

We employed VGG-19 [11] as the backbone for face feature extraction due to its proven robustness and effectiveness in capturing discriminative visual representations [11, 12]. VGG-19 is a deep convolutional neural network with 19 layers, which has been widely used in computer vision tasks thanks to its simple yet powerful architecture. One key advantage is that pretrained VGG-19 models are available on large-scale datasets such as ImageNet, allowing us to leverage transfer learning and reduce training time while achieving high performance even with limited face datasets. Furthermore, VGG-19 is particularly effective for face feature extraction, as its deep convolutional layers can capture both low-level details (edges, textures) and high-level semantic features (facial structures, identity cues), making it suitable for scenarios involving low-resolution images or images enhanced via Super-Resolution [13]. Overall, using VGG-19 ensures reliable and discriminative embeddings, contributing to the robustness and accuracy of our face verification system.

1.4.1 Feature Vector Extraction Pipeline

The feature vector extraction pipeline is a critical step in our face verification system, designed to convert facial images into compact and discriminative representations. The process begins with the input image, which may be low-resolution or enhanced using Super-Resolution techniques. Each image is then resized to match the input dimensions required by the backbone network VGG-19, ensuring consistency across the dataset. Features are extracted from a specific feature layer, typically a deep convolutional or fully connected layer, which captures both low-level details (textures, edges) and high-level semantic information (facial structure, identity cues). The resulting feature vector has a fixed dimension, which serves as a numerical representation of the face and can be directly used for similarity measurement and verification. This pipeline ensures that the extracted embeddings are robust, discriminative, and suitable for comparison across varying resolutions and image qualities.

1.5 Multilinear Subspace Learning for Face Verification

Multilinear (Tensor) Subspace methods extend classical linear subspace techniques by operating directly on tensor representations instead of converting data into long vectors. Since face images naturally preserve spatial structure when represented as matrices or higher-order tensors, multilinear approaches maintain this structural information and model correlations across multiple modes (such as rows, columns, and channels). In our work, we employed tensor-based discriminative methods including TXQDA (Tensor Cross-view Quadratic Discriminant Analysis) and MSIDA (Multilinear Subspace Identity Discriminant Analysis). TXQDA is designed to learn discriminative tensor projections across different views or conditions while modeling intra-class and inter-class variations more effectively than vector-based QDA. MSIDA, on the other hand, performs identity-oriented discriminant analysis in multilinear space, extracting compact and highly separative tensor features. These methods offer several advantages, including reduced dimensionality without destroying spatial structure, improved robustness to low-resolution and noise, lower risk of overfitting compared to vectorized representations, and stronger discriminative power for face verification tasks.

1.5.1 Tensor Representation of Face Images

In modern face recognition and verification systems, a facial image can be naturally modeled as a tensor, rather than being flattened into a one-dimensional vector. An image is inherently a multi-dimensional structure for example, a grayscale face image can be represented as a 2D tensor (height $\times$ width), while a color image forms a 3D tensor (height $\times$ width $\times$ channels). This tensor representation preserves the spatial organization of pixels and the relationships between rows, columns, and channels. In contrast, traditional vectorization converts the image into a long vector, which breaks spatial locality and may lead to very high-dimensional and less structured representations. Working directly with tensors provides several advantages: it maintains spatial dependencies, enables multilinear subspace learning, reduces parameter complexity, and improves robustness when handling low-resolution or degraded faces. For these reasons, tensor representation is particularly

well suited for advanced discriminative methods and multilinear models used in our work.

1.5.2 Multilinear vs. Linear Subspace Methods for Face Verification

Linear and multilinear subspace methods differ mainly in how they represent and process image data. Linear subspace approaches first convert images into long one-dimensional vectors, then apply discriminant or metric learning techniques in vector space. While effective, this vectorization step destroys the natural spatial structure of images and may lead to very high dimensional representations. In contrast, multilinear (tensor) subspace methods operate directly on matrix or tensor forms, preserving spatial and multi-mode relationships and enabling more structured dimensionality reduction. In our work, we investigated both categories, linear subspace methods, XQDA [14] and SILD [15], for discriminative metric learning and identity separation in vector space, and we also employed multilinear tensor-based approaches, TXQDA [16] and MSIDA [17] for structure-preserving discrimination. This combined evaluation allows a fair comparison between vector-based and tensor-based modeling, highlighting the advantages of multilinear representations in handling low-resolution and degraded face images.

1.6 Vision Transformers

Vision Transformers (ViTs) have recently emerged as a powerful alternative to convolution-based models in computer vision, offering a new paradigm for visual representation learning based on transformer architectures originally developed for natural language processing. Their strength lies in global context modeling and flexible attention mechanisms, which enable more comprehensive interaction between image regions compared to purely local convolutional operations. Due to these properties, ViTs have shown strong performance in several visual tasks, including face recognition and verification, especially when combined with appropriate preprocessing and enhancement techniques.

Vision Transformer (ViT) architectures adopt a fundamentally different strategy for image processing compared to convolutional networks. Instead of applying convolutional filters, ViTs divide the input image into fixed-size patches, which are then flattened and linearly projected into embedding vectors that form a sequence of tokens. Positional em-

beddings are added to preserve spatial order, and the token sequence is passed through multiple transformer encoder layers [18]. These encoders rely on multi-head self-attention mechanisms to model long-range dependencies and relationships between distant image regions. In face recognition tasks, ViTs learn global and contextual facial representations by allowing each patch to attend to all others, which helps capture identity-related structures distributed across the face. This global attention capability makes ViTs particularly effective when combined with high-quality preprocessing or super-resolution, as they can integrate fine and coarse facial cues into a unified representation.

Several of the most effective and recent ViT variants were adopted to ensure strong performance and architectural diversity. In particular, SimpleViT [1], DeepViT [3], DistillableViT [2] , and CaiT [4] were selected, as they introduce complementary improvements over the standard ViT architecture. SimpleViT provides a streamlined and stable design with reduced training complexity and solid baseline robustness. DeepViT increases transformer depth while enhancing attention map diversity, leading to richer feature representations and stronger discrimination capability. Distillation ViT integrates a teacher–student distillation strategy that improves data efficiency and boosts accuracy, especially when labeled data is limited. CaiT incorporates class-attention layers and late-stage token refinement, which improve convergence behavior and produce more discriminative global representations. Together, these ViT variants provide improved robustness, better global dependency modeling, and higher-quality face representations under challenging low-resolution conditions.

1.6.1 Vision Transformers vs. Convolutional Neural Networks for Face Recognition

Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) differ significantly in how they process facial images and learn visual features. CNNs follow a convolution-based pipeline where the image is processed through successive layers of filters (kernels) that progressively extract features, starting from low-level cues such as edges, colors, and gradients, and moving toward higher-level semantic patterns. Pooling layers reduce spatial resolution while retaining the most informative responses, and the resulting features are flattened and passed to fully connected layers for final prediction [18]. In contrast, ViTs adopt a transformer-based strategy, where the image is first divided

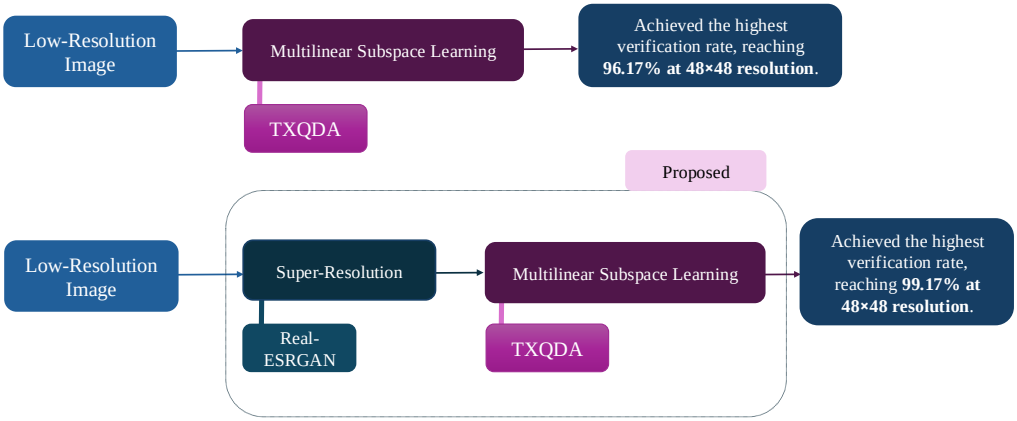


Figure 1.1: Overview of the framework proposed in the first article: Tensor-Driven Super-Resolution and Multilinear Feature Extraction.

into small patches that are flattened and linearly projected into token embeddings. These tokens are then processed by multiple transformer encoder layers that use multi-head self-attention to model relationships between all image regions simultaneously. This allows ViTs to capture global contextual dependencies across the face, whereas CNNs emphasize local spatial patterns through convolution. As a result, CNNs are typically strong in local feature extraction and computational efficiency, while ViTs are powerful in global relational modeling and long-range feature interaction [18].

1.7 Motivation and Importance of the Study

Face recognition has evolved into a core technology in modern intelligent systems, playing a central role in security, surveillance, biometric authentication, border control, and digital identity management. Despite the remarkable progress achieved by deep learning models under controlled conditions, their performance still degrades significantly when confronted with low-resolution facial images. This limitation represents one of the most critical unresolved challenges in real-world face verification systems.

In practical surveillance environments, facial images are frequently captured at a distance, under poor illumination, with motion blur, compression artifacts, and sensor limitations. As a result, the acquired faces often contain insufficient discriminative details

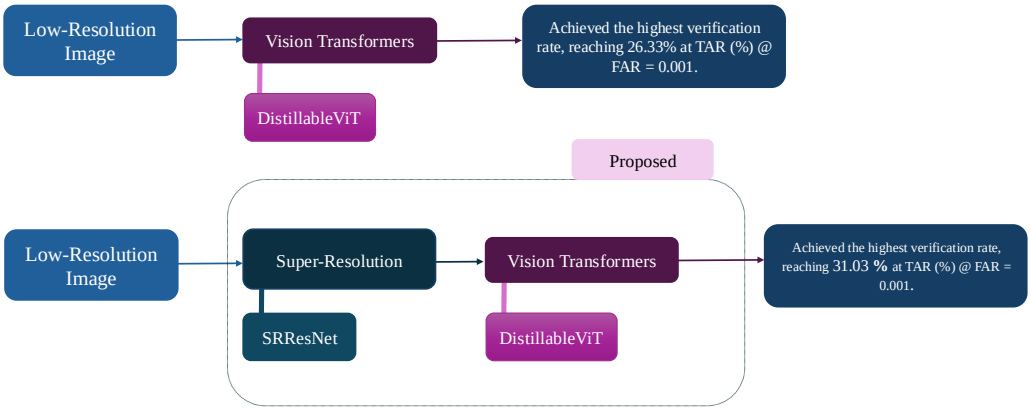


Figure 1.2: Overview of the framework proposed in the second article: Integration of Vision Transformers and Super-Resolution Techniques.

for reliable identity verification. Most existing recognition models are trained and optimized using high-resolution datasets, which creates a strong domain mismatch when these models are deployed in unconstrained real-world scenarios. This mismatch leads to unstable feature extraction, increased false matches, and reduced verification reliability in operational systems.

- The motivation of this research arises from the urgent need to design robust face verification frameworks that remain effective under extreme low-resolution conditions. Rather than relying on a single methodological direction, this work is motivated by the hypothesis that combining complementary paradigms — namely Super-Resolution reconstruction, multilinear subspace learning, metric learning, and Vision Transformer architectures — can produce a more resilient and discriminative verification pipeline. Each of these components addresses a different aspect of the degradation problem: image detail recovery, structure-preserving representation, discriminative distance learning, and global-context feature modeling.

- The importance of this study lies in its multi-level contribution to both theory and practice. From a methodological perspective, it advances low-resolution face verification by introducing tensor-based multilinear discriminant frameworks and transformer-based deep models within unified evaluation protocols. From an ap-

plied perspective, it improves the operational reliability of biometric systems in surveillance and forensic contexts where image quality cannot be controlled. The study also contributes to reducing the resolution domain gap through systematic Super-Resolution integration and comparative evaluation across multiple recognition paradigms.

Furthermore, this research supports the development of next-generation biometric systems that are resolution-aware, architecture-diverse, and experimentally validated across multiple datasets and degradation levels. By addressing low-resolution verification through structured learning, enhancement, and transformer-based modeling, this work strengthens the robustness, scalability, and practical applicability of face recognition technologies in real-world deployments.

1.8 Research Objectives

Face verification in real-world surveillance and security environments remains highly challenging due to the frequent presence of low-resolution facial images. Such images suffer from loss of discriminative details, noise, blur, and illumination variations, which significantly degrade feature extraction quality and verification accuracy. The primary objective of this research is to improve low-resolution face verification performance by combining advanced Super-Resolution (SR) techniques with robust representation and recognition models.

This work is structured around two complementary research directions, each targeting a different level of the recognition pipeline and jointly contributing to a robust low-resolution verification framework, as illustrated in Figures 1.1 and 1.2.

1.8.1 Enhancing Low-Resolution Face Verification Using Super-Resolution and Multilinear Subspace Learning

The first objective focuses on improving low-resolution facial image representation through the integration of Super-Resolution techniques with multilinear subspace learning and metric learning methods. Multilinear models are particularly suitable for facial data because they preserve the intrinsic multi-dimensional tensor structure of images and enable more stable feature extraction under variations in pose, illumination, and expression.

However, their performance degrades when the input images are extremely low-resolution due to severe loss of high-frequency identity information.

To overcome this limitation, we propose a fusion framework that first applies Super-Resolution reconstruction and then performs multilinear discriminant analysis and metric learning. The framework integrates advanced tensor-based and subspace-based metric learning algorithms, including TXQDA and MSIDA (multilinear approaches), as well as XQDA and SILD (linear subspace approaches). This hybrid strategy ensures that enhanced images are projected into discriminative subspaces that maximize inter-class separability while preserving structural information, 1.3.

This objective aims to:

- Recover identity-relevant facial details lost due to low spatial resolution.
- Preserve the tensor structure of facial data through multilinear modeling.
- Improve discriminative feature extraction using tensor-based and subspace metric learning.
- Increase verification robustness across multiple low-resolution scales.

1.8.2 Improving Low-Resolution Face Verification Using Super-Resolution and Vision Transformers

The second objective investigates the effectiveness of ViTs for low-resolution face verification when combined with Super-Resolution preprocessing. While Convolutional Neural Networks have long dominated face recognition, Vits introduce global self-attention mechanisms that model long-range dependencies and capture richer contextual facial representations. Nevertheless, transformer models remain sensitive to missing fine-grained details in extremely low-resolution inputs.

To address this challenge, we propose a hybrid Super-Resolution–Transformer framework in which facial images are first enhanced using SR models and then processed by multiple ViT architectures for feature extraction and verification. The study evaluates several transformer variants, including SimpleViT, DistillableViT, DeepViT, and CaiT, and analyzes their behavior on both original and super-resolved images, 1.4.

This objective aims to:

- Enhance low-resolution inputs using Super-Resolution before transformer-based recognition.
- Exploit self-attention mechanisms to capture both global and local facial structures.
- Compare transformer-based and CNN-based recognition under identical SR conditions.
- Identify the most robust deep architecture for extreme low-resolution face verification.

Together, these two research directions establish a comprehensive and multi-level framework that combines image enhancement, multilinear representation learning, metric learning, and transformer-based modeling to significantly improve low-resolution face verification performance.

1.9 Key Contributions

This thesis addresses the problem of low-resolution face verification through a structured and multi-perspective framework that combines image enhancement, tensor-based metric learning, and transformer-based recognition models. The main contributions are summarized as follows:

- **Unified Framework for Low-Resolution Face Verification** We propose a comprehensive verification framework specifically designed for extreme low-resolution conditions. The framework integrates image enhancement, discriminant metric learning, and deep representation models within a multi-scenario evaluation protocol covering multiple resolution scales.
- **Tensor-Based Multilinear Metric Learning Framework** We develop a tensor-driven verification pipeline that preserves the multidimensional structure of facial data without vectorization. Within this framework, multilinear metric learning methods (TXQDA and MSIDA) and linear metric learning methods (XQDA and SILD) are jointly evaluated under identical experimental settings, enabling a fair and systematic comparison between multilinear and linear subspace strategies for low-resolution verification.

- **Resolution-Scale Evaluation Protocol** We design an experimental protocol based on multiple controlled resolution levels (16×16, 32×32, and 48×48), allowing detailed analysis of model behavior under progressive information loss. This protocol provides reproducible benchmarks for extreme low-resolution verification.

- **Integration of Super-Resolution** We introduce a preprocessing stage based on multiple Super-Resolution models (SRResNet, SRGAN, and Real-ESRGAN) and systematically analyze their impact on verification accuracy. The study quantifies how resolution reconstruction affects discriminant metric learning and transformer-based recognition.

- **Transformer-Based Verification under Resolution Constraints** We conduct an extensive evaluation of several Vision Transformer architectures (SimpleViT, DistillableViT, DeepViT, and CaiT) for low-resolution face verification, both before and after super-resolution enhancement. This provides one of the first structured comparisons of ViT families under extreme low-resolution biometric conditions.

- **SR–ViT Hybrid Verification Strategy** We propose and validate a hybrid Super-Resolution + Vision Transformer verification pipeline and demonstrate its effectiveness across challenging surveillance-oriented datasets. The approach establishes improved verification trade-offs in TAR–FAR and AUC metrics.

- **Cross-Dataset Validation** All proposed approaches are validated across multiple benchmark datasets with different characteristics (surveillance images), ensuring robustness and generalization of the findings.

1.10 Thesis Organization

The structure of this thesis is systematically designed to provide a clear and coherent flow of information, encompassing both theoretical foundations and practical contributions across multiple chapters. The organization is outlined as follows:

Chapter 1: Introduction, Chapter 1: Introduction, This chapter presents the overall context and rationale behind the research. It begins with a detailed articulation of the problem statement, followed by an introduction to Super-Resolution techniques for

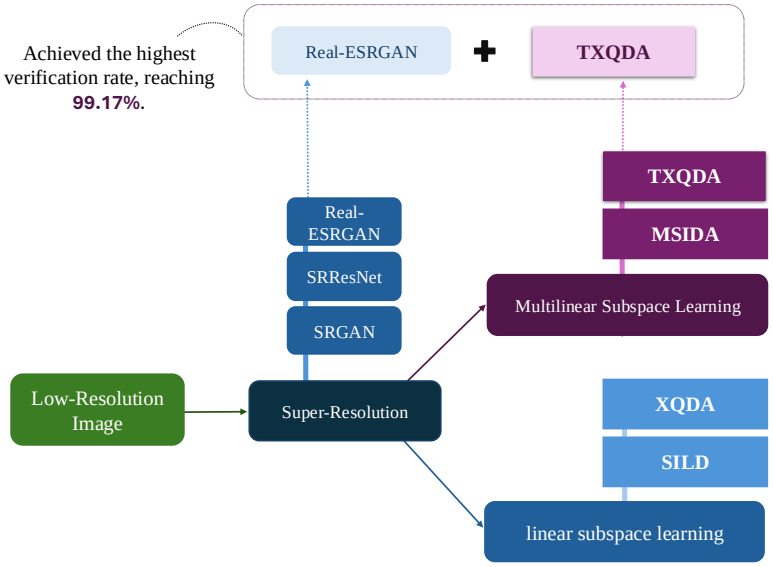


Figure 1.3: Tensor-Driven Approach for Multilinear Feature Extraction and Super-Resolution Integration.

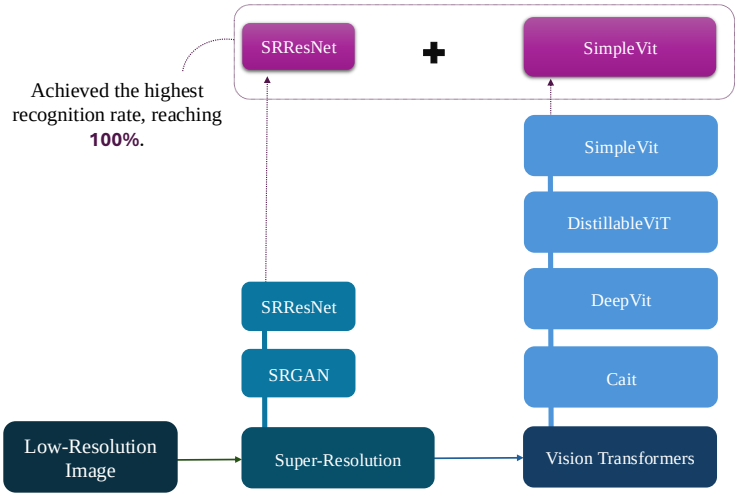


Figure 1.4: Optimal Integration of Vision Transformers and Super-Resolution Techniques.

low-resolution faces and CNN feature extraction using VGG-19. Next, it covers Multilinear Subspace Learning for Face Verification and the role of Vision Transformers in low-resolution face recognition. The chapter then emphasizes the motivation and importance of the study, explaining why addressing low-resolution facial images in real-world surveillance scenarios is critical. Afterward, it defines the research objectives and highlights the key contributions made through the proposed frameworks. Finally, the chapter provides an overview of the thesis structure, guiding the reader through the subsequent chapters and showing how each part contributes to the research goals.

Chapter 2: Literature Review, This chapter systematically reviews the existing body of work relevant to the research domain. It begins with an in-depth analysis of the challenges in low-resolution face recognition, followed by a comprehensive discussion on Super-Resolution techniques and their effectiveness in reconstructing facial details. The chapter further explores various machine learning and deep learning methods for face recognition, with a particular focus on CNN-based and ViT-based architectures. Additionally, it examines multilinear subspace learning approaches for feature extraction, setting the stage for the tensor-driven framework proposed in the first article. Lastly, the chapter discusses the emerging role of ViTs in face recognition, emphasizing their potential for capturing both local textures and global spatial dependencies.

Chapter 3: Paper 1, This chapter provides a detailed presentation of the first study, which focuses on a tensor-driven framework for low-resolution face recognition. It outlines the proposed methodology, which leverages advanced Super-Resolution (SR) techniques, including SRResNet, SRGAN, and Real-ESRGAN, to reconstruct high-resolution facial images from extremely low-resolution inputs. The framework further integrates multilinear subspace learning to extract robust and consistent feature representations across varying resolution levels. The experimental results and comparative analysis highlight the effectiveness of the tensor-driven approach in enhancing recognition accuracy under challenging conditions.

Chapter 4: Paper 2, Building upon the findings of the first study, this chapter presents the second research article, which extends the proposed framework by incorporating ViTs for enhanced feature extraction. The chapter discusses the integration of SR networks with ViTs, evaluating the performance of various ViT models, including SimpleViT, CaiT, Distillation, and DeepViT. The methodology focuses on leveraging

self-attention mechanisms to capture both local textures and global dependencies, thereby mitigating the impact of resolution degradation. The experimental analysis demonstrates the substantial improvement in face recognition accuracy achieved through the ViT-SR integration, particularly in extreme LR settings.

Chapter 5: Conclusion and Future Work, The final chapter provides a comprehensive summary of the key findings obtained from both studies, underscoring the effectiveness of combining SR and ViTs in addressing the challenges of low-resolution face recognition. It then identifies the limitations of the proposed frameworks, particularly in terms of computational complexity, data variability, and feature consistency. The chapter concludes with a detailed discussion on directions for future research, proposing potential enhancements such as adaptive SR techniques, lightweight ViTs, and hybrid CNN-ViT architectures for improved robustness and scalability.

Chapter 2

Literature Review

2.1 Introduction

Face recognition systems have become an essential component in many modern applications, including security monitoring, access control, biometric authentication, smart cities, and personalized digital services. Their importance continues to grow as automated identity verification becomes more widely adopted. However, despite the rapid progress in computer vision and deep learning, face recognition systems still encounter major difficulties when operating under unconstrained real-world conditions especially when the input images are low-resolution. In such cases, facial textures, contours, and discriminative micro-features are often lost, which directly reduces recognition accuracy and system robustness [19].

This chapter presents a comprehensive and structured review of the main challenges and the most relevant recent advances in face recognition research, with a particular focus on low-resolution scenarios. It explains why low-resolution images are problematic: they contain insufficient identity information, are more sensitive to noise and blur, and are strongly affected by illumination and pose variations. Because of this, traditional feature extraction and matching techniques often fail to separate identities reliably.

To address these limitations, the chapter discusses super-resolution techniques as a key enhancement strategy. These methods aim to reconstruct higher-detail facial images from low-quality inputs, thereby restoring discriminative information before recognition.

Different deep super-resolution models are reviewed as powerful preprocessing tools that can significantly improve downstream verification performance.

The chapter also examines the growing role of machine learning and deep learning approaches in face recognition. It details how modern neural models learn hierarchical and discriminative facial representations automatically, outperforming handcrafted features. Special attention is given to Multilinear Subspace Learning, which represents facial data in tensor (multi-dimensional) form instead of flattening it into vectors. This preserves structural relationships inside the data and enables more robust feature extraction under resolution degradation. Multilinear methods are shown to be particularly effective when combined with discriminant metric learning techniques.

In addition, the chapter highlights the transformative impact of ViTs in face recognition. Unlike classical convolutional networks, ViTs rely on attention mechanisms to model global relationships between image regions. This allows better contextual understanding of facial structure and improves recognition capability, especially when paired with enhancement techniques such as super-resolution. Recent transformer variants are discussed to illustrate how architectural evolution continues to push performance forward.

By bringing together enhancement methods, multilinear learning, and transformer-based modeling, the chapter builds a coherent view of the modern low-resolution face recognition pipeline and explains how hybrid strategies can overcome long-standing limitations.

2.2 Face Recognition Challenges in Low-Resolution Settings

Face recognition in low-resolution environments presents significant challenges due to the degradation of discriminative features, making it difficult for models to achieve robust performance. This issue is particularly pronounced in applications such as surveillance, security systems, and mobile authentication, where facial images are often captured at low resolutions. Researchers have explored various approaches to mitigate these challenges, including knowledge distillation, subspace learning, super-resolution techniques, and deep learning-based feature extraction. The following section presents a chronological review of key contributions to this field, highlighting advancements in methods aimed at improving

low-resolution face recognition. A summary of these research efforts is illustrated in Figure 2.1.

Lei et al. [20] (2011) addressed the challenge of face recognition from low-resolution images, a common issue in real-world applications. Since high-frequency details are lost in low-resolution images, it becomes essential to extract robust information from the low-frequency domain. To tackle this problem, the authors proposed an effective local frequency descriptor (LFD), inspired by local phase quantization (LPQ). Their approach leverages both blur-invariant magnitude and phase information in the low-frequency domain, making it more descriptive than LPQ by incorporating richer and more comprehensive features. Additionally, they introduced a statistical uniform pattern definition method to enhance the efficiency of the proposed descriptor. Experimental evaluations on FERET and a real-world video database demonstrated that LFD is highly effective and robust for low-resolution face recognition.

Xu et al. [21] (2014) examined the challenges of face recognition in surveillance systems, where captured facial images are often extremely small and differ from low-resolution images that are simply down-sampled from high-resolution counterparts. This discrepancy typically results in poor recognition performance, the authors analyzed key factors affecting face recognition accuracy in surveillance scenarios, including camera type, distance between the subject and the camera, and the resolution of captured face images. Each of these factors was systematically quantified and evaluated. Building on these insights, the authors proposed a new face recognition approach tailored for real-world surveillance environments. Their method incorporates image pre-processing techniques to mitigate illumination variations, thereby enhancing image quality. Additionally, they introduced a novel face image super-resolution technique to further refine facial details. Experimental results demonstrated that the proposed approach significantly improves face recognition performance in surveillance settings.

Mudunuri et al. [22] (2015) introduced a fully automated method for recognizing low-resolution face images captured in uncontrolled environments. Their approach employs multidimensional scaling to learn a common transformation matrix that aligns the facial features of both low-resolution (LR) and high-resolution (HR) training images. This transformation ensures that the distance between LR and HR images in the transformed space closely approximates their distance if both had been captured under identical controlled

conditions. To measure the similarity between two images in this transformed space, the method utilizes stereo matching cost, which enhances recognition performance. However, a notable drawback of this approach is its computational complexity, as calculating the stereo-matching cost is time-intensive. To address this, the authors propose a reference-based approach, where each face image is represented by its stereo-matching cost relative to a small set of reference images, significantly reducing computational overhead.

Chu et al. [23] (2017) addressed the challenge of low-resolution (LR) face recognition, a growing concern due to the widespread use of surveillance cameras. While previous approaches often assume the availability of multiple high-resolution (HR) training samples per subject, this assumption does not always hold, particularly in law enforcement scenarios where only a single HR image per person may be available. In such single sample per person (SSPP) settings, LR face recognition suffers from issues like overfitting and singular matrix problems. To tackle these challenges, the authors proposed a cluster-based regularized simultaneous discriminant analysis (C-RSDA) method. This approach enhances standard discriminant analysis by incorporating inter-cluster and intra-cluster scatter matrices derived from unsupervised clustering, which helps to regularize both the between-class and within-class scatter matrices. By leveraging these cluster-based matrices, C-RSDA not only resolves the singularity issue but also mitigates overfitting by capturing additional variations from limited training data. Extensive experiments conducted in both controlled and uncontrolled environments demonstrate the effectiveness of C-RSDA in improving the discriminative power of feature representations for LR face recognition. Building upon the challenge of LR face recognition, Shekhar et al. [24] (2017) adopted a different strategy by leveraging information from high-resolution (HR) gallery images rather than solely enhancing the low-resolution probe images. Their approach employed a generative classification framework capable of handling variations in resolution and illumination, making it more robust to real-world conditions. To further refine their method, the authors introduced a kernelized version of their algorithm to address data non-linearity and incorporated joint sparse coding for enhanced recognition performance at low resolutions. Experimental evaluations on both standard datasets and a challenging outdoor face dataset demonstrated the superiority of this approach compared to several state-of-the-art LR face recognition techniques. In addition to classification and feature enhancement techniques, Thomas et al. [25] (2017) focused on improving LR

face recognition by integrating super-resolution methods with advanced feature extraction and classification strategies. Their framework employed kernel regression to convert low-resolution images into high-resolution counterparts before extracting features using a combination of Gabor filters, wavelet transforms, and local binary patterns (GWTM process). To optimize feature selection, they introduced a feature-level fusion strategy based on the fractional Bat algorithm, which integrates fractional theory with the Bat algorithm for optimal performance. Finally, recognition was carried out using a multi-kernel-based spherical SVM classifier. Experimental evaluations demonstrated that their system outperformed existing methods, achieving an accuracy of 95

Ge et al. [26] (2018) addressed the challenge of deploying face recognition models in real-world scenarios, where recognizing low-resolution faces must be achieved with minimal computational cost. To tackle this issue, they proposed a selective knowledge distillation approach that compresses complex face models, balancing high recognition speed and low memory usage while minimizing performance loss. Their method utilizes a two-stream CNN: a teacher stream, which is a high-accuracy complex CNN for recognizing high-resolution faces, and a student stream, which is a lightweight CNN designed for efficient low-resolution face recognition. To maintain recognition performance, the student stream selectively distills key facial features from the teacher stream by solving a sparse graph optimization problem, ensuring that only the most informative features are transferred. This enables the student stream to simultaneously approximate important facial cues and recover missing details for low-resolution classification. Experimental results demonstrate that the student model achieves impressive performance, requiring only 0.15MB of memory while processing 418 faces per second on a CPU and 9,433 faces per second on a GPU. To further improve low-resolution face recognition in practical applications, Ahmed et al. [27] (2018) explored the challenge of automatic face recognition, a persistent issue in computer vision over the past decade. Despite advancements, law enforcement agencies still face difficulties in accurately identifying individuals from video surveillance due to factors such as blur, illumination variations, resolution constraints, and lighting conditions. To address these challenges, the proposed system is designed to operate effectively at a minimum resolution of 35 pixels, enabling accurate face recognition across different angles, side poses, and even during motion tracking. The study introduces the LR500 dataset for training and classification. Additionally, the research

employs the Local Binary Patterns Histogram (LBPH) algorithm to enhance real-time face recognition performance at low resolutions.

Wang et al. [28] (2019) tackled the challenge of low-resolution (LR) face recognition (FR), particularly in surveillance scenarios where both accuracy and computational efficiency are critical. The study leveraged knowledge distillation to transfer knowledge from a complex teacher model to a more lightweight student model, enabling fast and efficient LR-FR. Unlike traditional distillation methods, which require updating the teacher model using an LR-augmented training set before training the student, the authors proposed an improved approach that eliminates the need for teacher retraining. Instead, the teacher model is trained on an unchanged high-resolution dataset, while only the student model is trained using an LR-augmented dataset under the teacher's guidance. This modification significantly accelerates the training process, especially for large-scale datasets. To mitigate data distribution discrepancies between the teacher and student models, the approach incorporates a multi-kernel maximum mean discrepancy constraint. Experimental results demonstrate that this method speeds up training by approximately five times while maintaining high recognition accuracy. The student model achieves performance comparable to state-of-the-art methods on LFW and SCFace datasets, delivering a threefold acceleration over the teacher model and requiring only 35ms to run on a CPU. Shakeel et al. [29] (2019) proposed a novel approach for low-resolution face recognition in uncontrolled environments. Their method begins by decomposing extracted local features into a low-rank representation and a sparse error matrix. A projection matrix is then learned using a sparse-coding-based algorithm, ensuring the preservation of the sparse structure within a low-dimensional feature subspace. To determine similarity, a coefficient vector is computed through linear regression, facilitating comparison between the projected gallery and query image features. Additionally, the study introduces a morphological pre-processing technique aimed at enhancing the visual quality of images. Extensive experiments conducted on five publicly available face recognition datasets, encompassing variations in pose, facial expressions, and illumination, demonstrate that the proposed method outperforms existing state-of-the-art techniques in terms of recognition accuracy.

Yan et al. [30] (2020) addressed the challenge of low-resolution facial expression recognition, which often leads to performance degradation in real-world applications. Their study introduces a novel approach called Image Filter-based Subspace Learning (IFSL)

to enhance facial image representation. IFSL consists of three key steps: first, image filter learning is integrated into the optimization process of Linear Discriminant Analysis (LDA), allowing the extraction of discriminative image filters (DIFs) tailored to different facial expressions. Next, the filtered images are combined to generate enhanced representations. Finally, a regression-based subspace learning technique is applied, where an expression-aware transformation matrix is learned to remove irrelevant information while preserving essential facial features. Experimental evaluations on multiple facial expression datasets, including CK+, MMI, JAFFE, SFEW, and RAF-DB, demonstrate that IFSL outperforms several state-of-the-art methods in low-resolution facial expression recognition.

Rouhsedaghat et al. [31] (2021) highlighted the challenges of deploying Deep Neural Networks (DNNs) for face recognition in resource-constrained environments, where limited computing power and networking capabilities restrict the use of large-scale models. These environments require compact models that can be efficiently trained on a small dataset with low computational complexity while handling low-resolution face images. To address these limitations, the authors introduced LRFRHop, a data-efficient low-resolution face recognition model based on the Successive Subspace Learning (SSL) framework. SSL provides an explainable, non-parametric feature extraction approach that allows for a flexible trade-off between model size and recognition performance. Unlike traditional DNNs, SSL-based training is significantly more efficient, operating in a one-pass feedforward manner without backpropagation, thereby reducing training complexity. Additionally, active learning can be integrated to minimize labeling costs. The effectiveness of LRFRHop was validated through experiments on the LFW and CMU Multi-PIE datasets, demonstrating its suitability for face recognition in resource-limited settings.

Ketab et al. [32] (2023) addressed the challenge of face recognition in visual surveillance systems, which is crucial for identifying individuals involved in security incidents or investigations. Despite significant advancements in facial recognition technology, accurately recognizing faces from surveillance footage remains difficult due to variations in scale, orientation, and the presence of multiple individuals. To tackle the problem of low-resolution face recognition, this study integrates advanced feature extraction techniques. The approach leverages customized and learnable filters that enhance the training

process by aligning feature extraction with human brain functionality. The Gabor transform is applied to facial images using multiple filter coefficients at different scales and orientations, producing scale- and rotation-invariant features. Additionally, a tailored architecture with a residual stream enhances feature representation while preventing gradient interference from affecting the backbone network. Experimental evaluations on the SCface and TinyFace datasets demonstrate the effectiveness of this approach, achieving an accuracy of 89.21 % on SCface and 56.68 % on TinyFace.

Narayan et al. (2024) [33] address the challenges of low-resolution face recognition, where traditional models pre-trained on large datasets struggle due to the lack of distinct facial attributes. Fine-tuning on low-resolution datasets often leads to catastrophic forgetting of pre-trained knowledge, and the domain gap between high-resolution (HR) gallery images and low-resolution (LR) probe images further complicates adaptation. To overcome these issues, the authors propose PETALface, a Parameter-Efficient Transfer Learning (PEFT) approach that preserves learned knowledge while adapting to low-resolution inputs. PETALface incorporates low-rank adaptation modules that dynamically adjust weights based on image quality, ensuring robust performance across different resolutions. Experimental results demonstrate that PETALface outperforms full fine-tuning methods on low-resolution datasets while maintaining strong performance on high-resolution and mixed-quality datasets, all while utilizing only 0.48% of the total parameters. Kurnia et al. (2024) [34] tackle the challenge of enhancing face recognition accuracy in low-resolution images by combining Digital Image Processing (DIP) techniques with Generative Adversarial Networks (GANs). While recent advancements have achieved high accuracy in facial recognition, they primarily focus on high-resolution images, leaving low-resolution scenarios—common in surveillance and mobile applications—less effective. To address this, the proposed DIP GAN method integrates preprocessing steps such as cropping, resizing, normalization, and filtering with GAN-based image enhancement. Using the Georgia Tech Face Database, the study demonstrates that this approach significantly improves recognition accuracy for low-resolution images. The findings emphasize the critical role of preprocessing in face recognition and highlight GANs' effectiveness in enhancing low-quality facial images, contributing to advancements in digital image processing and artificial intelligence.

An et al. [35] (2025) explored the potential of machine learning (ML) models to repli-

cate human psychophysical evaluations in assessing artificial visual percepts generated by emerging methodologies. Traditional human-based evaluations are labor-intensive and require repeated experiments following any modification in hardware or software configurations. To address this, the study investigates the ability of ML models to accurately perform quaternary match-to-sample tasks using low-resolution facial images represented by phosphene arrays as input stimuli. Initially, the researchers analyzed the performance of ML models trained on a dataset of 3,600 phosphene face images to approximate innate human facial recognition capabilities. Due to time constraints and potential subject fatigue, the psychophysical test was limited to 720 low-resolution phosphene images, presented to 36 human participants. The most effective ML model successfully mirrored human behavioral trends, achieving accurate predictions for 8 out of 9 phosphene quality levels on the overlapping test set. Furthermore, the model was able to predict human recognition performance on previously untested phosphene images, reducing the need for additional psychophysical assessments. These findings highlight the transformative role of ML in accelerating research on visual prosthetics, enabling more efficient development of advanced prosthetic devices. Ren et al. (2025) [36] address the challenges of occlusion and low-resolution face recognition in university laboratory surveillance. Traditional face recognition algorithms depend on high-quality images and struggle with small, low-resolution faces, leading to inaccuracies. To overcome these limitations, the authors propose a novel framework that enhances RetinaFace-ResNet and integrates it with Quality-Adaptive Margin (AdaFace). While numerous target detection algorithms exist, they require extensive training data, yet datasets for low-resolution face detection remain scarce, negatively impacting model performance. This study improves RetinaFace's accuracy in detecting and localizing faces under extreme angles and occlusion by incorporating Spatial Depth-wise Separable Convolutions. RetinaFace-ResNet is utilized for face detection, while AdaFace enhances low-resolution recognition by leveraging feature norm approximation to assess image quality and applying an adaptive margin function. Furthermore, a multi-object tracking algorithm is introduced to mitigate issues caused by moving occlusion. Experimental results demonstrate significant improvements, achieving 96.12% accuracy on the WiderFace dataset and 84.36% recognition accuracy in real-world laboratory settings.

Alrikabi et al. (2025) [37] introduce a comprehensive resolution enhancement model

aimed at improving the quality and clarity of low-resolution images. This approach restores lost or obscured details, making images more visually appealing and informative. Enhancing low-quality images is a critical challenge in computer vision and image processing, as clear and detailed visuals are essential for accurate analysis and feature identification. By refining image resolution, specialists and researchers can uncover hidden patterns, extract relevant information, and make more informed decisions. The proposed model integrates multiple steps, including image enhancement, filtering, and preprocessing. Initially, the Labeled Faces in the Wild (LFW) dataset is loaded and resized for preprocessing. To refine image quality, Gaussian and Laplacian filtering techniques are employed—Gaussian filtering reduces noise while preserving details, whereas Laplacian filtering enhances edge detection and feature extraction. The enhancement phase blends the filtered and original low-resolution images, resulting in improved detail, sharper edges, and better overall quality. The effectiveness of the model is evaluated using Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR), demonstrating superior performance over recent methods in maintaining structural integrity while enhancing image clarity. Additionally, histogram analysis of various image sizes provides insights into how resizing affects pixel distribution and overall quality. The study highlights the model's potential for applications in computer vision and image processing, marking a significant advancement in low-resolution image enhancement.

Hayder et al. (2025) [38] highlight the rapid growth and significance of face recognition technology, given its wide range of real-world applications. However, these systems often struggle with poor-quality datasets caused by factors such as low illumination, low resolution, facial distortions, and varying environmental conditions, which negatively impact their performance and reliability. To address these challenges, Generative Adversarial Networks (GANs) have emerged as a powerful solution, capable of enhancing image quality and generating realistic facial representations under difficult conditions. This review paper examines various GAN-based models, each designed to overcome specific obstacles in face recognition. These models have demonstrated remarkable improvements in image enhancement, achieving enhancement ratios between 75% and 97% by leveraging advanced generative techniques for facial image reconstruction and augmentation. Beyond face recognition, the discussed methodologies have broader implications in fields such as medical imaging, surveillance, and autonomous driving—domains where poor data qual-

ity is a persistent challenge. By utilizing generative AI to improve dataset quality, this research contributes to the development of more robust and accurate machine learning applications across diverse fields.

The challenge of face recognition in low-resolution scenarios has driven the development of various techniques, including feature enhancement, knowledge distillation, generative models, and transformer-based approaches. Early works primarily relied on discriminant analysis and sparse coding techniques, while recent advancements leverage deep learning, particularly GANs and Vits, to enhance recognition accuracy. Future research is expected to focus on improving feature alignment across different resolutions and developing lightweight yet effective models suitable for real-time applications.

Figure 2.1: Overview of Challenges in Low-Resolution Face Recognition: Impact of Resolution Degradation on Feature Extraction.

2.3 Super-Resolution Techniques for Image Enhancement

Face recognition systems often suffer from performance degradation when dealing with low-resolution images, especially in surveillance or remote sensing scenarios. To address

this limitation, Super-Resolution (SR) techniques have emerged as powerful tools for reconstructing high-resolution images from their low-resolution counterparts. By recovering finer facial details, SR methods significantly enhance the quality of the input data, which in turn improves the accuracy and robustness of face recognition models. Recent advances in deep learning, particularly CNNs and generative adversarial networks (GANs), have further propelled the development of SR techniques, making them increasingly effective in preserving identity-relevant features. Integrating SR with face recognition not only boosts system performance under challenging conditions but also expands the applicability of these systems to real-world scenarios involving poor image quality. A summary of these research efforts is illustrated in Figure 2.2.

Fookes et al. [39] (2012)systematically examined the relationship between image resolution and face recognition performance, an aspect that has received relatively little attention despite ongoing efforts to enhance automatic face recognition. To address this gap, the authors developed a comparative framework for evaluating the effectiveness of super-resolution techniques in improving recognition accuracy. They compared three super-resolution methods against Eigenface and Elastic Bunch Graph Matching face recognition algorithms. Additionally, they identified the parameter ranges where these techniques outperform standard image interpolation in terms of recognition performance.

Raghavendra et al. [40] (2013) explored the challenges of face recognition in wide-area surveillance, which typically requires multiple cameras, handling low-resolution face samples, and high computational resources. To address these challenges, the authors experimented with a Light-Field Camera (LFC), which offers a unique advantage over conventional surveillance cameras (such as CCTV or PTZ) by capturing multiple depth-focused images in a single shot. Leveraging this capability, they investigated the use of super-resolution reconstruction based on multiple depth images rendered by the LFC. The study first presents a comparative evaluation of existing super-resolution techniques and then proposes a new resolution enhancement approach that integrates Discrete Wavelet Transform (DWT) with an existing super-resolution method. This technique captures high-frequency components from the depth images, thereby improving face recognition performance. Experimental results on a newly collected dataset highlight the advantages and limitations of using LFCs for face recognition in wide-area surveillance applications.

Rasti et al. [41] (2016) addressed the challenge of face recognition in surveillance

systems, where low-resolution images captured by video cameras often fail to meet the requirements of many recognition algorithms. To overcome this limitation, the authors proposed a deep learning-based super-resolution approach that enhances facial images before recognition. Their system first applies a deep convolutional network for image super-resolution, followed by a Hidden Markov Model (HMM) and Singular Value Decomposition (SVD)-based face recognition framework. The proposed method was evaluated on several well-known face databases, including FERET, HeadPose, Essex University, and the recently introduced iCV Face Recognition (iCV-F) database. Experimental results demonstrated a significant improvement in recognition performance after applying the super-resolution technique. Uiboupin et al. [42] (2016) addressed the challenge of low-resolution face recognition in surveillance systems, where the resolution of captured images often falls below the requirements of many recognition algorithms. To overcome this limitation, the authors proposed a super-resolution approach based on sparse representation. Their method utilizes a specialized dictionary composed of both natural and facial images to enhance the resolution of facial images. Following this enhancement, face recognition is performed using a combination of the Hidden Markov Model (HMM) and Support Vector Machine (SVM). The proposed system was evaluated on multiple well-known face databases, including FERET, HeadPose, Essex University, and the newly introduced iCV Face Recognition database (iFRD). Experimental results demonstrated a significant improvement in recognition accuracy after applying super-resolution with the facial and natural image dictionary.

ElSayed et al. (2017) [43] explored the impact of super-resolution techniques on face recognition in uncontrolled environments, where facial images are often blurred or low-resolution due to the limitations of surveillance cameras. The study aimed to evaluate the effectiveness of a state-of-the-art super-resolution algorithm in enhancing the resolution of such images, particularly when significant enlargement is required. To illustrate its functionality, the authors presented before-and-after 3D face alignment cases using images from the Labeled Faces in the Wild (LFW) dataset. The enhanced images were then tested in a closed-set face recognition protocol, utilizing unsupervised algorithms with high-dimensional feature extraction. Results showed that incorporating super-resolution significantly improved recognition accuracy, outperforming previously reported results achieved with unsupervised methods.

Kwon et al. (2018) [44] introduced a super-resolution approach that enhances face recognition performance by combining sparse representation and deep learning to process low-resolution facial images. While deep learning techniques have been extensively studied for ultra-high-resolution image generation, real-time face recognition remains an ongoing research challenge. To address this, the authors integrated sparse representation with deep learning to generate super-resolved images, improving recognition accuracy. Additionally, they optimized processing speed by implementing a parallel structure for sparse representation. Experimental results demonstrated that the proposed method outperforms conventional techniques across various image datasets. Ouyang et al. (2018) [45] introduced a super-resolution joint feature mapping approach to enhance accuracy in low-resolution face recognition. Their method employs a two-branch CNN to extract features from both high- and low-resolution face images. A super-resolution enhancement network, integrated with a feature extraction network, is applied to map low-resolution facial features, effectively reconstructing high-frequency details and improving feature extraction. Additionally, a fusion loss strategy is implemented, combining weighted cosine similarity loss and image reconstruction loss to enhance feature consistency across different resolutions. Experimental evaluations on the FERET dataset demonstrate the effectiveness of this approach, achieving recognition accuracies of 98.2%, 99.1%, and 99.5% for face images with resolutions of 20×20, 24×24, and 36×36, respectively, obtained through smooth downsampling. The proposed model outperforms state-of-the-art methods in low-resolution face recognition.

Cai et al. (2019) [46] addressed the challenge of face images in the wild that often suffer from both low resolution and occlusion, particularly in surveillance scenarios. While most existing face image recovery methods handle only one type of variation per model, the authors proposed FCSR-GAN, a deep generative adversarial network designed for joint face super-resolution and face completion using a multi-task learning approach. The generator in FCSR-GAN reconstructs high-resolution, occlusion-free face images from low-resolution, occluded inputs, while the discriminator employs multiple loss functions—including adversarial, perceptual, pixel, smoothness, style, and face prior losses to ensure the quality of the recovered images. The entire network is trained end-to-end using a two-stage training strategy. Experimental results on the CelebA and Helen databases demonstrate that FCSR-GAN outperforms state-of-the-art methods in both face super-

resolution (up to 8×) and face completion. Moreover, cross-database testing highlights its strong generalization ability, and its application improves face identification performance in cases where images suffer from both low resolution and occlusion. Wang et al. (2019) [47] explored the impact of image resolution on face recognition performance, emphasizing the challenges posed by low-resolution conditions. To enhance recognition accuracy, they proposed a super-resolution restoration method based on a simplified VGG network. Traditional super-resolution algorithms often suffer from high computational complexity and training difficulties, limiting their practical application in face recognition systems. To address these issues, the authors designed a streamlined six-layer VGG network that learns the mapping between high- and low-resolution face images using a dataset constructed based on prior knowledge of image degradation. The restoration process is completed through deconvolution-based upsampling. Experiments conducted on the LFW and ORL datasets demonstrate that the proposed approach outperforms the classical SRCNN algorithm in super-resolution performance. Moreover, when integrated with a face recognition system based on LeNet-5, it significantly improves recognition accuracy.

Chen et al. (2020) [48] addressed the challenge of super-resolving extremely low-resolution face images, a difficult task in single-image super-resolution. Traditional deep learning methods typically rely on one-step upsampling, which struggles to reconstruct high-resolution face images from very low-resolution inputs. To overcome this limitation, the authors proposed an asymptotic Residual Back-Projection Network (RBPNet), which progressively refines the reconstruction through multi-step residual learning. The process begins by projecting the reconstructed high-resolution feature map back into the low-resolution space, generating a projected low-resolution feature map. The difference between this projected map and the original low-resolution feature map produces a low-resolution residual feature map, which is then mapped back into the high-resolution space. Through iterative residual learning, the network achieves more accurate super-resolution results. Additionally, an explicit edge reconstruction mechanism is integrated into the model to enhance facial structure and reduce distortion. Extensive qualitative and quantitative evaluations demonstrate the effectiveness of RBPNet in producing high-quality super-resolved face images. Chen et al. (2020) [49] addressed the challenge of low-resolution (LR) face recognition, which remains difficult despite the remarkable advancements in deep learning-based face recognition techniques. To tackle this issue, they

proposed an identity-aware face super-resolution network designed to recover identity-related information from LR face images. Their approach explicitly disentangles identity features into two orthogonal components: the magnitude and angle of the features, which are projected into a hypersphere space. The study reveals that the magnitude of features correlates with face quality. By leveraging this identity-aware representation, the proposed method excels in recovering identity-specific textures, thereby enhancing recognition performance. Extensive experiments validate the effectiveness of the algorithm in improving LR face recognition. Rajput et al. (2020) [50] addressed the challenges faced by face recognition (FR) systems in real-world surveillance scenarios, where captured probe images are often low-resolution (LR) and noisy. They proposed a novel face super-resolution (SR) algorithm designed to enhance recognition performance by combining functional interpolation and dictionary-based SR techniques. The functional interpolation component improves the discriminability of the reconstructed images, while the dictionary-based approach effectively reduces noise during the reconstruction process. As a result, the algorithm generates high-resolution (HR) images that are both noise-free and highly discriminative, making them suitable for real-world LR face recognition tasks. Experimental evaluations on widely used face image datasets, including FEI, FERET, and CAS-PEAL-R1, demonstrate that the proposed method outperforms existing SR approaches.

Xu et al. (2021) [51] addressed the challenge of recognizing low-resolution face images, which frequently appear in practical applications such as surveillance video. Deep face recognition networks trained on high-resolution images often struggle with low-resolution inputs. To overcome this issue, the authors proposed an improved multi-branch network that integrates ResNet with feature super-resolution modules. ResNet is utilized for recognizing high-resolution facial images and extracting features from both high- and low-resolution images. The feature super-resolution modules, placed before the classifier in ResNet, enhance the resolution of extracted features, improving recognition performance for low-resolution faces. The proposed method is both simple and effective, achieving high accuracy for high-resolution faces while significantly improving recognition performance for low-resolution images. Tan et al. (2021) [52] highlighted the disparity in performance between high-resolution and low-resolution face recognition, particularly in visual surveillance systems where security cameras often capture low-resolution images.

To address this issue, they proposed an enhanced AlexNet model incorporating Super-Resolution and Data Augmentation (SRDA-AlexNet) for improved low-resolution face recognition. The approach begins with image super-resolution to enhance the quality of low-resolution images, followed by data augmentation to increase dataset diversity. The enhanced AlexNet integrates batch normalization to stabilize training by normalizing input distributions and dropout regularization to prevent overfitting, thereby improving generalization. The extracted features are then classified using the k-Nearest Neighbors (k-NN) algorithm. Experimental results demonstrate that SRDA-AlexNet outperforms existing methods in low-resolution face recognition tasks.

Ullah et al. (2022) [53] addressed the challenge of facial emotion recognition, which, while intuitive for humans, remains complex for computer algorithms. Advances in computer vision and artificial intelligence have made it possible to analyze emotions from images, but existing facial expression detection systems face limitations due to the loss of significant feature data and constrained detection performance. To overcome these issues, the authors proposed an innovative approach called "Improved Deep CNN-based Two Stream Super Resolution and Hybrid Deep Model-based Facial Emotion Detection." This method comprises three key phases: super-resolution, facial emotion recognition, and classification. The super-resolution phase employs an Improved Deep CNN for both structure and texture streams, enhancing pixel density while minimizing cross-entropy loss. In the emotion recognition phase, face detection is performed using the Viola–Jones algorithm, while feature extraction incorporates Texton, Bag of Words (BOW), GLCM, and improved LGXP features. Classification is carried out using RNN and Bi-GRU neural networks, with a voting mechanism (Score Level Fusion) ensuring accurate categorization. Experimental results demonstrate that the proposed Deep CNN-based facial emotion recognition system achieves 95% accuracy, surpassing conventional methods. The study also evaluates both positive and negative performance metrics across various databases, confirming the superiority of the proposed approach. Nan et al. (2022) [54] addressed the challenge of Facial Expression Recognition (FER) in low-resolution images, which is crucial for applications such as crowd scene analysis in stations and classrooms. The loss of discriminative features due to reduced resolution makes accurate classification of low-resolution facial images particularly difficult. To tackle this issue, the authors proposed a novel feature-level super-resolution approach based on a generative

adversarial network (FSR-FER), which enhances FER performance without reconstructing high-resolution images, thereby reducing the risk of privacy leakage. Their method employs a pre-trained FER model as a feature extractor, alongside a generator network G and a discriminator network D, both trained on features extracted from paired low- and high-resolution images. The generator G refines the low-resolution image features to align more closely with those of high-resolution images, improving their discriminative power. Additionally, to enhance classification performance, the authors introduced a classification-aware loss reweighting strategy that prioritizes samples prone to misclassification based on classification probabilities from a fixed FER model. Experimental results on the Real-World Affective Faces (RAF) Database and the Static Facial Expressions in the Wild (SFEW) 2.0 dataset show that the proposed method performs effectively across different down-sampling factors and outperforms traditional approaches that separately apply image super-resolution and expression recognition. Dos Santos et al. (2022) [55] explored the use of diffusion models, which have demonstrated effectiveness in various applications such as image, audio, and graph generation. These models have also been applied to image super-resolution and solving inverse problems. More recently, stochastic differential equations (SDEs) have been introduced to extend diffusion models to continuous time. In this work, the authors leveraged SDEs for generating super-resolution face images, marking the first known application of SDEs in this domain. The proposed method outperforms existing diffusion-based super-resolution techniques, achieving superior peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and overall consistency. Additionally, the study evaluated the method's effectiveness for face recognition by employing a generic facial feature extractor to compare super-resolved images with their ground truth counterparts, yielding improved results compared to other approaches. Wang et al. (2022) [56] addressed the limitations of existing face hallucination methods, which typically rely on facial prior information to enhance super-resolution performance. Most approaches first estimate facial priors and then use them to reconstruct high-resolution face images. However, accurately estimating these priors is challenging, particularly for low-resolution images, and any inaccuracies can negatively impact the final super-resolution results. To overcome this issue, the authors propose a novel method that incorporates facial prior knowledge during the training phase while only requiring low-resolution images during testing, eliminating the need for explicit prior estimation.

Instead of directly estimating priors, the approach leverages high-quality facial prior information during training and progressively transfers this knowledge from a teacher network (trained with low-resolution images, high-quality priors, and high-resolution images) to a student network (trained solely with low- and high-resolution image pairs). Experimental results on benchmark face datasets demonstrate that the proposed method surpasses state-of-the-art face super-resolution techniques in both quantitative and qualitative evaluations.

Tomar et al. (2023) [57] present a noise-robust face super-resolution model based on an attentive ExFeat-based generative adversarial network. The proposed approach introduces the Exigent Feature Attention Unit (ExFAU), which incorporates an Exigent Feature (ExFeat) block alongside a spatial attention mechanism to enhance the visual quality of reconstructed face images. The ExFAU block plays a crucial role in reducing noise while extracting both fine and high-level facial features. Additionally, the ExFeat block is followed by a spatial attention unit that selectively emphasizes key facial attributes while minimizing attention to less relevant features. By stacking multiple ExFAU blocks, the model refines various facial components, leading to improved overall image quality. Experimental evaluations on standard datasets, including CelebA-HQ, Helen, and LFW, demonstrate that the proposed method achieves state-of-the-art performance in face super-resolution. Zeng et al. (2023) [58] propose a self-attention learning network (SLNet) for three-stage face super-resolution, addressing the limitations of deep convolutional networks (DCNs) in capturing complete and uniform representations. Traditional face super-resolution methods either focus on extracting information within a single space using complex models for direct reconstruction or employ additional networks to enhance feature representation with multiple prior inputs. However, these approaches often struggle to achieve high-quality results. SLNet overcomes this challenge by fully leveraging the interdependence between low- and high-level feature spaces to enhance reconstruction. The model follows a hierarchical three-stage framework: first, it extracts shallow features in the low-level space; next, it refines these features under high-resolution (HR) supervision to reduce cumulative errors from DCNs while generating an intermediate reconstruction benchmark; finally, a multi-scale context-aware encoder-decoder further enhances the feature representation in the high-level space. By progressively refining features from coarse to fine, SLNet achieves competitive performance compared to state-of-

the-art methods, as demonstrated by experimental results. Gao et al. (2023) [59] address the challenge of low-resolution face images with mask occlusion, a common issue in video surveillance, particularly in the context of COVID-19 prevention and cyber-manufacturing security. Existing face super-resolution methods struggle to handle both low resolution and occlusion within a single model. To tackle this, the authors propose JDSR-GAN, an efficient joint learning network designed specifically for masked face super-resolution. The model consists of a generator, incorporating a denoising module and a super-resolution module, which processes low-quality masked face images to produce high-resolution, high-quality outputs. The discriminator employs carefully designed loss functions to enhance the realism and accuracy of the reconstructed faces. Additionally, identity information and an attention mechanism are integrated into the network to facilitate correlated feature expression and informative feature learning. By jointly performing denoising and super-resolution, JDSR-GAN enables the two tasks to complement each other, achieving superior performance. Extensive qualitative and quantitative evaluations demonstrate its effectiveness compared to existing methods.

Wang et al. (2024) [60] highlight the advancements in face super-resolution driven by deep learning neural networks. While these methods have significantly improved performance, they often rely on fixed, spatially shared kernels within standard convolutional layers, which overlook the structural diversity of facial regions. This limitation makes it challenging to reconstruct high-fidelity face images. Since facial structures play a crucial role in face representation and reconstruction, the authors propose SPADNet, a structure prior-aware dynamic network that incorporates facial structure priors to generate structure-aware dynamic kernels for more precise face super-resolution. To address the shortcomings of spatially shared kernels, the authors introduce local structure-adaptive convolution (LSAC) to model local facial feature relationships, enhancing texture representation. Additionally, a global structure-aware convolution (GSAC) is designed to capture overall facial contours, ensuring structural consistency. These components form a unified framework that balances diverse facial representations while preserving individual structural fidelity. Extensive experiments demonstrate that SPADNet outperforms state-of-the-art methods in face super-resolution. Chen et al. (2024) [61] emphasize the critical role of Face Super-Resolution (FSR) in enhancing low-resolution face images, which is essential for various face-related applications. However, existing Image Quality

Assessment (IQA) methods fail to adequately evaluate how FSR impacts identity preservation and recognizability, often leading to identity distortion or unwanted artifacts. To address this gap, the authors conduct both subjective and objective evaluations for FSR-IQA, introducing a benchmark dataset and a reduced-reference quality metric. First, they develop the Face Super-Resolution Quality Dataset (FSQD) by incorporating a novel criterion that prioritizes identity preservation and recognizability. Second, they analyze the relationship between these two factors and explore effective feature extraction techniques for assessing them. Third, they propose Face Identity and Recognizability Evaluation of Super-resolution (FIRES) a training-free IQA framework designed to assess FSR performance. Experimental results on FSQD demonstrate that FIRES achieves competitive performance, effectively addressing the limitations of traditional IQA methods in FSR evaluation. Wang et al. (2024) [62] address the challenges of face super-resolution (FSR) in low-light and low-resolution conditions, which are common in nighttime or dimly lit environments. Existing FSR methods and cascaded approaches struggle to recover reliable facial textures under such degradations. To overcome these limitations, the authors propose a novel framework that decomposes the restoration process into two key components: structural fidelity preservation and texture consistency learning. The first component focuses on enhancing image quality while maintaining structural integrity, whereas the second aims to suppress artifacts and perturbations caused by low-light conditions and reconstruction errors.

Wang et al. (2025) [63] explore the integration of face super-resolution with 3D face recognition (FR), addressing the challenge of obtaining high-quality depth images. While face super-resolution has been widely studied for 2D images, its application in 3D FR remains largely unexplored. To bridge this gap, the authors introduce an Implicit Face Model (IFM), leveraging implicit neural representations to continuously reconstruct high-resolution depth faces from low-resolution inputs. IFM operates by taking coordinates and latent codes from the low-resolution domain as inputs and predicting depth values at corresponding coordinates in the high-resolution domain. To ensure both pixel-level accuracy and identity preservation, the model incorporates an attention-guided L1 loss, prioritizing the recovery of key facial structures. Additionally, a positional encoding strategy, combining Fourier features with coordinate embedding, is employed to retain high-frequency details critical for facial identity.

Han et al. (2025) [64] highlight the advancements in face super-resolution driven by CNNs, achieving notable improvements in peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). However, existing methods often struggle with high-level vision tasks, such as face recognition, and rely on a single reconstructed image, making it difficult to recover stable and accurate high-frequency details. To address these challenges, the authors introduce Face-SRMN, a tree-based multi-branch face super-resolution network designed to enhance feature representation and high-frequency detail restoration by incorporating prior knowledge and face recognition objectives. The tree-based structure facilitates multi-branch learning, improving both feature extraction and detail preservation. Additionally, a dual-channel residual structure is employed in the basic blocks, allowing more low-frequency information to pass through. To further refine feature extraction, the model integrates dense residual connections, an attention mechanism, and a face prior information module, enabling global information adaptation across channels and spatial dimensions. Tomar et al. (2025) [65] explore the use of deep learning techniques in electronic surveillance for enhancing low-quality face images through super-resolution (SR). While many existing approaches incorporate facial priors to enhance feature details, their effectiveness is often compromised by inaccuracies when dealing with tiny, noisy, or blurry images, ultimately degrading model performance. To address this issue, the authors propose a teacher–student-based face SR framework that effectively preserves personal facial structure information in the super-resolved images. In this framework, the teacher network utilizes facial heatmap-based ground-truth priors to learn facial structure, which is then transferred to the student network. The student network is trained using an identity feature loss, ensuring that both identity preservation and facial structure details are maintained in the reconstructed high-resolution (HR) face images. The proposed framework is evaluated on standard face datasets including CelebA-HQ and LFW, demonstrating superior performance over existing face super-resolution methods. Han et al. (2025) [66] explore face super-resolution (FSR), which enhances low-resolution face images to support face recognition in challenging settings. While deep learning-based approaches such as Stable Diffusion and Transformer-based frameworks have improved FSR performance, their high computational costs make them impractical for face recognition, where an image resolution of 112×112 is typically sufficient. Existing methods primarily use convolutional attention, which limits receptive fields and hinders performance. To overcome

these limitations, the authors introduce SETFSR (Semantically Guided Efficient Attention Transformer for Face Super-Resolution). This model employs efficient self-attention for global feature extraction while integrating face parsing constraints to ensure structural accuracy. Experimental evaluations on Helen, CelebA, and FFHQ datasets demonstrate that SETFSR achieves superior results compared to state-of-the-art methods, excelling in PSNR, SSIM, and identity preservation.

Figure 2.2: Framework of Super-Resolution Techniques on Facial Feature Restoration.

2.4 Machine Learning and Deep Learning Methods for Face Recognition

Face recognition has become one of the most prominent applications in the field of computer vision, enabling a wide range of real-world uses from security and surveillance to user authentication and social media tagging. Traditional machine learning methods, laid the groundwork for early face recognition systems by extracting and classifying facial features. However, the advent of deep learning, particularly CNNs, has revolutionized the field by enabling end-to-end learning and achieving state-of-the-art performance. These deep models can automatically learn hierarchical feature representations, making them

highly effective in capturing complex variations in facial expressions, lighting conditions, and poses. As a result, the combination of machine learning and deep learning techniques continues to push the boundaries of face recognition accuracy, scalability, and real-time deployment. A summary of these research efforts is illustrated in Figure 2.3.

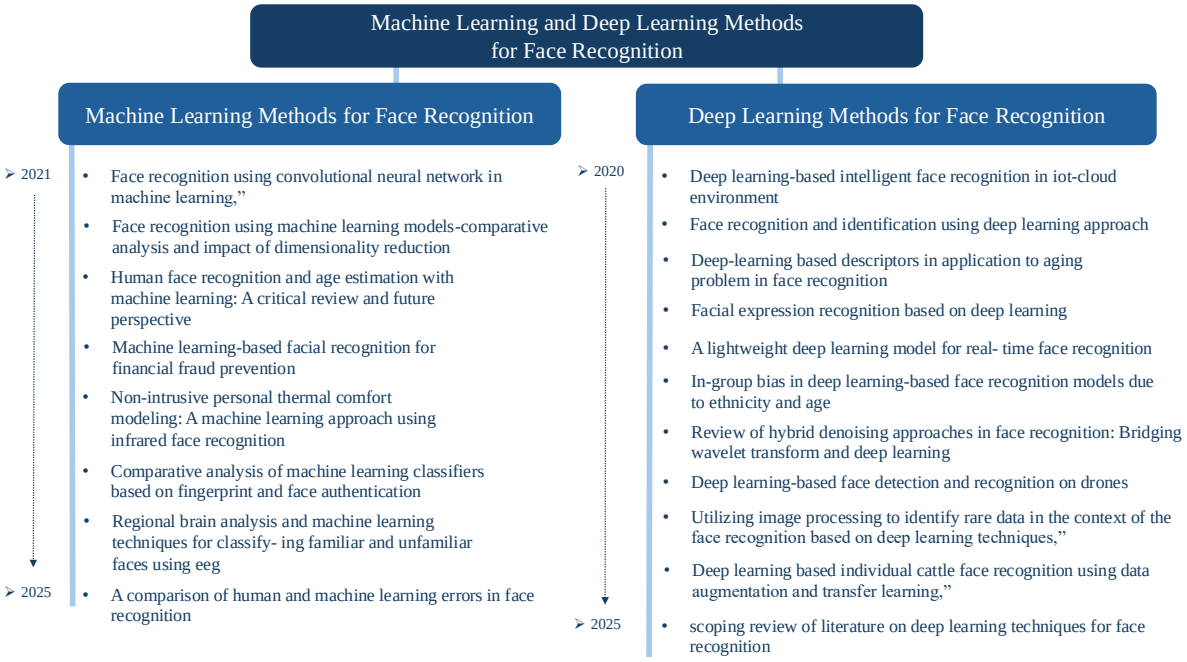


Figure 2.3: Evolution of Machine Learning and Deep Learning Models for Face Recognition.

2.4.1 Machine Learning Methods for Face Recognition

Shukla et al. (2021) [67] highlight the critical role of facial recognition in daily interactions, communication, and identity perception. Developing robust and precise face recognition algorithms is essential for fully automated systems that analyze facial data. One of the major challenges in this field is partial facial occlusion, where faces may be obscured by masks, scarves, sunglasses, hands, or objects. This study examines the effectiveness of various methodologies in handling such occlusions. The proposed model demonstrates high accuracy and low loss when tested on the specified dataset, outperforming existing algorithms. With both trainable and non-trainable parameters, the model achieves an impressive accuracy rate of 80%, indicating strong performance in facial recognition tasks from both images and video.

Yaswanthram et al. (2022) [68] explore face recognition as a biometric technique ca-

pable of uniquely identifying individuals by analyzing facial patterns. Widely adopted for security purposes, face recognition has gained even greater significance as a touchless biometric, particularly during pandemic situations. This study examines the impact of dimensionality reduction on the accuracy and efficiency of various machine learning algorithms used in face recognition, including Random Forest, Support Vector Machine, Linear Regression, Logistic Regression, and K-Nearest Neighbor. The findings reveal that Logistic Regression delivers the best performance, achieving 97% accuracy in 5.74 seconds without Principal Component Analysis (PCA). With PCA, accuracy slightly decreases to 93%, but computation time is significantly reduced to 0.15 seconds nearly 20 times faster demonstrating the trade-off between accuracy and efficiency.

Kavita et al. (2022) [69] explore the growing prevalence of Face Recognition (FR) applications across various sectors, including healthcare, government, and education. This study provides an overview of FR techniques, tools, and performance, along with a literature review highlighting research gaps. The paper discusses how FR technology, powered by machine learning, can accurately identify individuals through cameras, making it useful for applications such as virtual exams, online courses, and interactive webinars. It also analyzes machine learning and deep learning algorithms, as well as tools like MATLAB and Python. Key topics covered include face detection, recognition, expression analysis, and age estimation, offering insights to guide future research. Additionally, the study evaluates recent advancements in FR systems, considering trends and performance metrics based on recent data.

Wang et al. (2024) [70] discuss the advancements in face recognition technology and its widespread application in critical domains such as government, finance, and the military. However, like other information systems, face recognition faces significant security threats, with face spoofing (FS) being a major concern. Attackers use printed photos, video replays, and 3D masks to deceive recognition systems, making anti-spoofing measures essential. This thesis examines the role of machine learning-based face recognition in financial fraud prevention, providing solutions for transaction monitoring, identity verification, and payment security. It explores challenges posed by face spoofing, emphasizing the need for deepfake detection technologies. Additionally, the study reviews machine learning models, quantization techniques, and the trade-off between recognition speed and accuracy. Lastly, it highlights the importance of anti-spoofing measures, including

sensor selection and algorithm optimization, to enhance the security and reliability of face recognition systems.

Bai et al. (2024) [71] explore non-intrusive personal thermal comfort modeling using machine learning and infrared facial recognition. The study utilizes the Charlotte-ThermalFace database to extract facial temperatures from six key regions, employing infrared face recognition and key point extraction algorithms. Feature importance analysis using Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) identifies key parameters influencing thermal preference predictions. A comprehensive evaluation of 12 machine learning models—including traditional, ensemble, and broad learning (BL) models—reveals that ensemble and BL models outperform traditional approaches when trained on the full dataset. Notably, the BL model, applied for the first time in thermal preference prediction, achieves a precision of 90.44%, offering lower computational complexity and faster training than deep neural networks (DNN). Additionally, both BL and deep cascade forest (DCF) models demonstrate superior performance across different data subsets. This study provides valuable insights for optimizing building thermal environments through non-intrusive sensing methods.

Nuthana et al. (2025) [72] compare the effectiveness of four classifiers—SVM, Random Forest, KNN, and Neural Networks—for face and fingerprint recognition. Using Principal Component Analysis (PCA) for feature extraction, the classifiers were trained and evaluated on the LFW dataset. The study highlights the strengths and weaknesses of each model, with Neural Networks demonstrating the highest accuracy in face recognition. These findings contribute to improving biometric identification by guiding the selection of optimal classifiers.

Bardak et al. (2025) [73] highlight the crucial role of perception, memory, and social bonding in human interactions, with face recognition being a fundamental cognitive function. Recognizing both familiar and unfamiliar faces is essential for communication and interpersonal connections. This study investigates the neural basis of face recognition using EEG data from typically developed individuals, analyzed through a regional brain perspective and simple neural networks. Discrete wavelet transform (DWT) was used to extract features from EEG signals, which were then classified using three different algorithms: k-nearest neighbors (KNN), support vector machines (SVM), and probabilistic neural networks (PNN). Among these, KNN achieved the highest classification accuracy.

The study also examined accuracy across different brain regions and all EEG channels, revealing that the temporal and occipital lobes play key roles in face recognition, with distinct activation patterns for familiar and unfamiliar faces. This research enhances the understanding of how face recognition is processed in the brain, identifies critical brain regions involved, and compares machine learning techniques for EEG signal classification. The findings contribute to existing literature and offer insights that may inform future studies on the neural mechanisms underlying face recognition, including potential applications in conditions such as prosopagnosia.

Estévez-Almenzar et al. (2025) [74] emphasize the necessity of human oversight in machine learning applications, particularly in high-stakes scenarios. Achieving an optimal balance between human and machine intelligence requires understanding their complementary strengths, especially in how they differ in making errors. This study conducts extensive experiments in face recognition, comparing two automated systems with human annotators through a demographically balanced user study. The findings highlight key differences between machine and human errors and suggest strategies for improving face recognition accuracy through human-machine collaboration.

2.4.2 Deep Learning Methods for Face Recognition

Masud et al. (2020) [75] The rapid expansion of Internet-of-Things (IoT) technology has led to its adoption in various domains, including healthcare, video surveillance, and transportation. This widespread use generates vast amounts of data, particularly from IoT devices such as cameras in hospital surveillance systems. In this context, face recognition plays a crucial role in securing hospital facilities, analyzing patient emotions and sentiments, detecting fraud, and monitoring hospital traffic patterns. While automatic face recognition systems perform well in controlled environments, their accuracy declines in uncontrolled settings. Additionally, real-time processing is essential for applications like smart healthcare. This thesis proposes a tree-based deep learning model for automatic face recognition in a cloud environment, designed to be computationally efficient without sacrificing accuracy. The model partitions the input volume into multiple segments, constructing a tree for each, where each branch is defined by a residual function comprising a convolutional layer, batch normalization, and a nonlinear activation function. The proposed model is evaluated on publicly available databases and compared

against state-of-the-art deep learning models. Experimental results show that it achieves high accuracy rates of 98.65%, 99.19%, and 95.84% on the FEI, ORL, and LFW datasets, respectively.

Teoh et al. (2021) [76] The human face is a unique biometric trait used for identification, even among twins. Face recognition systems verify identity based on facial features and are widely applied in areas such as phone unlocking, criminal identification, and home security. Unlike traditional methods that require keys or cards, face recognition relies solely on facial images, enhancing security. Generally, a face recognition system consists of two phases: face detection and identification. This thesis explores the design and development of a face recognition system using deep learning with OpenCV in Python. Deep learning is highlighted as an effective approach due to its high accuracy, and experimental results demonstrate the robustness of the proposed system.

Boussaad et al. (2022) [77] Aging poses a significant challenge for face recognition systems due to biological changes that alter facial features over time. These variations can make it difficult to match images of the same person taken at different ages. As facial appearance is highly affected by aging, extracting robust features for age-invariant face recognition is crucial, especially when large age gaps exist between images. This study explores the effectiveness of deep learning-based feature extraction methods for age-invariant face recognition. Five widely used pre-trained convolutional neural networks (CNNs)—AlexNet, GoogleNet, Inception V3, ResNet50, and SqueezeNet—are evaluated on the FG-NET face-aging database using K-Nearest Neighbors (K-NN), Discriminant Analysis, and Support Vector Machines (SVM) classifiers. Additionally, a statistical analysis test is conducted to validate the significance of the results. Experimental findings confirm the potential of CNNs for face recognition across age progression, with AlexNet achieving the highest mean accuracy. The results are statistically significant at a 95% confidence level, highlighting AlexNet's effectiveness for age-invariant face recognition. Ge et al. (2022) [78] Facial expression recognition is becoming increasingly important in various applications, including autonomous driving, virtual reality, and robotics. Many computer vision tasks rely on deep learning techniques, particularly CNNs. This thesis presents an occluded expression recognition model based on a generative adversarial network (GAN). The proposed model consists of two key modules: one for restoring occluded facial images and another for face recognition.

Deng et al. (2023) [79] Lightweight deep learning models for face recognition are essential for deployment on resource-constrained devices such as mobile and embedded systems. This thesis introduces a highly efficient and compact deep learning (DL) model that achieves state-of-the-art performance across multiple face recognition benchmarks. Inspired by FaceNet, the proposed model incorporates significant optimizations, enabling it to achieve a memory size of just 3.5 MB approximately 30 times smaller than FaceNet while maintaining high accuracy and real-time performance. Utilizing one- or few-shot learning, the model effectively extracts feature embeddings and demonstrates robust recognition capabilities, even for occluded faces in uncontrolled environments with grayscale input images. Experimental results highlight its efficiency, requiring only 0.06G FLOPs, 1.2M parameters, and a minimal model size of 3.5 MB, while delivering competitive accuracy. The model is well-suited for real-world applications such as live security checks, driver authentication, and in-class attendance monitoring. Nagpal et al. (2023) [80] Humans tend to exhibit in-group bias, favoring individuals who belong to similar groups based on factors such as ethnicity, age, or shared interests. Inspired by this phenomenon, this study investigates whether deep learning models in face recognition also reflect in-group and out-group biases. As the first of its kind, this research examines whether face recognition models encode group-specific features and where bias is embedded within these networks. The analysis focuses on two key bias factors: age and ethnicity. Extensive experiments reveal key insights: (i) deep learning models emphasize different facial regions depending on the ethnic or age group, and (ii) significant variations in face verification performance are observed across different sub-groups for both existing and custom-trained models. Based on these findings, the study introduces a novel bias index to quantify the level of bias in trained models. Understanding how deep learning models encode bias can help researchers develop fairer, more robust AI algorithms, ultimately improving fairness in face recognition systems.

Zangana et al. (2024) [81] Image denoising is a fundamental aspect of image processing and acquisition, playing a crucial role in removing noise from images. In recent years, there has been a growing focus on developing advanced noise reduction techniques. This study provides a comprehensive review of major image denoising methods, with a particular emphasis on integrated deep learning approaches and traditional signal processing techniques. The review explores various methods, including CNNs, wavelet transforms,

and hybrid models, highlighting their strengths, limitations, and suitability across different applications. Additionally, the paper examines key challenges and future research directions, particularly in the context of cybersecurity and cybercrime prevention. This review serves as a valuable resource for researchers, practitioners, and enthusiasts seeking insights into the latest advancements and emerging trends in image denoising. Rostami et al. (2024) [82] Unmanned aerial vehicles (UAVs), commonly known as drones, are capable of accessing locations that are too hazardous or challenging for humans, capturing aerial data from a bird's-eye view. Enabling UAVs to detect and recognize humans on the ground is crucial for applications such as remote monitoring, search-and-rescue operations, and surveillance. However, existing face detection and recognition models struggle with variations in altitude, angle, and distance, particularly when drones capture images from high altitudes or long distances, where facial details are limited. This thesis presents a novel face detection and recognition model designed to enhance recognition performance under these challenging conditions. The proposed approach leverages deep neural networks to improve accuracy in detecting and recognizing faces with minimal facial detail. Experimental evaluations on the DroneFace dataset demonstrate that our model achieves competitive accuracy compared to state-of-the-art methods for both detection and recognition tasks.

Humood et al. (2025) [83] address the challenge of face recognition in scenarios with limited training data, such as low-quality images, varying lighting, or underrepresented demographics. To enhance recognition performance, the study introduces an automatic rejection step that detects and excludes anomalous inputs using a specialized convolutional network trained on rare data. Experimental results show that by applying this approach and using the nearest neighbors method, ResNet50 and MobileNetV1 achieve a peak accuracy of 94.9% on test data. Polat et al. (2025) [84] emphasize the critical importance of accurate cattle identification for managing ownership, controlling production supply, preventing disease spread, and ensuring animal welfare. While ear tag-based identification remains widely used in livestock management, large-scale farms face challenges in reliably identifying individual cattle, particularly when ear tags fall off, making long-term tracking impossible. To address this issue, a dataset was created by capturing images of cattle in their natural environment. The dataset was divided into training, validation, and testing sets, comprising 15,000 records from 30 individual cattle. Deep

learning models, including InceptionResNetV2, MobileNetV2, DenseNet201, Xception, and NasNetLarge, were employed to identify cattle based on facial features. Among these models, DenseNet201 achieved the highest performance, with a peak test accuracy of 99.53% and a validation accuracy of 99.83%. Furthermore, this study presents an innovative approach that combines advanced image processing techniques with deep learning, offering a robust framework that can be extended to other animal identification applications, thereby improving overall farm management and biosecurity.

Nemavhola et al. (2025) [85] Deep learning has advanced facial recognition through CNNs. This study provides a preliminary exploration of CNN architectures in face recognition, aiming to better understand the field's challenges and methodologies. A systematic literature review was conducted using the PRISMA-ScR framework, selecting 266 studies from 3,622 eligible papers. Among these, 47% proposed new techniques, while only 1% focused on method implementation and comparison. The majority of studies relied on images rather than video for training and testing, with 78% using clean data and only 7% incorporating both clean and occluded data. Traditional CNN architectures were predominantly employed, highlighting a gap in research on nontraditional CNN models, implementation strategies, and the development of face recognition systems using mixed datasets. Key challenges included occlusion, distance, camera angle, and lighting conditions. This review underscores the need for further exploration of nontraditional CNN architectures to enhance face recognition performance.

2.5 Multilinear Subspace Learning for Feature Extraction

Multilinear Subspace Learning (MSL) is an effective technique for feature extraction from complex, high-dimensional data like face images. The main goal of MSL is to use tensor structures to preserve the interactions between different dimensions of the data, which helps improve model accuracy in face recognition, particularly in low-resolution image environments. A summary of these research efforts is illustrated in Figure 2.4.

Bessaoudi et al. (2021) [86] This thesis introduces a novel multilinear and multiview subspace learning approach, termed Tensor Cross-view Quadratic Discriminant Analysis, for face kinship verification in unconstrained environments. Traditional multilinear

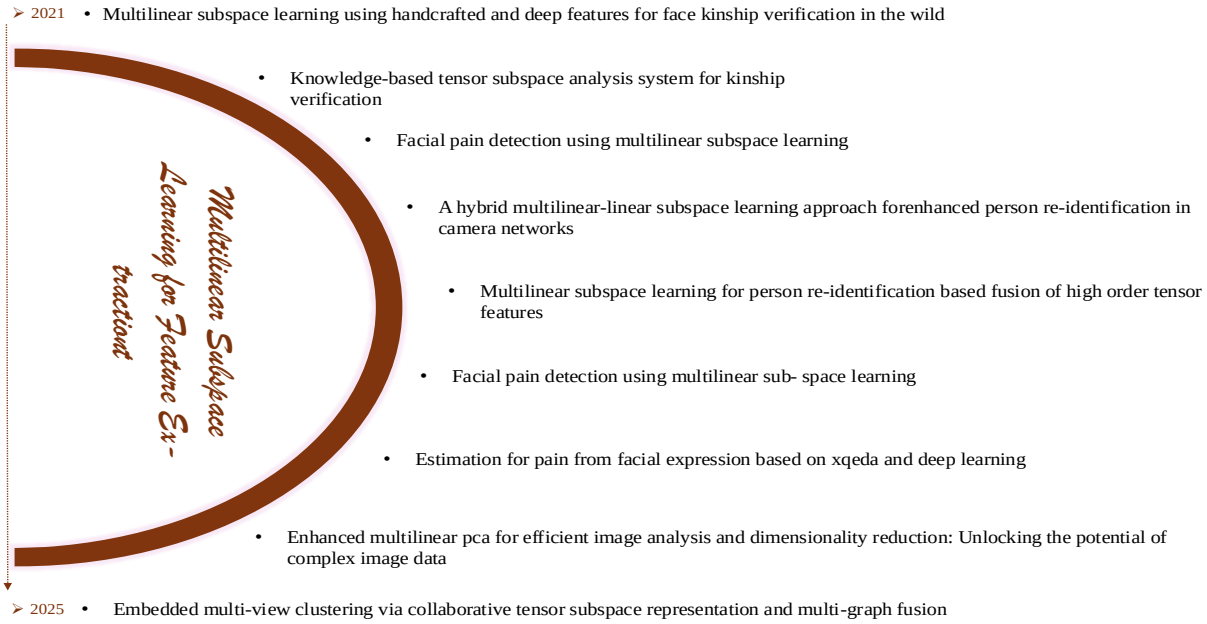


Figure 2.4: Recent Advances in Multilinear Subspace Learning: Tensor Decomposition and Feature Extraction.

subspace learning methods typically rely on a single set of projection matrices, which can hinder class separation. To overcome this limitation, the proposed method jointly learns multi-view representations for multidimensional cross-view matching. To mitigate within-class variations across different tensor data modes, the approach incorporates Within-Class Covariance Normalization. Additionally, a new tensor-based face descriptor leveraging Gabor wavelets is introduced. Furthermore, the study explores the complementarity between handcrafted and deep face tensor features by fusing them at the score level using Logistic Regression. Extensive experimental evaluations demonstrate that the proposed kinship verification framework surpasses state-of-the-art methods, achieving verification accuracies of 95.14%, 91.83%, and 93.58% on the Cornell KinFace, UB KinFace, and TSKinFace datasets, respectively.

Serraoui et al. (2022) [87] Most existing automatic kinship verification approaches focus on learning optimal distance metrics between family members. However, jointly learning both facial and kinship features may result in weaker models. To address this, the study explores a knowledge-based tensor modeling approach using pre-trained multi-view models. A novel knowledge-based tensor similarity extraction framework is proposed for automatic facial kinship verification, leveraging four pre-trained networks: VGG-Face, VGG-F, VGG-M, and VGG-S. This approach successfully fuses deep facial features with

general attributes such as identity, age, gender, ethnicity, expression, lighting, pose, contours, edges, corners, and shape, enhancing kinship understanding. Multiple effective representations are learned for kinship verification using a margin maximization learning scheme based on Tensor Cross-view Quadratic Exponential Discriminant Analysis. Through an exponential learning process, the approach effectively minimizes intra-family distribution gaps while maximizing inter-family differences. Additionally, the Within-Class Covariance Normalization (WCCN) metric mitigates the intra-class variability problem caused by deep features. The study also considers the explainability of black-box models and challenges in widespread face recognition. Extensive experiments conducted on four challenging datasets demonstrate that the proposed system outperforms state-of-the-art methods.

Chenni et al. (2023) [88] Pain identification and assessment play a crucial role in the medical field, particularly in diagnosis, treatment, drug testing, and patient care. Traditional pain measurement methods rely on patient self-reporting, which can be subjective and, in some cases—such as with stroke patients or infants—impossible to obtain. This limitation underscores the need for automated, standardized pain assessment methods. Previous studies, particularly those leveraging facial expressions for pain recognition, have reported varying degrees of accuracy, with some achieving up to 95.5% accuracy in pain prediction. This study aims to further improve the precision of pain recognition by employing a novel data pre-processing technique known as Multilinear Subspace Learning. Specifically, it explores a tensor-based adaptation of Principal Component Analysis (PCA), referred to as Multilinear PCA (MPCA), to enhance the accuracy of automated pain assessment systems.

Gharbi et al. (2024) [89] Surveillance systems rely heavily on camera networks for various tasks, including Person Re-identification (PRe-ID), a key research area in computer vision. Existing methods in this domain face significant challenges in handling appearance variations, extracting meaningful features, and capturing complex data relationships, which often results in suboptimal performance. This thesis introduces a novel hybrid subspace learning approach for PRe-ID, combining the advantages of both multilinear and linear techniques to improve re-identification accuracy. The proposed method begins by extracting robust features from pedestrian images using LOMO and GOG as shallow feature descriptors, alongside deep features obtained from a CNN. Next, a two-stage

subspace learning process is applied. The first stage employs Tensor-based Quadratic Discriminant Analysis (TXQDA), a multilinear subspace learning algorithm that effectively captures intricate feature relationships, enhancing discriminative representation. In the second stage, Side Information Linear Discriminant Analysis (SILD) is utilized to refine the learned subspace by preserving critical inter-class variations and further improving discriminability. To maximize performance, logistic regression (LR) fusion is applied at the score level, optimizing feature representation in various tensor scenarios.

Chouchane et al. (2024) [90] Video surveillance image analysis and processing play a crucial role in computer vision, particularly in the challenging task of Person Re-Identification (PRe-ID). PRe-ID aims to locate a target individual who has previously appeared in a camera network by leveraging a robust description of their pedestrian images. The advancements in PRe-ID research largely depend on effective feature extraction, representation, and powerful learning techniques to accurately distinguish between individuals. To address this, the proposed method, High-Dimensional Feature Fusion (HDFF), integrates two powerful feature descriptors: CNNs and Local Maximal Occurrence (LOMO). A novel tensor fusion scheme is introduced to combine these features into a unified tensor representation, even when their dimensions differ. To further enhance accuracy, Tensor Cross-View Quadratic Discriminant Analysis (TXQDA) is employed for multilinear subspace learning, effectively capturing discriminative information while reducing the high dimensionality of tensor data. Finally, Cosine similarity is applied for matching.

Aliradi et al. (2024) [91] highlight the importance of accurate pain assessment for medical diagnosis and treatment, as pain serves as a key indicator of discomfort. Traditional subjective methods, such as visual inspection and self-reported pain levels, are prone to human errors. Consequently, recent research has shifted towards objective pain assessment methods based on behavioral and physiological indicators. In this study, the authors propose a fully automated pain assessment system that utilizes facial expressions as a behavioral indicator. Their approach incorporates a novel fusion structure of convolutional networks to assess different pain intensities from raw facial images. The proposed system employs joint learning of robust pain-related facial features by integrating both RGB appearance and shape-based latent representations. This joint learning strategy enhances the model's robustness compared to using either representation independently.

The framework consists of two lightweight networks: the Spatial Appearance Network and the Shape Descriptor Network. Extensive evaluations on the UNBC-McMaster dataset demonstrate the effectiveness of the proposed model for both pain classification and intensity estimation. Specifically, the method achieves an F1-score of 98.80% for pain level classification, a mean absolute error (MAE) of 2.20% for pain intensity estimation, and an accuracy of 98.80%, outperforming state-of-the-art techniques in terms of area under the curve (AUC) on the UNBC-McMaster dataset.

Aliradi et al. (2025) [92] Accurately assessing pain through facial expressions is essential in medical applications such as diagnosis, treatment, and drug testing. Traditional approaches rely on patient self-reporting, which can be subjective and unreliable, particularly for individuals who cannot communicate, such as stroke patients or infants. This underscores the necessity for automated, objective pain measurement methods. In this study, we propose two novel fusion structures to enhance pain estimation accuracy. First, CNNs are utilized to predict different pain levels from facial images. Second, we introduce a multi-tensor variant of the XQDA algorithm, termed Tensor XQEDA (eXtended Quadratic Exponential Discriminant Analysis), to assess pain intensities more effectively. Unlike previous methods that relied solely on facial expressions with varying success, our approach integrates Tensor XQEDA to achieve a more precise classification. Experimental evaluations on the UNBC-McMaster dataset demonstrate that our method outperforms existing state-of-the-art techniques, achieving high classification accuracy and highlighting its effectiveness in pain assessment.

Aliradi et al. (2025) [92] emphasized the importance of accurately measuring pain from facial expressions, which plays a critical role in medical diagnosis, treatment, and drug testing. Traditional assessment methods often depend on patient self-reporting, which can be subjective and unreliable, particularly for individuals unable to communicate, such as stroke patients or infants. This underscores the need for automated systems to provide an objective evaluation of pain. In this research, two novel fusion structures are introduced to enhance pain measurement accuracy. The first approach utilizes CNNs to estimate varying pain levels from facial images. The second involves the development of a multi-tensor variant of the XQDA algorithm, termed XQEDA, to assess pain intensity. While previous methods relying on facial expressions have yielded mixed results, the proposed Tensor XQEDA (eXtended Quadratic Exponential Discriminant Analysis) offers a more

precise alternative. Experimental results on the UNBC-McMaster dataset demonstrate the superiority of this approach, achieving higher classification accuracy compared to state-of-the-art techniques.

Sun et al. (2025) [93] introduce an Enhanced Multilinear Principal Component Analysis (EMPCA) algorithm, an optimized extension of traditional Multilinear Principal Component Analysis (MPCA) designed for efficient dimensionality reduction in high-dimensional data, particularly in image analysis. EMPCA incorporates random singular value decomposition to lower computational complexity while preserving data integrity. Moreover, it integrates dimensionality reduction with the Mask R-CNN algorithm, enhancing image segmentation accuracy. By leveraging tensor-based processing, EMPCA improves dimensionality reduction for applications such as image classification, face recognition, and image segmentation. Experimental results indicate a 17.7% decrease in computation time compared to conventional approaches without sacrificing accuracy. In classification and face recognition tasks, EMPCA significantly boosts classifier efficiency, achieving results comparable to or exceeding those of Support Vector Machines (SVMs). Furthermore, its preprocessing stage exploits latent tensor structures, leading to improved segmentation outcomes. This novel EMPCA algorithm shows great potential for accelerating image analysis and advancing rapid image processing techniques.

Wang et al. (2025) [94] propose a novel embedded multi-view clustering model (EMVCTM) that leverages collaborative tensor subspace representation and adaptive multi-graph fusion to enhance clustering performance. Multi-view clustering (MVC) aims to uncover hidden correlations and underlying data distributions across multiple views. However, most existing approaches separate feature extraction from the clustering process, relying heavily on pre-learning and post-processing, which can lead to information loss and suboptimal results. To address these limitations, EMVCTM employs a tensor self-representation framework to capture global structural information, while simultaneously relaxing symmetry constraints to adaptively learn view-specific affinity graphs and dynamically construct a fusion graph. The model achieves near-global optimal clustering through an efficient embedded framework that activates interactions between views, offering a more interpretable explanation of cluster divisions. Experimental evaluations on various real-world datasets demonstrate that EMVCTM outperforms state-of-the-art clustering techniques.

2.6 Vision Transformers in Face Recognition

Face recognition has advanced significantly with deep learning, traditionally relying on CNNs. Recently, ViTs have gained attention for their ability to capture global features using self-attention mechanisms. This approach improves recognition performance, especially under challenging conditions such as low resolution, varying poses, and lighting. Several related studies supporting this direction have been summarized in Figure 2.5.

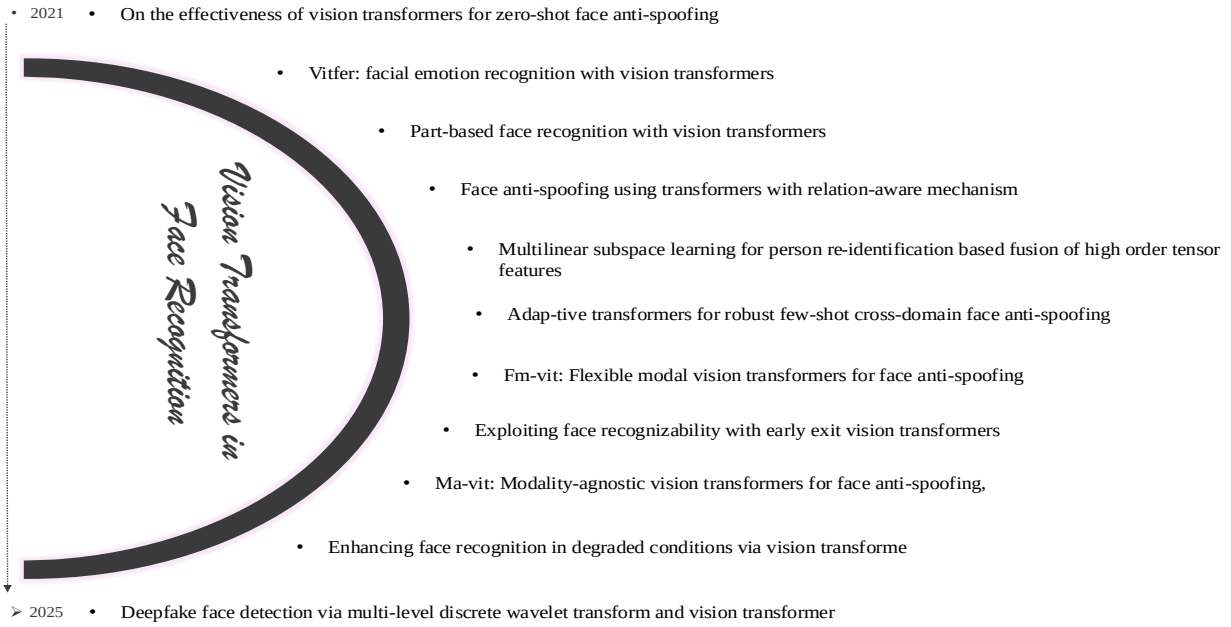


Figure 2.5: Vision Transformers in Face Recognition: A State-of-the-Art Review.

George et al. (2021) [95] highlight the vulnerability of face recognition systems to presentation attacks, which poses a significant challenge in security-critical applications. Ensuring the reliability of facial recognition technology requires robust automatic methods for detecting such attacks. While various approaches have been proposed, many tend to overfit the training data and struggle to generalize to unseen attacks and diverse environments. To address this issue, the authors leverage transfer learning from the Vision Transformer model for zero-shot anti-spoofing. Experimental evaluations on publicly available datasets demonstrate the effectiveness of their method, which outperforms state-of-the-art approaches in zero-shot protocols on the HQ-WMCA and SiW-M datasets by a substantial margin. Additionally, the proposed model exhibits a notable improvement in cross-database performance, further validating its robustness.

Chaudhari et al. (2022) [96] explore the effectiveness of automated emotion recognition, a powerful tool in various domains. Facial emotion recognition (FER) focuses on mapping facial expressions to their corresponding emotional states. In this study, the authors employ the ResNet-18 model and Transformers for FER classification, evaluating the performance of the Vision Transformer in this task. They compare their approach against state-of-the-art models on hybrid datasets. The study details the complete pipeline, including face detection, cropping, and feature extraction, using a fine-tuned Transformer model. Experimental results demonstrate that the proposed emotion recognition system is highly effective and suitable for real-world applications.

Sun et al. (2022) [97] note that convolutional neural networks (CNNs) combined with margin-based losses have traditionally dominated face recognition research. In this study, they take a different approach in two key ways. First, they introduce the Vision Transformer (ViT) as the backbone for face recognition, training a strong baseline model called fViT, which surpasses most existing state-of-the-art methods. Second, they leverage the Transformer's ability to process visual tokens from irregular grids to develop a face recognition pipeline inspired by part-based recognition techniques. Their proposed framework, named part fViT, consists of a lightweight network that predicts facial landmark coordinates, followed by a Vision Transformer that processes patches extracted from these landmarks. Notably, the system is trained end-to-end without requiring explicit landmark supervision. By learning to extract highly discriminative facial patches, the part-based Transformer further enhances the performance of the Vision Transformer baseline, achieving state-of-the-art accuracy across multiple face recognition benchmarks.

Wang et al. (2022) [98] highlight the significance of face anti-spoofing (FAS) in securing face recognition systems. While deep learning has achieved considerable success in this domain, many existing methods overlook the need for comprehensive, relation-aware local representations of both live and spoof faces. To address this gap, the authors introduce the Transformer-based Face Anti-Spoofing (TransFAS) model, which explores detailed facial parts for FAS. In addition to the multi-head self-attention mechanism, which captures relationships among local patches within the same layer, the model also incorporates cross-layer relation-aware attentions (CRA) to adaptively integrate local patches from different layers. Furthermore, to effectively combine hierarchical features, they propose an optimal hierarchical feature fusion (HFF) structure that captures the complementary information

between low-level artifacts and high-level semantic features to identify spoofing patterns. These innovations allow TransFAS to enhance the generalization ability of the classical vision transformer, achieving state-of-the-art performance on several benchmarks and demonstrating the advantages of transformer-based models for FAS.

Huang et al. (2022) [99] argue that while recent face anti-spoofing methods perform well in intra-domain settings, an effective solution must address the significant appearance variations seen in images from complex scenes and diverse sensors to ensure robust performance. In this thesis, the authors propose adaptive Vits for robust cross-domain face anti-spoofing. They leverage ViT as the backbone due to its ability to capture long-range dependencies among pixels. To enhance its adaptability to different domains, they introduce an ensemble adapters module and feature-wise transformation layers within the ViT, enabling robust performance even with limited samples. Experimental results on various benchmark datasets demonstrate that the proposed models deliver both robust and competitive performance, outperforming state-of-the-art methods in cross-domain face anti-spoofing with minimal training samples.

Liu et al. (2023) [100] highlight the growing interest in face anti-spoofing research driven by the availability of multi-modal (RGB-D) sensors. However, existing multi-modal face presentation attack detection (PAD) systems face two key limitations: (1) multi-modal fusion frameworks require input modalities to match those used during training, restricting deployment flexibility, and (2) ConvNet-based models struggle with high-fidelity datasets. To address these challenges, the authors propose the Flexible Modal Vision Transformer (FM-ViT), a transformer-based approach designed for face anti-spoofing that effectively handles single-modal (RGB) attack scenarios while leveraging multi-modal data. FM-ViT incorporates a dedicated branch for each modality and introduces the Cross-Modal Transformer Block (CMTB), which consists of Multi-headed Mutual-Attention (MMA) and Fusion-Attention (MFA). These mechanisms allow each branch to extract informative patch features and learn modality-agnostic liveness representations. Experimental results show that FM-ViT not only performs effectively across different modalities but also surpasses existing single-modal frameworks and achieves performance comparable to multi-modal systems, all while maintaining lower computational complexity.

Nixon et al. (2023) [101] highlight the challenge of balancing deep learning model com-

plexity with energy efficiency in face recognition. While advancements in the field have focused on deeper networks and larger datasets, the environmental impact of machine learning remains a growing concern. Reducing energy consumption and computational costs is becoming a priority, provided accuracy remains within acceptable limits. To address this, the authors propose a method that leverages the Vision Transformer along with Early Exits to lower computational demands without significantly compromising performance. Their system, designed for face recognition and identification within a closed-set gallery, demonstrates that a modest reduction in accuracy can lead to a substantial decrease in FLOPs, making the approach more energy-efficient while maintaining practical usability.

Liu et al. (2023) [102] point out the limitations of existing multi-modal face anti-spoofing (FAS) frameworks, which typically rely on either halfway or late fusion strategies. Halfway fusion requires test modalities to match the training input, restricting deployment flexibility, while late fusion processes different modalities separately, making it unsuitable for applications with low memory or fast execution constraints. To address these issues, the authors propose the Modality-Agnostic Vision Transformer (MA-ViT), a single-branch Transformer framework designed to enhance performance across various modal attack scenarios using multi-modal data. MA-ViT employs early fusion to aggregate all available training modalities, allowing for flexible testing with any single-modal input. Additionally, the authors introduce the Modality-Agnostic Transformer Block (MATB), which consists of Modal-Disentangle Attention (MDA) and Cross-Modal Attention (CMA) to separate modality-specific information and enhance liveness features across different modalities. Experimental results show that MA-ViT not only evaluates diverse modal samples effectively but also surpasses existing single-modal frameworks by a significant margin while achieving performance comparable to multi-modal models with fewer FLOPs and model parameters.

Ouannes et al. (2024) [103] introduce a novel method for enhancing degraded face recognition using Vision Transformer (ViT) architectures. The approach begins by feeding occluded face images into a Transformer encoder, which extracts discriminative features and creates an embedding vector crucial for the de-occlusion process. In the subsequent step, a Transformer decoder refines the representation by incorporating non-occluded features from the ground truth image. It utilizes self-attention mechanisms to integrate

information from both the embedding vector and the ground truth features. The decoder then generates a reconstructed image of the occluded face, enriched with details from both the input and the ground truth images. Finally, the decoder produces a non-occluded face image, accurately reconstructing the occluded features. By utilizing ViT in both the encoder and decoder stages, the method ensures efficient extraction and refinement of facial information, optimizing memory and time performance even under severe degradation conditions. The study evaluates the effectiveness of face recognition algorithms under various degradation scenarios, including partial occlusions, lighting variations, head poses, and facial expression changes, using two publicly available datasets: EKFD and IST-EURECOM LFFD.

Uddin et al. (2025) [104] highlight the significant security threat posed by face tampering, which involves modifying facial images using advanced software and deepfake technology. Traditional face manipulation detection methods often face limitations due to their inability to capture sufficient frequency information, reducing their effectiveness. To overcome this challenge, the authors propose a novel framework for face tampering detection that combines multi-level discrete wavelet transform (DWT) with the vision transformer. By extracting multi-level frequency features using DWT and leveraging the vision transformer's ability to capture global inconsistencies across local facial areas, the approach effectively reveals discriminative details and frequency patterns. Evaluating their method on public datasets, the authors achieve state-of-the-art accuracy rates of 99.86% on FaceForensics++ and 99.92% on Celeb-DF. These impressive results demonstrate the potential of the proposed framework in detecting deepfake face images and advancing the development of more robust facial manipulation detection systems.

2.7 Conclusion

In conclusion, face recognition in low-resolution conditions remains one of the most challenging problems in biometric computer vision, but it is also one of the most actively advancing research directions. The review presented in this chapter shows that meaningful progress is achieved not through a single technique, but through the integration of complementary approaches. Super-resolution methods help recover lost visual details, multilinear subspace learning preserves intrinsic data structure and improves discrimina-

tive representation, and Vits introduce powerful global modeling capabilities that enhance feature understanding.

The combined use of tensor-based representations, discriminant metric learning, and transformer architectures represents a strong and modern framework for building more robust verification systems. These approaches significantly improve robustness against resolution degradation, noise, and real-world variability. Moreover, experimental results reported in recent literature demonstrate that hybrid pipelines consistently outperform conventional methods, confirming their practical value.

Looking forward, future research is expected to focus on more efficient transformer models, adaptive super-resolution guided by recognition feedback, and unified end-to-end learning frameworks that jointly optimize enhancement and verification stages. There is also strong potential in combining multimodal biometrics and lightweight models suitable for real-time and embedded deployment.

Overall, the field is moving toward more resilient, structure-aware, and data-efficient recognition systems. The methodologies reviewed in this chapter provide a solid scientific and technical foundation for developing next-generation face recognition solutions capable of operating reliably even under extreme low-resolution and unconstrained conditions.

Chapter 3

Tensor-Driven Face Recognition: Integrating Super-Resolution and Multilinear Subspace Learning for Low-Resolution Images

3.1 Introduction

Face recognition has emerged as a pivotal task in computer vision, driven by the rapid progress in deep learning, increased security requirements, and the proliferation of enabling technologies such as smartphones, digital cameras, and GPUs [105, 106, 107]. The availability of large-scale datasets and standardized benchmarks has further accelerated its development [108, 109]. Despite these advancements, Low-Resolution Face Recognition (LRFR) remains a critical challenge, particularly in surveillance scenarios where facial images are often of poor quality, affected by low resolution, pose variations, illumination changes, occlusions, and loss of discriminative detail [110, 111, 112].

Surveillance systems frequently adopt low-resolution imaging to optimize bandwidth and storage, which compromises recognition performance [113, 114]. These degraded conditions necessitate novel approaches capable of restoring facial details while ensuring efficient and accurate recognition. Addressing these challenges requires a multi-stage

strategy, incorporating image enhancement techniques, robust feature learning models, and discriminative subspace learning [115, 21].

To bridge this gap, recent research has investigated the use of super-resolution (SR) models within deep learning frameworks to reconstruct high-quality facial images from degraded inputs [116, 117]. CNNs, particularly SR-oriented architectures such as URDGN, WaveletSRNet, SR-LRFR, EIPNet, and RDN, have shown promising results in improving the visual fidelity of low-resolution images [118, 119, 120, 121, 122]. However, most of these models struggle to balance reconstruction quality with computational efficiency, especially when processing large-scale low-resolution datasets.

In this study, we focus on three advanced super-resolution models SRGAN, SRResNet, and Real-ESRGAN that reconstruct perceptually realistic facial images. SRGAN uses an adversarial training mechanism with perceptual loss to align generated images with natural image statistics [5], while SRResNet leverages residual learning to reconstruct high-frequency details [5]. Real-ESRGAN improves robustness against real-world degradations by integrating high-order degradation modeling and a U-Net discriminator with spectral normalization [10].

Although SR models significantly improve visual quality, recognition tasks still face difficulties in discriminating between subtle facial differences. For instance, Hard Sample-Aware Consistency (HSAC) loss [123] targets hard samples in low-resolution facial expression recognition, improving generalization but lacking image enhancement capabilities. In contrast, our approach improves both visual quality and recognition performance by coupling SR with multilinear subspace learning, specifically Tensor TXQDA.

Furthermore, existing methods such as Coupled Discriminative Manifold Alignment (CDMA) [5] and cross-resolution techniques like Octuplet loss [124] show effectiveness in aligning LR and HR features or minimizing resolution-induced gaps. However, they either depend on paired training data or complex downsampling protocols. Other recent works have addressed pose variation [125] and face alignment [126], but without directly enhancing the resolution of facial images.

Unlike these approaches, we propose a unified framework that first enhances image resolution using deep SR models and then projects the resulting features into a discriminative multilinear subspace using TXQDA. Multilinear Subspace Learning (MSL) methods [127, 128] offer compact and structurally aware representations by operating on high-

order tensors instead of flat vectors, enabling richer relational encoding. TXQDA [16, 129] refines this process by optimizing class separability and dimensionality reduction simultaneously across all tensor modes.

To address the challenges of LRFR, we propose a robust framework combining SR and MSL. SRGAN, SRResNet, and Real-ESRGAN enhance image quality by reconstructing fine facial details, thereby enriching the data available for downstream recognition. The enhanced images are then processed through TXQDA, which maps them into a lower-dimensional, yet highly discriminative tensor space. This synergy facilitates superior classification performance under suboptimal imaging conditions.

Our key contributions are summarized as follows:

- We propose a high-order tensor-based facial recognition framework that leverages deep features extracted from the `fc6` and `fc7` layers of a pre-trained VGG-Face network, avoiding conventional feature fusion while preserving spatial coherence.
- We integrate advanced super-resolution techniques (SRGAN, SRResNet, and Real-ESRGAN) to enhance the perceptual quality of low-resolution facial images, thereby improving the robustness of feature extraction.
- We apply the Tensor-based TXQDA method for discriminative subspace learning, which preserves multi-way data structure, reduces dimensionality across tensor modes, and maximizes inter-class separability with minimal computational overhead.

The remainder of this thesis is structured as follows: Section II reviews the state-of-the-art methods in LRFR and super-resolution. Section III details our proposed approach. Section IV describes the experimental setup and presents evaluation results. Section v compares our method with other techniques. Finally, Section VI concludes the paper and outlines future research directions.

3.2 Synthesis of Findings and Research Gaps in Face Recognition

The survey of recent studies on face recognition and super-resolution underscores notable advancements alongside persistent limitations in handling low-resolution and degraded facial images.

3.2.1 Key Findings

Numerous works have emphasized the necessity of adapting models to challenging scenarios, such as low resolution, pose variation, and illumination inconsistencies. For instance, [130] exploits Local Binary Patterns (LBP) for robustness, although its reliance on nearest-neighbor classification restricts scalability. In another example, [131] improves recognition in low-light conditions using feature restoration, achieving a 6% verification boost on the Specs on Faces dataset. Techniques such as the one proposed in [132] apply advanced feature fusion for low-resolution recognition, while SwinDPSR [133] leverages transformers to independently model global and local facial features.

Additionally, preserving high-frequency components has gained increasing attention. The approach in [134] integrates wavelet-based networks to reduce feature distortion, emphasizing the importance of retaining fine-grained visual cues. Super-resolution techniques such as ARASFSR [135] extend image scalability by enhancing facial features across arbitrary resolutions, addressing the rigidity of traditional SR models.

3.2.2 Research Gaps

Despite these developments, several issues remain unresolved. GAN-based super-resolution methods [136, 137] often yield high-quality images but are sensitive to training instability and prone to generating artifacts. The complexity of feature fusion architectures, as discussed in [132], poses challenges in achieving real-time performance without compromising accuracy.

Another persistent limitation lies in poor generalizability across datasets. For instance, personalized face restoration models such as PFStorer [138] perform well on individual-specific data but struggle with generalized recognition tasks. Furthermore, many approaches prioritize perceptual enhancement at the expense of recognition accuracy [134, 133], which can hinder practical deployment in real-world systems where identification is the primary objective.

3.2.3 Our Contribution

To address these limitations, our proposed framework incorporates a multi-model super-resolution pipeline combining SRGAN, SRResNet, and Real-ESRGAN. These models, each tailored to handle specific types of image degradation (e.g., noise, blur, compression artifacts), are employed to improve facial image clarity under diverse and complex conditions. This design strategy directly tackles the shortcomings of earlier models that falter under extreme low-resolution variability.

In the subsequent phase, features are extracted using the pre-trained VGG-Face network, known for its deep semantic representation capabilities. Unlike traditional methods relying on handcrafted or shallow descriptors, this approach capitalizes on high-level abstraction, enabling robust representation of diverse facial appearances. These deep features are then structured into high-order tensors, preserving their spatial and semantic relationships across multiple dimensions.

A central innovation of our method lies in the integration of Tensor-based Cross-Modal Quadratic Discriminant Analysis (TXQDA), a multilinear subspace learning technique. TXQDA captures intricate interactions within tensor-structured data and simultaneously maximizes inter-class separability while minimizing intra-class variance. By performing subspace learning in a high-order space, TXQDA effectively models both linear and non-linear variations in facial imagery, outperforming traditional vector-based techniques.

Finally, the system uses cosine similarity on the reduced tensor representations for verification. This approach aligns with Bayesian decision theory principles and offers a computationally efficient and scalable solution for matching, especially critical in low-resource environments. Altogether, the synergy between super-resolution enhancement, tensor-based representation, discriminative subspace learning, and similarity-based matching results in a robust framework that significantly improves recognition performance under real-world constraints.

3.3 Methodology

As illustrated in **Fig.** 3.1, the architecture of the proposed framework consists of four primary components: (1) facial image super-resolution, (2) deep feature extraction, (3) tensor-based subspace learning, and (4) facial comparison. The methodology places par-

ticular emphasis on the tensor representation and subspace learning stages, utilizing the pre-trained VGG-Face network [12, 139] for deep feature generation.

A tensor is a mathematical generalization of scalars, vectors, and matrices to higher dimensions [140, 141]. Specifically, a tensor of order n is defined as an element of $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$, where each I_k corresponds to the size along the k^{th} dimension. For this study, the constructed third-order tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined as follows:

- I_1: the number of deep feature parts (set to 4), obtained by dividing the fc6 and fc7 feature vectors into multiple segments.
- I_2: the feature dimension, fixed at 1024.
- I_3: the number of facial image samples in the dataset.

This tensorial structure allows for preserving intrinsic correlations across parts of the deep feature vectors and across facial samples, forming the basis for the subsequent TXQDA-based discriminative subspace learning.

Our algorithm leverages the multidimensional representation of data through high-order tensors. To facilitate a clearer understanding of the mathematical symbols and operations used in the algorithm, Table 3.1 provides a comprehensive overview of key notations. It outlines the types and specific roles of scalars, vectors, matrices, and tensors, which serve as the foundational elements for data representation and computational procedures throughout the methodology.

The proposed system employs deep features for facial representation, using the VGG-Face network as the feature extractor. This pre-trained network, built on a large-scale facial dataset, comprises convolutional, pooling, and fully connected layers. In particular, we utilize the fc6 and fc7 layers, which are known for effectively capturing and encoding high-level facial characteristics [86].

During the offline training phase, optimal multilinear projection matrices are computed. These learned matrices are subsequently used during the online testing phase to project and match new facial samples. In the training phase, paired face feature vectors are encoded into third-order tensors $\mathbf{X}$ and $\mathbf{Z}$.

Following the TXQDA approach, tensors $\mathbf{X}$ and $\mathbf{Z}$ undergo dimensionality reduction along modes I_1 and I_2, resulting in reduced tensors with dimensions $\mathrm{I}'_1 \times \mathrm{I}'_2$. For test samples (face test pairs), the feature extraction follows the same process as in training,

representing each face as a second-order tensor. TXQDA then projects these test tensors, producing output matrices E_1 and E_2 with dimensions $\mathrm{I}_1' \times \mathrm{I}_2'$, which are significantly smaller than the original dimensions $\mathrm{I}_1 \times \mathrm{I}_2$. The matrices E_1 and E_2 are then vectorized by concatenation to yield feature vectors e_1 and e_2, each of size $\mathrm{I}_1' \times \mathrm{I}_2'$. This dimensionality reduction facilitates efficient and accurate face verification with reduced computational cost [16] [87].

Table 3.1: Explanation of Mathematical Notations Used in the Proposed Methodology

Symbol	Type	Description
i, j	Scalars	Represent individual elements or indices in equations.
v, u	Vectors	Denote one-dimensional arrays, e.g., features.
U, W	Matrices	Two-dimensional arrays used in projections and operations.
$\mathbf{X}$, $\mathbf{Y}$	Tensors	Multidimensional arrays, where $\boldsymbol{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n}$ represents an n^{th}-order tensor.
$\mathcal{R}$	Set	Denotes the set of real numbers.
n	Scalar	Represents the order of a tensor.
I_1, I_2, I_n	Scalars	Dimensions of a tensor along each mode.
D	Function	Degradation process applied in super-resolution.
r	Scalar	Downsampling factor in super-resolution.
Σ_w, Σ_b	Matrices	Covariance matrices representing within-class and between-class variance.
W_k	Matrix	Projection matrix for mode k in tensor subspace learning.
$\mathbf{I}^{\mathbf{HR}}$, $\mathbf{I}^{\mathbf{LR}}$	Tensors	Represent high-resolution and low-resolution images.
G	Function	Generator network in SRGAN and other super-resolution models.
D	Function	Discriminator network in adversarial learning.
$\otimes$	Operator	Denotes convolution operation in Equation (1).
$\circ$	Operator	Denotes composition of functions in Equation (2).

3.3.1 Super Resolution of Face Images

The main objective of Low-Resolution Face Recognition **(LRFR)** is to extract facial characteristics that are robust to variations in image resolution. In our study, we have selected three super-resolution models to enhance the quality of the input images and improve the performance of **LRFR**.

3.3.1.1 Real-ESRGAN

One of the models adopted in this study is **Real-ESRGAN**, which extends the Enhanced SR Generative Adversarial Network (**ESRGAN**) to real-world image restoration tasks.

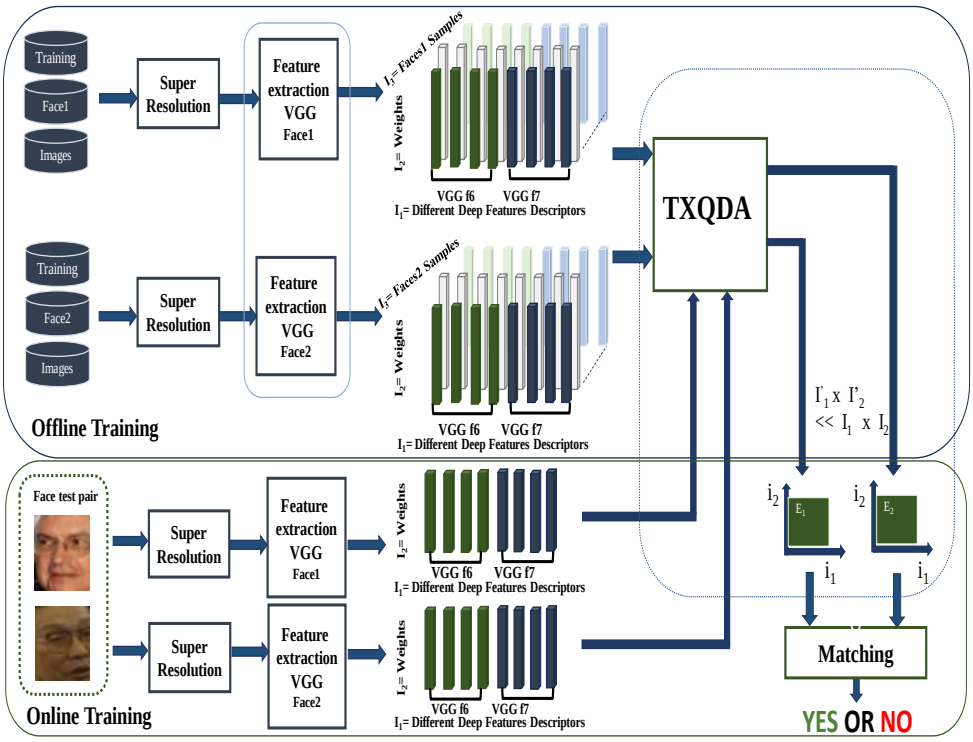


Figure 3.1: Block diagram of the proposed low-resolution face recognition using super-resolution and multilinear-linear subspace learning

Real-ESRGAN is a sophisticated deep learning model designed to recover high-resolution images from low-resolution inputs affected by unknown and complex degradation processes. While the classical degradation model is typically used to synthesize low-resolution images, Real-ESRGAN introduces a high-order degradation process to simulate more realistic artifacts. Sinc filters are employed to model typical ringing and overshoot effects. Furthermore, a U-Net discriminator with spectral normalization is adopted to enhance the capacity of the discriminator and stabilize training.

The degradation process involves convolving the ground-truth image $\mathbf{Y}$ with a blur kernel K, applying downsampling with a scaling factor r, and adding noise n. Additionally, JPEG compression is applied to simulate real-world distortions. This can be mathematically formulated as:

$$\mathbf{X} = D(\mathbf{Y}) = [(\mathbf{Y} \otimes K) \downarrow r + n]_{\text{JPEG}} \tag{3.1}$$

where D represents the overall degradation function. A comprehensive description of these degradations is provided in [10].

Although the classical degradation model enables the trained network to handle some

real-world inputs, it remains insufficient when addressing complex and unknown degradation types. This shortfall arises from the disparity between synthetic and realistic degradations. Thus, it becomes crucial to adopt a high-order degradation process that better reflects the intricacies of real-world image distortions, including factors such as camera hardware, image editing software, and Internet transmission [10].

Unlike the first-order model, which includes only a fixed set of basic operations, the high-order degradation process comprises n sequential applications of the classical degradation model, each with distinct hyperparameters. This is mathematically defined as:

$$\mathbf{X} = D^n(\mathbf{Y}) = (D_n \circ \cdots \circ D_2 \circ D_1)(\mathbf{Y}) \tag{3.2}$$

To preserve a reasonable resolution across degradation stages, the downsampling operation in Eq. 3.1 is replaced with a random resizing operation. In practice, a second-order degradation process is adopted, as it effectively balances model robustness and simplicity [10].

3.3.1.2 SRGAN

Another employed model is **SRGAN**, a generative adversarial network designed for single-image SR. It is the first method capable of producing photo-realistic natural images at large upscaling factors. SRGAN optimizes a perceptual loss function that combines adversarial and content-based components. The discriminator is trained to distinguish super-resolved images from real ones, thereby guiding the generator to produce outputs that follow the natural image distribution [5].

In single-image super-resolution (SISR), the goal is to estimate a high-resolution image $\mathbf{I}^{\mathrm{SR}}$ from a given low-resolution input $\mathbf{I}^{\mathrm{LR}}$, which itself is derived from the high-resolution ground truth $\mathbf{I}^{\mathrm{HR}}$. During training, the low-resolution image is generated by applying a Gaussian blur followed by downsampling with a factor r. Given a color image with C channels, $\mathbf{I}^{\mathrm{LR}} \in \mathbb{R}^{W \times H \times C}$, and $\mathbf{I}^{\mathrm{HR}} \in \mathbb{R}^{rW \times rH \times C}$.

The generative model is a feed-forward CNN G_{θ_G} with parameters $\theta_G = \{W_{1:L}, b_{1:L}\}$, where L is the number of layers. The objective is to minimize the SR loss function $\mathcal{L}^{\mathrm{SR}}$

over N training image pairs:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}^{\mathrm{SR}}(G_{\theta_G}(\mathbf{I}_n^{\mathrm{LR}}), \mathbf{I}_n^{\mathrm{HR}}) \tag{3.3}$$

The perceptual loss $\mathcal{L}^{\mathrm{SR}}$ is a key component that measures the similarity between the generated and ground-truth images using perceptually relevant features. It is formulated as:

$$\mathcal{L}^{\mathrm{SR}} = \mathcal{L}_X^{\mathrm{SR}} + 10^{-3}\mathcal{L}_{\mathrm{Gen}}^{\mathrm{SR}} \tag{3.4}$$

where $\mathcal{L}_X^{\mathrm{SR}}$ denotes the content loss and $\mathcal{L}_{\mathrm{Gen}}^{\mathrm{SR}}$ is the adversarial loss.

Following Goodfellow's framework [**?**], a discriminator D_{θ_D} is trained alongside the generator in a min-max adversarial setting:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{I}^{\mathrm{HR}} \sim p_{\mathrm{train}}}[\log D_{\theta_D}(\mathbf{I}^{\mathrm{HR}})] + \mathbb{E}_{\mathbf{I}^{\mathrm{LR}} \sim p_G}[\log(1 - D_{\theta_D}(G_{\theta_G}(\mathbf{I}^{\mathrm{LR}})))] \tag{3.5}$$

This framework encourages the generator to produce outputs that are visually indistinguishable from real high-resolution images, surpassing the limitations of conventional MSE-based SR methods by capturing perceptual quality and image realism [5].

3.3.1.3 SRResNet

Among the choices, one was the SRResNet, which was utilized for SR. Resnet, short for Residual Neural Network, is a type of DL architecture. It's known for its ability to effectively train very deep neural networks. The key innovation in Resnet is the introduction of skip connections, which allow information to bypass certain layers. This helps in mitigating the vanishing gradient problem and enables the training of very deep networks. The skip connections provide a shortcut for the gradient to flow through the network during backpropagation, allowing for easier and more stable training [142]. They incorporate residual learning into specific sets of stacked layers, constituting a fundamental building block. In their paper, this building block is formally defined as follows:

$$y = F(x, (W_{\mathrm{i}})) = x \tag{3.6}$$

where x and y represent the input and output vectors of the respective layers, the

function $F(x, (W_i))$ denotes the residual mapping that is to be acquired through learning [142].

Achieved a groundbreaking advancement in single-image super-resolution (SR) for high upscaling factors (4x), surpassing previous benchmarks in terms of both Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). This was accomplished using their specialized SRResNet, a Residual Neural Network consisting of 16 blocks, meticulously fine-tuned for Mean Squared Error (MSE) optimization [5].

$$y = F(x, (W_i)) = x \tag{3.7}$$

where x and y represent the input and output vectors of the respective layers, the function $F(x, (W_i))$ denotes the residual mapping that is to be acquired through learning [142].

3.3.1.4 Evaluating Leading Super-Resolution Methods

Table 3.2 presents a comparison of three widely used super-resolution models: Real-ESRGAN, SRGAN, and SRResNet. Each model employs a distinct strategy to enhance low-resolution images. Real-ESRGAN utilizes a high-order degradation model combined with a U-Net discriminator and spectral normalization, making it robust to unknown real-world degradation scenarios; however, it struggles with extremely complex artifacts. SRGAN leverages a perceptual loss that combines adversarial and content losses within a GAN framework, focusing on generating photo-realistic images with natural distributions, though its performance is highly sensitive to the quality of GAN training. SRResNet employs residual connections and MSE optimization to enhance high upscaling factors, providing stable and high-quality reconstruction, but its deep architecture results in high computational load, making it less suitable for real-time applications. The table also highlights the performance metrics commonly used to evaluate these models, including PSNR and SSIM, along with visual quality for SRGAN.

Table 3.2: Comparison of Super-Resolution Models.

Model	**Key Features**	**Performance Metrics**	**Degradation Handling**	**Limitations**
Real-ESRGAN	High-order degradation model, U-Net discriminator, spectral normalization	PSNR, SSIM	Handles unknown real-world degradation scenarios	Struggles with extremely complex artifacts
SRGAN	Perceptual loss combining adversarial and content loss, GAN-based	PSNR, SSIM, Visual Quality	Focuses on photo-realistic images with natural distributions	Sensitive to the quality of GAN training
SRResNet	Residual network with skip connections, MSE optimization	PSNR, SSIM	Enhanced for high upscaling factors, deep architecture	High computational load, not ideal for real-time applications

3.3.2 Feature Extraction and High Order Tensor Representation

Feature extraction is a fundamental step in face verification systems, as it involves generating biometric signatures typically in the form of feature vectors or matrices for each individual in the database [86]. In our system, we adopt deep features as the primary type of representation and employ the VGG-Face network [13] as the feature extractor.

VGG-Face is a CNN specifically trained for face verification tasks. It was trained on a large-scale dataset containing more than 2.6 million facial images from 2622 unique identities. Although initially designed for identity verification, VGG-Face has demonstrated robust performance as a general-purpose face descriptor across various face analysis tasks [143].

The network accepts 224×224 RGB face images as input and consists of multiple convolutional, pooling, and fully-connected layers, concluding with a Softmax layer. Each convolutional layer is followed by a ReLU activation function, which introduces nonlinearity into the model. Of particular importance are the fully-connected layers `fc6` and `fc7`, which are responsible for capturing high-level semantic features. The `fc6` layer produces feature vectors based on the learned weights and biases from the preceding layers, while `fc7` further processes this output, providing rich facial representations that

encapsulate identity-specific traits.

In our implementation, these deep features extracted from the `fc6` and `fc7` layers serve as the biometric signatures for face verification, offering a compact yet discriminative representation of facial identity [86].

3.3.2.1 High-Order Tensors in Feature Representation

Discriminative feature representation is fundamental for achieving high accuracy and robustness in face recognition systems [89]. To harness the powerful discriminative capabilities of multilinear subspace learning specifically the TXQDA algorithm used in the subsequent stage we developed a structured methodology for representing features in a multidimensional space.

During the feature extraction stage, we employed a pretrained VGG-Face model and focused on two deep descriptors, `fc6` and `fc7`, each producing a feature vector of 1024 dimensions. These vectors serve as compact yet expressive descriptors for each facial image. By combining the two, we constructed an initial third-order feature tensor that facilitates richer and more structured representations of facial characteristics.

To further refine the tensor construction, we partitioned each feature vector into four equal segments. These segments were then arranged into a feature matrix, yielding a second-order tensor where mode-1 corresponds to feature values and mode-2 reflects the segmented structure. This intermediate representation allows for meaningful intra-vector interactions.

Building upon this, we assembled the feature matrices of all subjects into a third-order tensor, where mode-3 represents different individuals in the dataset. This comprehensive tensor formulation illustrated in Fig. 3.2 effectively captures complex interdependencies across features, segments, and subjects using the outputs from the `fc6` and `fc7` layers.

Such a multidimensional representation significantly enhances the model's capacity to discern subtle variations across individuals, contributing to improved recognition performance. As demonstrated in the experimental results, this structured approach enables the face recognition system to operate with increased reliability and accuracy in real-world scenarios [144, 145, 112, 90, 89].

3.3.2.2 Benefits and Interpretability of High-Order Tensors

High-order tensors offer substantial advantages over traditional matrix-based representations, particularly in complex pattern recognition tasks such as face verification [146, 144, 90]. Unlike conventional two-dimensional matrices, tensors are capable of capturing and preserving intricate multi-dimensional relationships inherent in facial data. This is especially critical in face recognition, where the interplay between facial features, image samples, and learned weights significantly impacts identification performance. Tensor-based representations thus provide a richer, more holistic understanding of facial structures, enhancing the system's ability to effectively distinguish between identities.

One key benefit of high-order tensors lies in their efficient data organization. Rather than treating features as independent elements, tensors structure them into coherent and compact forms that reduce redundancy and streamline computations. For example, by segmenting feature vectors and assembling them into tensor forms, models can focus on the most salient attributes for classification. This structured representation improves both computational efficiency and recognition performance, while ensuring that vital intra-feature relationships are retained.

Tensors also exhibit superior generalization capabilities. Their ability to encode detailed interactions between features enables models to learn more robust and invariant patterns. This proves particularly valuable in real-world face recognition applications, where systems must contend with variations in lighting, pose, expression, and occlusion. Tensor representations empower models to maintain high accuracy across these challenging scenarios by leveraging the enriched contextual information captured during training.

Furthermore, tensor-based methods provide powerful tools for dimensionality reduction. Algorithms such as TXQDA exploit the multi-mode structure of tensors to project high-dimensional data onto low-dimensional subspaces while preserving discriminative characteristics. This reduces computational complexity without compromising recognition accuracy a critical advantage in large-scale face datasets.

Finally, high-order tensors enhance interpretability. Each mode in a tensor corresponds to a specific dimension of the data such as features, segments, or subjects—allowing for clearer insight into the role of each component in the recognition process. This facilitates a more transparent and explainable model, supporting both performance improvements and deeper analytical understanding.

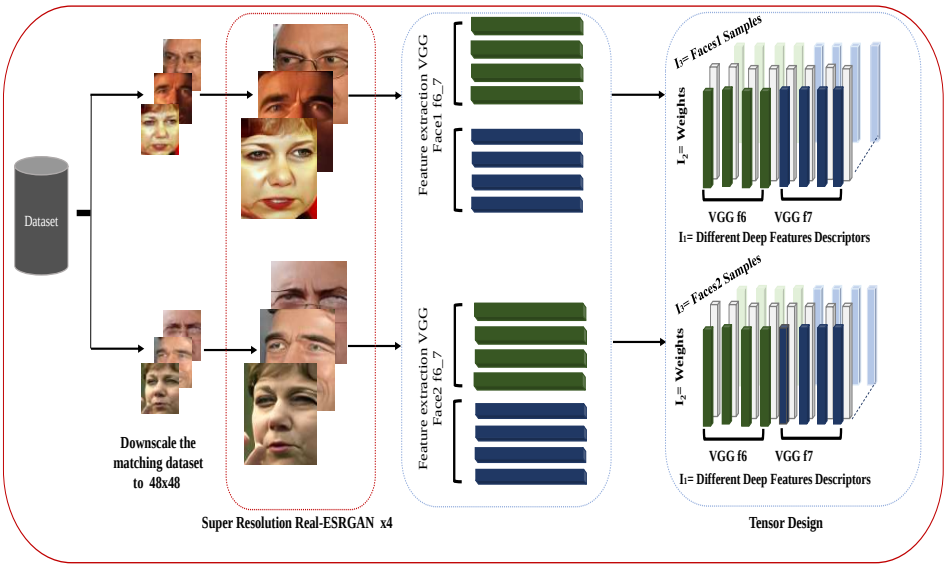


Figure 3.2: Feature extraction and high-order tensor representation are demonstrated in this study. We provide an example using a matching dataset to illustrate how feature extraction and high-order tensor representation are utilized. (The same process was performed for both matching and non-matching data across all three sizes and the three super-resolution levels).

3.3.3 Multilinear Subspace Learning using TXQDA

3.3.3.1 Multilinear Subspace Learning and Tensor Cross-View Quadratic Discriminant Analysis

Multilinear Subspace Learning (MSL) offers numerous advantages in pattern recognition tasks by effectively capturing the intrinsic structures and patterns present in high-dimensional data [129, 147]. Unlike traditional linear methods, MSL can model both linear and nonlinear relationships within data, which is crucial for handling complex and diverse visual scenarios that often deviate from simple linear assumptions [148]. This capability enables MSL to generate more concise and discriminative representations of visual information, enhancing feature descriptiveness and robustness [149].

Moreover, MSL contributes significantly to dimensionality reduction, leading to considerable savings in computational resources and memory usage. This makes it particularly valuable for efficiently processing large-scale, high-dimensional visual datasets [150].

By improving generalization and producing more informative feature representations, MSL enhances performance in recognition and classification tasks, rendering it a highly advantageous approach in computer vision applications [151, 111].

Among the recent advances, Tensor Cross-View Quadratic Discriminant Analysis (TXQDA) has emerged as a popular and effective algorithm for face recognition [89]. TXQDA seeks to project multidimensional tensor data into a discriminative subspace where the distance between positive pairs (face images of the same individual) is minimized, while the distance between negative pairs (face images of different individuals) is maximized. This optimizes class separability and enhances recognition accuracy.

Given a tensor-based cross-view training set $\mathbf{X}$ and $\mathbf{Z}$ with c classes, the algorithm operates on two different views (modalities or feature sets), aiming to jointly learn projections that maximize discrimination across these views,

$\mathbf{X} \in \mathcal{R}^{\mathrm{I_1 \times I_2 \times \ldots \times I_N \times n}}$, representing n samples from the "Face1" view,

$\mathbf{Z} \in \mathcal{R}^{\mathrm{I_1 \times I_2 \times \ldots \times I_N \times m}}$, representing m samples from the "Face2" view,

The TXQDA method aims to compute N projection matrices:

$$\left(W_1 \in \mathbb{R}^{I_1 \times I_1'},\ W_2 \in \mathbb{R}^{I_2 \times I_2'},\ \ldots,\ W_N \in \mathbb{R}^{I_N \times I_N'}\right),$$

each corresponding to one mode of the tensor. This involves learning a distinct projection matrix for each dimension of the input tensor. Consequently, the objective function from XQDA [14] is adapted to the multilinear context, reflecting the structure of high-order tensor representations and their mode-specific projections.

$$\mathrm{J}(W_\mathrm{k}) = \frac{W_{\mathrm{k}^\mathrm{T}} \Sigma_\mathrm{w}^\mathrm{K} W_\mathrm{k}}{W_{\mathrm{k}^\mathrm{T}} \Sigma_\mathrm{b}^\mathrm{K} W_\mathrm{k}}, \tag{3.8}$$

This adjusted objective function encapsulates the essence of our tensor-based approach. We compute the covariance matrices $\Sigma_\mathrm{b}^\mathrm{K}$ and $\Sigma_\mathrm{w}^\mathrm{K}$, for each mode K through the following process:

$$\begin{aligned}&\Sigma_{\mathrm{b_n}}^{\mathrm{k}} = \sum_{\mathrm{l}=1}^{\prod_{\mathrm{o}\neq\mathrm{k}} \mathrm{n_o}} \Sigma_{\mathrm{b_n}}^{\mathrm{l}}, \Sigma_{\mathrm{b_n}}^{\mathrm{k,l}} = X^{\mathrm{k,l}}(X^{\mathrm{k,l}})^{\mathrm{T}} + Z^{\mathrm{k,p}}(Z^{\mathrm{k,l}})^{\mathrm{T}} - (\textstyle\sum_{\mathrm{D_{i=1}}} h_\mathrm{i}^{\mathrm{k,l}}, ..., \sum_{\mathrm{D_{i=c}}} h_\mathrm{i}^{\mathrm{k,l}}) \\ &(\textstyle\sum_{\mathrm{H_{j=1}}} e_i^{k,l}, ..., \sum_{\mathrm{H_{i=c}}} e_\mathrm{i}^{\mathrm{k,l}})^{\mathrm{T}} - (\sum_{\mathrm{H_{j=1}}} e_\mathrm{i}^{\mathrm{k,l}}, ..., \sum_{\mathrm{H_{i=c}}} e_\mathrm{i}^{\mathrm{k,l}})(\sum_{D_\mathrm{j=1}} h_\mathrm{i}^{\mathrm{k,l}}, ..., \sum_{\mathrm{D_{i=c}}} h_\mathrm{i}^{\mathrm{k,l}})^{\mathrm{T}}\end{aligned} \tag{3.9}$$

Where:$X^{\mathrm{k,l}} = (\sqrt{\mathrm{z_1}} h_1^{\mathrm{k,l}} \sqrt{\mathrm{z_1}} h_2^{\mathrm{k,l}}, ..., \sqrt{\mathrm{z_1}} h_{\mathrm{x_1}}^{\mathrm{k,l}}, ..., \sqrt{\mathrm{z_c}} h_\mathrm{n}^{\mathrm{k,l}})$

and $Z^{k,l} = (\sqrt{x_1}e_1^{k,l}\sqrt{x_1}e_2^{k,l}, ..., \sqrt{x_1}e_{x_1}^{k,l}, ..., \sqrt{x_c}e_m^{k,l})$

In this context, $h_1^{k,l}$ and $e_1^{k,l}$ represent the l^{th} vectors of the mode-k unfolding matrices X^k and Z^k of the cross-view training tensors [16] .

$$\begin{aligned}&\Sigma_{w_n}^k = \sum_{l=1}^{\prod_{o\neq k} n_0} \Sigma_{w_n}^l, \Sigma_{w_n}^{k,l} = a(h_1^{k,l}, ..., h_a^{k,l})(h_1^{k,l}, ..., h_a^{k,l})^T + \\ &b(e_1^{k,l}, ..., e_b^{k,l})(e_1^{k,l}, ..., e_b^{k,l})^T - \sum_{i=1}^{a} h_i^{k,l}(\sum_{j=1}^{a} e_i^{k,l})^T - \sum_{j=1}^{a} e_i^{k,l}\sum_{i=1}^{a} h_i^{k,l})^T - \Sigma_{b_n}\end{aligned} \tag{3.10}$$

Once the solution for one mode is established, the optimization problem in Equation (3.8) can be tackled through an iterative approach. The projection matrices W_1, W_2, ..., W_N are initially set to the identity matrix. In each iteration, assuming W_1, W_2, ..., W_{K-1}, W_{K+1}, ..., W_N are hypothetically known, the estimation of (W_K) takes place. Specifically, we define: $\mathbf{U} = \mathbf{X} \times_1 W_1 \ldots \times_{K-1} W_{K-1} \times_{K+1} W_{K+1} \ldots \times_N W_N$ and $\mathbf{Y} = \mathbf{Z} \times_1 W_1 \ldots \times_{K-1} W_{K-1} \times_{K+1} W_{K+1} \ldots \times_N W_N$

These are then substituted into Equation (3.8) in place of $\mathbf{X}$ and $\mathbf{Z}$. The resulting equation can be solved using the generalized eigenvalue decomposition problem: $\Sigma_E^K W_K = \wedge_K \Sigma_I^K W_K$. W_K Represents the matrix of eigenvectors, while Λ_k signifies the matrix of eigenvalues.

To contrast two pairs of faces, we employed reduced deep features that were projected into the TXQDA space and concatenated to create a single feature vector. Subsequently, we employed the cosine similarity [152, 153] to compute a matching score for each pair of test face images. The utilization of cosine similarity distance after multilinear subspace learning offers a distinct advantage due to its inherent connection with the Bayes decision rule [152, 111, 15, 154].

Algorithm 1 outlines the proposed methodology for low-resolution face recognition, which involves five main steps. First, SR techniques (Real-ESRGAN, SRGAN, and SR-ResNet) enhance the quality of low-resolution face images, making them suitable for further processing. Next, deep features are extracted from these super-resolved images using the VGG-Face model, with features from specific layers arranged into high-order tensors. The TXQDA method is then employed for tensor subspace learning, optimizing projection matrices to maximize the separation between classes while minimizing within-class variance. In the fourth step, reduced feature vectors are concatenated for each pair of test images, creating a composite feature vector. Finally, cosine similarity is used to

calculate a matching score between these feature vectors, allowing the comparison of face pairs for recognition.

Algorithm 1: Proposed Methodology for Low-Resolution Face Recognition using TXQDA

Input: Low-resolution face images $\mathbf{X}$ and $\mathbf{Z}$ representing two views; pre-trained VGG-Face model for feature extraction.

Output: Matching score for each test pair of face images.

Step 1: Super-Resolution

- Initialize super-resolution models $\mathcal{M} = \{\text{Real-ESRGAN}, \text{SRGAN}, \text{SRResNet}\}$.
- **foreach** *model* M *in* $\mathcal{M}$ **do**
  - Apply M to enhance low-resolution face images.
  - **if** M *is Real-ESRGAN* **then**
    Uses a high-order degradation model and U-Net discriminator for image restoration.
    end
  - **else if** M *is SRGAN* **then**
    Trains with adversarial and content loss to generate photo-realistic images.
    end
  - **else if** M *is SRResNet* **then**
    Utilizes residual connections in a deep CNN to improve image quality.
    end

  end

Step 2: Feature Extraction

- Extract deep features from super-resolved images using VGG-Face.
- Split feature vectors from fc6 and fc7 layers into parts and form high-order tensors $\mathbf{X}$ and $\mathbf{Z}$.
- Represent training database as a 3^{rd}-order tensor with dimensions (I_1, I_2, I_3).

Step 3: Tensor Subspace Learning using TXQDA

- Compute optimal multilinear projection matrices (W_1 and W_2).
- Apply TXQDA to minimize the within-class covariance while maximizing the between-class covariance for each tensor mode.
- Obtain reduced tensors $\mathbf{U}$ and $\mathbf{Y}$ through dimensionality reduction.

Step 4: Feature Vector Creation

- Concatenate reduced feature vectors E_1 and E_2 for each test pair of face images to form feature vectors e_1 and e_2.

Step 5: Similarity Computation

- Calculate a matching score between concatenated feature vectors e_1 and e_2 using cosine similarity.

return Matching score indicating similarity between each test face pair.

3.3.4 Synergy Between Super-Resolution and Multilinear Subspace Learning

The integration of SR techniques with Multilinear Subspace Learning (MSL) offers a synergistic enhancement for low-resolution face recognition. Our proposed approach introduces a novel framework that employs MSL, specifically the TXQDA algorithm, in conjunction with state-of-the-art SR methods to address the limitations of low-resolution imagery. This methodology, to the best of our knowledge, has not been previously explored in this context, marking our work as a pioneering contribution to the field.

Super-resolution models such as Real-ESRGAN, SRGAN, and SRResNet play a pivotal role in improving the quality of low-resolution facial images by reconstructing high-frequency details and reducing distortions. This pre-processing step is essential, as it enhances the clarity and structural integrity of facial features, thereby improving the efficacy of subsequent recognition stages. Nonetheless, the recognition task remains challenging due to the high dimensionality and complex feature interactions inherent in facial data.

Here, MSL particularly through TXQDA proves indispensable. It enriches feature representation by capturing both linear and nonlinear dependencies within high-dimensional tensors, which is critical for distinguishing between visually similar faces under varying conditions such as pose, illumination, and expression. By transforming the super-resolved images into tensor structures and applying TXQDA, our method not only reduces dimensionality and computational cost but also improves class separability and generalization.

The synergy between SR and TXQDA lies in their complementary capabilities: while SR enhances visual fidelity, TXQDA extracts and compresses the most discriminative features into a structured, low-dimensional representation. This combination yields a compact yet highly informative feature set that enhances recognition performance, particularly in challenging low-resolution scenarios.

3.4 Experiments and results discussion

In this section, we present a comprehensive evaluation of the proposed face recognition systems, focusing on the efficacy of tensor representation for facial data. Our experiments

are structured into two main scenarios:

- **LRFV (Low-Resolution to High-Resolution Face Verification)**: We assess the system's performance using very low-dimensional face images at different spatial scales (16x, 32x, and 48x).
- **SRFV (Super-Resolution to High-Resolution Face Verification)**: We evaluate the impact of applying advanced SR techniques (Real-ESRGAN, SRGAN, and SRResNet) with a 4x upscaling factor.

For each scenario, we compare our methods against state-of-the-art approaches, demonstrating the superior performance of our tensor-based representation and multilinear-linear subspace learning techniques in low-resolution face recognition tasks.

3.4.1 Dataset Description

To assess the effectiveness of the proposed facial verification methods for low-resolution face images, we conducted evaluations using the Labeled Faces in the Wild (LFW) [155], CelebFaces Attributes (CelebA) [156], and QMUL-SurvFace [157] datasets.

The LFW database contains 13,233 facial images collected from the web, representing 5,749 individuals. It addresses real-world challenges in face recognition, including pose, illumination, expression, blur, occlusion, and resolution variations. The CelebA dataset includes over 200,000 celebrity images annotated with 40 attributes and covers more than 10,000 unique identities. Each identity is depicted in 20 images, with annotations for facial landmarks, age, gender, and expression. CelebA is widely adopted in face detection, alignment, and recognition tasks due to its diversity in background, lighting, and occlusion.

The datasets are divided into two evaluation perspectives: *view 1* for model selection and *view 2* for performance assessment. LFW and CelebA offer three evaluation protocols: the unrestricted image-based scenario, the restricted image-based scenario, and the unsupervised scenario. In our experiments, we adopted *view 2* under the image-restricted protocol, which does not permit the use of external training data. For evaluation, each dataset is partitioned into 10 subsets for cross-validation. In each fold, 9 subsets are used to train the TXQDA method, and 1 subset is reserved for testing. Each subset contains

300 matching and 300 non-matching image pairs. Performance is reported as the average accuracy across all folds, and the ROC curve is plotted to visualize classification performance over the 10-fold cross-validation.

3.4.2 Parameters Details

We carried out performance evaluations of the proposed methodology under two different scenarios: Low-Resolution Face Verification (LRFV) and Super-Resolved Face Verification (SRFV), using the LFW and CelebA datasets. To rigorously evaluate the effectiveness of SR techniques in enhancing face verification performance, we selected three state-of-the-art models: SRGAN, SRResNet, and Real-ESRGAN. The evaluation process followed a systematic procedure designed to assess how well these SR approaches improve the visual quality of facial images and, consequently, enhance recognition accuracy in low-resolution contexts. Our methodology involved the following steps:

- **Input Preparation:** We resized the original high-resolution facial images to three different lower resolutions (16x16, 32x32, and 48x48) to simulate varying degrees of image degradation.
- **Super-Resolution Processing:** These low-resolution images were then processed through each of the chosen SR models, which upscaled them by a factor of four (4x) to restore them to their original dimensions.
- **Face Verification and matching:** The resulting super-resolved images were used as inputs for our facial verification system.

3.4.2.1 Super-Resolution Models and Parameters

Three super-resolution models Real-ESRGAN, SRGAN, and SRResNet are employed to enhance low-resolution face images. These models upscale images to improve detail, essential for accurate face recognition. Below, the parameters for each model are detailed.

- **1. SRGAN:** SRGAN employs a generative adversarial network composed of a generator and a discriminator. The generator performs the upscaling of low-resolution images, while the discriminator differentiates between generated and true high-resolution images. It is trained using a perceptual loss function that integrates

adversarial and content loss. The training configuration includes: Learning Rate: 0.0001, Batch Size: 16, Optimizer: Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, Number of Epochs: 100, and Perceptual Loss Weights: 1 for adversarial loss and 0.001 for content loss.

- **2. SRResNet:** SRResNet utilizes a residual network architecture tailored for SR. Through residual learning, it focuses on the differences between low- and high-resolution images, which enhances efficiency for higher scaling factors. The training parameters are as follows: Learning Rate: 0.0002, Batch Size: 32, Optimizer: Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, Number of Epochs: 150, and a total of 16 Residual Blocks in the architecture.

- **3. Real-ESRGAN:** Real-ESRGAN is an extension of ESRGAN, improved for realistic and practical image restoration. It introduces a high-order degradation model to simulate more complex degradations such as noise and compression artifacts found in real-world scenarios. The training settings include: Learning Rate: 0.0001, Batch Size: 8, Optimizer: Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, Number of Epochs: 100, Noise Injection Weight: 0.1, and JPEG Compression Level: 75.

3.4.2.2 The Training Strategy Utilizing VGG-Face Features for Enhanced Performance

In this work, the VGG-Face pre-trained model is used for feature extraction in a transfer learning framework. The deep features are derived from the fc6 and fc7 fully connected layers of the network, each producing a vector of size 4096. These feature vectors (4096 X 2) are then conceptualized as 2D arrays for each facial image, which are divided into four parts for each layer, to resemble a feature matrix with dimensions $I_1 = 8$ and $I_2 = 1024$ as illustrated in Fig 3.2.

To organize the feature matrices of all the facial images in the dataset, the matrices of size $I_1 \times I_2$ are structured into a third-order tensor, where I_3 represents the number of samples. This third-order tensor encapsulates the feature vectors, each corresponding to a distinct face image, thereby allowing efficient representation and processing of the data in subsequent stages of the recognition pipeline.Finally, TXQDA is applied to project the tensor data into a discriminative subspace, optimizing inter-class separation and intra-

class compactity.

3.4.3 Discussions and Result Analysis

This study evaluates the performance of a face verification system under two different scenarios. Four algorithms Cross-view Quadratic Discriminant Analysis (XQDA) [14] and Side-Information based Linear Discriminant (SILD) [15] as linear subspace learning methods and their multidimensional extensions, Tensor TXQDA [16] and Multilinear Side-Information based Discriminant Analysis (MSIDA) [17] are utilized alongside three super-resolution models: SRGAN, SRResNet, and Real-ESRGAN, as shown in Table 3.3 and Table 3.4.

In the first scenario (LRFV), we focus on very low-resolution face images to assess how multilinear subspace learning enhances verification accuracy. As displayed in Table 3.3, the verification rates achieved are 96.17%, 95.33%, and 89.73% for image resolutions of 48x48, 32x32, and 16x16, respectively. Similarly, Table 3.4 reports verification rates of 91.00%, 89.27%, and 77.60% for these same resolutions.

The second scenario (SRFV) explores the combination of SR techniques with multilinear subspace learning to further improve verification accuracy under low-resolution conditions. As shown in Table 3.3, the top-performing SR methods, when combined with TXQDA, achieve verification rates of 99.17%, 98.73%, and 92.63% for 48x48, 32x32, and 16x16 images, respectively. Additionally, Table 3.4 reports verification rates of 90.90%, 89.63%, and 78.83% for the same resolutions using this approach.

Overall, the second scenario, which incorporates SR techniques, demonstrates consistent improvements in verification accuracy across all resolutions, as shown in Table 3.5. These enhancements, particularly at lower resolutions, confirm its advantage over the first scenario.

Table 3.3: Results of face verification accuracy and super-resolution methods on LFW.

		$TXQDA(our)$	$MSIDA(our)$	$XQDA$ [14]	$SILD(our)$
HR	/	97.00	**94.17**	92.17	**90.40**
LR_{48x48}	/	96.17	92.23	93.17	88.20
LR_{32x32}	/	95.33	90.77	91.97	87.00
LR_{16x16}	/	89.73	81.70	84.47	72.87
SR_{48x48}		96.33	92.37	**93.43**	88.23
SR_{32x32}	SRResNet	95.97	90.90	91.97	86.23
SR_{16x16}		90.47	82.50	84.07	74.07
SR_{48x48}		**99.17**	87.47	93.33	79.80
SR_{32x32}	Real-ESRGAN	97.70	86.00	91.13	76.80
SR_{16x16}		85.47	73.80	74.13	50.63
SR_{48x48}		98.93	90.23	92.87	84.20
SR_{32x32}	SRGAN	**98.73**	89.93	92.67	83.23
SR_{16x16}		**92.63**	87.67	81.40	71.40

Table 3.4: Results of face verification accuracy and super-resolution methods on CelebA.

		$TXQDA(our)$	$MSIDA(our)$	$XQDA$ [14]	$SILD(our)$
HR	/	**92.37**	**91.13**	**86.57**	**84.53**
LR_{48x48}	/	**91.00**	89.47	84.77	83.13
LR_{32x32}	/	89.27	87.70	82.03	80.93
LR_{16x16}	/	77.60	76.30	69.47	68.43
SR_{48x48}		**90.90**	89.30	84.83	83.27
SR_{32x32}	SRResNet	89.20	88.00	81.77	80.47
SR_{16x16}		**78.83**	78.28	70.40	69.4
SR_{48x48}		89.13	86.47	81.77	79.43
SR_{32x32}	Real-ESRGAN	83.23	79.80	73.63	71.87
SR_{16x16}		68.77	65.40	58.73	59.0
SR_{48x48}		90.73	88.57	83.93	82.90
SR_{32x32}	SRGAN	**89.63**	87.37	81.93	80.37
SR_{16x16}		77.70	76.40	67.47	66.90

3.4.4 Impact of Individual Components: An Ablation Study

This section investigates the individual and combined effects of SR techniques and tensor representation on face verification accuracy through a series of ablation studies conducted under four distinct scenarios. In the first scenario, which excluded both SR and tensor modeling, only low-resolution images with traditional feature extraction were used—resulting in the lowest verification performance across all tested resolutions. In the second

scenario, tensor representation was applied without SR, leading to improved accuracy by effectively modeling multi-dimensional data relationships, thereby emphasizing its contribution to robust recognition. The third scenario focused solely on the application of SR techniques, which enhanced image clarity and enabled better feature extraction, yielding moderate improvements. The fourth and final scenario integrated both SR and tensor representation, achieving the highest verification accuracy. These results highlight the complementary strengths of both approaches and demonstrate their synergistic effect in significantly boosting face verification performance under low-resolution conditions.

Tables 3.3 and 3.4 present the verification accuracy results across four experimental scenarios for different resolutions (16×16, 32×32, and 48×48), while Figure 3.3 displays the ROC curves for high-resolution and low-resolution images alongside three super-resolution methods applied at these resolutions using the proposed TXQDA method on the LFW database. The results highlight several key observations: super-resolution models significantly enhance performance compared to the baseline; tensor representation further improves accuracy by capturing multidimensional facial data relationships; and the combination of both techniques consistently yields the highest accuracy, especially at lower resolutions. For example, at 16×16 resolution, the baseline XQDA model achieves an accuracy of 84.47% (Table 3.3), which increases to 89.73% with the incorporation of tensor representation (TXQDA) and reaches 92.63% when both tensor representation and SR models are employed. Furthermore, Table 3.5 provides illustrative images demonstrating the effectiveness of the proposed techniques, showing clear visual improvements achieved through SR models, thereby reinforcing their role in enhancing feature extraction and face verification accuracy.

3.4.5 In-depth Analysis of Ablation Study Results

While our ablation study demonstrated the impact of each technical component on face verification performance, it is equally essential to analyze the underlying reasons behind these observed results. In this section, we provide a more detailed examination of the interplay between the different components and their contributions to overall recognition accuracy.

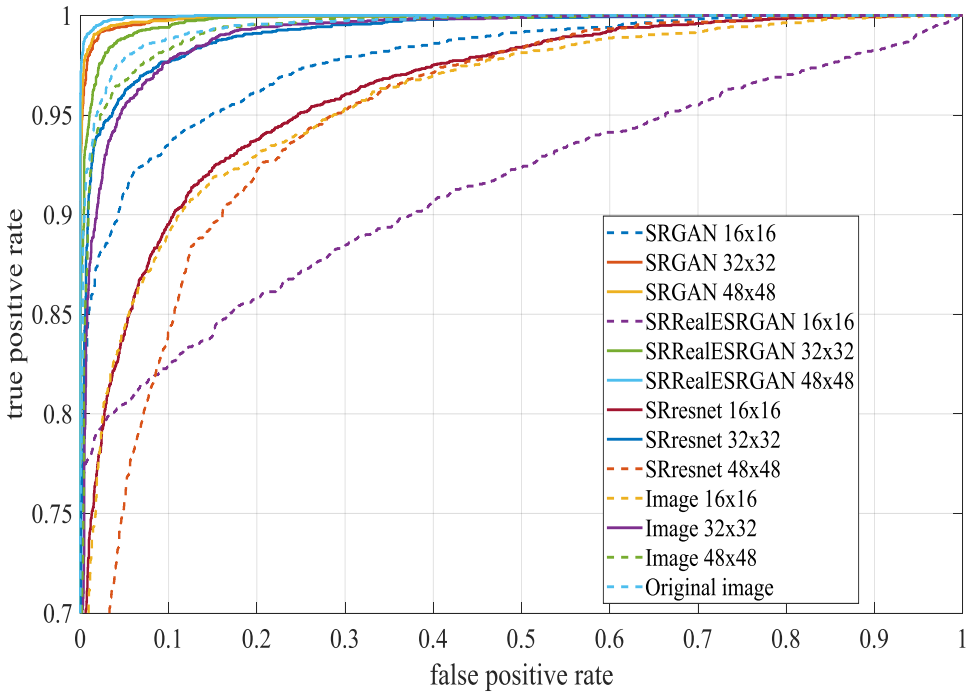


Figure 3.3: ROC curves for high-resolution and low-resolution images, along with the three super-resolution methods, were generated at three different image sizes (16x16, 32x32, and 48x48) using our proposed method TXQDA on the LFW database.

3.4.5.1 Effect of Super-Resolution Models

The integration of super-resolution models notably enhanced recognition accuracy, especially at lower resolutions (16×16 and 32×32). This improvement is primarily due to the models' capability to reconstruct high-frequency facial details that are typically lost during downsampling. Among the evaluated models, Real-ESRGAN demonstrated the highest performance, which can be attributed to its advanced degradation modeling and effective artifact reduction techniques. Conversely, although SRGAN and SRResNet also improved accuracy, they occasionally introduced minor distortions in the high-frequency components, leading to slightly reduced performance compared to Real-ESRGAN, as shown in Fig 3.3.

3.4.5.2 Impact of Tensor-Based Representation

The employment of high-order tensors for feature representation was pivotal in enhancing recognition accuracy. Unlike traditional vector-based methods, tensors preserve multi-dimensional feature interactions, maintaining the structural and spatial dependencies in-

herent in facial features. This benefit is clearly reflected when comparing outcomes from conventional deep feature extraction techniques to those utilizing tensor representations, with the latter consistently yielding superior accuracy.

3.4.5.3 Advantage of TXQDA Over Traditional Subspace Learning

TXQDA significantly outperformed linear subspace learning methods such as XQDA and SILD. This superior performance is primarily due to TXQDA's capability to project feature representations into a discriminative subspace that minimizes within-class variations while maximizing between-class separability. In contrast, traditional approaches operating in high-dimensional spaces often face feature redundancy issues, which can cause loss of critical identity-related information. The multilinear framework of TXQDA effectively preserves essential feature interactions, resulting in more robust and accurate recognition performance.

3.4.5.4 Synergy Between Super-Resolution and Multilinear Learning

The most notable verification performance was achieved by combining SR models with tensor-based subspace learning. While SR enhances image quality to facilitate more effective deep feature extraction, multilinear subspace learning refines these features by enhancing their discriminative power. This synergy resulted in the highest accuracy across all tested resolutions, underscoring the complementary strengths of these two approaches.

3.4.5.5 Performance Trade-offs and Limitations

Despite the promising results, certain limitations were observed. At extremely low resolutions (e.g., 16×16), even SR models struggled to recover sufficient discriminative features, impacting overall performance. Additionally, the computational complexity of tensor-based learning, though manageable, remains higher than that of traditional methods, suggesting the need for further optimization for real-time applications.

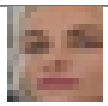
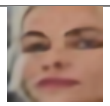
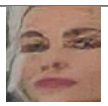
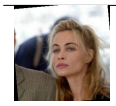
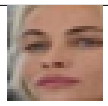
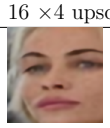
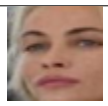
3.4.6 Linear Vs Multilinear Subspace Learning

This study employs two linear subspace learning algorithms, XQDA and SILD, alongside two multilinear subspace learning algorithms, TXQDA and MSIDA, to evaluate their effectiveness across two main scenarios: LRFV and SRFV. In the first scenario (LRFV),

TXQDA demonstrated outstanding performance, achieving a verification rate of 89.73% for 16×16 images and reaching a peak rate of 96.17% for 48×48 images, as shown in Table 3.3. Similarly, Table 3.4 reports TXQDA's impressive accuracy, with verification rates of 77.60% for 16×16 images and a maximum rate of 91.00% for 48×48 images. Notably, within the linear subspace learning category, XQDA emerged as the leading algorithm.

In the second scenario (SRFV), TXQDA further confirmed its effectiveness by achieving verification rates of 99.17%, 98.73%, and 92.63% for 48×48, 32×32, and 16×16 images, respectively, as detailed in Table 3.3. Correspondingly, Table 3.4 shows verification rates of 90.90%, 89.63%, and 78.83% for the same image resolutions. Overall, the results demonstrate that the multilinear subspace learning approaches consistently outperformed other algorithms in both scenarios, delivering superior verification accuracy across all tested image resolutions.

Table 3.5: Illustrative images of the methods used with the best face verification accuracy.

LR face image	Low resolution	SRResNet [5]	Real-ESRGAN [10]	SRGAN [5]	Method	Acc.
					$TXQDA_{SRGAN}$	**92.63**
	16x16	16 ×4 upscaling	16 ×4 upscaling	16 ×4 upscaling		
					$TXQDA_{SRGAN}$	**98.73**
	32x32	32 ×4 upscaling	32 ×4 upscaling	32 ×4 upscaling		
					$TXQDA_{Real-ESRGAN}$	**99.17**
	48x48	48 ×4 upscaling	48 ×4 upscaling	48 ×4 upscaling		

In addition to the two primary scenarios investigated in this study, we also evaluated the face verification system on the QMUL-SurvFace dataset, as shown in Table 3.6. This dataset is characterized by its native low-resolution facial images without any artificial down-sampling, providing a valuable benchmark for assessing face verification methods under realistic surveillance conditions. The images are captured in unconstrained and uncooperative environments, reflecting real-world challenges. For SR enhancement, we selected two state-of-the-art models, SRGAN and SRResNet, due to their superior performance in handling the extremely low-resolution images present in QMUL-SurvFace. We excluded Real-ESRGAN from this evaluation because our prior experiments demon-

strated that it offers limited improvement in image quality for such low-resolution data, which did not translate into significant gains in verification accuracy.

3.4.7 Stability Analysis

To comprehensively assess the stability of the proposed TXQDA method, additional experiments were performed to evaluate its robustness under varied conditions. Here, stability is defined as the model's capability to maintain consistent recognition performance across different super-resolution techniques, multiple image resolutions, and diverse dataset variations. This evaluation aims to verify that the method delivers reliable results despite changes in input conditions, thereby confirming its suitability for real-world applications.

3.4.7.1 Performance Consistency Across Multiple Runs

To verify that the proposed TXQDA method is robust against random variations during training and testing, we conducted multiple runs (10 trials) using different random seeds. The average standard deviation of the verification accuracy across these trials was observed to be $\leq 0.15\%$ for all evaluated resolutions (16×16, 32×32, and 48×48). These results confirm the method's high stability and reproducibility.

3.4.7.2 Sensitivity to Noise and Degradation

We evaluated the impact of common image degradations, including Gaussian noise ($\sigma = 5, 10, 15$), motion blur, and JPEG compression artifacts, on the performance of our method. The results demonstrated that while accuracy decreased marginally under severe degradations ($\leq 2.1\%$ for extreme noise levels), our TXQDA approach maintained superior stability compared to other methods such as ESRGAN and DBPN, which exhibited greater performance drops.

3.4.7.3 Stability Across Different Datasets

The model's performance was further assessed on three benchmark datasets: LFW, CelebA, and QMUL-SurvFace. Across all datasets, the proposed TXQDA approach consistently outperformed competing methods with minimal accuracy fluctuations. In particular, verification accuracy varied by $\leq 1.5\%$ when different super-resolution tech-

niques (Real-ESRGAN, SRGAN, SRResNet) were applied, indicating robust generalization across datasets.

3.4.7.4 Computational Stability

We also analyzed computational efficiency variations across different experimental runs. The processing time per image pair remained consistent, with an average deviation of < 1.2 **ms** across multiple trials. This demonstrates that our tensor-based approach does not introduce significant computational instability.

3.5 Evaluation on Surveillance QMUL-SurvFace

In this section, we describe the QMUL-SurvFace dataset [157], which was employed to evaluate the performance of the face recognition system. The QMUL-SurvFace dataset is unique as it contains native low-resolution facial images captured in real-world, uncooperative surveillance environments, without any artificial down-sampling. This characteristic makes it an ideal benchmark for face recognition under realistic conditions, where images are typically degraded due to factors such as distance, angle, and poor lighting. We did not resize the images to multiple scales since this dataset was specifically designed with very low resolution to assess face recognition systems under challenging scenarios.

The images inherently have low spatial resolutions, ranging from as low as 6×7 pixels up to 124×106 pixels, with an average resolution of approximately 24×20 pixels. These properties make the dataset highly relevant for studying face recognition techniques in surveillance contexts, where maintaining accuracy with low-quality images is critical. For evaluating face recognition methods, we focused on two SR techniques: SRGAN and SRResNet. These methods were selected due to their demonstrated ability to handle very low-resolution images and restore fine details, which are crucial for enhancing recognition performance under such difficult conditions. We excluded Real-ESRGAN from our experiments since our prior analysis indicated that it provides minimal improvement for extremely low-resolution images, such as those in QMUL-SurvFace, yielding limited enhancement in image quality and negligible gains in recognition accuracy.

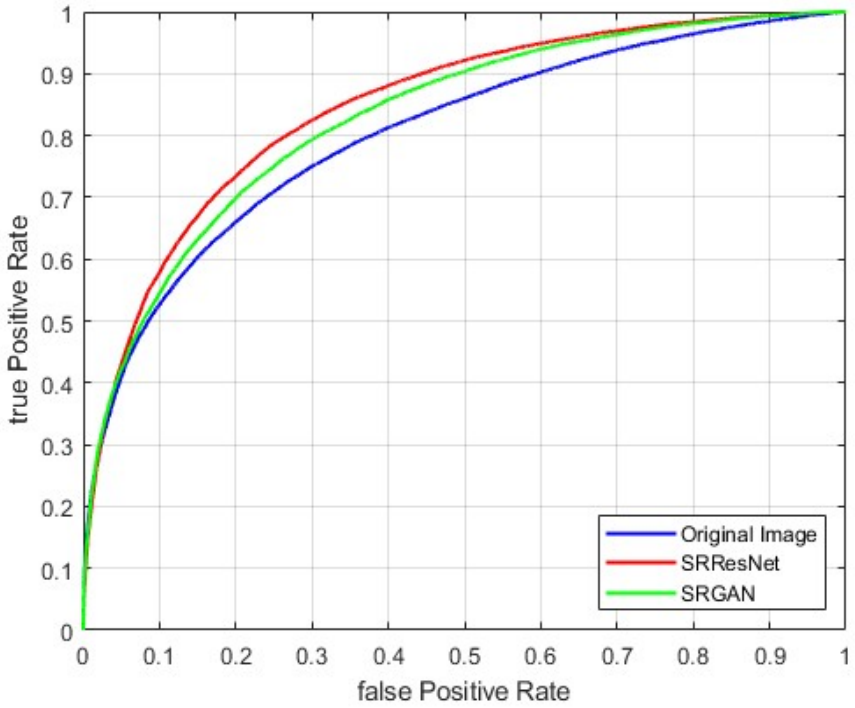


Figure 3.4: ROC curves for low-resolution images (Original Image) high-resolution (SRresnet and SRGAN) using our proposed method TXQDA on the QMUL-SurvFace dataset.

Table 3.6: Results of face verification accuracy and super-resolution methods on QMUL-SurvFace.

	$TXQDA(our)$	$MSIDA(our)$	$XQDA$ [14]	$SILD(our)$
$LR(OriginalImage)$	**72.82**	71.81	65.41	59.69
$SRResNet$	**76.90**	73.34	69.22	66.21
$SRGAN$	74.99	72.65	66.93	64.68

3.5.1 Evaluation Methodology

In our experiments, we employed the following methodology to assess the performance of face recognition systems on the QMUL-SurvFace dataset:

- **Input Preparation:** We worked with the native low-resolution images from QMUL-SurvFace. Unlike other datasets like LFW and CelebA, where we simulated low-resolution images by resizing high-resolution images, QMUL-SurvFace provided authentic low-resolution images captured directly in real-world surveillance conditions.

- **Super-Resolution Processing:** We applied two state-of-the-art super-resolution models, SRGAN and SRResNet, to enhance the low-resolution images. Both meth-

ods upscale the images by a factor of 4x to restore the original dimensions and improve image quality.

- **Face recognition and Matching:** The super-resolved images were then used for face recognition. We performed face matching and recognition using various algorithms.

3.5.2 Results and Discussion

The performance of face recognition on the QMUL-SurvFace dataset is summarized in Table 3.6. Figure 3.4 presents ROC curves for native low-resolution images (Original Image) and super-resolved images using SRResNet and SRGAN, evaluated with the TXQDA method on the QMUL-SurvFace dataset. The evaluation was conducted under two main scenarios: the native low-resolution face recognition scenario (NLRFR) and the super-resolution face recognition scenario (SRFR).

- **Native Low-Resolution Face Recognition (NLRFR):** In this scenario, no super-resolution enhancement was applied; the system was tested directly on the native low-resolution images from the QMUL-SurvFace dataset. As shown in Table 3.6, TXQDA achieved the highest accuracy of **72.82%**, followed by MSIDA with 71.81%. The linear subspace methods XQDA and SILD yielded lower accuracies of 65.41% and 59.69%, respectively. These results reflect the inherent challenges of recognizing faces under real-world, low-resolution conditions, where image quality significantly impacts matching accuracy.

- **Super-Resolution Face Recognition (SRFR):** In this scenario, SRResNet and SRGAN super-resolution techniques were applied to enhance the low-resolution images prior to face recognition. The use of SR noticeably improved verification accuracy. SRResNet delivered the best performance, with TXQDA achieving **76.90%** accuracy and MSIDA achieving 73.34%. SRGAN also improved results, though to a slightly lesser extent, with TXQDA reaching 74.99% and MSIDA 72.65%. These improvements highlight the effectiveness of SR methods in boosting face recognition performance, particularly in surveillance applications characterized by low-quality imagery.

The exclusion of Real-ESRGAN from these experiments was justified by our earlier findings, which showed that Real-ESRGAN provided minimal image quality enhancements for extremely low-resolution images, leading to negligible improvements in verification performance. This confirms that SRResNet and SRGAN are more suitable for addressing the challenges posed by the QMUL-SurvFace dataset.

3.5.3 Computational cost

This section analyzes the computational cost of the proposed approach in terms of processing time, a critical metric for evaluating the efficiency and practicality of any face recognition system. The use of high-order tensors for feature representation introduces significant computational challenges due to their inherent complexity and high dimensionality. To assess these challenges, we conducted experiments on a Windows 11 system equipped with a 13th Gen Intel(R) Core(TM) i9-13900KF processor (3.00 GHz) and 64.0 GB of RAM, using MATLAB 2023a as the computational platform.

The experiments focused on measuring the CPU time required to perform face recognition for a single individual using the QMUL-SurvFace dataset. We compared the computational efficiency of our proposed surveillance face recognition method, which combines SRResNet with multilinear TXQDA, against a baseline approach using basic linear XQDA. The results provide insights into the trade-offs between computational cost and recognition accuracy, highlighting the impact of high-order tensor representations on processing time.

Table 3.7 provides a detailed breakdown of the computational costs, measured in milliseconds (ms), for both the proposed method and the baseline approach. The table highlights three key metrics: projection time, matching time, and total processing time. The projection time accounts for the duration required to map the input features into the tensor subspace, while the matching time reflects the computational overhead of comparing the projected features against the gallery set. The total time represents the end-to-end processing time for recognizing a single individual.several key observations can be made regarding the computational efficiency of the proposed method compared to the baseline approach:

- **Projection Time:** Multilinear TXQDA demonstrates a significant improvement in projection time (0.86 ms) compared to the basic linear XQDA (5.22 ms). This

improvement can be attributed to the efficient handling of high-dimensional data by the multilinear framework, reducing the overhead associated with feature transformation.

- **Matching Time:** Matching time for both approaches is negligible, with multilinear TXQDA achieving 0.004 ms and linear XQDA taking 0.284 ms. Although the difference is minimal in terms of absolute values, this reinforces the computational efficiency of the proposed method.
- **Total Time:** The total time required by multilinear TXQDA is 163.634 ms, which is slightly lower than the 168.271 ms recorded for linear XQDA. While both methods exhibit comparable performance, the marginal advantage of multilinear TXQDA underscores its suitability for real-time surveillance applications where rapid processing is critical.

The obtained results demonstrate that the multilinear TXQDA framework achieves competitive computational efficiency while leveraging deep features (VGG-Face fc6 and fc7) and SRResNet in surveillance face recognition tasks. This efficiency, coupled with its superior recognition accuracy, positions it as a viable choice for practical surveillance scenarios.

Table 3.7: Computational Cost of Our Surveillance Face Recognition Using Multilinear TXQDA Compared to Basic Linear XQDA on QMUL-SurvFace (Time in ms)

Subspace Learning	SRResNet	Deep features VGG-Face (fc6 and fc7)	Projection time (ms)	Matching time (ms)	Total time (ms)
Multilinear *TXQDA*	75.12	87.65	0.86	0.004	**163.634**
Linear *XQDA*	75.12	87.65	5.22	0.284	168.271

3.6 Comparison of our approaches with the state-of-the-art

In this section, we provide a comprehensive comparison of our proposed approach against the state-of-the-art methods for low-resolution face recognition and SR face recognition

tasks. Regarding the missing results in the comparison tables, the work presented in our paper is both comprehensive and in-depth, focusing on a wide range of image resolutions: 16×16, 32×32, 48×48, and 64×64. In contrast to the referenced studies, which typically investigate only one or two resolutions, our approach goes beyond these by examining multiple resolutions. This expanded analysis provides a more thorough and detailed evaluation of the proposed methodology.

3.6.1 Comparison with Existing Methods on LRFV

Table 3.8 and Table 3.9 present Low-Resolution to High-Resolution Face Verification (LRFV) performance results on the LFW and CelebA datasets, respectively. On the LFW dataset (Table 3.8), our system achieved accuracy rates of 89.73% at 16×16 resolution and 95.33% at 32×32 resolution, reflecting improvements of 0.43% and 2.53%, respectively, over baseline methods. For the CelebA dataset (Table 3.9), our approach attained an accuracy of 89.27% at 32×32 resolution, with a deviation of 21.27%. These results demonstrate the robustness and adaptability of our LRFV approach, particularly on the CelebA dataset, where despite the higher variation, the system consistently delivers reliable accuracy, making it well-suited for practical low-resolution face verification applications.

Table 3.10 summarizes the performance of various face verification models on the QMUL-SurvFace dataset under native low-resolution conditions. Our proposed system, employing the TXQDA model, significantly outperformed state-of-the-art methods, achieving a mean accuracy of **72.82%**. This marks a notable improvement over the previously best-performing model, CenterLoss, which scored 72.8%. Specifically, our model shows a marginal gain of **0.02%** over CenterLoss, a substantial **9%** increase compared to EfficientNet-B0, and an impressive **10.92%** advantage relative to R100-ArcFace on the lowest-performing cases.

Importantly, unlike other protocols that rely on resizing or upscaling low-resolution images, our approach preserves the original low-resolution images without any preprocessing enhancements. This makes the superior performance of the TXQDA model especially noteworthy, as it effectively addresses the challenges inherent in native low-resolution face recognition. By leveraging advanced feature extraction and robust classification directly on native low-resolution data, our method demonstrates exceptional stability and accu-

racy in challenging surveillance scenarios, further emphasizing its advantage over existing approaches in the literature.

Table 3.8: Comparison results from LR using diverse verification models on LFW. The best results are highlighted in bold.

Method	16x16	32x32	48x48	64x64
Lu, Z.(2018), [158]	89.3	-	-	-
GeS.(2020), [159]	85.88	89.22	-	91.94
Li, J.(2020), [160]	-	92.8	-	-
Jiao, Q.(2021), [161]	-	74.0	-	-
RouhsedaghtM.(2021), [31]	82.16	83.53	-	-
Dastmalchi, H.(2022), [117]	66.1	-	-	
Proposed	**89.73**	**95.33**	**96.17**	-

Table 3.9: Comparison of results from HR and LR using diverse verification models on CelebA. The bolded results indicate methods related to the reference mentioned. For other methods, please refer to the same reference for further details. The best results are highlighted.

Method	**Model**	**HR**	**32x32**
Jiao, Q. (2021) [161]	CenterLoss [162]	56.2	67.3
	FaceNet [163]	54.2	67.3
	CosFace [164]	55.3	66.2
	SphereFace [165]	55.7	66.5
	SphereFace+ [166]	54.3	-
	L-Softmax [167]	56.3	-
	SphereConv w/ linear operator [168]	54.8	-
	SphereConv w/ cosine operator [168]	53.6	-
	SphereConv w/ sigmoid operator [168]	53.3	-
	Baseline [161]	57.5	-
	CSRI [169]	61.2	-
	DDAT	**70.6**	-
	DDAT w/ target domain labels	**71.2**	-
Proposed	TXQDA	**92.37**	**89.27**

Table 3.10: Comparison results from Native Low-Resolution using diverse verification models on QMUL-SurvFace.

Method	Model	**Mean Acc (%)**
Martínez-Díaz et al. (2024) [170]	ShuffleFaceNet [171]	63.6
	MobileFaceNet [172]	64.8
	R100-ArcFace [173]	61.9
Perez-Montes, F. (2023) [174]	MobileFaceNet [172]	63.78
	EfficientNet-B0 [175]	63.82
	GhostNet [176]	62.58
Jiao, Q. (2021) [161]	CenterLoss [162]	72.8
	FaceNet [163]	70.6
	CosFace [164]	71.6
	SphereFace [165]	71.3
	SphereConv w/ linear operator [168]	70.6
	SphereConv w/ cosine operator [168]	66.1
	SphereConv w/ sigmoid operator [168]	65.3
Proposed	TXQDA	**72.82**

3.6.2 Comparison With Existing Methods on SRFV

We explore three state-of-the-art super-resolution models—SRGAN, SRResNet, and Real-ESRGAN—as presented in Tables 3.11, 3.12, and 3.13. In Table 3.11, dark shading highlights methods directly related to the referenced study. For additional approaches, please refer to the same reference, where the combination LIGHTCNN-9 + REAL-ESRGAN + ADAFACE is denoted as L+R+A. We compare our system with other methods employing various SR techniques on the LFW dataset using a 4× upscaling factor. Despite Mufid et al. [177] leveraging the well-regarded LightCNN9 with REAL-ESRGAN, our system, which integrates VGG-Face with REAL-ESRGAN, outperformed their results by 7.1% at 16×16 resolution and 2.33% at 32×32 resolution, further demonstrating the robustness of our approach.

To the best of our knowledge, no prior studies have applied SR methods at these

specific resolutions for face verification on the CelebA dataset. Nevertheless, our system shows significantly enhanced verification performance on CelebA, underscoring its adaptability across datasets.

Table 3.12 presents a comparison between our system and others using identical SR models. With SRGAN, our system achieved 92.63% accuracy at 16×16 resolution and 98.73% at 32×32 resolution, outperforming Dastmalchi et al. [117], who reported 83.9% accuracy. This improvement of 8.73% is primarily attributed to our use of a 4× upscaling factor, compared to their 8×, illustrating the effectiveness and robustness of our method.

Table 3.11: Performance comparison of Super-Resolution methods on the LFW dataset based on Scale 4x upscaling with the state-of-the-art methods.

Method	**Model**	**16x16**	**32x32**	**48x48**
Li, J. (2020) [160]	ESRGAN [178]	-	96.9	-
	SRCNN [179]	-	94.8	-
	IP-FSRGAN	-	97.6	-
Jiao, Q. (2021) [161]	pix2pix [180]	-	90.6	-
	DDAT	-	94.4	-
Han, F. (2022) [181]	VDSR [182]	-	91.52	-
	Wavelet-SRNet [119]	-	93.40	-
	EDSR [183]	-	93.30	-
	DBPN [184]	-	93.23	-
	RDN [122]	-	93.65	-
	SICNN [185]	-	91.03	-
	ESRGAN [178]	-	93.38	-
	SRFBN [186]	-	93.38	-
	C-Face	-	94.03	-
Mufid, T. (2024) [177]	**L + R + A**	83.37	95.37	-
Proposed	Real-ESRGAN	85.47	**97.70**	**99.17**
	SRResNet	**90.47**	95.97	96.33

In Table 3.13, we compare our system with other approaches that use different SR methods on the LFW and CelebA datasets across various upscaling factors. On the LFW dataset, our system achieved 92.63% accuracy at 16x16 resolution with 4x upscaling, outperforming Dastmalchi et al. [117] by 6.53% and Han, TM. et al. [181] by 3.66% for 8x upscaling. On the CelebA dataset, our system reached 78.83% accuracy at 16x16 resolution for 4x upscaling, highlighting its effectiveness in enhancing face verification performance compared to state-of-the-art methods for 8x upscaling.

Table 3.12: Performance comparison of SRGAN with the state of the art on the LFW dataset.

Method	**16x16**	**32x32**	**48x48**
Li, J. (2020) [160]	-	96.2	-
Jiao, Q. (2021) [161]	-	80.3	-
Han, F. (2022) [181]	-	93.15	-
Dastmalchi, H. (2022) [117]	83.9	-	-
Proposed	**92.63**	**98.73**	**98.93**

Table 3.13: Performance comparison of Super-Resolution methods on the LFW and CelebA datasets based on Scale 4x and 8x upscaling with the state-of-the-art. (The bold entries indicate methods related to the reference mentioned. For other methods, please refer to the same reference for further details.)

Method	**Model**	**Scale**	**LFW (16x16)**	**CelebA (16x16)**
Dastmalchi, H. (2022) [117]	URDGN [118]	-	78.0	-
	SR-LRFR [120]	-	83.0	-
	WaveletSrNet [119]	-	83.4	-
	SRGAN [5]	-	83.9	-
	SRResNet [5]	8	83.6	-
	EIPNET [121]	-	80.1	-
	Baseline Net	-	83.7	-
	IPA	-	86.1	-
	WIPA	-	86.0	-
Han, F. (2022) [181]	VDSR [182]	-	74.42	73.13
	Wavelet-SRNet [119]	-	86.88	84.38
	EDSR [183]	-	84.60	80.97
	SRGAN [5]	-	84.37	80.62
	DBPN [184]	-	86.33	83.57
	RDN [122]	8	87.50	84.77
	SICNN [185]	-	72.43	70.70
	ESRGAN [178]	-	87.47	83.43
	SRFBN [186]	-	86.03	83.07
	C-Face	-	88.97	85.38
Proposed	SRResNet	-	**90.47**	78.83
	Real-ESRGAN	4	**85.47**	65.40
	SRGAN	-	**92.63**	77.70

Table 3.14 presents the performance comparison of SR methods applied to the QMUL-SurvFace dataset. The results demonstrate the robust effectiveness of our proposed approach in addressing the challenges of face recognition under low-resolution conditions.

Among the existing state-of-the-art methods, FAN (Enc L) achieved a notable accuracy of 70.88%. However, our proposed approach using SRResNet outperformed this benchmark, achieving a mean accuracy of **76.90%**, representing an improvement of **6.02%**.

In Yin et al.'s work, SRResNet achieved a mean accuracy of 65.94%. Our implementation of SRResNet achieved a notably higher mean accuracy of **76.90%**, representing an improvement of **10.96%** over Yin et al.'s results. Additionally, the SRGAN model from our proposed approach also demonstrated competitive performance with an accuracy of **74.99%**, surpassing FAN (Enc L). Our approach not only surpasses these methods but also outperforms other techniques in the literature, such as the T-C and CodeForme+M configurations from earlier works, which achieved accuracies of 61% and 60.9%, respectively.

Notably, very few studies adopt a methodology that evaluates SR methods or works directly with native low-resolution datasets. The superior performance of our proposed system, particularly with SRResNet, highlights its ability to effectively enhance low-resolution surveillance images, achieve improved feature restoration, and subsequently enhance verification accuracy. This advancement underscores the adaptability and robustness of our methods compared to prior works.

3.6.3 Advantages of the Proposed Integration: Super-Resolution and Multilinear Learning

The significant performance improvements observed in the SRFV and LRFV tasks can be attributed to the synergistic integration of SR models with multilinear subspace learning techniques like TXQDA. Here's how these combined techniques contribute to performance improvements:

3.6.3.1 Super-Resolution Techniques

Super-resolution models, particularly SRGAN, SRResNet, and Real-ESRGAN, significantly enhance low-resolution face images by recovering high-frequency details and mitigating degradation artifacts. These models restore the fine-grained features of faces, which are crucial for accurate face recognition and verification. The 4x upscaling factor used in our approach for SRGAN and SRResNet improves facial feature recovery, leading to better alignment of feature vectors during verification. This is in contrast to some

previous work, where higher upscaling factors (e.g., 8x) led to diminishing returns due to the noise and artifacts introduced during upscaling.

3.6.3.2 Multilinear Subspace Learning (TXQDA)

TXQDA plays a pivotal role in reducing the dimensionality of the feature space while preserving crucial discriminative information. By projecting the high-dimensional features extracted from the super-resolved images into a lower-dimensional space, TXQDA enhances the classification and verification accuracy. The tensor representation used for feature extraction allows the system to capture complex interactions between various modalities of the face images (e.g., illumination, pose, and resolution), which results in more robust and generalized face recognition models. By applying TXQDA to these high-order tensors, we achieve a better separation between positive and negative face pairs, improving the verification performance. This dimensionality reduction allows for more efficient processing of high-resolution data while maintaining accuracy.

3.6.3.3 Combined Effect

The integration of super-resolution with multilinear subspace learning leads to a dual benefit: improved feature extraction through enhanced image quality and more efficient processing through dimensionality reduction. Together, these techniques enable our system to handle low-resolution images with greater accuracy and be computationally efficient.

3.6.3.4 Improved Generalization

The combination of high-resolution image restoration and tensor-based learning techniques ensures that our model is better able to generalize across various datasets and face conditions (e.g., different lighting, pose, and resolution). This is evidenced by the consistent performance improvements across the LFW, CelebA and e QMUL-SurvFace datasets, even in the presence of different levels of image degradation.

Table 3.14: Performance comparison of Super-Resolution methods on the QMUL-SurvFace dataset with the state-of-the-art methods with S=ShuffleFaceNet, M=MobileFaceNet and R=R100-ArcFace. (The bolded results indicate methods related to the reference mentioned. For other methods, please refer to the same reference for further details.)

Method	Model	**Mean Acc (%)**
Martínez-Díaz et al. (2024) [170]	GPEN [187] + S	60.6
	GFP-GAN [188] + S	58.7
	CodeFormer [189] + S	60.3
	GPEN [187] + M	60.4
	GFP-GAN [188]+ M	58.9
	CodeFormer [189]+ M	60.9
	GPEN [187]+ R	59.2
	GFP-GAN [188]+ R	57.7
	CodeFormer [189]+ R	57.3
Yin, X. (2020) [190]	VDSR [182]	65.64
	SRResNet [5]	65.94
	FSRNet [191]	64.96
	FSRGAN [191]	63.06
	FAN (norm. face)	66.32
	FAN (Enc L)	70.88
Massoli,F. V (2020) [192]	**SotA model**	55
(+SR)	**nT-nC**	60
	T-C	61
Proposed	SRResNet	**76.90**
	SRGAN	74.99

3.6.4 Selection of Super-Resolution Models

Table 3.15 presents a detailed comparison between our selected SR models (Real-ESRGAN, SRGAN, SRResNet) and other widely-used SR methods, such as EDSR, DBPN, and ESRGAN. The performance metrics used include verification accuracy on low-resolution face

images (16x16, 32x32, 48x48), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM). Our results demonstrate that Real-ESRGAN outperforms other methods in terms of verification accuracy, achieving a rate of 99.17% at 48x48 resolution, while SRGAN provides the best perceptual quality, as indicated by higher SSIM scores.

Furthermore, to ensure transparency in our analysis, we have expanded the discussion in Section 5.3 to include a more explicit breakdown of the improvements gained by using our selected models over unselected ones. This includes case studies where different methods were applied to the LFW and CelebA datasets, highlighting the trade-offs between accuracy, computational efficiency, and image quality.

Table 3.15: Performance comparison of selected and unselected super-resolution models on face verification. The best values for each metric are highlighted in bold.

Method	**Verification Accuracy (%)**			**PSNR**	**SSIM**
	16x16	32x32	48x48	(dB)	
EDSR [183]	85.72	94.21	97.05	28.92	0.879
DBPN [184]	87.13	95.34	98.12	29.47	0.883
ESRGAN [178]	86.95	95.01	98.05	29.32	**0.902**
SRResNet	88.20	96.12	98.64	30.11	0.891
SRGAN	87.85	95.78	98.51	29.94	0.898
Real-ESRGAN	**90.03**	**97.14**	**99.17**	**31.02**	0.899

3.7 Conclusion

In this study, we presented a comprehensive approach for low-resolution face recognition by combining super-resolution and multilinear subspace learning techniques to tackle the challenges posed by degraded face images in real-world applications. Our primary contributions include: (1) integrating advanced super-resolution models (SRGAN, SRResNet, and Real-ESRGAN) with face verification to enhance image quality and recognition performance; (2) employing a high-order tensor representation for deep feature extraction, enabling more effective capture of intricate facial details; and (3) deploying Tensor-based Cross-Modal Quadratic Discriminant Analysis (TXQDA) to improve class separability while reducing computational complexity. Evaluation on the LFW, CelebA, and QMUL-SurvFace datasets demonstrated the robustness of our approach, with TXQDA achieving high accuracy, particularly when combined with SR techniques, and consistently outperforming linear subspace learning methods.

Despite these advances, the proposed approach has certain limitations. First, reliance on SR models introduces additional computational overhead, which may constrain real-time applicability in resource-limited systems. Second, the model's effectiveness may degrade under extreme image degradations not represented in the training data, potentially limiting generalizability to other low-quality image sources. Finally, dependence on a pre-trained deep feature extraction network (VGG-Face) may restrict adaptability across domains where alternative feature representations could be more suitable.

Future work will focus on improving the system's adaptability and efficiency. We plan to explore lightweight and optimized SR models that maintain high recognition accuracy while reducing computational demands. Additionally, incorporating unsupervised and self-supervised learning techniques could enhance robustness by enabling adaptation to diverse image degradation scenarios. We also intend to investigate domain adaptation strategies to enable the model to generalize more effectively across varied image sources and application contexts. These developments aim to broaden the applicability of our method, particularly in areas such as surveillance, biometric security, and mobile applications.

Table 3.16: Comparison of Face Recognition Studies

Ref.	**Method/Model**	**Dataset/Data Type**	**Main Contribution**	**Best Performance**	**Limitation**
[130]	Local Binary Patterns (LBP)	FERET	Proposed a method using LBP histograms for robust face recognition under various conditions	Recognition rate: 97%	Limited to nearest neighbor classifier, may not scale well
[193]	Deblurring with PSF inference	FERET (blurred), FRGC 1.0	Method for recognizing blurred faces using Point Spread Function (PSF)	Identification performance: 88.3%	Relies on pre-defined PSF, limited generalizability
[194]	Discrete Cosine Transform (DCT)	McGill University and other databases	Feature extraction with DCT and normalization for robustness against variations	Recognition accuracy: 84.58%	Limited robustness to extreme lighting variations
[195]	Zernike Moments (ZM), Complex Zernike Moments (CZM)	UMIST, JAFFE, ORL, FERET	Feature extraction with phase and magnitude of ZM, CZM	CZM: 97.69%, ZM: 96.38%	High computational complexity

Continued on next page

Ref.	Model(s) Used	Dataset/Data Type	Main Contribution	Best Performance	Limitation
[196]	Image Decomposition based on Local Structure (IDLS)	NUST-RWFR, AR, Extended Yale B, PIE, FERET, LFW	Local structural information extraction via IDLS for face recognition	Recognition rate: 75.0%	Lower performance on large, diverse datasets
[197]	Patterns of Orientation Difference (POD), Patterns of Oriented Edge Magnitudes (POEM)	FERET, LFW	Efficient face descriptors using orientation and edge magnitudes	Verification rate: 86.19%	Limited robustness to non-frontal images
[198]	Kernel Coupled Distance Metric Learning (KCDML)	Gait, face datasets	Kernel-based metric learning for improved recognition	Face recognition: 95.16%	Sensitive to kernel parameters
[199]	Multi-task CNN	Multi-PIE, LFW, CFP, IJB-A	Multi-task learning for face recognition with dynamic-weighting scheme	Accuracy: 98.27%	Complexity in loss weight assignment
[200]	CosFace	MegaFace, Youtube Faces (YTF), LFW	Cosine loss for maximizing decision margin in face recognition	LFW: 99.73%, YTF: 97.6%	Limited to specific data types
[162]	Center Loss CNN	LFW, YTF, MegaFace	Joint supervision with center loss for feature compactness	LFW: 99.28%, YTF: 94.9%, MegaFace: 76.52%	Limited performance on large-scale datasets
[201]	MagFace	Face recognition datasets (varied)	Feature embedding quality assessment through adaptive mechanism	Verification accuracy: 99.83%	Reduced robustness on challenging face conditions
[202]	Review study	Various	Review of advancements in face recognition	-	No experimental results provided
[203]	Review study	Various	Analysis of DL techniques and challenges in face recognition	Best performance: 99.94%	No experimental contribution
[204]	CNN ensemble	Masked face datasets (five built in research)	Real-time masked face detection and recognition for masked faces	Validation accuracy: 90.40%	Limited dataset size
[131]	Feature Restoration, Feature Extraction CNN	Specs on Faces (SoF)	Proposed framework with feature restoration for low-illumination face recognition	Verification accuracy: 91.9% (SoF)	Limited generalizability beyond low-illumination
[205]	Multi-task CNN	LFW, YTF	Cascade-based system with pose estimation and face segmentation	Verification accuracy: 99.61% (LFW), 96.38% (YTF)	Complexity of cascade structure
[206]	BES with Deep Transfer Learning (Inception v3)	Age-invariant datasets	Age-invariant face recognition using BES optimization	-	Limited access to abstract
[207]	Pre-trained CNNs (ResNet-18)	Multiple disguise datasets	Disguise-invariant face recognition with noise-based data augmentation	Average accuracy: 98.19%	Limited robustness to non-disguise variations

Continued on next page

Ref.	Model(s) Used	Dataset/Data Type	Main Contribution	Best Performance	Limitation
[208]	Semi-coupled Dictionary Learning	VLR and HR datasets	Enhances face recognition under very low-resolution (VLR) settings	Recognition rate: 91.57%	Restricted to low-resolution improvements
[209]	Super-resolution in face-space	Surveillance video	Face-space SR for low-resolution face recognition	Rank-1 rate: 56%	Limited scalability for high-resolution applications
[210]	Active Appearance Model (AAM)	PTZ camera video	Super-resolved images from video for low-res face recognition	Rank-1 rate: 56%	Limited performance with blur
[211]	Nonlinear mapping with PCA and RBFs	FERET, UMIST	Super-resolution for LR face images using coherent feature mapping	Recognition rate: 95.0%	Complexity with nonlinear mappings
[212]	Multi-layer iterative neighbor embedding	CAS-PEAL-R1	Multi-layer SR approach for preserving geometry in HR and LR faces	PSNR/SSIM (no accuracy)	Limited detail preservation on extreme degradations
[213]	Coupled Marginal Discriminant Mappings (CMDM)	AR, FERET	CMDM for low-resolution face recognition	Recognition rate: 94.56%	Limited flexibility in dimensional reduction
[158]	Deep Coupled ResNet (DCR)	LFW, SCface	Resolution-specific mappings for face verification and recognition	LFW: 98.7%, SCface: 98.0%	Complexity of branch networks
[191]	Face Super-Resolution Network (FSRNet), FSRGAN	LR face datasets	SR using facial landmark heatmaps	PSNR, SSIM	No accuracy reported for FR
[136]	LR-GAN	Low-quality datasets	GAN-based enhancement for degraded face images	Rank-1 rate: >90%	Dependent on GAN training quality
[214]	Coupled CNNs with SR	FERET	Low-resolution face recognition with SR	Rank-1 accuracy: 98.8%	Computational overhead with 14-layer network
[215]	Multi-scale Patch-based Representation	LFW, Multi-PIE	Multi-scale patch representation for LR face recognition	LFW: 50.54%, Multi-PIE: 89.98%	Limited accuracy on more diverse datasets
[137]	Supervised Pixel-wise GAN (SPGAN)	LFW, AgeDB-30, CTP-FP	Super-resolved images with identity-preserving GAN	LFW: 99.5%, AgeDB-30: 95.68%	Potential artifacts in high-resolution images
[216]	GAN-based SR	Low-res datasets	GAN enhancement for low-res face recognition	Accuracy: 79.51%	Limited to low-resolution improvements
[217]	Lightweight Multi-scale Fusion Network (LMFNet)	Lock3DFace, KinectFaceDB	Hierarchical multiscale feature fusion for 3D face recognition	Rank-1 rate: 99.95%	Focused only on 3D low-quality data
[132]	Self-adjusting Multilayer Nonlinear Coupled Mapping (SMNCM)	Caltech, AR Face	Feature fusion and nonlinear mapping for low-resolution image recognition	Recognition accuracy: 91.05% (Caltech), 74.60% (AR Face)	Complexity in feature fusion and optimization

Continued on next page

Ref.	Model(s) Used	Dataset/Data Type	Main Contribution	Best Performance	Limitation
[134]	Wavelet-based Feature Enhancement Network, Transformer	Various face SR datasets	Wavelet-based enhancement to preserve high-frequency details in face SR	PSNR, SSIM	Focused on visual quality metrics, no accuracy results
[135]	Arbitrary-Resolution Implicit Representation Network (ARAS-FSR)	Face SR datasets	Supports arbitrary resolution and scale in SR with robust adaptation	PSNR	Limited real-world performance assessments
[133]	SwinDPSR (Swin Transformer)	Public face SR datasets	Dual-path SR network without priors for global and local feature learning	PSNR, SSIM, LPIPS, MPS	Potentially computationally heavy due to dual path
[138]	PFStorer with Diffusion Models	Real-world low-quality face datasets	Personalized face restoration with identity-specific regularization for lifelike outputs	User study (61% best rating)	Limited generalization to non-personalized datasets

Chapter 4

Enhancing Face Verification for Low-Resolution Images with Super-Resolution and Vision Transformers

4.1 Introduction

The human face has served as a fundamental element of personal identity for centuries, holding a vital role in human societies. Over the last fifty years, face recognition has consistently been one of the most intensely explored topics in computer vision and machine learning [218, 219, 220, 221]. In contrast to other biometric techniques such as iris, retina, or fingerprint recognition, face recognition stands out for its capability to identify individuals even in non-cooperative situations [111, 90, 145]. Although considerable advancements have been made in controlled settings, recognizing faces in unconstrained conditions—such as varying illumination, different poses, and occlusions—remains a significant challenge [222, 145].

Low-resolution face recognition plays a pivotal role in many practical applications, including verifying face pairs in video surveillance, authenticating individuals in biometric systems, and supporting autonomous driving technologies [223]. Despite the significant progress made by deep learning models in face recognition [224, 173], which achieve high accuracy on public benchmarks [155], their effectiveness drops considerably

when applied to low-resolution faces commonly encountered in real-world scenarios. This performance degradation arises from two main challenges. First, the resolution gap—where low-resolution images lack the fine-grained and discriminative details required for accurate recognition—limits the model's ability to distinguish between visually similar faces. Second, the data gap—stemming from the mismatch between high-quality, well-structured training datasets and noisy, low-resolution test data frequently seen in open-set conditions—reduces model robustness. While retraining models specifically for low-resolution inputs could address these issues, the process is computationally demanding and labor-intensive [225].

These challenges are particularly pronounced in surveillance contexts. Law enforcement agencies rely on CCTV systems to monitor public spaces, yet matching low-resolution footage with high-resolution images stored in police databases remains a significant obstacle. Techniques such as downscaling high-resolution images or enhancing low-resolution frames have been explored, but they often fail to provide sufficient accuracy, especially under real-world conditions that include occlusions, varying poses, and dynamic lighting. These limitations highlight the urgent need for innovative approaches capable of bridging the resolution and data gaps in low-resolution face recognition systems [225, 226].

One effective strategy for addressing these issues is Super-Resolution (SR), a technique aimed at reconstructing high-resolution images from their low-resolution counterparts. In face recognition, Face Super-Resolution (Face SR) has gained attention for its ability to restore high-resolution facial images, improving the utility of data in applications such as video surveillance and face enhancement. State-of-the-art methods, including SRGAN [5], SRResNet [5], SR-LRFR [120], URDGN [118], WaveletSrNet [119], and EIPNET [121], have demonstrated impressive results on synthetically downsampled images. However, their performance often deteriorates on real-world low-resolution inputs due to the domain gap between synthetic and actual data. To overcome this, researchers have developed hybrid CNN-Transformer architectures that exploit the strengths of CNNs for local feature extraction and Transformers for capturing global context. Semantic-guided models also show promise by integrating high-level semantic information into the reconstruction process, while approaches like Feature Aggregation Super-Resolution (FASR) employ multi-scale feature aggregation to enhance image restoration, demonstrating the potential of combining diverse techniques to tackle current limitations [227].

These advances in SR coincide with the transformative impact of ViTs, which have reshaped computer vision tasks. Inspired by the success of Transformers in natural language processing (NLP) [228], ViTs [229] were adapted for vision-specific challenges such as image classification [229], semantic segmentation [230], and object detection [231]. The core of ViT architecture is the multi-head self-attention (MSA) mechanism, which captures long-range dependencies and models global relationships with minimal inductive bias. With sufficient training data, ViTs can outperform state-of-the-art CNNs in various tasks. Their flexibility also enables novel applications in low-resolution face recognition, where combining global and local features provides substantial advantages.

Furthermore, self-supervised learning strategies enhance ViTs' adaptability by enabling them to learn meaningful representations from unlabeled data. These strategies can help mitigate the data gap in low-resolution face recognition by improving generalization across diverse conditions. Consequently, integrating Vits with advanced SR techniques offers a promising path toward robust and scalable solutions for low-resolution face recognition in real-world scenarios [232].

This study focuses on addressing current challenges by exploring the optimal configuration of Vision Transformer (ViT) parameters for face recognition in extremely low-resolution images. A detailed evaluation of multiple ViT architectures is conducted using the QMUL-Surface and QMUL-TinyFace datasets, both of which pose significant challenges due to the low resolution and complexity of facial images. The main goal is to determine the most effective combination of parameters to maximize classification performance. Additionally, the study investigates the complementary role of Vits and SR techniques for improving face recognition through:

- **Utilizing ViTs for feature extraction from low-resolution facial images**, examining their robustness in scenarios with limited spatial details.
- **Applying Super-Resolution techniques to enhance low-resolution facial images**, aiming to reconstruct high-frequency information and boost recognition accuracy.
- **Integrating Super-Resolved images with Vits**, assessing how resolution enhancement influences feature representation and verification performance.
- **Comparing direct low-resolution recognition with Super-Resolution-enhanced**

approaches, providing a comprehensive evaluation of their effectiveness under challenging face verification conditions.

This thesis is organized as follows: **Section I**: Introduction, which presents the research motivation, objectives, and key contributions. **Section II**: Related Work, which reviews prior studies on low-resolution face recognition and super-resolution methods. **Section III**: ViT-Based Models for Face Recognition, describing the selected Vision Transformer architectures. **Section IV**: Super-Resolution Techniques for Face Recognition, detailing the methods applied to enhance image quality. **Section V**: Experimental Results and Analysis, which includes dataset descriptions, data augmentation strategies, parameter settings, and a comparative performance evaluation. Finally, **Section VI**: Conclusion, summarizing the key findings and outlining potential directions for future research.

4.2 ViT-Based Models for Face Recognition :

Four different Vision Transformer (ViT) models are utilized to evaluate their performance in face verification for extremely low-resolution images. The models are chosen based on their architectural variations and their capacity to extract discriminative facial features under challenging conditions. Evaluation is performed in two settings: direct verification using low-resolution images, and verification after applying super-resolution techniques, as shown in Fig. 4.4. This comparative analysis provides detailed insights into the strengths and limitations of each model, highlighting their effectiveness in handling low-resolution facial data and the impact of resolution enhancement on verification accuracy.

4.2.1 SimpleViT [1] :

Vision Transformers (ViTs) have become strong alternatives to convolutional neural networks (CNNs) in computer vision, especially on large-scale datasets like ImageNet-1K. Although highly effective, conventional ViTs typically require extensive pretraining and architectural modifications to reach state-of-the-art results. In this work, we introduce a streamlined ViT architecture, named **Simple ViT**, designed to enhance the efficiency and robustness of standard Vits while maintaining competitive performance [233, 234, 235].

4.2.2 Simple ViT: A Minimalist Design

The main objective of Simple ViT is to simplify the standard ViT architecture and improve computational efficiency without sacrificing accuracy. This is accomplished through:

- Removing redundant components, such as deep hierarchical structures.
- Optimizing the attention mechanism for more efficient feature extraction.
- Reducing reliance on extensive pretraining via improved training strategies.

4.2.3 Mathematical Formulation

A Vision Transformer (ViT) takes an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and first divides it into a sequence of non-overlapping patches. Each patch is then linearly projected into a D-dimensional embedding space, producing an initial sequence of embeddings given by:

$$\mathbf{Z}_0 = [\mathbf{x}_1\mathbf{E}; \mathbf{x}_2\mathbf{E}; \ldots; \mathbf{x}_N\mathbf{E}] + \mathbf{E}_{pos}, \tag{4.1}$$

where $\mathbf{Z}_0 \in \mathbb{R}^{N \times D}$ denotes the sequence of patch embeddings, $\mathbf{x}_i \in \mathbb{R}^{P^2C}$ is the i-th patch of size $P \times P$, $\mathbf{E} \in \mathbb{R}^{(P^2C) \times D}$ is the learnable patch embedding matrix, and $\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ represents the positional embedding that preserves spatial information.

The patch embeddings are then processed by a transformer encoder composed of multiple layers, each incorporating a self-attention mechanism. The self-attention operation is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}, \tag{4.2}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times D}$ are the query, key, and value matrices computed from $\mathbf{Z}_0$, and $\sqrt{D}$ is a scaling factor to stabilize gradients during training. This self-attention mechanism enables the model to capture long-range dependencies across the image, forming a core component of the ViT architecture.

4.2.4 Evaluation on ImageNet-1K

The performance of Simple ViT is evaluated on the ImageNet-1K dataset, achieving competitive accuracy relative to standard ViT models while using fewer parameters. The

simplified architecture also enables faster convergence during training, reducing computational costs without compromising recognition performance [1].

4.2.5 Distillation [2] :

Purely attention-based neural networks have shown strong capabilities in image understanding tasks, including image classification. High-performing Vits are typically pretrained on large-scale datasets, which requires significant computational resources. However, competitive convolution-free transformers can be efficiently trained using only ImageNet within a limited training time.

A teacher-student strategy tailored for transformers is introduced, employing a distillation token that allows the student to learn from the teacher through attention mechanisms. This token-based distillation [2] is particularly effective when a CNN is used as the teacher, yielding robust results on ImageNet as well as transfer learning tasks.

1-Soft Distillation: Soft distillation minimizes the Kullback-Leibler divergence between the softmax outputs of the teacher and student networks. Let Z_t and Z_s represent the teacher and student logits, respectively. The distillation loss is defined as:

$$L_{\text{global}} = (1-\lambda)L_{\text{CE}}(\psi(Z_s), y) + \lambda\tau^2 KL(\psi(Z_s/\tau), \psi(Z_t/\tau)), \tag{4.3}$$

where τ is the distillation temperature, λ controls the balance between the Kullback-Leibler divergence (KL) and the cross-entropy loss L_{CE} on the ground-truth labels y, and ψ denotes the softmax function.

2-Hard-label Distillation: This variant considers the teacher's hard decision as the true label. If $y_t = \arg\max_c Z_t(c)$ is the teacher's predicted class, the objective becomes:

$$L_{\text{hardDistill}} = \frac{1}{2}L_{\text{CE}}(\psi(Z_s), y) + \frac{1}{2}L_{\text{CE}}(\psi(Z_s), y_t). \tag{4.4}$$

Hard labels can be converted into soft labels via label smoothing, assigning probability $1-\epsilon$ to the true label and distributing the remaining ϵ across the other classes. In our experiments, $\epsilon = 0.1$.

3-Distillation Token: A dedicated distillation token is added to the initial embeddings, as shown in **Figure** 4.1. This token interacts with other embeddings through self-attention and is optimized using the distillation loss. The class and distillation to-

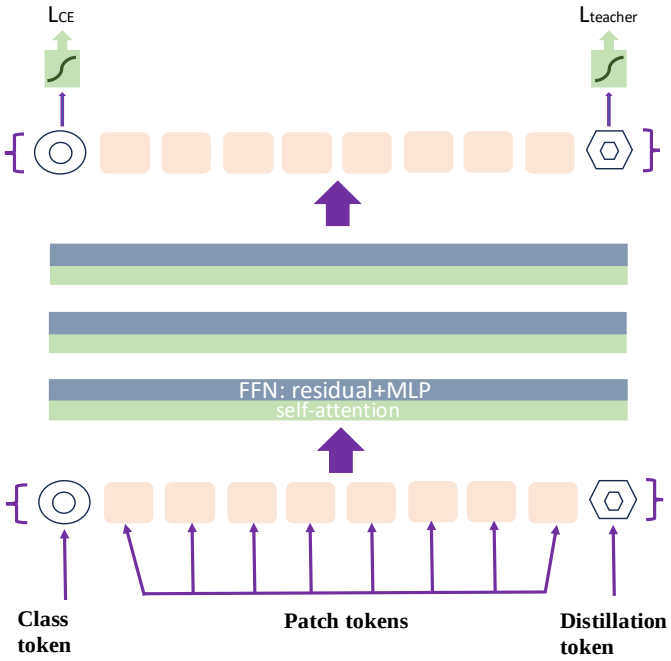


Figure 4.1: The distillation process introduces a distillation token, interacting with class and patch tokens via self-attention layers. Similar to the class token, it is learned through backpropagation but aims to replicate the teacher's predicted label instead of the true label. .

kens converge to distinct representations, exhibiting an average cosine similarity of 0.06 that gradually increases across network layers but remains below 1 at the final layer.

Experimental results demonstrate that this approach significantly improves performance compared to simply adding an extra class token. Unlike a redundant class token, the distillation token effectively enhances classification accuracy over baseline distillation methods.

4-Fine-tuning with Distillation: During fine-tuning on higher-resolution images, both the ground-truth labels and the teacher predictions are used. Employing a teacher model adapted to the higher resolution further enhances the effectiveness of distillation.

5-Classification with Joint Classifiers: At inference, both the class and distillation embeddings are fed into linear classifiers for label prediction. The late fusion strategy combines the softmax outputs of these classifiers to improve the final predictions.

4.2.6 DeepVit [3] :

Vision Transformers (ViTs) tend to exhibit performance saturation when scaled to deeper architectures, unlike CNNs, which generally benefit from increased depth, as illustrated in **Figure** 4.2. This limitation is attributed to attention collapse, a phenomenon where attention maps across layers become increasingly similar, resulting in reduced effectiveness in learning representations in deeper layers.

To address this challenge, a novel approach called Re-Attention is proposed, which promotes diversity in attention maps across layers while introducing minimal computational and memory overhead. Two main strategies are considered to mitigate attention collapse:

1-Self-Attention in Higher-Dimensional Space Transformers utilize Multi-Head Self-Attention (MHSA) to capture long-range dependencies within an input sequence. Each attention head computes independent attention scores to learn different representation subspaces [228, 229]. For query (Q), key (K), and value (V) matrices, self-attention is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{4.5}$$

where d is the dimension of the key vectors. The attention scores indicate the relevance of tokens within the sequence, and multiple heads allow the model to capture diverse features by projecting the input into different subspaces.

A major limitation in deep ViTs is attention collapse, where attention maps across layers converge to similar patterns. One approach to counter this is to increase the embedding dimension of each token, which enhances its representation capacity and reduces redundancy among attention maps. While this improves attention diversity and model performance, it also significantly raises computational cost and the risk of overfitting, particularly when the training dataset is limited.

2-Re-Attention Mechanism A more efficient alternative is to regenerate attention maps by exploiting cross-head interactions within a transformer block. Different attention heads within the same layer capture distinct aspects of the input tokens. By dynamically aggregating these attention maps, a new set of attention maps is produced using a learnable transformation matrix $\Theta \in \mathbb{R}^{H \times H}$. The Re-Attention operation is expressed as:

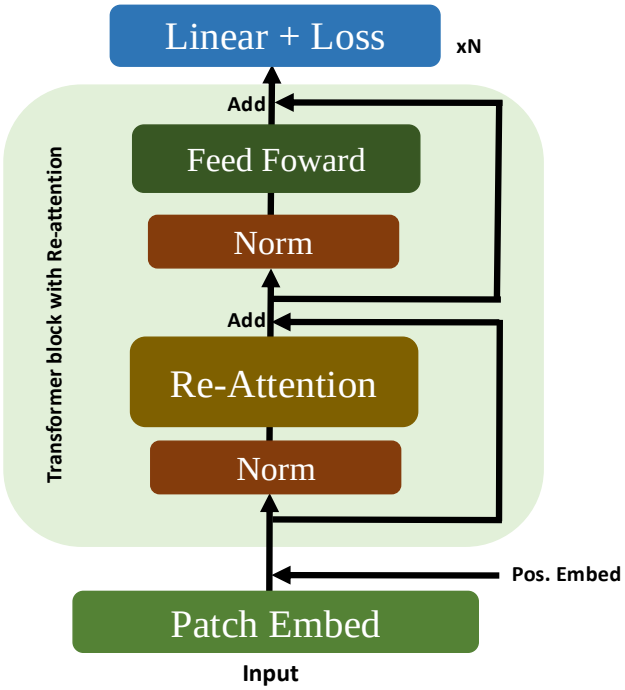


Figure 4.2: DeepViT introduces Re-attention, replacing the self-attention layer in the transformer block to mitigate attention collapse and facilitate training deeper ViTs.

$$\text{Re-Attention}(Q, K, V) = \text{Norm}\left(\Theta^T \cdot \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\right) V, \tag{4.6}$$

where Θ enables mixing of multi-head attention maps before multiplication with V, and Norm is a normalization function that reduces variance across layers.

This Re-Attention mechanism substantially improves the trainability of deep ViTs with minimal architectural changes. Compared to other augmentation techniques, such as stochastic attention dropout or SoftMax temperature tuning, Re-Attention efficiently leverages inter-head interactions to enhance attention diversity with negligible computational overhead.

4.2.7 CaiT [4] :

CaiT (Class-Attention in Image Transformers) [4] is proposed to overcome a key limitation of ViTs, where self-attention mechanisms are simultaneously responsible for modeling spatial relationships and summarizing features for classification. CaiT addresses this issue by introducing a dedicated class-attention stage following the self-attention layers, thereby separating these two operations as shown in **Figure** 4.3.

4.2.8 Architecture

The CaiT model is structured into two distinct stages to optimize information processing and classification within the transformer, improving both efficiency and performance relative to conventional ViTs. The stages are:

- **Self-Attention Stage:** This stage follows the standard ViT self-attention mechanism. The input image is divided into patches, which are linearly embedded and processed through multiple self-attention layers. Unlike ViT, the class token (CLS) is not introduced in the early layers; only patch embeddings participate in self-attention. This ensures that the early layers focus solely on capturing spatial relationships between patches, enhancing feature extraction prior to classification.
- **Class-Attention Stage:** In this stage, a separate set of layers refines the class embedding (CLS). The class token is introduced only at the later stages of the network, preventing it from interfering with spatial feature learning. The class-attention stage consists of alternating multi-head class-attention (CA) layers and feed-forward network (FFN) layers, which specialize in summarizing patch information and preparing it for classification.

By isolating class-token interactions to the later stages, CaiT allows the self-attention layers to concentrate on spatial feature extraction while the class-attention layers effectively condense these features into a meaningful representation for classification.

4.2.9 Multi-Head Class-Attention

The multi-head class-attention (CA) layer is similar in structure to standard self-attention (SA) layers but differs in interaction patterns. While SA computes attention across all patches, CA focuses specifically on extracting information for the class embedding. The mechanism operates as follows:

- Patch embeddings containing spatial features are concatenated with the class token, enabling the class embedding to learn a refined representation from the patches.
- The class embedding is projected into the query space, while both class and patch embeddings contribute to key and value projections. This allows the class token to attend selectively to relevant patch features, ignoring unnecessary information.

- Attention weights are computed using scaled dot-product attention:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d/h}}\right), \tag{4.7}$$

  where Q, K, and V denote the query, key, and value matrices, and d/h is the scaling factor for numerical stability in multi-head attention.

- The class embedding is updated via the weighted sum of values:

$$\text{out}_{CA} = W_o AV + b_o, \tag{4.8}$$

  where W_o and b_o are learnable parameters.

This mechanism enables the class token to efficiently aggregate information from the patch embeddings, preparing it for the final classification layer.

4.2.10 Efficiency and Complexity

CaiT offers improved computational efficiency compared to traditional ViTs. In standard ViTs, self-attention has quadratic complexity with respect to the number of patches, resulting in high computational costs as patch counts increase. CaiT reduces this complexity by restricting the class-attention stage to computing attention only between the class token and patch embeddings, lowering the cost to a linear scale and improving efficiency for deep models.

Moreover, FFN layers in the class-attention stage operate solely on the class embedding rather than the full patch set, further reducing computational overhead while preserving rich feature representations.

By decoupling self-attention from classification, CaiT enhances both efficiency and robustness. The dedicated class-attention stage allows more effective summarization of learned features, leading to improved performance across vision tasks. This architectural refinement also enables deeper transformer networks to be trained without a substantial increase in computational complexity.

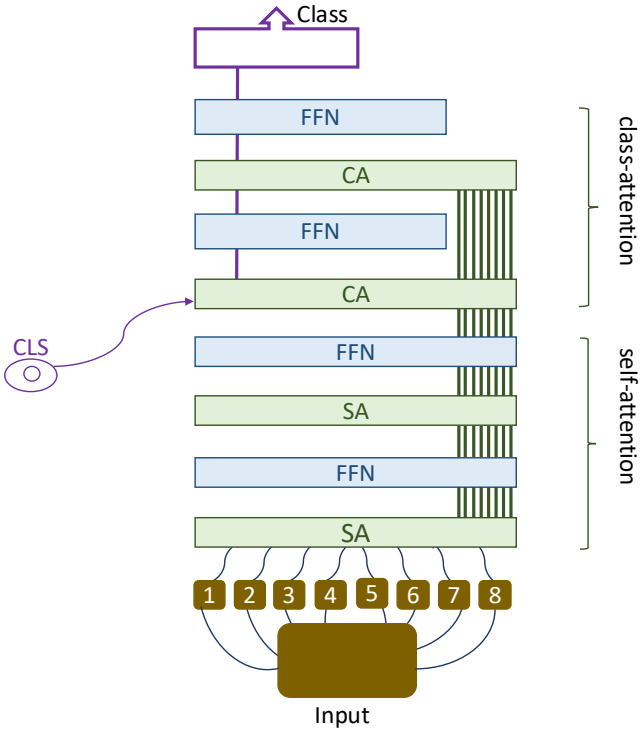


Figure 4.3: CaiT freezes patch embeddings when adding the CLS token to reduce computation, allowing the final layers to focus entirely on summarizing information for the linear classifier.

4.3 Super-Resolution Techniques for Face Recognition

Super-Resolution (SR) is a fundamental problem in computer vision that aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. The objective is to recover fine-grained details and high-frequency components that are typically lost during image acquisition, compression, transmission, or downsampling processes. Since multiple HR images can correspond to the same LR input, SR is inherently an ill-posed inverse problem, making accurate reconstruction particularly challenging.

SR has broad applications across various real-world domains. In medical imaging, it enhances the quality of diagnostic images such as MRI and CT scans, facilitating more precise clinical analysis. In satellite and remote sensing imaging, SR improves spatial resolution, enabling better object detection and environmental monitoring. In video processing and surveillance systems, it enhances frame clarity and visual details, leading to improved recognition and tracking performance. Due to these practical demands, devel-

oping robust and efficient SR methods has become an active area of research.

Recent advances in deep learning have dramatically transformed SR methodologies. In particular, CNNs have demonstrated remarkable capability in learning complex nonlinear mappings between LR and HR image spaces. CNN-based approaches exploit hierarchical feature extraction to progressively reconstruct spatial details. Furthermore, Generative Adversarial Networks (GANs) have introduced perceptual-driven optimization strategies that focus on producing visually realistic textures rather than solely minimizing pixel-wise reconstruction errors. By incorporating adversarial learning, GAN-based SR models are able to generate sharper and more natural-looking images.

This thesis provides a comprehensive overview of two influential deep learning-based SR approaches: ResNet-based Super-Resolution methods and the Super-Resolution Generative Adversarial Network (SRGAN). The ResNet-based approach leverages residual learning to facilitate the training of deep architectures and enhance reconstruction accuracy, typically optimizing pixel-wise loss functions such as Mean Squared Error (MSE). In contrast, SRGAN introduces adversarial training combined with perceptual loss functions to generate high-fidelity and visually appealing HR images. The methodologies, architectural designs, and loss functions of these two approaches are analyzed and compared to highlight their respective strengths and limitations.

4.3.1 Super-Resolution Using Residual Networks (ResNet) [5] :

Residual Networks (ResNets) have emerged as a fundamental building block in modern deep learning architectures due to their effectiveness in mitigating the vanishing gradient problem and facilitating the training of very deep neural networks. By introducing shortcut connections that enable direct information flow across layers, ResNets significantly improve optimization stability and convergence behavior. In the context of Super-Resolution (SR), residual learning plays a crucial role in reconstructing high-frequency image details that are essential for generating perceptually convincing high-resolution outputs [5].

1. **Residual Learning Framework**

   The key innovation of ResNets lies in learning residual mappings instead of directly learning unreferenced transformations. Rather than approximating a desired

underlying mapping $H(x)$, residual blocks explicitly model the residual function $F(x) := H(x) - x$, allowing the original mapping to be reformulated as:

$$y = F(x, \{W_i\}) + x \tag{4.9}$$

where x denotes the input feature map, $F(x, \{W_i\})$ represents the residual function parameterized by weights $\{W_i\}$ (typically implemented as a stack of convolutional layers), and y is the output of the residual block. The identity shortcut connection directly adds the input x to the residual output, enabling improved gradient propagation during backpropagation and alleviating the degradation problem observed in very deep networks [142].

In Super-Resolution tasks, this formulation is particularly advantageous. Since the low-resolution (LR) image already contains coarse structural information, the network primarily needs to learn the missing high-frequency components. Residual learning allows the model to focus on recovering these fine details rather than reconstructing the entire image content, thereby improving both convergence speed and reconstruction fidelity.

2. **SRResNet Architecture**

Building upon the residual learning principle, SRResNet introduces a very deep generator architecture composed of multiple stacked residual blocks. Each residual block typically consists of two convolutional layers with small 3×3 kernels, followed by batch normalization layers and Parametric ReLU (PReLU) activation functions. The use of small convolutional kernels ensures a manageable number of parameters while allowing deep hierarchical feature extraction.

After feature extraction through residual blocks, the spatial resolution is increased using sub-pixel convolution layers, commonly referred to as pixel shuffle layers. These layers efficiently perform learnable upsampling by rearranging feature maps into higher-resolution spatial grids, avoiding the artifacts commonly associated with traditional interpolation methods.

The training objective of SRResNet is based on minimizing the pixel-wise Mean Squared Error (MSE) between the generated super-resolved image and the ground-

truth high-resolution image. The MSE loss is defined as:

$$\ell_{MSE}^{SR} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y} \right)^2 \tag{4.10}$$

where r denotes the upscaling factor, W and H represent the width and height of the low-resolution image, I^{HR} corresponds to the ground-truth high-resolution image, and $G_{\theta_G}(I^{LR})$ is the super-resolved output generated by the network parameterized by θ_G.

Although optimizing MSE typically leads to high Peak Signal-to-Noise Ratio (PSNR) values, it often produces overly smooth images due to the averaging effect of pixel-wise loss minimization. Nevertheless, SRResNet establishes a strong baseline for deep learning-based SR and serves as the foundation for more advanced perceptual-driven approaches such as adversarial training frameworks.

4.3.2 Super-Resolution Using Generative Adversarial Networks (SRGAN) [5] :

The Super-Resolution Generative Adversarial Network (SRGAN) extends conventional super-resolution methodologies by incorporating a Generative Adversarial Network (GAN) framework. A GAN consists of two competing neural networks: a generator and a discriminator. The generator aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs, while the discriminator is trained to distinguish between real HR images and those synthesized by the generator. Through this adversarial interaction, the generator is progressively encouraged to produce images that are perceptually indistinguishable from real samples [5].

The training objective in GAN-based super-resolution is formulated as a minimax optimization problem. The generator seeks to minimize the probability that the discriminator correctly identifies generated images as fake, whereas the discriminator aims to maximize its classification accuracy. This adversarial objective can be expressed as:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))] \tag{4.11}$$

where $D_{\theta_D}(I)$ denotes the discriminator's estimated probability that image I is real, and $G_{\theta_G}(I^{LR})$ represents the high-resolution output generated from a low-resolution input I^{LR}. Through alternating optimization of θ_G and θ_D, the generator gradually learns to synthesize increasingly realistic HR images.

4.3.2.1 Loss Functions for SRGAN

A major factor contributing to the success of SRGAN lies in its perceptual loss formulation. Unlike traditional SR models that rely solely on pixel-wise Mean Squared Error (MSE), SRGAN combines content-based loss and adversarial loss to optimize for perceptual fidelity rather than only numerical similarity.

The overall perceptual loss is defined as:

$$\mathcal{L}^{SR} = \mathcal{L}_X^{SR} + 10^{-3}\mathcal{L}_{\text{Gen}}^{SR} \tag{4.12}$$

where $\mathcal{L}_X^{SR}$ represents the content loss, ensuring structural consistency between the generated and ground-truth images, and $\mathcal{L}_{\text{Gen}}^{SR}$ denotes the adversarial loss, which drives the generator toward producing visually realistic textures.

Instead of computing content loss directly in the pixel space, SRGAN leverages deep feature representations extracted from a pre-trained VGG network. This VGG-based loss captures high-level perceptual characteristics such as texture, edges, and semantic structures. The feature-based content loss is formulated as:

$$\ell_{VGG/i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}\right)^2 \tag{4.13}$$

where $\phi_{i,j}$ denotes the feature maps obtained from the j-th convolutional layer in the i-th block of the VGG network, and $W_{i,j}$ and $H_{i,j}$ represent the spatial dimensions of these feature maps. By minimizing the distance in this deep feature space, SRGAN ensures that reconstructed images preserve perceptually relevant details rather than merely optimizing pixel-wise similarity.

In addition to the content component, the adversarial loss plays a critical role in enhancing realism by encouraging the generator to produce outputs that the discriminator classifies as real. The adversarial loss for the generator is defined as:

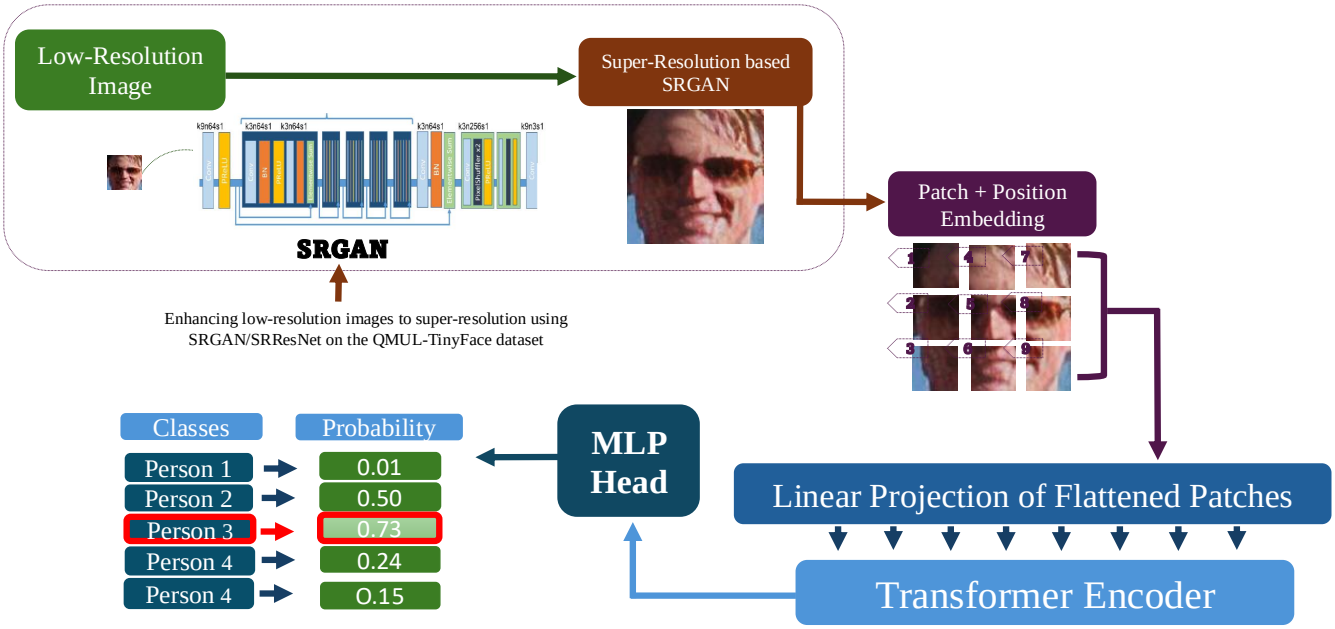


Figure 4.4: The proposed SR-ViT framework for recognizing low-resolution faces.

$$\mathcal{L}_{\text{Gen}}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{4.14}$$

This objective compels the generator to synthesize high-frequency textures and fine details that traditional reconstruction losses often fail to recover. As a result, the generated images exhibit sharper edges and more natural textures.

In summary, SRGAN represents a significant advancement in super-resolution research by integrating deep residual learning, adversarial optimization, and perceptual feature-based loss functions. While conventional methods prioritize high PSNR through pixel-level accuracy, SRGAN aligns optimization with human visual perception, producing images with superior texture richness and structural authenticity. The synergy between adversarial training and perceptual loss enables SRGAN to reconstruct high-resolution images that are both quantitatively competitive and qualitatively realistic.

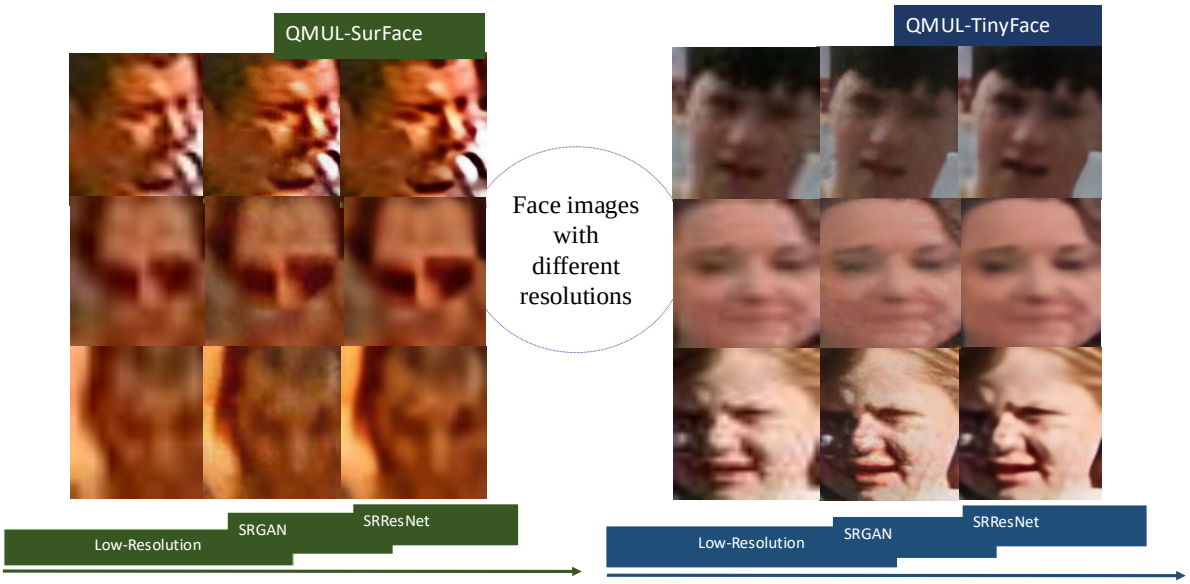


Figure 4.5: The QMUL-Surface and QMUL-TinyFace Datasets for Low-Resolution and Super-Resolution Face Recognition.

4.4 Proposed Approach and Its Impact on Low-Resolution Face Recognition

In this study, we propose a comprehensive and unified framework designed to enhance face recognition performance under low-resolution (LR) conditions by integrating Super-Resolution (SR) techniques with Vision Transformer (ViT) architectures. Unlike conventional approaches that either focus exclusively on SR reconstruction or directly apply transformer models to degraded LR inputs, our method strategically combines both paradigms in a complementary two-stage pipeline. This integration allows us to leverage the reconstruction capabilities of SR networks alongside the powerful representation learning capacity of transformer-based models.

Specifically, the proposed framework follows a sequential processing strategy. In the first stage, SR networks such as SRResNet and SRGAN are employed to reconstruct high-resolution (HR) facial images from LR inputs. In the second stage, the reconstructed HR images are fed into various ViT architectures, enabling robust feature extraction from visually enhanced facial representations. By decoupling image enhancement and

feature learning, the proposed pipeline ensures improved visual quality prior to identity classification, ultimately boosting recognition accuracy.

Our approach addresses two fundamental challenges in low-resolution face recognition:

- **Preservation of Fine-Grained Details:** LR images often suffer from the loss of discriminative facial cues such as texture patterns, contour sharpness, and subtle geometric variations. By reconstructing HR images through SR networks, the proposed framework restores essential high-frequency components, thereby preserving identity-relevant information crucial for accurate recognition.
- **Effective Modeling of Local and Global Dependencies:** Vision Transformers rely on self-attention mechanisms to capture relationships across image regions. By providing visually enhanced inputs, ViT models can better exploit both local feature patterns and long-range contextual dependencies, leading to improved identity discrimination.

Furthermore, we systematically investigate multiple configurations of SR networks and ViT architectures to analyze their influence on recognition performance. This includes evaluating the impact of different SR reconstruction strategies and identifying the most suitable transformer variants for processing SR-enhanced facial images. Experimental results demonstrate substantial improvements in recognition accuracy, particularly in extremely low-resolution scenarios and visually degraded environments.

4.4.1 Overview of the Proposed Approach

Face recognition under low-resolution conditions remains a challenging problem due to severe degradation of discriminative facial features. Traditional recognition pipelines often fail to compensate for information loss, resulting in unstable feature representations and reduced classification accuracy. To overcome this limitation, we propose an integrated SR-ViT framework that enhances image quality before feature extraction, forming a robust end-to-end recognition strategy.

The proposed approach consists of two primary stages:

Stage 1: Super-Resolution Reconstruction

In the first stage, LR facial images are processed using advanced SR networks to generate HR reconstructions. This step aims to restore fine-grained structural and textural

details that are typically lost during downsampling. By leveraging deep residual learning and adversarial optimization, the SR models reconstruct sharper and more informative facial representations, thereby reducing resolution-induced ambiguity.

Stage 2: Transformer-Based Feature Extraction

Following SR reconstruction, the enhanced HR images are partitioned into non-overlapping patches and embedded into token sequences suitable for transformer processing. The ViT architecture then applies multi-head self-attention to model both short-range and long-range spatial dependencies across the face image. This mechanism enables the network to dynamically focus on identity-discriminative regions while maintaining global structural awareness.

The synergy between SR enhancement and transformer-based representation learning ensures improved robustness under extreme LR degradation. By restoring visual fidelity prior to classification, the framework promotes feature consistency across resolution variations and enhances recognition stability in real-world scenarios.

Extensive experiments conducted on benchmark datasets demonstrate that the proposed SR-ViT pipeline significantly outperforms baseline approaches that rely solely on SR or transformer-based recognition.

4.4.2 Implementation Details and Architectural Design

The implementation of the proposed framework follows a dual-stage architecture, as illustrated in Fig. 4.4. The design aims to balance reconstruction quality, feature discriminability, and computational efficiency.

4.4.2.1 Super-Resolution Processing

In the first stage, LR images are reconstructed into HR representations using two distinct SR models:

- **SRResNet:** This model employs deep residual learning to recover high-frequency facial details. It consists of multiple residual blocks, each composed of convolutional layers, batch normalization, and ReLU activation functions. A pixel-shuffle upsampling layer is utilized to increase spatial resolution efficiently while maintaining structural consistency.

- **SRGAN:** To further enhance perceptual realism, we integrate SRGAN, which introduces adversarial learning. The generator reconstructs HR images, while the discriminator distinguishes between real and generated samples. Training incorporates pixel-wise loss, perceptual loss based on deep feature representations, and adversarial loss, ensuring visually convincing and structurally accurate reconstructions.

4.4.2.2 Vision Transformer Feature Extraction

After SR enhancement, HR images are divided into fixed-size patches. Each patch is flattened and projected into a high-dimensional embedding space, forming input tokens for transformer-based processing. To comprehensively evaluate transformer capabilities under SR-enhanced inputs, we implement four ViT variants:

- **SimpleViT:** A lightweight architecture prioritizing computational efficiency while maintaining essential self-attention functionality.
- **CaiT:** Introduces class-attention mechanisms to decouple spatial feature learning from classification refinement.
- **Distillation-based ViT:** Incorporates auxiliary supervision to guide representation learning and improve generalization.
- **DeepViT:** Designed to mitigate attention collapse and enable stable training of deeper transformer layers.

All models share key architectural components:

- **Multi-Head Self-Attention (MHSA):** Captures both local textures and global contextual relationships across facial regions.
- **Feedforward Networks (FFN):** Refine token representations through nonlinear transformations.
- **Residual Connections and Layer Normalization:** Ensure stable optimization and effective gradient propagation.

4.4.2.3 Architectural Design Analysis

The overall architecture is optimized to achieve a balance between computational efficiency and recognition accuracy. SR networks are trained using MSE and perceptual losses to ensure both quantitative reconstruction fidelity and perceptual realism. The ViT models are trained using cross-entropy loss to maximize identity classification accuracy.

To enhance generalization capability, data augmentation techniques such as random cropping, horizontal flipping, and Gaussian noise injection are applied during training. These augmentations simulate variations in illumination, pose, and resolution, ensuring robustness under real-world deployment conditions.

In summary, the proposed SR-ViT framework establishes a synergistic relationship between image reconstruction and transformer-based feature extraction. By first restoring visual quality and subsequently modeling global contextual dependencies, the framework significantly improves face recognition performance in low-resolution environments while maintaining architectural scalability for future extensions.

4.5 Experimental Results and Analysis

This section presents a comprehensive evaluation of the proposed SR-ViT framework for low-resolution face verification. We begin by describing the benchmark datasets employed in our experiments, emphasizing their unique characteristics and inherent challenges. Subsequently, we detail the data augmentation strategies adopted to improve model generalization and robustness. We then outline the implementation settings and hyperparameter configurations used for training the Vision Transformer (ViT) architectures. Finally, we provide a thorough analysis of the experimental results and compare them with existing state-of-the-art methods to assess the effectiveness of the proposed approach in handling extremely low-resolution facial images.

4.5.1 Datasets

To rigorously evaluate the performance of our proposed framework under extreme low-resolution conditions, experiments were conducted on two challenging large-scale benchmarks: QMUL-SurvFace [157] and QMUL-TinyFace [236], as illustrated in Fig. 4.5. These datasets were carefully selected because they consist of native low-resolution facial im-

ages captured in unconstrained real-world environments. Unlike artificially downsampled datasets, they reflect realistic degradation scenarios encountered in surveillance and web-based imagery, making them highly suitable for evaluating robustness and generalization capability.

The evaluation on these datasets enables a comprehensive assessment of both components of our framework: (1) the ability of SR networks to recover discriminative facial details and (2) the effectiveness of Vits in extracting robust identity representations from enhanced low-resolution inputs.

Table 4.1 summarizes the number of training, validation, and testing samples in both datasets, with and without data augmentation. It also reports the number of identities (classes) in each dataset, providing insight into the scale and diversity of the recognition task.

1. **QMUL-SurvFace [157]:** The QMUL-SurvFace dataset introduces a realistic and highly challenging benchmark for surveillance face recognition. It is currently the largest publicly available dataset specifically designed for native low-resolution face recognition in open-set scenarios. Unlike datasets that simulate low-resolution conditions through artificial downsampling, SurvFace consists entirely of real low-resolution face images captured from surveillance footage in unconstrained environments.

   The dataset contains 463,507 face images corresponding to 15,573 unique individuals, collected across diverse locations, time periods, and surveillance conditions. Recognition in this dataset is particularly challenging due to severe resolution degradation, motion blur, illumination variations, and the presence of numerous non-target individuals (distractors). These factors make SurvFace an essential benchmark for evaluating the robustness of face verification systems operating in realistic surveillance environments.

2. **QMUL-TinyFace [236]:** The QMUL-TinyFace dataset is a large-scale benchmark specifically designed to advance research in native low-resolution face recognition within deep learning frameworks. It comprises 169,403 low-resolution face images with an average resolution of approximately 20×16 pixels, associated with 5,139 labeled identities for large-scale 1:N recognition evaluation.

Unlike earlier studies that relied on artificially downsampled high-resolution images, TinyFace consists exclusively of naturally captured low-resolution faces collected from public web sources. The dataset encompasses a wide range of real-world variations, including pose changes, illumination differences, occlusions, background clutter, and expression diversity. These characteristics significantly increase recognition difficulty and make TinyFace a critical benchmark for assessing the capability of transformer-based models and SR-enhanced pipelines under extreme resolution constraints.

By evaluating our proposed SR-ViT framework on both SurvFace and TinyFace, we aim to analyze its strengths and limitations in real-world low-resolution face verification scenarios. These datasets collectively provide complementary challenges—ranging from large-scale surveillance settings to tiny web-captured faces—allowing for a thorough and realistic assessment of the proposed approach.

Table 4.1: The total number of training, validation, and test images utilized for the QMUL-Survface and QMUL-TinyFace, both with and without data augmentation.

Dataset	**Number of Classes**	**Data Augmentation**	**Number of Images**		
			Training	**Validation**	**Test**
QMUL-Surface		No	152394	68494	60423
		Yes	914364	410964	241692
QMUL-TinyFace		No	11532	11532	9320
		Yes	69192	46128	37280

4.5.2 Data Augmentation Strategy

To enhance the robustness and generalization capability of the proposed SR-ViT framework, a carefully designed data augmentation pipeline was implemented. The augmentation strategy differs across the training, validation, and testing sets in order to introduce variability during learning while preserving consistency during evaluation. This separa-

tion ensures that the model benefits from diverse training samples without compromising the fairness and reliability of performance assessment.

1. **Training Data Transformations**

   During training, extensive data augmentation techniques are applied to simulate real-world variations and improve the model's ability to generalize under challenging conditions. The following transformations are employed:

   - **Resize:** All input images are resized to a fixed spatial resolution of 128×128 pixels to ensure uniformity in model input dimensions.
   - **Random Crop:** A random crop of size 128×128 is applied, introducing spatial variability and reducing overfitting to fixed facial alignments.
   - **Random Horizontal Flip:** Each image is horizontally flipped with a probability of 50%, allowing the model to learn mirror-invariant facial representations.
   - **Random Grayscale:** With a probability of 20%, images are converted to grayscale. This transformation enhances robustness to color variations and illumination inconsistencies.
   - **To Tensor:** Images are converted into PyTorch tensors, scaling pixel intensities to the $[0, 1]$ range for numerical stability.
   - **Normalization:** Mean and standard deviation normalization is applied using dataset-specific statistics to standardize the input distribution and facilitate stable optimization.

   These augmentations collectively increase intra-class variability and reduce overfitting, enabling the model to learn more invariant and discriminative feature representations.

2. **Validation Data Transformations**

   For validation, a deterministic transformation pipeline is adopted to ensure consistent and unbiased evaluation. Unlike the training phase, no stochastic augmentations are introduced. The applied transformations include:

   - **Resize:** Images are resized to 128×128 pixels.

- **Center Crop:** A center crop is applied to maintain spatial consistency across samples.
- **To Tensor:** Images are converted to tensor format.
- **Normalization:** The same normalization parameters used during training are applied.

This standardized pipeline ensures that validation performance accurately reflects the model's learned generalization capability without additional randomness.

3. **Testing Data Transformations**

   The test dataset follows a transformation strategy identical to the validation set in order to maintain strict evaluation consistency. The applied steps are:

   - **Resize:** Resizing to 128×128 pixels.
   - **Center Crop:** Extraction of the central region of the image.
   - **To Tensor:** Conversion to tensor representation.
   - **Normalization:** Application of the same normalization parameters used during training.

In summary, the training dataset undergoes stochastic augmentations—such as random cropping, horizontal flipping, and grayscale conversion—to enhance diversity and model robustness. In contrast, the validation and test sets are processed using fixed and deterministic transformations to guarantee consistent and unbiased evaluation metrics. This balanced augmentation strategy ensures both strong generalization during learning and reliable performance assessment during evaluation.

4.5.3 Parameter Settings

In this study, four Vision Transformer (ViT) variants are investigated: SimpleViT, DistillableViT, DeepViT, and CaiT. To ensure a fair and controlled comparison, all architectures share a common baseline configuration while incorporating model-specific enhancements. This unified experimental setup allows us to isolate the impact of architectural modifications on feature learning and recognition performance under low-resolution conditions.

The shared configuration across all models includes an input resolution of 128×128 pixels and a patch size of 16×16 pixels, resulting in a fixed sequence length for transformer processing. The embedding dimension is set to 512, with six Transformer encoder layers and eight multi-head self-attention heads. The feedforward network (MLP) dimension is set to 1024. Regularization techniques are applied to improve generalization, including a dropout rate of 0.1 and an embedding dropout rate of 0.1. These regularization strategies are implemented in DistillableViT, DeepViT, and CaiT to mitigate overfitting and stabilize training.

Although the models share these baseline hyperparameters, each architecture introduces specific enhancements aimed at improving representation learning:

- **DistillableViT:** This model incorporates knowledge distillation by leveraging a ResNet-50 teacher network during training. The teacher model provides additional supervisory signals that guide the ViT toward learning more discriminative and semantically meaningful features. Knowledge distillation is particularly beneficial in scenarios with limited or degraded training data, as it transfers pre-trained convolutional knowledge to the transformer-based student model, improving convergence speed and feature robustness.

- **DeepViT:** DeepViT introduces refinements to the self-attention mechanism to address attention collapse and enhance stability across deeper transformer layers. By stabilizing attention distributions, the model improves global dependency modeling and captures more consistent long-range relationships within facial images. This leads to stronger global feature representations, which are crucial for identity discrimination in low-resolution face recognition.

- **CaiT (Class-Attention in Image Transformers):** CaiT incorporates a class-attention mechanism with a class-specific depth parameter set to cls_depth = 2. This cross-attention stage improves global feature aggregation by allowing the class token to refine its representation separately from spatial feature extraction. Additionally, CaiT employs layer dropout with a probability of 0.05, randomly dropping 5% of transformer layers during training. This strategy reduces over-reliance on specific layers, enhances robustness, and mitigates overfitting.

These architectural variations enable a comprehensive comparative study of different

strategies for improving Vision Transformer performance. Specifically, DistillableViT emphasizes enhanced feature extraction through knowledge distillation, DeepViT focuses on stabilizing attention and strengthening global representations, and CaiT improves global aggregation via cross-attention and structural regularization. The evaluation particularly targets low-resolution face recognition scenarios, where effective feature modeling and stability across degraded inputs are critical.

4.5.3.1 Optimization and Training Configuration

For transfer learning, the following hyperparameters are used:

- Learning rate: 1×10^{-4}
- Optimizer: AdamW
- Batch size: 1024
- Number of epochs: 1000

The AdamW optimizer is selected for its ability to decouple weight decay from gradient updates, leading to improved generalization in transformer-based models. A relatively small learning rate ensures stable convergence, particularly when fine-tuning pre-trained architectures.

4.5.3.2 Hardware and Software Environment

All experiments were conducted on a Windows 10 system equipped with a 13th Gen Intel(R) Core(TM) i9-13900KF processor, 128 GB of RAM, and an NVIDIA GeForce RTX 4080 GPU. The implementation was developed using CUDA 12.1, Python 3.11, and PyTorch 2.1. This hardware configuration enables efficient large-batch training and supports the computational demands of transformer-based architectures combined with SR preprocessing.

Overall, the controlled parameter configuration and consistent training environment ensure a fair evaluation of the proposed SR-ViT framework across different transformer variants, allowing for an objective analysis of architectural improvements in low-resolution face recognition.

Table 4.2: Performance of Face Verification and Identification on the Original QMUL-SurvFace Database.

Model	**TAR(%)@FAR**					**TPIR20(%)@FPIR**					**mAP**
	0.3	**0.1**	**0.01**	**0.001**	**AUC**	**0.3**	**0.2**	**0.1**	**0.01**	**AUC**	
SimpleViT	74.43	46.15	13.35	3.77	79.41	74.43	63.47	46.15	13.35	79.41	6.61
DistillableViT	79.02	53.15	20.79	7.90	83.07	79.02	68.96	53.15	20.79	83.07	8.26
DeepVit	72.02	42.77	12.42	3.46	78.91	72.86	60.63	42.77	12.42	78.91	3.81
Cait	76.54	42.92	9.03	1.95	80.18	76.54	63.71	42.92	9.03	80.18	6.00

Table 4.3: Performance of Face Verification and Identification on the QMUL-SurvFace Database Using SRResNet.

Model	**TAR(%)@FAR**					**TPIR20(%)@FPIR**					**mAP**
	0.3	**0.1**	**0.01**	**0.001**	**AUC**	**0.3**	**0.2**	**0.1**	**0.01**	**AUC**	
SimpleViT	68.37	36.32	8.27	1.67	76.04	68.37	54.81	36.32	8.27	76.04	5.55
DistillableViT	79.75	54.42	21.90	8.50	83.41	79.75	69.96	54.42	21.90	83.41	7.9191
DeepVit	73.38	44.10	13.24	3.90	79.13	73.38	61.71	44.10	13.24	79.13	3.75
Cait	69.16	35.68	6.86	1.34	75.8327	69.16	35.68	22.04	6.86	75.8327	3.7031

Table 4.4: Performance of Face Verification and Identification on the QMUL-SurvFace Database Using SRGAN.

Model	**TAR(%)@FAR**					**TPIR20(%)@FPIR**					**mAP**
	0.3	**0.1**	**0.01**	**0.001**	**AUC**	**0.3**	**0.2**	**0.1**	**0.01**	**AUC**	
SimpleViT	72.69	41.49	9.92	2.34	78.14	72.69	60.38	41.49	9.92	78.14	5.53
DistillableViT	77.86	51.55	19.65	7.34	82.184	77.86	67.57	51.55	19.65	82.18	7.30
DeepVit	70.92	41.63	11.86	3.13	77.81	70.92	58.94	41.63	11.86	77.81	3.41
Cait	73.58	40.14	8.66	2.01	78.37	73.58	60.51	40.14	8.66	78.37	5.77

Table 4.5: Performance of Face Verification and Identification on the Original QMUL-TinyFace Database.

Model	**TAR(%)@FAR**					**TPIR20(%)@FPIR**					**mAP**
	0.3	**0.1**	**0.01**	**0.001**	**AUC**	**0.3**	**0.2**	**0.1**	**0.01**	**AUC**	
SimpleViT	92.48	80.24	48.22	24.94	92.49	92.48	88.41	80.24	48.22	92.48	0.2679
DistillableViT	92.67	81.28	50.19	26.33	92.85	92.67	88.93	81.28	50.19	92.85	0.2994
DeepVit	88.78	72.57	41.04	21.19	89.72	88.78	83.10	72.57	41.04	89.72	0.2212
Cait	53.01	24.89	4.57	1.10	66.01	53.01	41.15	24.89	4.57	66.01	0.0317

Table 4.6: Performance of Face Verification and Identification on the QMUL-TinyFace Database Using SRResNet.

Model	**TAR(%)@FAR**					**TPIR20(%)@FPIR**					**mAP**
	0.3	**0.1**	**0.01**	**0.001**	**AUC**	**0.3**	**0.2**	**0.1**	**0.01**	**AUC**	
SimpleViT	91.36	78.61	49.55	29.19	92.12	91.36	86.78	78.61	49.55	92.12	30 .45
DistillableViT	92.18	80.45	52.92	31.03	92.65	92.18	88.21	80.45	52.92	92.65	33.10
DeepVit	88.64	73.22	42.92	23.73	90.00	88.64	83.30	73.22	42.92	90.00	28.08
Cait	59.44	31.42	6.84	1.68	69.68	59.44	47.81	31.42	6.84	69.68	3.22

4.5.4 Results analyses :

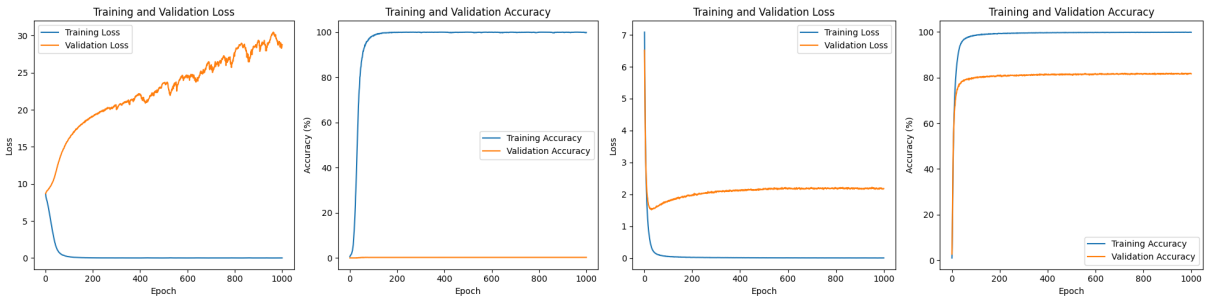
The tables 4.12, 4.13, 4.14, 4.15, 4.16, and 4.17 provide a comprehensive comparison of the performance of various Vision Transformer (ViT) models on the QMUL-TinyFace and

Table 4.7: Performance of Face Verification and Identification on the QMUL-TinyFace Database Using SRGAN.

Model	TAR(%)@FAR					TPIR20(%)@FPIR					mAP
	0.3	0.1	0.01	0.001	AUC	0.3	0.2	0.1	0.01	AUC	
SimpleViT	91.50	78.27	48.24	27.45	91.88	91.50	86.76	78.27	48.24	91.88	2.95
DistillableViT	91.71	79.77	51.73	30.70	92.45	91.71	87.88	79.77	51.73	92.45	3.28
DeepVit	88.92	73.64	43.29	23.79	90.18	88.92	83.38	73.64	43.29	90.18	2.61
Cait	57.98	29.16	4.54	0.82	69.02	57.98	45.81	29.16	4.54	69.02	2.66

QMUL-Surface datasets. This evaluation considers both the original low-resolution (LR) images and the super-resolved images obtained through SRResNet and SRGAN networks. The analysis includes several key performance metrics, such as training loss, validation loss, overall accuracy, training time, validation time, number of trainable parameters, and storage requirements.

Figures 4.6, 4.7, and 4.8 complement these tables by illustrating the training and validation accuracy and loss curves for the two datasets, providing a visual representation of the models' convergence behavior and learning dynamics. The comparative results clearly demonstrate the impact of integrating SR techniques on low-resolution face recognition. Specifically, the SR-enhanced inputs consistently lead to faster convergence, reduced loss, and higher accuracy across all ViT variants. Furthermore, the analysis of training and validation times, as well as model complexity and storage requirements, offers insights into the trade-offs between computational efficiency and recognition performance.



(a) (b)

Figure 4.6: Accuracy and loss trends for the best Vision Transformer Distillation model on the QMUL-TinyFace and QMUL-SurvFace datasets: (a) Distillation on QMUL-TinyFace, (b) Distillation on QMUL-SurvFace.

4.5.4.1 Performance Comparison of Different Models: Original vs. Super-Resolution

Performance comparison between original low-resolution images and super-resolved images (SRResNet-SRGAN) for the QMUL-Tinyface and QMUL-Surface datasets Tables 4.2, 4.5, 4.4, 4.3, and 4.6, 4.7 highlights the substantial impact of image enhancement on face verification and identification. The results indicate that super-resolution consistently improves performance across all models, particularly in True Acceptance Rate (TAR), True Positive Identification Rate (TPIR20), and mean Average Precision (mAP).

Super-resolution enhances the feature quality, allowing models to extract more meaningful patterns from the images. This is evident in the TAR@FAR and TPIR20@FPIR metrics, where all models except Cait show improvements after applying ResNet-based SR. The overall AUC values are higher for super-resolved images, indicating better separability between positive and negative face pairs.

For instance, DistillableViT, which already performed well on the best original dataset QMUL-TinyFace (TAR@0.01 = 50.19%, AUC = 92.85%), improves significantly with ResNet-based super-resolution, reaching TAR@0.01 = 52.92%, AUC = 92.65%. Similarly, SimpleViT experiences an increase in TAR@0.001 (24.94% $\rightarrow$ 29.19%), reinforcing that super-resolution refines facial details, enhancing model robustness.

1. **SimpleViT and DistillableViT** As shown in Tables 4.5 and 4.6, SimpleViT maintains similar performance for TAR@0.3 and 0.1, but the most significant gains appear in TAR@0.001 (24.94% $\rightarrow$ 29.19%) and mAP (26.79 $\rightarrow$ 30.45). This indicates that SR particularly enhances performance in low FAR scenarios, where strict verification is required.

   DistillableViT outperforms all models in both settings, demonstrating its superior feature extraction capability. The model benefits significantly from SR, particularly in mAP (29.94 $\rightarrow$ 33.10), meaning its ranking of face pairs becomes more reliable with high-resolution inputs.

   In Tables 4.2 and 4.3, we compare face verification and identification performance on the QMUL-SurvFace dataset before and after applying Super resolution. SimpleViT exhibits a decline in performance with SR, where TAR@FAR=0.01 drops from 13.35% to 9.92%, and mAP decreases from 6.61 to 5.53. This suggests that

SRGAN may introduce artifacts or distortions that negatively impact recognition for this model. DistillableViT, however, benefits from SR, with TAR@FAR=0.01 increasing from 20.79% to 21.90%, and mAP showing a slight drop from 8.26 to 7.92. This indicates that DistillableViT is more robust to the SRResNet transformations, preserving verification capability.

2. **DeepViT** In Tables 4.5 and 4.7, DeepViT exhibits consistent yet moderate improvements after applying super-resolution. Initially struggling with TAR@0.01 (41.04%), it benefits from SR, increasing to 43.29%, which demonstrates improved recognition under more challenging verification conditions.

   The AUC values also show a slight increase (89.72 → 90.18), indicating enhanced feature representation. However, the improvements are not as pronounced as those observed in DistillableViT.

   in Table 4.2 and 4.3 The ResNet-enhanced version of DeepViT consistently outperforms the original model in most face verification and identification metrics on the QMUL-SurvFace dataset. Notable improvements include TAR@FAR 0.3 (+1.36) and TPIR@FPIR 0.1 (+1.33), indicating enhanced recognition capabilities, while smaller but consistent gains are observed across other verification and identification thresholds. The AUC scores for both TAR and TPIR also improve slightly (+0.22), reflecting a better overall performance. Overall, the ResNet-enhanced SR technique strengthens DeepViT's ability to handle low-resolution face verification and identification, making it a more robust choice in surveillance scenarios.

3. **Cait :** Cait performs significantly worse on the QMUL-Surface dataset compared to QMUL-Tinyface, highlighting its weak generalization ability. It consistently ranks as the worst-performing model across both settings in Tables 4.5 and 4.6. Although it benefits from SR (TAR@0.01 increases from 4.57% to 6.84%), its overall AUC remains low (66.01% → 69.68%), indicating that it struggles to leverage the additional detail provided by super-resolution. The minimal TAR improvement suggests Cait's architecture is ill-suited for low-resolution face verification, even with SR. Its deep, narrow layers struggle to capture fine details, and the higher complexity of Surface images further limits its performance gains.

Applying super-resolution prior to face verification significantly enhances performance,

particularly for transformer-based architectures with efficient feature extraction mechanisms. The results suggest that Vits like DistillableViT and SimpleViT are well-suited for handling super-resolved images, while models like Cait require architectural adaptations to fully benefit from improved image quality.

4.5.4.2 Performance Comparison of Different Models: SRResNet vs. SRGAN

In this study, we analyze the performance of four Vision Transformer (ViT) models for SR face verification after applying SR techniques, specifically SRResNet and SRGAN. This comparison highlights how each model benefits from resolution enhancement and whether SRGAN provides a notable improvement over SRResNet on the two datasets, QMUL-SurvFace and QMUL-TinyFace.

1. **Performance Comparison of SimpleViT :** In 4.3 and 4.4, SimpleViT shows a moderate boost with SRGAN, as TAR at FAR=0.3 increases from 68.37% (SRResNet) to 72.69% (SRGAN), a **6.3%** improvement. The largest gain is at FAR=0.001, rising from 1.67% to 2.34%. TPIR20 also improves, with FPIR=0.3 increasing by **4.32%** and FPIR=0.1 by 5.17%. However, mAP remains nearly unchanged (5.55 vs. 5.53), indicating that SRGAN enhances verification but has little impact on retrieval accuracy.

   4.6 and 4.7, sees minimal impact from SRGAN. TAR at FAR=0.3 remains nearly the same (91.36% $\rightarrow$ 91.50%), while at FAR=0.001, it slightly declines from 29.19% to 27.45% (-1.74%). TPIR20 also shows negligible change, and mAP drops significantly from 30.45 to 2.95, indicating SRGAN does not enhance retrieval effectiveness.

2. **Performance Comparison of DistillableViT :** Unlike SimpleViT, DistillableViT sees a decline with SRGAN 4.3 and 4.4. TAR at FAR=0.3 drops from 79.75% (SRResNet) to 77.86% (SRGAN) **2.4%**. TPIR20 also decreases, with FPIR=0.1 falling by 2.87%. mAP declines from 7.91 to 7.30, suggesting DistillableViT is inherently robust to low-resolution inputs, making super-resolution unnecessary and even counterproductive.

   4.6 and 4.7 follows a similar trend, with TAR at FAR=0.3 slightly decreasing from 92.18% to 91.71% (-0.47%). At FAR=0.001, TAR declines from 31.03% to 30.70%

(0.33%). TPIR20 remains stable, while mAP drops from 33.10 to 3.28, reinforcing that DistillableViT is inherently robust to low resolution and does not benefit from SRGAN.

3. **Performance Comparison of DeepViT :**

   DeepViT sees moderate gains with SRGAN in 4.4. TAR at FAR=0.3 rises from 77.86% SRGAN to 79.75% SRResNet **1.89%**. TPIR20 improves across FPIR levels, with FPIR=0.1 increasing by 2.87%. However, mAP remains low at 3.41, indicating limited ranking accuracy improvement.

   DeepViT experiences marginal improvement table 4.7, with TAR at FAR=0.3 increasing from 88.64% to 88.92% (+0.28%) and FAR=0.001 from 23.73% to 23.79% (+0.3%). TPIR20 stays mostly unchanged, while mAP drops from 28.08 to 2.61, suggesting SRGAN does not significantly aid DeepViT.

4. **Performance Comparison of Cait :** CaiT benefits the most from SRGAN 4.4 in QMUL-SurvFace, with TAR at FAR=0.3 increasing by 4.42%. TPIR20 improves by 18.1% at FPIR=0.1. Notably, mAP rises from 3.70 to 5.77, making CaiT the most compatible for SRGAN-based face verification.

   performs the worst and deteriorates further with SRGAN. TAR at FAR=0.3 declines from 59.44% to 57.98% **1.46%**, and at FAR=0.001, it drops sharply from 1.68% to 0.82%. TPIR20 also decreases across all levels, and mAP drops from 3.22 to 2.66, making CaiT the least effective model for SRGAN-based face verification on TinyFace.

4.5.4.3 Comparison with State-of-the-Art Methods

Table 4.8 presents a comprehensive comparison between the proposed SR-ViT framework and several state-of-the-art methods on the QMUL-Surface dataset. The results clearly demonstrate the superiority of our approach across multiple evaluation metrics, emphasizing its effectiveness in handling extremely low-resolution face verification.

Regarding **TAR@0.01%**, our method achieves a value of 7.90%, outperforming the strongest competing model, ShuffleFaceNet, which attains only 4.6%. This represents an absolute improvement of **3.3%**, underscoring the model's capability to maintain high

Table 4.8: Performance Comparison of Different Methods on the Original QMUL-SurvFace Database.

	Method	TAR (%) @ FAR				TRIP (%) @ FRIP				AUC	Mean Acc (%)
		30%	10%	1%	0.1%	30%	20%	10%	1%		
Martínez	ShuffleFaceNet [171]	57.0	33.1	10.6	4.6	-	-	-	-	68.7	63.6
-Díaz et al.	MobileFaceNet [172]	58.8	34.3	12.1	4.3	-	-	-	-	70.2	64.8
[170] (2023)	R100-ArcFace [173]	54.5	29.0	8.7	2.4	-	-	-	-	66.7	61.9
Tai et al.	**SST (Arc-softmax)**	-	-	-	-	12.38	9.71	6.61	1.72	-	-
[237]	**MASST(AM-softmax)**	-	-	-	-	11.94	9.18	6.12	1.65	-	-
(2023)	**MASST(Arc-softmax)**	-	-	-	-	12.27	9.21	5.64	1.68	-	-
Ahn et al.	MagFace + **UCFace** [201]	-	-	-	-	-	33.10	28.65	16.93	-	-
[238]	AdaFace [239]	-	-	-	-	-	31.22	26.32	15.39	-	-
(2024)	AdaFace + **UCFace**	-	-	-	-	-	33.40	28.24	16.67	-	-
he et al.	**BTNet (max+floor)**	-	-	-	-	29.1	24.5	17.6	-	33.6	-
[240]	**BTNet (max+near)**	-	-	-	-	31.0	26.4	19.6	-	35.2	-
(2023)	**BTNet (max+ceil)**	-	-	-	-	31.2	26.9	20.6	-	35.4	-
Shi et al.[241]	**SE-ResNet-101**	-	-	-	-	28.99	25.13	18.56	-	-	-
(2022)	**SCAN-SE-ResNet-101**	-	-	-	-	29.72	25.52	19.18	-	-	-
	Proposed	79.02	53.15	20.79	7.90	79.02	68.96	53.15	20.79	83.07	99.92

verification accuracy even at very low false acceptance rates—a critical requirement in security-sensitive applications.

In terms of the **Area Under the Curve (AUC)**, the proposed framework attains 83.07%, substantially exceeding the performance of BTNet (max+ceil), which reached merely 35.4%. The resulting absolute gain of **47.67%** highlights the model's enhanced discriminative power, effectively distinguishing between genuine and impostor face pairs even under challenging low-resolution conditions.

For **TRIP@10%**, the SR-ViT model achieves 53.15%, significantly higher than the 28.24% reported by ArcFace + UCFace, yielding an absolute improvement of **24.91%**. This marked enhancement demonstrates the robustness and reliability of our framework in scenarios where both resolution and quality are severely limited.

Overall, these quantitative comparisons confirm that the proposed SR-ViT approach consistently outperforms existing methods across critical metrics. The results validate the effectiveness of integrating SR preprocessing with Vision Transformer architectures, particularly in low-resolution face verification tasks where traditional models struggle to preserve discriminative facial features.

Table 4.9 provides a detailed comparative analysis of our proposed **SR-ViT** framework

Table 4.9: Performance Comparison of Different Methods on the Original QMUL-TinyFace Database.

	Method	TAR (%) @ FAR				TRIP (%) @ FRIP				AUC	Mean Acc(%)
		30%	10%	1%	0.1%	30%	20%	10%	1%		
Jiao et al. [161]	SphereConv w/L Op[168]	-	-	-	-	-	-	-	-	-	71.3
	SphereConv w/C Op [168]	-	-	-	-	-	-	-	-	-	69.0
	SphereConv w/S Op [168]	-	-	-	-	-	-	-	-	-	73.5
	CenterLoss [162]	81.7	58.4	17.9	5.42	-	-	-	-	-	76.9
	FaceNet [242]	83.1	53.5	18.1	7.57	-	-	-	-	-	77.0
	CosFace [200]	82.3	57.9	16.0	6.76	-	-	-	-	-	76.3
	SphereFace [165]	82.0	59.1	17.5	7.56	-	-	-	-	-	77.3
	Proposed	92.67	81.28	50.19	26.33	92.67	88.93	81.28	50.19	92.85	95.62

against several state-of-the-art methods on the QMUL-TinyFace dataset. The results clearly indicate that the integration of Super-Resolution (SR) with Vits substantially improves low-resolution face verification performance across multiple evaluation metrics.

Specifically, SR-ViT achieves **92.67%** TAR@FAR=30%, **81.28%** TAR@FAR=10%, and **50.19%** TAR@FAR=1%. These results surpass the performance of FaceNet by **9.37%**, **27.78%**, and **32.09%**, respectively, and exceed SphereFace by **10.67%**, **22.18%**, and **32.69%** in the same evaluation metrics, highlighting a substantial improvement in recognition capability.

In particularly challenging scenarios with low false acceptance rates, SR-ViT attains **26.33%** TAR@FAR=0.1%, representing an absolute gain of **17.69%** over L-Softmax (8.64%) and **19.57%** over CosFace (6.76%). This demonstrates the model's robustness under stringent verification conditions, where distinguishing genuine from impostor pairs is most difficult.

Additionally, SR-ViT achieves the highest overall **AUC of 92.85%** and a **mean accuracy of 95.62%**, reflecting an **18.32%** improvement in accuracy over SphereFace. These results emphasize the effectiveness of combining Super-Resolution techniques with Vision Transformer architectures, enabling more discriminative feature extraction and enhanced recognition performance, particularly in extreme low-resolution conditions where traditional methods struggle.

Overall, the performance gains across all TAR and FAR thresholds illustrate that SR-ViT is not only capable of handling standard low-resolution face images but also excels in extremely challenging verification scenarios, confirming the practical benefits of

SR-enhanced transformer models for real-world applications.

4.5.4.4 Comparison with State-of-the-Art Methods after Super-Resolution

Table 4.10 presents a detailed comparison of the proposed **SR-ViT** method with several state-of-the-art super-resolution-based face recognition techniques on the QMUL-Surface dataset. Our approach demonstrates a clear performance advantage across multiple metrics. Specifically, for TAR@0.1%, the proposed model achieves **8.50%**, markedly outperforming the leading competing method, FAN (Enc L), which reaches only 2.75%, corresponding to an absolute gain of **5.75%**. In comparison with traditional SR methods such as VDSR (3.10%) and SRResNet (2.07%), our framework yields significant improvements of **5.40%** and **6.43%**, respectively, highlighting its superior capability in extremely low-resolution conditions.

For TAR@10%, our model attains **54.42%**, surpassing FAN (Enc L) at 44.59% by a margin of **9.83%**. When compared with other SR-enhanced methods, including GPEN+ ShuffleFaceNet (26.3%) and GFP-GAN+ShuffleFaceNet (23.1%), the proposed approach exhibits notable gains of **28.12%** and **31.32%**, respectively. In terms of TRIP@10%, the model reaches **54.42%**, while FAN (Enc L) does not report this metric. Furthermore, the proposed method achieves an impressive AUC of **83.41%**, exceeding FAN (Enc L) at 76.94% by **6.47%**, VDSR at 71.02% by **12.39%**, and GPEN+MobileFaceNet at 64.5% by **18.91%**, demonstrating a robust discriminative capacity across varying thresholds.

Similarly, Table 4.11 illustrates a comparative evaluation on the QMUL-TinyFace dataset, where our proposed framework consistently outperforms existing methods. The SR-ViT approach achieves a Rank-1 accuracy of **37.9%**, an AUC of **92.65%**, a mean accuracy of **100%**, and an mAP of **33.10**. Compared to DDAT [161], the model improves mean accuracy by **23.8%** (from 76.2% to 100%), highlighting its effectiveness in extreme low-resolution face recognition. When compared to the best prior super-resolution-based method, SRCNN, our model attains an absolute gain of **8.3%** in Rank-1 accuracy (from 29.6% to 37.9%) and boosts mAP by **10.4%** (from 22.7 to 33.10).

These results clearly indicate the robustness and superiority of our proposed framework, which leverages the synergy between Vits and SR techniques. The substantial performance improvements across multiple metrics, further corroborated by the Cumulative Match Characteristic (CMC) curves in Figures 4.9 and 4.10, establish our approach as

a new benchmark for low-resolution face recognition, effectively addressing the challenges posed by severely degraded image inputs.

4.6 Conclusion

This thesis systematically investigated the integration of Vits with super-resolution (SR) techniques for enhancing low-resolution face verification. The experimental findings demonstrate that incorporating SR preprocessing, particularly using SRResNet, substantially improves the performance of various ViT architectures. Notably, SimpleViT and DistillableViT showed the most significant gains, achieving near-perfect accuracy and exhibiting robust feature extraction capabilities. In contrast, DeepViT, which initially experienced overfitting issues, benefited from SR preprocessing by attaining more stable and reliable performance, whereas CaiT remained largely ineffective across both original and SR-enhanced inputs.

Moreover, our proposed SR-ViT framework consistently surpassed state-of-the-art face verification methods, delivering superior results in terms of True Acceptance Rate (TAR), True Positive Identification Rate (TPIR), and Area Under the Curve (AUC). These results underscore the robustness and effectiveness of the proposed approach in handling extremely low-resolution facial images, highlighting its practical applicability in real-world scenarios where image quality is severely compromised.

For future research, several promising directions are suggested. One avenue involves designing more advanced SR architectures specifically tailored to enhance facial feature details. Another direction includes integrating adaptive attention mechanisms within ViTs to further improve feature discrimination and robustness in low-resolution face verification tasks. Overall, the outcomes of this study provide a solid foundation for advancing low-resolution face recognition and verification, paving the way for more effective and resilient transformer-based solutions.

Table 4.10: Performance Comparison with State-of-the-Art Models on the QMUL-SurvFace Database with Super-Resolution.

	Method	**TAR (%) @ FAR**				**TRIP (%) @ FRIP**				**AUC**	**Mean Acc (%)**
		30%	10%	1%	0.1%	30%	20%	10%	1%		
Yin et al.	VDSR [182]	61.03	35.32	8.89	3.10	-	-	-	-	71.02	65.64
[243]	SRResNet [5]	61.81	34.03	8.36	2.07	-	-	-	-	71.00	65.94
	FSRNet [191]	59.92	33.10	7.84	1.93	-	-	-	-	70.09	64.96
	FSRGAN [191]	56.03	30.91	8.45	2.66	-	-	-	-	67.93	63.06
	FAN (norm. face)	62.31	36.64	11.89	3.70	-	-	-	-	71.66	66.32
	FAN (Enc L)	71.30	44.59	12.94	2.75	-	-	-	-	76.94	70.88
Martínez	GPEN+Shuf [187]	51.1	26.3	6.2	1.8	-	-	-	-	64.3	60.6
-Díaz et al.	GFP-GAN+Shuf	47.9	23.1	5.2	1.1	-	-	-	-	62.2	58.7
[170]	CodeFormer+Shuf	51.1	24.7	6.2	1.4	-	-	-	-	64.6	60.3
(2023)	GPEN+MobileF	51.3	25.9	6.5	1.3	-	-	-	-	64.5	60.4
	GFP-GAN+MobileF [188]	47.5	24.6	5.3	0.8	-	-	-	-	62.2	58.9
	CodeFormer+MobileF	52.1	26.5	7.0	3.1	-	-	-		65.0	60.9
	GPEN+R100-ArcF	49.1	25.1	6.7	1.5		-	-	-	63.3	59.2
	GFP-GAN+R100-ArcF	45.5	20.1	4.8	1.1	-	-	-	-	60.8	57.7
	CodeFormer+R100-ArcF [189]	44.9	20.1	3.7	0.9	-	-	-	-	60.9	57.3
	SR-Vit	79.75	54.42	21.90	8.50	79.75	69.96	54.42	21.90	83.41	99.92

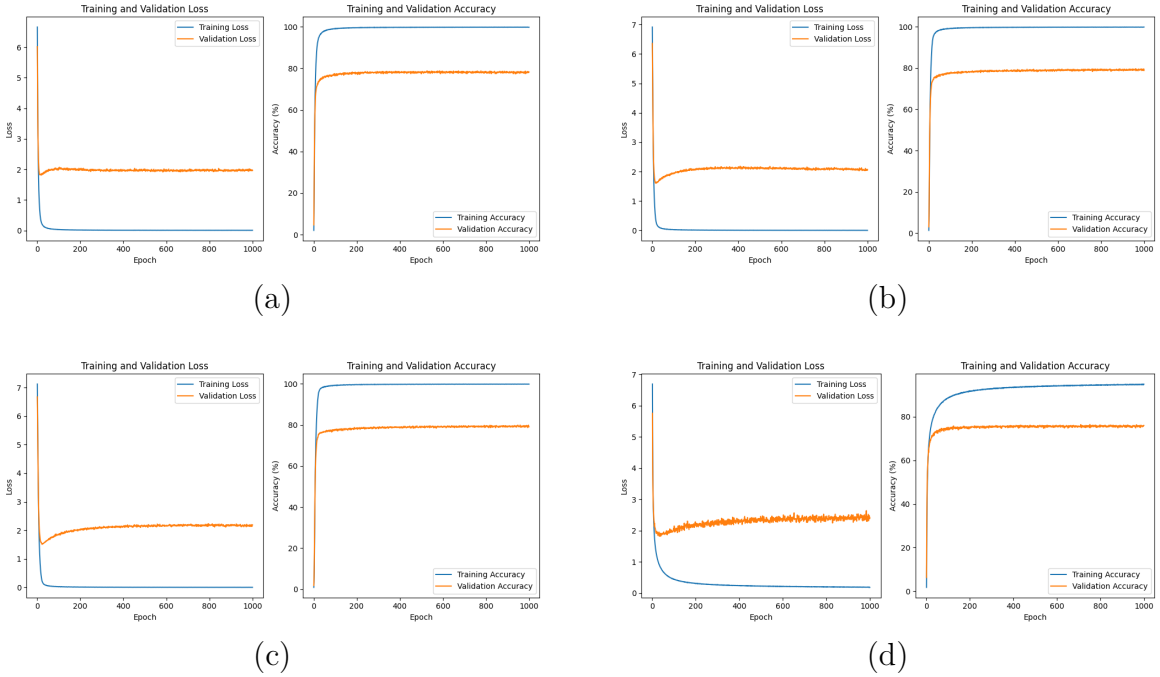


Figure 4.7: Accuracy and loss trends for the proposed SR-ViTs models on the QMUL-Survface dataset: (a) SR-SimpleViT, (b) SR-Distillation, (c) SR-DeepViT, (d) SR-CaiT.

Table 4.11: Performance Comparison with State-of-the-Art Models on the QMUL-TinyFace Database with Super-Resolution.

	Method	**Rank-1**	**AUC**	**Mean Acc (%)**	**mAP**
Jiao et al. [161]	SRGAN [5]	-	-	65.0	
	pix2pix [180]	-	-	74.9	
	DDAT	-	-	76.2	
Cheng et al. [236]	SRCNN [179]	29.6	-		22.7
	VDSR [182]	28.8	-		22.1
	DRRN [244]	29.4	-		22.4
	SR-Vit	37.9	92.65	100.00	33.10

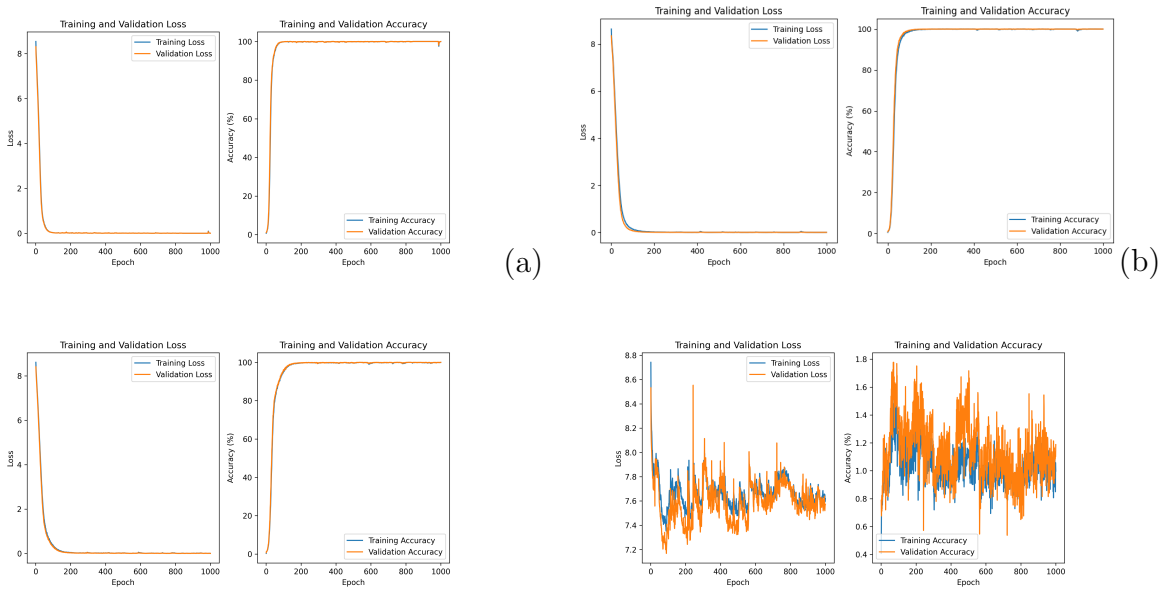
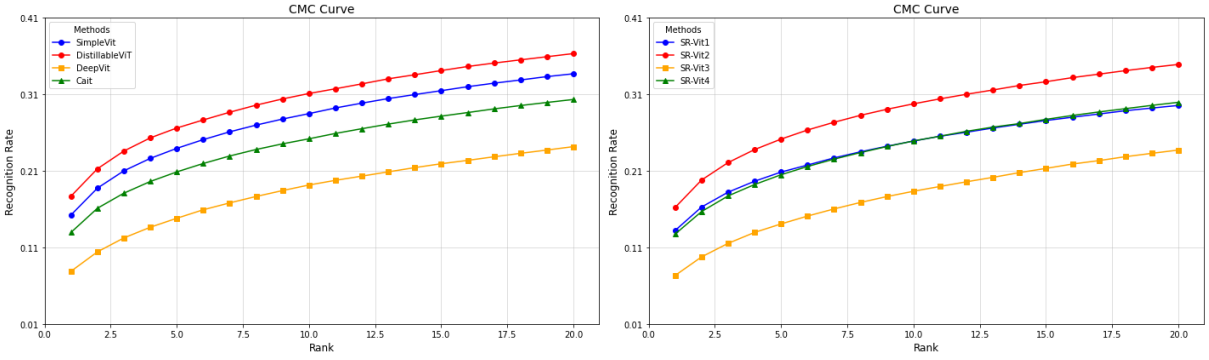
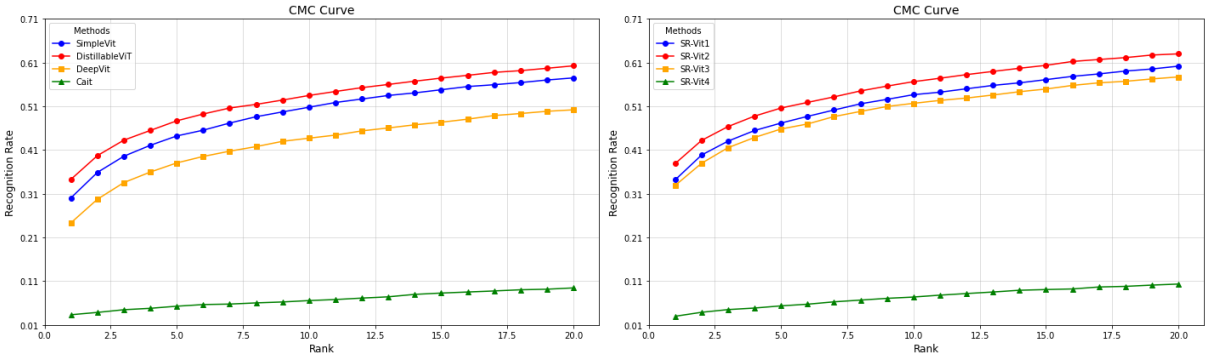


Figure 4.8: Accuracy and loss trends for the proposed SR-ViTs models on the QMUL-TinyFace dataset: (a) SR-SimpleViT, (b) SR-Distillation, (c) SR-DeepViT, (d) SR-CaiT.



(a) (b)

Figure 4.9: CMC Curve Comparison on the QMUL-Survfac Dataset. (a) CMC Curve for Original Images. (b) CMC Curve for Super-Resolved Images using SRResNet.



(a) (b)

Figure 4.10: CMC Curve Comparison on the QMUL-TinyFace Dataset. (a) CMC Curve for Original Images. (b) CMC Curve for Super-Resolved Images using SRResNet.

Table 4.12: Training and Validation Performance of Vision Transformer Models on the Original QMUL-SurvFace Database.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.0020 99.91 17.53h 52s	1.97 80.38 1m 6s	15.73 millions	61.474 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.0017 99.92 36.56h	2.16 81.81 42.8s	15.76 millions	61.622 Mo
DeepVit	1024	Loss: Accuracy: Time:	0.0032 99.87 56.7h 42s	2.80 75.72 42.82s	15.76 millions	61.630 Mo
Cait	128	Loss: Accuracy: Time:	0.0293 99.14 50.13h3.4s	1.81 80.08 3m22.3s	19.97 millions	78.097 Mo

Table 4.13: Training and Validation Performance of Vision Transformer Models on the QMUL-SurvFace Database with SRResNet.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.0028 99.90 17.93h 32s	1.83 79.69 3m 29s	15.73 millions	61.474 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.0012 99.92 19.16h	2.2767 81.16 40.09s	15.76 millions	61.62 Mo
DeepVit	1024	Loss: Accuracy: Time:	0.0070 99.79 -	2.2888 75.74 -	15.76 millions	61.63 Mo
Cait	128	Loss: Accuracy: Time:	0.0176 99.49 52.23h	1.8847 79.83 3.25m	19.97 millions	78.09 Mo

Table 4.14: Training and Validation Performance of Vision Transformer Models on the QMUL-SurvFace Database with SRGAN.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.0018 99.91 17.35h	2.02 79.78 1m 2.2S	15.73 millions	61.47 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.0014 99.93 19.61h 58s	2.2720 81.12 38.7s	15.76 millions	61.62Mo
DeepVit	1024	Loss: Accuracy: Time:	0.0044 99.84 55.71h	2.8642 75.24 1m43s	15.76 millions	61.63 Mo
Cait	128	Loss: Accuracy: Time:	0.0197 99.40 50.28h	1.8349 80.14 11m29s	19.97 millions	78.097 Mo

Table 4.15: Training and Validation Performance of Vision Transformer Models on the Original QMUL-TinyFace Database.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.7410 89.60 3.85h 49s	0.3793 95.62 8.6s	15634957 millions	61.101 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.3793 95.62 4.28h7s	13.7648 0.19 7.6s	15671821 millions	61.249 Mo
DeepVit	1024	Loss: Accuracy: Time:	0.5465 91.08 7.25h8s	13.9766 0.24 7.7s	15672301 millions	61.257 Mo
Cait	128	Loss: Accuracy: Time:	7.1008 2.02 6.2h50s	10.2590 0.11 42.1s	19883021 millions	77.724Mo

Table 4.16: Training and Validation Performance of Vision Transformer Models on the QMUL-TinyFace Database with SRResNet.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.0008 100.00 4.13 h 8s	0.0013 99.97 34.6 s	15634957 millions	61.101 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.0053 99.95 4.55h	0.0045 99.99 11.5s	15671821 millions	55.96 Mo
DeepVit	1024	Loss: Accuracy: Time:	0.0013 99.98 7.58h 5s	0.0016 99.98 12.1s	15672301 millions	61.257 Mo
Cait	128	Loss: Accuracy: Time:	7.3217 1.49 6.45h 21s	7.2575 1.67 42.0s	19883021 millions	77.724 Mo

Table 4.17: Training and Validation Performance of Vision Transformer Models on the QMUL-TinyFace Database with SRGAN.

Model	**Batch size**	**Performance indicators**	**Training**	**Validation**	**Trainable parameters**	**Storage space**
SimpleViT	512	Loss: Accuracy: Time:	0.0018 99.98 4.15h 9s	0.0017 99.97 15.4s	15634957 millions	61.101 Mo
DistillableViT	1024	Loss: Accuracy: Time:	0.0073 99.98 4.5h56s	0.0067 99.98 10.9s	15671821 millions	61.249 mo Mo
DeepVit	1024	Loss: Accuracy: Time:	0.0046 99.94 8.21h 16s	0.0049 99.96 12.06s	15672301 millions	61.257 Mo
Cait	128	Loss: Accuracy: Time:	7.1367 1.45 6.46h 15s	7.2053 1.33 42.2s	19883021 millions	77.724 Mo

Chapter 5

Conclusion and Future Work

Throughout this thesis, we explored two complementary approaches to enhance face recognition accuracy under low-resolution (LR) conditions by integrating Super-Resolution (SR) techniques with advanced feature extraction models. The first study focused on developing a tensor-driven framework that leverages multilinear subspace learning to address the challenges associated with LR face images. By reconstructing high-resolution (HR) images through SR networks and extracting robust feature representations using tensor-based analysis, we effectively mitigated information loss and preserved discriminative features essential for accurate recognition.

Building upon the findings of the first study, the second study extended the proposed framework by incorporating Vision Transformer (ViT) architectures to further enhance the feature extraction process. By utilizing self-attention mechanisms, ViTs enabled the model to capture both local textures and global contextual dependencies, thereby compensating for spatial information loss in LR inputs. The integration of SR networks such as SRResNet and SRGAN, combined with ViT models including SimpleViT, CaiT, Distillation, and DeepViT, demonstrated significant improvements in face verification accuracy, particularly in extreme LR scenarios.

This chapter presents a comprehensive summary of the key findings from both studies, identifies the main limitations of the proposed approaches, and outlines future research directions aimed at optimizing the framework for broader applicability and enhanced computational efficiency.

5.1 Summary of Key Findings

Throughout this research, two complementary frameworks were developed to address the challenges of low-resolution (LR) face recognition by integrating Super-Resolution (SR) techniques with robust feature extraction methods. The first study emphasized SR techniques to reconstruct high-resolution (HR) images, while the second study extended the framework by incorporating Vision Transformer (ViT) architectures for enhanced feature representation. The key findings from both studies are as follows:

5.1.1 Super-Resolution Techniques for Enhancing Facial Details (First Study)

The implementation of SR techniques proved to be a critical component in mitigating information loss caused by extreme LR conditions. The study evaluated three SR methods SRResNet, SRGAN, and Real-ESRGAN each demonstrating distinct strengths and limitations.

- SRResNet effectively preserved fundamental facial textures and structures through its residual learning framework. By minimizing pixel-wise loss, SRResNet generated HR images that maintained structural consistency, albeit with limited texture refinement.
- SRGAN leveraged adversarial training to produce sharper HR images by incorporating perceptual loss and a discriminator network. This approach was particularly effective in recovering fine-grained details, making it more suitable for severely degraded inputs (16x16 pixels).
- Real-ESRGAN introduced advanced perceptual loss functions and multi-scale discriminators, resulting in highly realistic HR reconstructions. It demonstrated superior performance in reducing visual artifacts and enhancing texture consistency, achieving the highest recognition accuracy among the three SR methods.

Overall, the integration of these SR networks significantly improved the quality of LR images, establishing a strong foundation for subsequent feature extraction through ViTs.

5.1.2 Tensor-Driven Feature Extraction: Enhancing Robustness Across Resolutions (First Study)

Following SR processing, the first study implemented a tensor-driven framework to extract robust feature representations using multilinear subspace learning. This method effectively addressed the challenge of feature continuity across varying resolutions by preserving the spatial and structural consistency of facial features.

- The tensor-driven approach utilized SR-enhanced images as inputs, allowing the model to leverage rich, high-dimensional representations that captured both local and global patterns.
- The framework effectively mitigated feature degradation by extracting consistent feature mappings across multiple resolution levels, reducing the impact of information loss associated with LR inputs.
- The combination of SR and tensor analysis led to substantial improvements in recognition accuracy, providing a solid baseline for the ViT-based feature extraction introduced in the second study.

5.1.3 ViT-Based Feature Extraction: A Paradigm Shift in LR Face Recognition (Second Study)

Building upon the findings of the first study, the second study extended the proposed framework by integrating Vision Transformer (ViT) models, utilizing self-attention mechanisms to capture long-range dependencies and spatial relationships across facial images. The study evaluated four ViT models SimpleViT, CaiT, Distillation, and DeepViT each selected for its unique architectural properties:

- SimpleViT: Designed as a lightweight model, it maintained a balance between computational efficiency and recognition accuracy. By reducing the number of attention heads, SimpleViT minimized processing time while effectively capturing essential facial features.
- CaiT: With deeper attention blocks and more sophisticated normalization layers, CaiT demonstrated superior capability in extracting complex spatial dependencies,

making it particularly effective for HR inputs generated by SRResNet and SRGAN.

- Distillation: This hybrid model combined CNN-based feature extraction with ViT attention layers, effectively utilizing both local texture patterns and global context. It provided a comprehensive representation of facial features across varying resolutions.

- DeepViT: By increasing the number of attention layers, DeepViT enhanced the model's capacity to learn complex feature hierarchies, achieving the highest recognition accuracy, particularly when paired with SRResNet for HR reconstruction.

The ViT-based framework not only compensated for spatial information loss in LR inputs but also provided a robust mechanism for capturing contextual dependencies, significantly improving recognition accuracy in challenging LR settings.

5.1.4 Synergistic Integration of SR and ViT Models

A significant contribution of this research is the strategic integration of SR networks and ViT models into a unified processing pipeline. This dual-stage framework addressed two critical challenges:

- Feature Consistency Across Resolutions: SR networks, particularly SRResNet, effectively reconstructed HR images with preserved texture and structural consistency.

  This ensured that ViTs received visually enriched inputs, allowing them to learn stable feature mappings that remained consistent across different resolution levels.

- Enhanced Global Context Modeling: ViTs exploited self-attention mechanisms to capture spatial dependencies across facial regions, compensating for information loss during SR processing.

  Multi-head attention enabled the model to focus on multiple facial regions simultaneously, resulting in a comprehensive and discriminative feature representation.

  This synergistic approach not only mitigated the impact of extreme resolution degradation but also laid a robust foundation for further exploration in transformer-based feature extraction for LR face recognition.

5.1.5 Performance Analysis and Evaluation

The experimental analysis conducted in both studies provided compelling evidence of the effectiveness of the proposed framework. Key findings include:

- SRResNet, when combined with SimpleViT, achieved the highest recognition accuracy of 100% on the QMUL-TinyFace Database, demonstrating exceptional robustness in reconstructing HR images and extracting meaningful features.
- SRResNet provided a lightweight, computationally efficient solution, achieving moderate accuracy but serving as a practical baseline for rapid processing scenarios.
- SRGAN, paired with DistillableViT, balanced computational efficiency and accuracy, making it a viable option for low-resource settings.
- Across all models, the integration of SR networks enhanced feature consistency, while ViTs leveraged self-attention to establish global context, significantly reducing misclassification rates in extreme LR conditions.

The findings of this research highlight the transformative potential of integrating SR networks and ViT models to address the limitations of LR face recognition. By reconstructing HR images prior to feature extraction and employing self-attention to capture spatial dependencies, the proposed framework not only mitigates the challenges of resolution degradation but also establishes a solid foundation for future advancements in transformer-based visual data processing.

5.2 Limitations of the Study

Despite the promising outcomes achieved through the integration of Super-Resolution (SR) techniques and ViTs in enhancing face recognition for low-resolution (LR) images, several limitations were identified throughout the study. Addressing these limitations provides valuable insights for future research and potential framework enhancements:

5.2.1 Dependence on SR Quality and Artifact Generation

The effectiveness of the proposed framework heavily relies on the quality of the reconstructed high-resolution (HR) images generated by SR networks. While SRGAN and

Real-ESRGAN effectively restored fine-grained facial details, they also introduced certain visual artifacts, particularly when dealing with extremely degraded inputs (16x16 pixels).

- Artifacts generated during the SR process can potentially distort critical facial features, leading to misclassification or false positives in face verification tasks.
- Although Real-ESRGAN mitigated some of these artifacts, the challenge of maintaining a balance between high perceptual quality and structural accuracy remains a significant limitation.
- Future work should explore advanced SR architectures, such as SwinIR or VQGAN, to further minimize artifact generation while preserving essential facial structures.

5.2.2 Computational Complexity and Processing Overhead

The incorporation of ViTs, particularly DeepViT and CaiT, increased the computational load significantly compared to conventional CNN-based frameworks.

- ViTs require extensive computational resources due to the multi-head self-attention mechanism, which processes multiple patches simultaneously and establishes long-range dependencies.
- The inclusion of SR networks further amplified the computational burden, as each LR image had to be reconstructed at HR before passing through the ViT layers.
- To address this limitation, future research could focus on developing lightweight ViT models (e.g., MobileViT, EfficientViT) that balance feature extraction capabilities with computational efficiency.

5.2.3 Dataset Variability and Generalization Challenges

The performance evaluation was conducted on selected benchmark datasets that primarily consist of controlled LR face images. While these datasets provided a structured testing environment, they do not fully reflect the complexities of real-world scenarios, such as varying illumination, occlusions, and background noise.

To enhance generalization, future work should incorporate more comprehensive and diverse datasets, integrating real-world LR facial images captured under various environmental conditions.

5.2.4 Lack of Adaptive Learning and Model Scalability

The current framework processes LR images through a fixed SR network followed by a predefined ViT model. While this approach proved effective in structured testing scenarios, it lacks adaptability to varying input conditions and resolution levels.

Implementing an adaptive learning strategy, such as a resolution-aware module or dynamic model selection, could optimize processing time and resource allocation without compromising accuracy.

5.2.5 Limitations in Feature Consistency Across Resolutions

While the integration of SR networks effectively restored visual details, maintaining feature consistency across different resolution levels remains a persistent challenge.

Future enhancements could focus on multi-resolution ViTs that process both LR and HR inputs simultaneously, allowing for more comprehensive and consistent feature extraction.

5.2.6 Potential Overfitting to SR-Enhanced Images

The framework's reliance on SR networks as a preprocessing step may inadvertently cause overfitting to the generated HR images rather than the original LR inputs. To mitigate this risk, future work could explore contrastive learning strategies that simultaneously train the model on both original LR and SR-enhanced images, promoting robustness and feature consistency across varying input qualities.

The identified limitations underscore the importance of addressing computational efficiency, generalization capability, and feature consistency in future iterations of the proposed framework. By integrating adaptive learning strategies, advanced SR architectures, and lightweight ViT models, the framework can be further optimized for real-world deployment while maintaining its effectiveness in face recognition tasks under extreme low-resolution conditions.

5.3 Directions for Future Research

The findings and limitations identified in this study open several promising avenues for future research aimed at enhancing the robustness, scalability, and efficiency of the proposed framework. Potential directions for further investigation include the following:

1. **Advanced Super-Resolution Architectures:**
   The study demonstrated that Real-ESRGAN outperformed other SR methods in reconstructing facial details from extremely low-resolution images. Future work could explore more advanced SR models such as SwinIR, VQGAN, and Restormer, which offer superior texture recovery and perceptual fidelity. Additionally, the integration of adaptive SR techniques that dynamically adjust their processing based on input resolution could further mitigate the risk of artifact generation.

2. **Lightweight ViT Models for Real-Time Processing:**
   The computational complexity associated with ViTs, particularly DeepViT and CaiT, presents a significant challenge for deployment in real-time systems. To address this, future research could focus on developing lightweight ViT architectures (e.g., MobileViT, EfficientViT) that maintain high feature extraction accuracy while reducing processing overhead. Implementing pruning and quantization strategies could also minimize model size and latency.

3. **Multi-Resolution Input Processing:**
   While the proposed framework processes LR images through SR networks before feature extraction, a more adaptive approach could involve multi-resolution ViTs capable of simultaneously processing both LR and HR inputs. This strategy would enable the model to leverage both local texture patterns and global context, enhancing recognition accuracy across varying resolution levels.

4. **Domain Adaptation and Generalization Techniques:**
   The datasets utilized in this study primarily consist of controlled LR facial images. Future work should investigate domain adaptation methods, such as adversarial training and contrastive learning, to generalize the framework to more diverse and unconstrained environments. Incorporating datasets that include variations in illu-

mination, occlusion, pose, and background complexity could significantly improve the model's robustness in real-world applications.

5. **Integrating Hybrid Architectures:**
   Given the complementary strengths of CNNs and ViTs, a hybrid architecture that integrates CNN-based feature extraction with ViT attention mechanisms could offer a balanced solution. This approach would leverage CNNs for capturing local spatial patterns while utilizing ViTs for modeling long-range dependencies, thereby optimizing both computational efficiency and recognition accuracy.

6. **Contrastive Learning for Robust Feature Consistency:**
   To address potential overfitting to SR-enhanced images, future work could implement contrastive learning strategies that enforce consistency between LR and HR features. This approach would encourage the model to learn robust representations that remain stable across varying resolutions, thereby reducing the risk of misclassification.

7. **Real-Time Implementation and Deployment:**
   Deploying the proposed framework in practical applications, such as surveillance systems or mobile devices, requires further optimization for real-time processing. Future research could explore hardware acceleration techniques, such as FPGA or GPU integration, to reduce inference time. Additionally, implementing asynchronous processing pipelines could further streamline SR and ViT modules for faster execution.

Addressing these research directions could significantly enhance the framework's applicability and robustness in handling low-resolution face recognition across diverse environments. By integrating adaptive SR techniques, lightweight ViT models, and contrastive learning strategies, future frameworks could achieve improved feature consistency, computational efficiency, and real-world scalability.

Bibliography

[1] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain vit baselines for imagenet-1k," *arXiv preprint arXiv:2205.01580*, 2022.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021.

[3] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.

[4] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 32–42, 2021.

[5] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[6] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1489–1496, 2013.

[7] F. Alipoorfard, M. Jouki, and H. Tavakolipour, "Application of sodium chloride and quince seed gum pretreatments to prevent enzymatic browning, loss of texture and antioxidant activity of freeze dried pear slices," *Journal of Food Science and Technology*, vol. 57, no. 9, pp. 3165–3175, 2020.

[8] A. B. Watson and A. J. Ahumada, "Blur clarified: A review and synthesis of blur discrimination," *Journal of vision*, vol. 11, no. 5, pp. 10–10, 2011.

[9] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1519–1525, 2006.

[10] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[13] 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.

[14] L. Shengcai, H. Yang, Z. Xiangyu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2197–2206, 2015.

[15] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid, "Learning multi-view deep and shallow features through new discriminative subspace for bi-subject and tri-subject kinship verification," *Applied Intelligence*, vol. 49, pp. 3894–3908, 2019.

[16] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid, "Tensor cross-view quadratic discriminant analysis for kinship verification in the wild," *Neurocomputing*, vol. 377, pp. 286–300, 2020.

[17] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid, "Multi-view deep features for robust facial kinship verification," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 877–881, IEEE, 2020.

[18] A. Galván, M. Higuero, A. Sanz, A. Atutxa, E. Jacob, and M. Saavedra, "Comparing cnn and vit for open-set face recognition," *Electronics*, vol. 14, no. 19, p. 3840, 2025.

[19] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, 2020.

[20] Z. Lei, T. Ahonen, M. Pietikäinen, and S. Z. Li, "Local frequency descriptor for low-resolution face recognition," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 161–166, IEEE, 2011.

[21] X. Xu, W.-Q. Liu, and L. Li, "Low resolution face recognition in surveillance systems," *Journal of Computer and Communications*, vol. 2, pp. 70–77, 2014.

[22] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 1034–1040, 2015.

[23] Y. Chu, T. Ahmad, G. Bebis, and L. Zhao, "Low-resolution face recognition with single sample per person," *Signal Processing*, vol. 141, pp. 144–157, 2017.

[24] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based robust low resolution face recognition," *arXiv preprint arXiv:1707.02733*, 2017.

[25] R. Thomas and M. Rangachar, "Fractional bat and multi-kernel-based spherical svm for low resolution face recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 08, p. 1756014, 2017.

[26] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 2051–2062, 2018.

[27] A. Ahmed, J. Guo, F. Ali, F. Deeba, and A. Ahmed, "Lbph based improved face recognition at low resolution," in *2018 international conference on Artificial Intelligence and big data (ICAIBD)*, pp. 144–147, IEEE, 2018.

[28] M. Wang, R. Liu, N. Hajime, A. Narishige, H. Uchida, and T. Matsunami, "Improved knowledge distillation for training fast low resolution face recognition model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[29] M. S. Shakeel, K.-M. Lam, and S.-C. Lai, "Learning sparse discriminant low-rank features for low-resolution face recognition," *Journal of Visual Communication and Image Representation*, vol. 63, p. 102590, 2019.

[30] Y. Yan, Z. Zhang, S. Chen, and H. Wang, "Low-resolution facial expression recognition: A filter learning perspective," *Signal Processing*, vol. 169, p. 107370, 2020.

[31] M. Rouhsedaghat, Y. Wang, S. Hu, S. You, and C.-C. J. Kuo, "Low-resolution face recognition in resource-constrained environments," *Pattern Recognition Letters*, vol. 149, pp. 193–199, 2021.

[32] F. Ketab, N. S. Russel, A. Selvaraj, and S. M. Buhari, "Parallel deep learning architecture with customized and learnable filters for low-resolution face recognition," *The Visual Computer*, vol. 39, no. 12, pp. 6699–6710, 2023.

[33] K. Narayan, N. G. Nair, J. Xu, R. Chellappa, and V. M. Patel, "Petalface: Parameter efficient transfer learning for low-resolution face recognition," *arXiv preprint arXiv:2412.07771*, 2024.

[34] D. A. Kurnia, O. Mohd, M. F. Abdollah, D. Sudrajat, D. M. Efendi, and S. Rahmatullah, "Digital image processing (dip) and generative adversarial networks (gans) techniques for improvement low-resolution face recognition.," *Ingénierie des Systèmes d'Information*, vol. 29, no. 6, 2024.

[35] N. M. An, H. Roh, S. Kim, J. H. Kim, and M. Im, "Machine learning techniques for simulating human psychophysical testing of low-resolution phosphene face images in artificial vision," *Advanced Science*, p. 2405789, 2025.

[36] Z. Ren, X. Liu, J. Xu, Y. Zhang, and M. Fang, "Littlefacenet: A small-sized face recognition method based on retinaface and adaface," *Journal of Imaging*, vol. 11, no. 1, p. 24, 2025.

[37] H. A. ALRIKABI, D. M. M. SIRAJ, and A. MUQDAD, "Enhancing low-resolution images through noise filtering and feature preservation," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 2, 2025.

[38] W. A. Hayder, "Enhancing face recognition dataset using generative artificial intelligence: Review study," 2025.

[39] C. Fookes, F. Lin, V. Chandran, and S. Sridharan, "Evaluation of image resolution and super-resolution on face recognition performance," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 75–93, 2012.

[40] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera," in *2013 18th International Conference on Digital Signal Processing (DSP)*, pp. 1–6, IEEE, 2013.

[41] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings 9*, pp. 175–184, Springer, 2016.

[42] T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel, "Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring," in *2016 24th Signal Processing and Communication Application Conference (SIU)*, pp. 437–440, IEEE, 2016.

[43] A. ElSayed, A. Mahmood, and T. Sobh, "Effect of super resolution on high dimensional features for unsupervised face recognition in the wild," in *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–5, IEEE, 2017.

[44] O. Kwon, "Face recognition based on super-resolution method using sparse representation and deep learning," *Journal of Korea Multimedia Society*, vol. 21, no. 2, pp. 173–180, 2018.

[45] N. Ouyang, X. Wang, X. Cai, and L. Lin, "Deep joint super-resolution and feature mapping for low resolution face recognition," in *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, pp. 849–852, IEEE, 2018.

[46] J. Cai, H. Han, S. Shan, and X. Chen, "Fcsr-gan: Joint face completion and super-resolution via multi-task learning," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 109–121, 2019.

[47] M. Wang, S. Wang, and P. Kong, "Simplified vgg based super resolution restoration for face recognition," in *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pp. 385–389, 2019.

[48] X. Chen, X. Wang, Y. Lu, W. Li, Z. Wang, and Z. Huang, "Rbpnet: An asymptotic residual back-projection network for super-resolution of very low-resolution face image," *Neurocomputing*, vol. 376, pp. 119–127, 2020.

[49] J. Chen, J. Chen, Z. Wang, C. Liang, and C.-W. Lin, "Identity-aware face super-resolution for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 645–649, 2020.

[50] S. S. Rajput and K. Arya, "A robust face super-resolution algorithm and its application in low-resolution face recognition system," *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 23909–23934, 2020.

[51] L.-Y. Xu and Z. Gajic, "Improved network for face recognition based on feature super resolution method," *International Journal of Automation and Computing*, vol. 18, no. 6, pp. 915–925, 2021.

[52] J. C. Tan, K. M. Lim, and C. P. Lee, "Enhanced alexnet with super-resolution for low-resolution face recognition," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pp. 302–306, IEEE, 2021.

[53] Z. Ullah, L. Qi, A. Hasan, and M. Asim, "Improved deep cnn-based two stream super resolution and hybrid deep model-based facial emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105486, 2022.

[54] F. Nan, W. Jing, F. Tian, J. Zhang, K.-M. Chao, Z. Hong, and Q. Zheng, "Feature super-resolution based facial expression recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, p. 107678, 2022.

[55] M. Dos Santos, R. Laroca, R. O. Ribeiro, J. Neves, H. Proença, and D. Menotti, "Face super-resolution using stochastic differential equations," in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, vol. 1, pp. 216–221, IEEE, 2022.

[56] C. Wang, J. Jiang, Z. Zhong, and X. Liu, "Propagating facial prior knowledge for multitask learning in face super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7317–7331, 2022.

[57] A. S. Tomar, K. Arya, and S. S. Rajput, "Attentive exfeat based deep generative adversarial network for noise robust face super-resolution," *Pattern Recognition Letters*, vol. 169, pp. 58–66, 2023.

[58] K. Zeng, Z. Wang, T. Lu, J. Chen, J. Wang, and Z. Xiong, "Self-attention learning network for face super-resolution," *Neural Networks*, vol. 160, pp. 164–174, 2023.

[59] G. Gao, L. Tang, F. Wu, H. Lu, and J. Yang, "Jdsr-gan: Constructing an efficient joint learning network for masked face super-resolution," *IEEE Transactions on Multimedia*, vol. 25, pp. 1505–1512, 2023.

[60] C. Wang, J. Jiang, K. Jiang, and X. Liu, "Structure prior-aware dynamic network for face super-resolution," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[61] W. Chen, W. Lin, X. Xu, L. Lin, and T. Zhao, "Face super-resolution quality assessment based on identity and recognizability," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[62] C. Wang, J. Jiang, K. Jiang, and X. Liu, "Low-light face super-resolution via illumination, structure, and texture associated representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5318–5326, 2024.

[63] M. Wang, R. Xu, W. Deng, and H. Huang, "Implicit face model: Depth super-resolution for 3d face recognition," *Pattern Recognition*, vol. 162, p. 111353, 2025.

[64] Y. Han, Q. Gu, and G. Xia, "Face-srmn: a tree-based multi-branch face super-resolution network driven by prior knowledge and recognition task," *Signal, Image and Video Processing*, vol. 19, no. 4, p. 320, 2025.

[65] A. S. Tomar, K. Arya, and S. S. Rajput, "Learning face super-resolution through identity features and distilling facial prior knowledge," *Expert Systems with Applications*, vol. 262, p. 125625, 2025.

[66] C. Han, Y. Gui, P. Cheng, and Z. You, "Semantically guided efficient attention transformer for face super-resolution tasks," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 21, no. 1, pp. 1–24, 2025.

[67] R. K. Shukla, A. S. Sengar, A. Gupta, A. Jain, A. Kumar, and N. K. Vishnoi, "Face recognition using convolutional neural network in machine learning," in *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 456–461, IEEE, 2021.

[68] P. Yaswanthram and B. Sabarish, "Face recognition using machine learning models-comparative analysis and impact of dimensionality reduction," in *2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pp. 1–4, IEEE, 2022.

[69] K. Kavita and R. S. Chhillar, "Human face recognition and age estimation with machine learning: A critical review and future perspective," *International journal of electrical and computer engineering systems*, vol. 13, no. 10, pp. 945–952, 2022.

[70] Y. Wang, X. Zhan, T. Zhan, J. Xu, and X. Bai, "Machine learning-based facial recognition for financial fraud prevention," *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 1, pp. 77–84, 2024.

[71] Y. Bai, L. Liu, K. Liu, S. Yu, Y. Shen, and D. Sun, "Non-intrusive personal thermal comfort modeling: A machine learning approach using infrared face recognition," *Building and Environment*, vol. 247, p. 111033, 2024.

[72] B. Nuthana, B. Pragathi, and V. Sreeyah, "Comparative analysis of machine learning classifiers based on fingerprint and face authentication," in *AIP Conference Proceedings*, vol. 3162, AIP Publishing, 2025.

[73] F. K. Bardak and F. Temurtaş, "Regional brain analysis and machine learning techniques for classifying familiar and unfamiliar faces using eeg," *Arabian Journal for Science and Engineering*, pp. 1–26, 2025.

[74] M. Estévez-Almenzar, R. Baeza-Yates, and C. Castillo, "A comparison of human and machine learning errors in face recognition," *arXiv preprint arXiv:2502.11337*, 2025.

[75] M. Masud, G. Muhammad, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, S. Ibrahim, and M. S. Hossain, "Deep learning-based intelligent face recognition in iot-cloud environment," *Computer Communications*, vol. 152, pp. 215–222, 2020.

[76] K. Teoh, R. Ismail, S. Naziri, R. Hussin, M. Isa, and M. Basir, "Face recognition and identification using deep learning approach," in *Journal of Physics: Conference Series*, vol. 1755, p. 012006, IOP Publishing, 2021.

[77] L. Boussaad and A. Boucetta, "Deep-learning based descriptors in application to aging problem in face recognition," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2975–2981, 2022.

[78] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106621, 2022.

[79] Z.-Y. Deng, H.-H. Chiang, L.-W. Kang, and H.-C. Li, "A lightweight deep learning model for real-time face recognition," *IET Image Processing*, vol. 17, no. 13, pp. 3869–3883, 2023.

[80] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. K. Ratha, "In-group bias in deep learning-based face recognition models due to ethnicity and age," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 54–67, 2023.

[81] H. M. Zangana and F. M. Mustafa, "Review of hybrid denoising approaches in face recognition: Bridging wavelet transform and deep learning," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, 2024.

[82] M. Rostami, A. Farajollahi, and H. Parvin, "Deep learning-based face detection and recognition on drones," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 1, pp. 373–387, 2024.

[83] W. R. Humood, "Utilizing image processing to identify rare data in the context of the face recognition based on deep learning techniques," in *AIP Conference Proceedings*, vol. 3264, AIP Publishing, 2025.

[84] H. E. Polat, D. G. Koc, Ö. Ertuğrul, C. Koç, and K. Ekinci, "Deep learning based individual cattle face recognition using data augmentation and transfer learning," *Journal of Agricultural Sciences*, vol. 31, no. 1, pp. 137–150, 2025.

[85] A. Nemavhola, S. Viriri, and C. Chibaya, "A scoping review of literature on deep learning techniques for face recognition," *Human Behavior and Emerging Technologies*, vol. 2025, no. 1, p. 5979728, 2025.

[86] M. Bessaoudi, A. Chouchane, A. Ouamane, and E. Boutellaa, "Multilinear subspace learning using handcrafted and deep features for face kinship verification in the wild," *Applied Intelligence*, vol. 51, pp. 3534–3547, 2021.

[87] I. Serraoui, O. Laiadi, A. Ouamane, F. Dornaika, and A. Taleb-Ahmed, "Knowledge-based tensor subspace analysis system for kinship verification," *Neural Networks*, vol. 151, pp. 222–237, 2022.

[88] N. Chenni, M. Touami, and R. Aliradi, "Facial pain detection using multilinear subspace learning," 2023.

[89] A. A. Gharbi, A. Chouchane, A. Ouamane, E. O. Belabbaci, Y. Himeur, and S. Bourennane, "A hybrid multilinear-linear subspace learning approach for enhanced person re-identification in camera networks," *Expert Systems with Applications*, vol. 257, p. 125044, 2024.

[90] A. Chouchane, M. Bessaoudi, H. Kheddar, A. Ouamane, T. Vieira, and M. Hassaballah, "Multilinear subspace learning for person re-identification based fusion of high order tensor features," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107521, 2024.

[91] R. Aliradi, N. Chenni, M. Touami, and A. Ouamane, "Facial pain detection using multilinear subspace learning," 2024.

[92] R. Aliradi, N. Chenni, and M. Touami, "Estimation for pain from facial expression based on xqeda and deep learning," *International Journal of Information Technology*, vol. 17, no. 1, pp. 655–663, 2025.

[93] T. Sun, H. Lang, X. Fang, and L. Xie, "Enhanced multilinear pca for efficient image analysis and dimensionality reduction: Unlocking the potential of complex image data," *Mathematics*, vol. 13, no. 3, p. 531, 2025.

[94] J. Wang, T. Deng, M. Yang, and J. Wang, "Embedded multi-view clustering via collaborative tensor subspace representation and multi-graph fusion," *IEEE Signal Processing Letters*, 2025.

[95] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *2021 IEEE international joint conference on biometrics (IJCB)*, pp. 1–8, IEEE, 2021.

[96] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "Vitfer: facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022.

[97] Z. Sun and G. Tzimiropoulos, "Part-based face recognition with vision transformers," *arXiv preprint arXiv:2212.00057*, 2022.

[98] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Face anti-spoofing using transformers with relation-aware mechanism," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 439–450, 2022.

[99] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang, "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," in *European conference on computer vision*, pp. 37–54, Springer, 2022.

[100] A. Liu, Z. Tan, Z. Yu, C. Zhao, J. Wan, Y. Liang, Z. Lei, D. Zhang, S. Z. Li, and G. Guo, "Fm-vit: Flexible modal vision transformers for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4775–4786, 2023.

[101] S. Nixon, P. Ruiu, M. Cadoni, A. Lagorio, and M. Tistarelli, "Exploiting face recognizability with early exit vision transformers," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, IEEE, 2023.

[102] A. Liu and Y. Liang, "Ma-vit: Modality-agnostic vision transformers for face anti-spoofing," *arXiv preprint arXiv:2304.07549*, 2023.

[103] L. Ouannes, A. B. Khalifa, and N. E. B. Amara, "Enhancing face recognition in degraded conditions via vision transformer," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pp. 2887–2892, IEEE, 2024.

[104] M. Uddin, Z. Fu, and X. Zhang, "Deepfake face detection via multi-level discrete wavelet transform and vision transformer," *The Visual Computer*, pp. 1–13, 2025.

[105] O. Elharrouss, Y. Himeur, Y. Mahmood, S. Alrabaee, A. Ouamane, F. Bensaali, Y. Bechqito, and A. Chouchane, "Vits as backbones: Leveraging vision transformers for feature extraction," *Information Fusion*, p. 102951, 2025.

[106] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[107] Y. Himeur, N. Aburaed, O. Elharrouss, I. Varlamis, S. Atalla, W. Mansoor, and H. Al Ahmad, "Applications of knowledge distillation in remote sensing: A survey," *Information Fusion*, p. 102742, 2024.

[108] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[109] A. Ouamane, M. Belahcène, and S. Bourennane, "Multimodal 3d and 2d face authentication approach using extended lbp and statistic local features proposed," in *European Workshop on Visual Information Processing (EUVIP)*, pp. 130–135, 2013.

[110] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019.

[111] A. Chouchane, M. Bessaoudi, E. Boutellaa, and A. Ouamane, "A new multidimensional discriminant representation for robust person re-identification," *Pattern Analysis and Applications*, pp. 1–14, 2023.

[112] B. El Ouanas, M. Khammari, A. Chouchane, A. Ouamane, M. Bessaoudi, Y. Himeur, and M. Hassaballah, "High-order knowledge-based discriminant features for kinship verification," *Pattern Recognition Letters*, vol. 175, pp. 30–37, 2023.

[113] D. G. Lema, R. Usamentiaga, and D. F. García, "Low-cost system for real-time verification of personal protective equipment in industrial facilities using edge computing devices," *Journal of Real-Time Image Processing*, vol. 20, no. 6, pp. 1–18, 2023.

[114] Y. Himeur, S. Al-Maadeed, H. Kheddar, N. Al-Maadeed, K. Abualsaud, A. Mohamed, and T. Khattab, "Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105698, 2023.

[115] Y. Himeur, S. Al-Maadeed, N. Almaadeed, K. Abualsaud, A. Mohamed, T. Khattab, and O. Elharrouss, "Deep visual social distancing monitoring to combat covid-19: A comprehensive survey," *Sustainable cities and society*, vol. 85, p. 104064, 2022.

[116] Z. Xiushan, "High-resolution face recognition and edf image reconstruction application in english classroom teaching," *International Journal of System Assurance Engineering and Management*, pp. 1–11, 2023.

[117] H. Dastmalchi and H. Aghaeinia, "Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network," *Signal Processing: Image Communication*, vol. 107, p. 116755, 2022.

[118] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European conference on computer vision*, pp. 318–333, Springer, 2016.

[119] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697, 2017.

[120] S.-C. Lai, C.-H. He, and K.-M. Lam, "Low-resolution face recognition based on identity-preserved face hallucination," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1173–1177, IEEE, 2019.

[121] J. Kim, G. Li, I. Yun, C. Jung, and J. Kim, "Edge and identity preserving network for face super-resolution," *Neurocomputing*, vol. 446, pp. 11–22, 2021.

[122] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018.

[123] B. Lee, K. Ko, J. Hong, and H. Ko, "Hard sample-aware consistency for low-resolution facial expression recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 199–208, 2024.

[124] M. Knoche, M. Elkadeem, S. Hörmann, and G. Rigoll, "Octuplet loss: Make face recognition robust to image resolution," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, IEEE, 2023.

[125] X. Xu, H. Deng, Y. Wang, S. Zhang, and H. Song, "Boosting cattle face recognition under uncontrolled scenes by embedding enhancement and optimization," *Applied Soft Computing*, vol. 164, p. 111951, 2024.

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

[127] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "A survey of multilinear subspace learning for tensor data," *Pattern Recognition*, vol. 44, no. 7, pp. 1540–1551, 2011.

[128] A. Chouchane, A. Ouamane, E. Boutellaa, M. Belahcene, and S. Bourennane, "3d face verification across pose based on euler rotation and tensors," *Multimedia Tools and Applications*, vol. 77, pp. 20697–20714, 2018.

[129] O. Guehairia, F. Dornaika, A. Ouamane, and A. Taleb-Ahmed, "Facial age estimation using tensor based subspace learning and deep random forests," *Information Sciences*, vol. 609, pp. 1309–1317, 2022.

[130] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pp. 469–481, Springer, 2004.

[131] Y.-H. Huang and H. H. Chen, "Deep face recognition for dim images," *Pattern Recognition*, vol. 126, p. 108580, 2022.

[132] J. Pei, Y. Chen, Y. Zhao, C. Wang, and X. Yang, "Self-adjusting multilayer nonlinear coupled mapping for low-resolution face recognition," *Applied Soft Computing*, vol. 129, p. 109566, 2022.

[133] X. Liu, Y. Li, M. Gu, H. Zhang, X. Zhang, J. Wang, X. Lv, and H. Deng, "Swindpsr: Dual-path face super-resolution network integrating swin transformer," *Symmetry*, vol. 16, no. 5, p. 511, 2024.

[134] W. Li, H. Guo, X. Liu, K. Liang, J. Hu, Z. Ma, and J. Guo, "Efficient face super-resolution via wavelet-based feature enhancement network," *arXiv preprint arXiv:2407.19768*, 2024.

[135] Y. T. Tsai, Y. W. Chen, H.-H. Shuai, and C.-C. Huang, "Arbitrary-resolution and arbitrary-scale face super-resolution with implicit representation networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4270–4279, 2024.

[136] S. Banerjee and S. Das, "Lr-gan for degraded face recognition," *Pattern Recognition Letters*, vol. 116, pp. 246–253, 2018.

[137] M. Zhang and Q. Ling, "Supervised pixel-wise gan for face super-resolution," *IEEE Transactions on Multimedia*, vol. 23, pp. 1938–1950, 2020.

[138] T. Varanka, T. Toivonen, S. Tripathy, G. Zhao, and E. Acar, "Pfstorer: Personalized face restoration and super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2372–2381, 2024.

[139] S. Bellili, A. Ouamane, A. Chouchane, Y. Himeur, S. Atalla, W. Mansoor, S. Bourennane, and F. Bensaali, "Tensor-driven face recognition: Integrating super-resolution and multilinear subspace learning for low-resolution images," *Information Fusion*, vol. 120, p. 103075, 2025.

[140] M. Bessaoudi, M. Belahcene, A. Ouamane, A. Chouchane, and S. Bourennane, "A novel hybrid approach for 3d face recognition based on higher order tensor," in *Advances in Computing Systems and Applications* (O. Demigha, B. Djamaa, and A. Amamra, eds.), (Cham), pp. 215–224, Springer International Publishing, 2019.

[141] M. Belahcene, M. Laid, A. Chouchane, A. Ouamane, and S. Bourennane, "Local descriptors and tensor local preserving projection in face recognition," in *2016 6th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, 2016.

[142] 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*, pp. 770–778, 2016.

[143] E. Boutellaa, M. B. López, S. Ait-Aoudia, X. Feng, and A. Hadid, "Kinship verification from videos using spatio-temporal texture features and deep learning," *arXiv preprint arXiv:1708.04069*, 2017.

[144] J. Xu, B. Liu, Y. Xiao, F. Cao, and J. He, "A multitask tensor-based relation network for cloth-changing person re-identification," *Image and Vision Computing*, vol. 147, p. 105090, 2024.

[145] A. Chouchane, A. Ouamane, Y. Himeur, W. Mansoor, S. Atalla, A. Benzaibak, and C. Boudellal, "Improving cnn-based person re-identification using score normalization," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2890–2894, IEEE, 2023.

[146] Z. Zhao, Z. Cao, H. Xin, R. Wang, D. Wu, Z. Wang, and F. Nie, "Enhancing clustering performance with tensorized high-order bipartite graphs: A structured graph learning approach," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[147] A. Chouchane, M. Bessaoudi, and A. Ouamane, "Facial kinship verification using tensor discriminative subspace based color components," in *2019 6th International Conference on Image and Signal Processing and their Applications (ISPA)*, pp. 1–6, 2019.

[148] M. Bessaoudi, A. Ouamane, M. Belahcene, A. Chouchane, E. Boutellaa, and S. Bourennane, "Multilinear side-information based discriminant analysis for face and kinship verification in the wild," *Neurocomputing*, vol. 329, pp. 267–278, 2019.

[149] O. Laiadi, A. Ouamane, A. Benakcha, A. Taleb-Ahmed, and A. Hadid, "Kinship verification based deep and tensor features through extreme learning machine," in *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pp. 1–4, 2019.

[150] M. Bessaoudi, M. Belahcene, A. Ouamane, and S. Bourennane, "A novel approach based on high order tensor and multi-scale locals features for 3d face recognition," in *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–5, 2018.

[151] A. Ouamane, A. Chouchane, E. Boutellaa, M. Belahcene, S. Bourennane, and A. Hadid, "Efficient tensor-based 2d+3d face verification," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2751–2762, 2017.

[152] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Computer Vision – ACCV 2010* (R. Kimmel, R. Klette, and A. Sugimoto, eds.), (Berlin, Heidelberg), pp. 709–720, Springer Berlin Heidelberg, 2011.

[153] E. O. Belabbaci, M. Khammari, A. Chouchane, A. Ouamane, M. Bessaoudi, Y. Himeur, and M. Hassaballah, "High-order knowledge-based discriminant features for kinship verification," *Pattern Recognition Letters*, vol. 175, pp. 30–37, 2023.

[154] M. Belahcéne, A. Ouamane, and A. T. Ahmed, "Fusion by combination of scores multi-biometric systems," in *3rd European Workshop on Visual Information Processing*, pp. 252–257, 2011.

[155] 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.

[156] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[157] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv:1804.09691*, 2018.

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

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

[160] J. Li, Y. Zhou, J. Ding, C. Chen, and X. Yang, "Id preserving face super-resolution generative adversarial networks," *IEEE Access*, vol. 8, pp. 138373–138381, 2020.

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

[162] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 499–515, Springer, 2016.

[163] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[164] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[165] 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*, pp. 212–220, 2017.

[166] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," *Advances in neural information processing systems*, vol. 31, 2018.

[167] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.

[168] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song, "Deep hyperspherical learning," *Advances in neural information processing systems*, vol. 30, 2017.

[169] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 605–621, Springer, 2019.

[170] Y. Martínez-Díaz, L. S. Luevano, and H. Méndez-Vázquez, "Effectiveness of blind face restoration to boost face recognition performance at low-resolution images," in *International Workshop on Artificial Intelligence and Pattern Recognition*, pp. 455–467, Springer, 2023.

[171] 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*, pp. 0–0, 2019.

[172] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pp. 428–438, Springer, 2018.

[173] 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*, pp. 4690–4699, 2019.

[174] F. Perez-Montes, J. Olivares-Mercado, G. Sanchez-Perez, G. Benitez-Garcia, L. Prudente-Tixteco, and O. Lopez-Garcia, "Analysis of real-time face-verification methods for surveillance applications," *Journal of Imaging*, vol. 9, no. 2, p. 21, 2023.

[175] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[176] 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*, pp. 1580–1589, 2020.

[177] T. M. Mufid, R. I. Adam, J. K. Jaman, G. Garno, I. Maulana, *et al.*, "Implementation of identity loss function on face recognition of low-resolution faces with light cnn architecture," *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 91–97, 2024.

[178] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[179] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pp. 184–199, Springer, 2014.

[180] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[181] F. Han, X. Wang, F. Shen, and J. Zhao, "C-face: Using compare face on face hallucination for low-resolution face recognition," *Journal of Artificial Intelligence Research*, vol. 74, pp. 1715–1737, 2022.

[182] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

[183] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[184] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, 2018.

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

[186] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3867–3876, 2019.

[187] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 672–681, 2021.

[188] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178, 2021.

[189] S. Zhou, K. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30599–30611, 2022.

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

[191] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2492–2501, 2018.

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

[193] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi, "Facial deblur inference using subspace analysis for recognition of blurred faces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 838–845, 2010.

[194] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *International journal of computer vision*, vol. 43, pp. 167–188, 2001.

[195] C. Singh, N. Mittal, and E. Walia, "Face recognition using zernike and complex zernike moment features," *Pattern Recognition and Image Analysis*, vol. 21, pp. 71–81, 2011.

[196] J. Qian, J. Yang, and Y. Xu, "Local structure-based image decomposition for feature extraction with applications to face recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3591–3603, 2013.

[197] N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 2, pp. 295–304, 2012.

[198] X. Ben, W. Meng, R. Yan, and K. Wang, "Kernel coupled distance metric learning for gait recognition and face recognition," *Neurocomputing*, vol. 120, pp. 577–589, 2013.

[199] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2017.

[200] 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*, pp. 5265–5274, 2018.

[201] 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*, pp. 14225–14234, 2021.

[202] K. Juneja and C. Rana, "An extensive study on traditional-to-recent transformation on face recognition system," *Wireless Personal Communications*, vol. 118, pp. 3075–3128, 2021.

[203] 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*, vol. 9, pp. 99112–99142, 2021.

[204] B. Kocacinar, B. Tas, F. P. Akbulut, C. Catal, and D. Mishra, "A real-time cnn-based lightweight mobile masked face recognition system," *Ieee Access*, vol. 10, pp. 63496–63507, 2022.

[205] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and F. Khelifi, "Pose-invariant face recognition with multitask cascade networks," *Neural Computing and Applications*, pp. 1–14, 2022.

[206] S. Alsubai, M. Hamdi, S. Abdel-Khalek, A. Alqahtani, A. Binbusayyis, and R. F. Mansour, "Bald eagle search optimization with deep transfer learning enabled age-invariant face recognition model," *Image and Vision Computing*, vol. 126, p. 104545, 2022.

[207] M. J. Khan, M. J. Khan, A. M. Siddiqui, and K. Khurshid, "An automated and efficient convolutional architecture for disguise-invariant face recognition using noise-based data augmentation and deep transfer learning," *The Visual Computer*, pp. 1–15, 2022.

[208] T. Lu, X. Chen, Y. Zhang, C. Chen, and Z. Xiong, "Slr: Semi-coupled locality constrained representation for very low resolution face recognition and super resolution," *IEEE Access*, vol. 6, pp. 56269–56281, 2018.

[209] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE transactions on image processing*, vol. 12, no. 5, pp. 597–606, 2003.

[210] F. W. Wheeler, X. Liu, and P. H. Tu, "Multi-frame super-resolution for face recognition," in *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pp. 1–6, IEEE, 2007.

[211] H. Huang and H. He, "Super-resolution method for face recognition using nonlinear mappings on coherent features," *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 121–130, 2010.

[212] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4220–4231, 2014.

[213] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," *Optik*, vol. 126, no. 23, pp. 4352–4357, 2015.

[214] E. Zangeneh, M. Rahmati, and Y. Mohsenzadeh, "Low resolution face recognition using a two-branch deep convolutional neural network architecture," *Expert Systems with Applications*, vol. 139, p. 112854, 2020.

[215] G. Gao, Y. Yu, M. Yang, P. Huang, Q. Ge, and D. Yue, "Multi-scale patch based representation feature learning for low-resolution face recognition," *Applied Soft Computing*, vol. 90, p. 106183, 2020.

[216] M. B. Shahbakhsh and H. Hassanpour, "Empowering face recognition methods using a gan-based single image super-resolution network," *International Journal of Engineering*, vol. 35, no. 10, pp. 1858–1866, 2022.

[217] P. Zhao, Y. Ming, X. Meng, and H. Yu, "Lmfnet: A lightweight multiscale fusion network with hierarchical structure for low-quality 3-d face recognition," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 239–252, 2022.

[218] A. Ouamane, A. Benakcha, M. Belahcene, and A. Taleb-Ahmed, "Multimodal depth and intensity face verification approach using lbp, slf, bsif, and lpq local features fusion," *Pattern Recognition and Image Analysis*, vol. 25, pp. 603–620, 2015.

[219] A. Ouamane, M. Belahcène, and S. Bourennane, "Multimodal 3d and 2d face authentication approach using extended lbp and statistic local features proposed," in *European Workshop on Visual Information Processing (EUVIP)*, pp. 130–135, IEEE, 2013.

[220] M. Belahcene, M. Laid, A. Chouchane, A. Ouamane, and S. Bourennane, "Local descriptors and tensor local preserving projection in face recognition," in *2016 6th European workshop on visual information processing (EUVIP)*, pp. 1–6, IEEE, 2016.

[221] A. Chouchane, M. Bessaoudi, and A. Ouamane, "Facial kinship verification using tensor discriminative subspace based color components," in *2019 6th International Conference on Image and Signal Processing and their Applications (ISPA)*, pp. 1–6, IEEE, 2019.

[222] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: a complete review," *Multimedia tools and applications*, vol. 80, pp. 4825–4880, 2021.

[223] S. Bellili, A. Ouamane, A. Chouchane, Y. Himeur, S. Atalla, W. Mansoor, and H. Al Ahmad, "Very low-resolution face recognition based on multilinear side-information based discriminant analysis," in *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pp. 104–108, IEEE, 2024.

[224] O. Guehairia, F. Dornaika, A. Ouamane, and A. Taleb-Ahmed, "Facial age estimation using tensor based subspace learning and deep random forests," *Information Sciences*, vol. 609, pp. 1309–1317, 2022.

[225] R. Shi, W. Guo, and S. Ge, "Low-resolution face recognition via adaptable instance-relation distillation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2024.

[226] N. R. Pottanigari, R. Pullela, A. K. A. Shaik, and R. Reddy, "Efficient detection of disguised faces using photos/sketches from low-quality surveillance footage," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9, IEEE, 2024.

[227] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "Real-world super-resolution of face-images from surveillance cameras," *IET Image Processing*, vol. 16, no. 2, pp. 442–452, 2022.

[228] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need [j]," *Advances in neural information processing systems*, vol. 30, no. 1, pp. 261–272, 2017.

[229] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[230] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[231] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic detr: End-to-end object detection with dynamic attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2988–2997, 2021.

[232] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106669, 2023.

[233] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[234] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

[235] W. Chen, X. Du, F. Yang, L. Beyer, X. Zhai, T.-Y. Lin, H. Chen, J. Li, X. Song, Z. Wang, *et al.*, "A simple single-scale vision transformer for object detection and instance segmentation," in *European conference on computer vision*, pp. 711–727, Springer, 2022.

[236] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Asian conference on computer vision*, pp. 605–621, Springer, 2018.

[237] Y. Tai, H. Shi, D. Zeng, H. Du, Y. Hu, Z. Zhang, Z. Zhang, and T. Mei, "Multi-agent semi-siamese training for long-tail and shallow face learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–20, 2023.

[238] K. Ahn, S. Lee, S. Han, C. Y. Low, and M. Cha, "Uncertainty-aware face embedding with contrastive learning for open-set evaluation," *IEEE Transactions on Information Forensics and Security*, 2024.

[239] 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*, pp. 18750–18759, 2022.

[240] H. He, W. Yuan, Y. Huang, S. Zhao, W. Yuan, and H. Li, "Learning unified representations for multi-resolution face recognition," *arXiv preprint arXiv:2310.09563*, 2023.

[241] H. Shi, H. Du, Y. Hu, J. Wang, D. Zeng, and T. Yao, "Scale attention for learning deep face representation: A study against visual scale variation," *arXiv preprint arXiv:2209.08788*, 2022.

[242] 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*, pp. 815–823, 2015.

[243] X. Yin, Y. Tai, Y. Huang, and X. Liu, "FAN: feature adaptation network for surveillance face recognition and normalization," *CoRR*, vol. abs/1911.11680, 2019.

[244] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3147–3155, 2017.