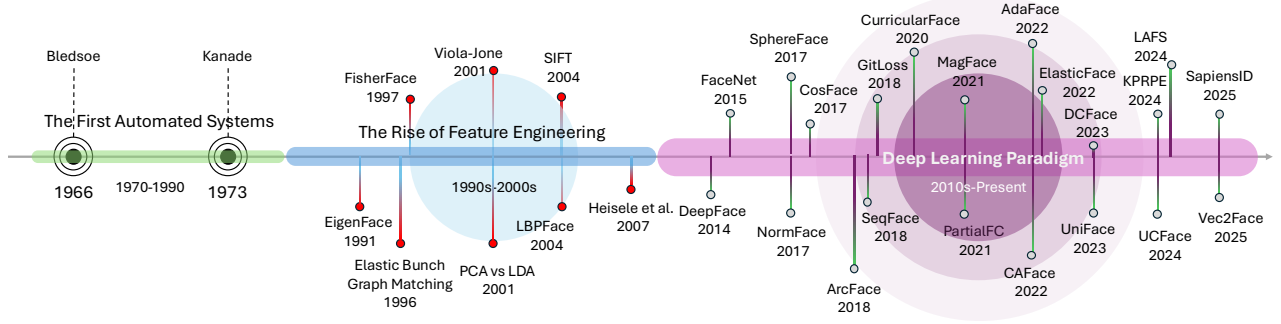# 50 Years of Automated Face Recognition

Minchul Kim, Anil Jain, Xiaoming Liu
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI, 48824
{kimminc2, jain, liuxm}@cse.msu.edu

Fig. 1: Timeline of face recognition evolution, tracing the transition from hand-crafted features to deep learning-based approaches. Circle sizes reflect annual publication volume for each era.

Abstract—Over the past 50 years, automated face recognition has evolved from rudimentary, handcrafted systems into sophisticated deep learning models that rival and often surpass human performance. This paper chronicles the history and technological progression of FR, from early geometric and statistical methods to modern deep neural architectures leveraging massive real and AI-generated datasets. We examine key innovations that have shaped the field, including developments in dataset, loss function, neural network design and feature fusion. We also analyze how the scale and diversity of training data influence model generalization, drawing connections between dataset growth and benchmark improvements. Recent advances have achieved remarkable milestones: state-of-the-art face verification systems now report False Negative Identification Rates of 0.13% against a 12.4 million gallery in NIST FRVT evaluations for 1:N visa-to-border matching. While recent advances have enabled remarkable accuracy in high- and low-quality face scenarios, numerous challenges persist. While remarkable progress has been achieved, several open research problems remain. We outline critical challenges and promising directions for future face recognition research, including scalability, multi-modal fusion, synthetic identity generation, and explainable systems.

Index Terms—Face recognition, biometrics, computer vision, deep learning, synthetic data, loss function

1 Introduction

For half a century, the dream of machines 'seeing' and recognizing faces has captivated researchers and fueled imaginations, leaping from the realm of science fiction to become a pervasive reality. What began as a computationally intractable problem, requiring painstaking manual feature engineering, has blossomed into a cornerstone of modern security, convenience, and even social interaction. However, this rapid ascent has not been without its complexities. The journey from Kanade's pioneering work [1] to today's deep learning behemoths reveals not just a story of algorithmic innovation, but a shifting landscape of ethical considerations, data dependencies, and the ever-present challenge of defining 'identity' itself. This paper chronicles that 50-year evolution, examining the pivotal breakthroughs, the persistent hurdles, and the emerging frontiers that will shape the future of automated face recognition (FR).

FR has become one of the most prevalent biometric modalities employed today, largely mirroring how humans themselves identify each other [2], [3]. See Fig. 2 for representative applications. Several factors contribute to this widespread adoption. Faces can be identified at a distance, offering a non-contact and less intrusive method compared to other biometrics like fingerprints or iris [4]. Face acquisition can be achieved using low-cost cameras, making it accessible and scalable across diverse applications [3], [4]. The non-contact nature of FR offers hygienic advantages, especially salient in a post-pandemic era [5], [6]. Furthermore, face recognition can be performed covertly using ubiquitous surveillance cameras, and benefits from the existence of extensive legacy databases containing facial images such as passports, visas, mugshots and driver's licenses [3].

Notably, even prior to deep model based FR, automated face recognition systems demonstrated the potential to surpass human capabilities in certain scenarios. Studies conducted in 2007 indicate that algorithms could outperform average human performance in matching face pairs, particularly in simpler cases [7]. Further research in 2010 reveals that a specific algorithm exceeded the accuracy of thousands
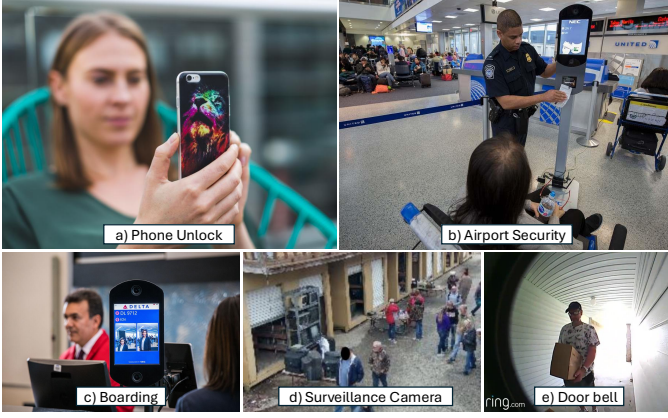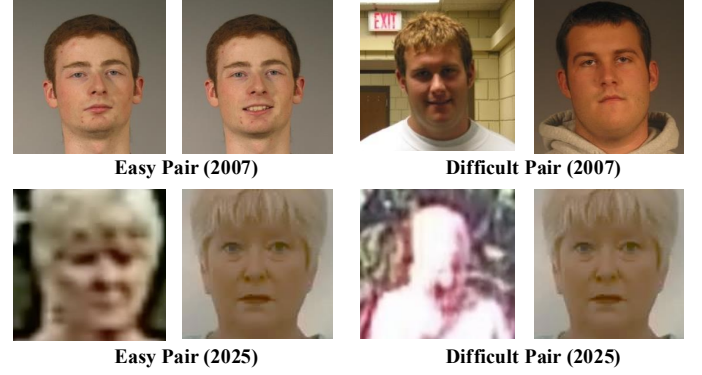
Fig. 2: Examples of real-world face recognition (FR) applications: (a) cellphone unlocking via facial authentication, (b) identity verification at airport security checkpoints, (c) FR for boarding pass verification, (d) public surveillance with facial analysis, and (e) smart doorbells employing FR for home security. These use cases highlight the ubiquity and versatility of FR systems across personal, commercial, and governmental domains.

Fig. 3: Visualization of easy and difficult face pairs for algorithms in 2007 (top) and 2025 (bottom), where difficulty is defined by the pairs that State-of-the-Art (SoTA) models of the time struggle to correctly identify [7], [31].

of customs inspectors when dealing with straightforward facial comparisons in operational settings [8]. Some examples of challenging pairs are given in Fig. 3. While these early successes are significant, the field has undergone a dramatic transformation since then, fueled by advancements in deep networks and the availability of large-scale datasets. This paper will revisit this critical juncture, exploring how the field has evolved and the extent to which current face recognition technology has surpassed human abilities across a wider range of challenging conditions.

The progress in FR has been driven by multiple factors: advances in computing power, the increasing availability of large-scale datasets, and a fundamental shift in our understanding of how to represent and compare facial imagery. Early approaches rely on handcrafted features designed to capture specific aspects of facial structure, such as holistic approaches of Eigenfaces [9] and neighborhood approaches like Local Binary Patterns (LBP) [10]. The introduction of machine learning techniques, including statistical models and cascaded classifiers [11], further refined the accuracy and robustness of FR systems. However, the advent of deep learning marks a paradigm shift, enabling algorithms to learn discriminative facial features directly from a large collection of representative training data, as demonstrated by seminal works such as DeepFace [12], DeepID [13], and FaceNet [14]. These advancements lead to unprecedented improvements in FR performance (both accuracy and efficiency) but also introduce new challenges, including concerns about bias in training data, achieving higher accuracies even for low-quality face images in presence of pose, expression and illumination variations and occlusion.

This paper will trace the historical development of face recognition techniques, beginning with the foundational work in feature extraction and pattern matching, progressing through the statistical methods that dominated the field for decades, and culminating in the transformative impact of deep learning. We will focus on key innovations in network architecture [15], [16], loss function design [17]–[22], and the utilization of increasingly large and diverse datasets [23]–[30]. Furthermore, we will address the emerging role of synthetic data generation as a means to overcome data limitations and mitigate privacy concerns.

Finally, we will discuss the remaining challenges, including adaptation to low-quality images, surpassing human recognition capability, multimodal fusion (such as face and gait), and enhancing the interpretability of complex deep learning models. By providing a comprehensive overview of the field's past, present, and future, this paper aims to inform both researchers and practitioners and to stimulate further innovation in this critical area of computer vision and artificial intelligence.

While several valuable surveys [32]–[35] on FR have emerged in recent years, often providing detailed catalogs of contemporary techniques or in-depth explorations of particular sub-domains, this paper offers a distinct perspective; our work spans the full 50-year evolution of the field, providing a comprehensive historical narrative that contextualizes the current SoTA within its broader trajectory. Recent surveys have also addressed specialized topics within the broader face analysis domain, such as 3D face recognition [36], demographic bias [37] and face anti-spoofing [38], offering deep dives into these critical subfields. In contrast, our survey maintains a strict focus on face recognition (FR) itself, distinct from related tasks such as expression recognition, age estimation, spoof detection, and deepfake identification, which are not covered here. We hope the readers can find insight in connecting the field's historical evolution to its present state and outlining the critical challenges and emerging paradigms that will shape its future.

2 FACE RECOGNITION FRAMEWORK

"This recognition problem is made difficult by the great variability in head rotation and tilt, lighting intensity and angle, facial expression, aging, etc."
— Bledsoe, Chan and Bisson (1966)

A modern face recognition system fundamentally operates in two distinct phases: *enrollment* and *recognition*[1]. The

---

1. We use *comparison* and *recognition* interchangeably. Recognition involves applying a threshold to a comparison (similarity) score.
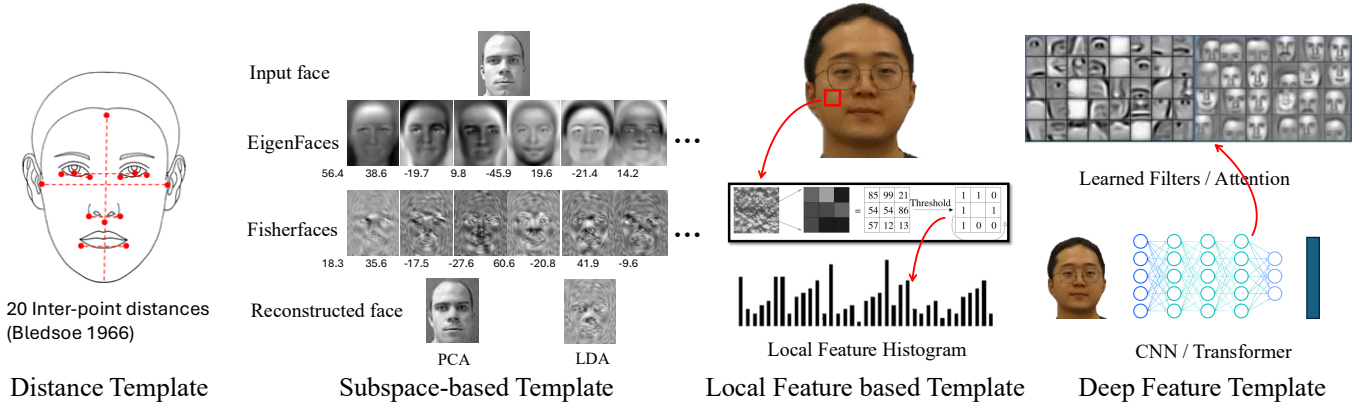
Fig. 4: The evolution of face templates over time. Early systems relied on geometric distances between landmarks and linear subspaces (*e.g.*, PCA, LDA), followed by local texture-based features like LBP. The modern FR era leverages deep learning, where CNNs and Transformers learn highly discriminative feature embeddings directly from training data.
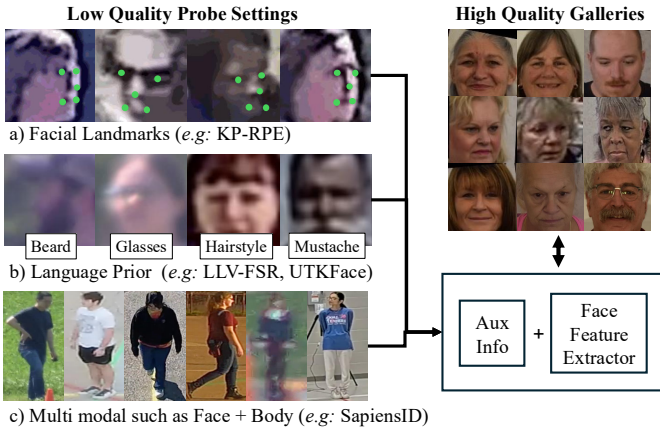
Fig. 5: Illustration of template enhancement by incorporating auxiliary information such as facial landmarks (*e.g.*, KP-RPE [39]), language priors (*e.g.*, LLV-FSR [40]), and multi-modal cues (*e.g.*, SapiensID [41]).

enrollment stage establishes a baseline by acquiring and storing facial data, while the recognition stage leverages this stored information to either confirm a claimed identity or determine an unknown identity. This process can be broadly categorized into two primary tasks; **Verification**: a one-to-one comparison confirming if a presented face matches a specific individual. **Identification**: a one-to-many search to determine the identity of a face from a gallery of known individuals.

While identification serves the broader goal of determining who someone is, it unfolds in two distinct forms: **closed retrieval** and **open search**. *Closed Retrieval* assumes the probe image belongs to a known individual, ranking potential matches within a predefined set, a method long relied upon in forensic investigations and archival systems. *Open Search*, on the other hand, acknowledges the unknown, forcing the system to not only rank candidates but also reject impostors when necessary. This distinction, subtle yet profound, underpins the challenge of building FR systems that are both inclusive and discerning, ensuring that recognition is not merely about finding similarities but also knowing when to say, "this face does not belong in a dataset

of interest, *e.g.* a watchlist." As size of the face databases continue to grow, FR algorithms need to be scaled for higher accuracy and speed. The largest known face database is reported to have 50 billion faces [42].

The enrollment process begins with capturing a digital representation of a face (the 'gallery' or 'target' image). This raw data is then processed through a series of steps, beginning with quality assessment to ensure reliability. A crucial component is the *feature extraction* stage, where salient characteristics are distilled from the facial image, creating a compact and informative 'template'. This template, rather than the raw image itself, is stored in a database for efficient comparison. Fig. 4 summarizes the evolution of face templates. In some applications, the face image is also stored in addition to the template for manual adjudication. Furthermore, as depicted in Fig. 5, auxiliary information such as facial landmarks, language priors and multi-modal face-body cues can be integrated to enrich the template. During recognition, a feature set is extracted from the input face and compared against the stored templates using a *matching function* to generate a similarity score. A decision (acceptance or rejection for verification, ranking for identification) is then made based on this score.

However, achieving robust and accurate face recognition is inherently challenging. The appearance of a face is remarkably variable, influenced by a multitude of factors. These *intra-class or intra-person* variations encompass changes in lighting conditions, head pose, facial expression, age progression, and the presence of occlusions such as glasses, hats, or masks [26], [28]–[30]. Variations in image quality (*e.g.*, resolution, blur, and noise) further exacerbate the problem [23]–[25]. Early face recognition systems often struggled to address these challenges, necessitating carefully controlled imaging environments with constrained pose and illumination conditions [9], [10].

Yet, intra-class variations cannot be examined in isolation. FR systems must also achieve high variance for different subjects. This means contending with inter-class similarities, even when different individuals exhibit highly similar facial features. This includes biological cases such as identical twins and familial resemblances, where genetic similarities result in closely matching facial structures [43].

Moreover, non-related individuals may coincidentally look alike (so-called doppelgangers) further complicating the discrimination task [44]. Both intra-class variation and inter-class similarity must be jointly addressed to design FR systems that are both robust and discriminative.

Modern FR systems strive to achieve invariance to intra-class variations and variance to inter-class differences [13], [14], [17], [19]–[22]. This has been achieved through increasingly sophisticated algorithms, moving from hand-engineered features to learned representations via machine learning and, most recently, deep learning. The ability to effectively manage these sources of variation remains a central focus of ongoing research, driving the development of more resilient and reliable FR technologies. The following sections will detail the evolution of techniques used to address these challenges, from the earliest approaches to the cutting-edge methods employed today.

The human face encodes a wide spectrum of information. As illustrated in Fig. 6, a single image can reveal identity, demographic traits, physical attributes, and social cues. However, deep learning-based FR systems typically do not treat these aspects independently. Instead, they amalgamate all visible cues—be it scars, expressions, or age—into a compact, high-dimensional embedding. While this approach has driven significant performance gains, it often comes at the cost of interpretability. The resulting features are highly discriminative but opaque, making it difficult to disentangle what specific attributes are contributing to a match decision. As FR systems become more pervasive, improving the transparency and explainability of these learned representations remains an important area of ongoing research.

3 HISTORY OF FACE RECOGNITION

Early Precursors. The history of face recognition is intertwined with the broader need for reliable personal identification, initially driven by law enforcement and security concerns. The enactment of the Habitual Criminals Act in 1869 in the UK marked an early attempt to formalize the identification of repeat offenders [45]. This period also saw the rise of fingerprinting, with pioneers like Henry Faulds, Francis Galton and Edward Henry recognizing the uniqueness and potential of minutiae points for individual identification [46].

The First Automated Systems (1970s-1990s) Early attempts at automated face recognition emerges in the 1960s, notably with the work of Bledsoe, but these systems relied heavily on manual landmark identification, limiting their practicality [47]. A significant breakthrough came with Takeo Kanade's development of the first fully automated face recognition system in 1973 [1].

The Rise of Feature Engineering (1990s–2000s). The 1990s mark a paradigm shift in face recognition, moving from handcrafted geometric features to holistic, appearance-based representations. A seminal contribution is the Eigenfaces by Turk and Pentland [9], which leverages principal component analysis (PCA) to represent faces as linear combinations of orthogonal basis images. This approach enables more compact and discriminative facial representations. However, Eigenfaces exhibit limitations in handling variations in lighting and facial expression. To address this,
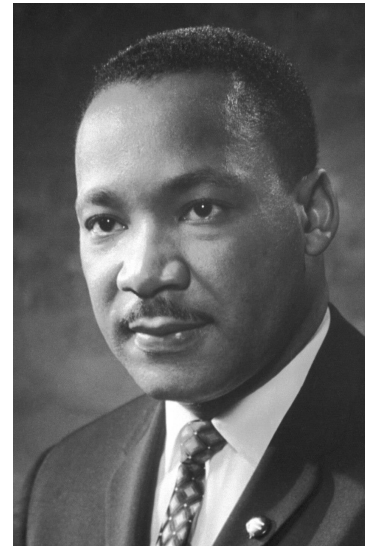
Identity: Martin

Demographics
Age: Late 30s
Gender: Male
Race: Black

Attributes
Hair: Trimmed
Mole: No
Beard: No
Mustache: Yes
Scar: No

Social Cues
Expression: Neutral

Fig. 6: A visual breakdown of the various types of information that can be extracted from a human face. These include identity-specific features, demographic traits, soft biometrics (e.g., presence of beard, mustache, scar), and high-level social cues such as emotion or expression. Image source.

Belhumeur *et al.* [48] introduce Fisherfaces, which apply linear discriminant analysis (LDA) to better separate individuals in a lower-dimensional subspace, improving robustness under varying illumination. This PCA-versus-LDA debate is further explored by Martinez and Kak [49], who highlight the strengths and weaknesses of both in practical scenarios.

Model-based techniques such as Elastic Bunch Graph Matching [50] provide pose-invariant recognition by encoding facial landmarks through a graph-based structure, bridging the gap between rigid appearance models and deformable representations. Complementing these efforts are texture-based descriptors such as Local Binary Patterns (LBP) [10] and Scale-Invariant Feature Transform (SIFT) [51], which mitigate sensitivity to lighting and expression by capturing local structural patterns.

A critical breakthrough in face detection in images emerges with the Viola-Jones algorithm [11], enabling real-time detection by Haar-like features and boosting. This work opens doors for practical applications in surveillance and consumer electronics. Around the same period, Heisele *et al.* [52] propose a component-based framework, integrating part-based local features to enhance robustness against occlusion and pose variation, thereby reinforcing the shift toward modular and discriminative feature engineering.

Deep Learning Paradigm (2010s-Present). The advent of deep learning [15], [16] in the 2010s revolutionizes the field. This paradigm shift is fueled by innovations in neural network architectures and the availability of large-scale datasets for training and evaluating the networks. Landmark papers like AlexNet [15] and ResNet [16] demonstrate the power of convolutional networks for image recognition, paving the way for their adoption in FR. The ImageNet dataset [15] provides a crucial resource for pre-training these large models, which were then fine-tuned for FR tasks.

Early pioneering works like DeepFace [12] and FaceNet [14] demonstrate the potential of deep learning for face
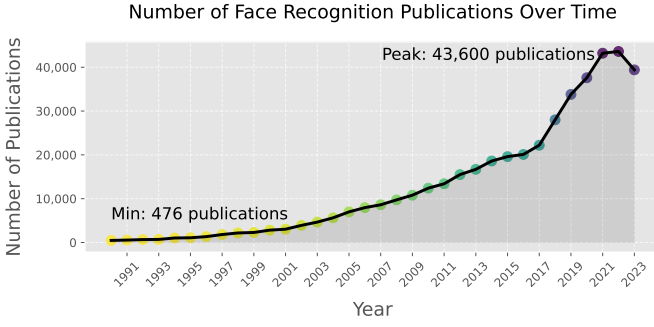
Fig. 7: Number of FR publications over time. Research activity in the field grew steadily until the early 2010s, followed by an explosive increase coinciding with the rise of deep learning. The peak in 2022 reflects the technology's mainstream adoption, though recent years suggest a slight cooling-off period in publications. However, in terms of deployments, FR continues to gain momentum. The global FR market size was valued at USD 7.73 billion in 2024. The market is projected to grow from USD 8.83 billion in 2025 to USD 24.28 billion by 2032 [56].

Fig. 8: Comparison of loss function paradigms in deep FR. Left: Triplet loss in contrastive learning reduces intra-class distance (pull positive training samples closer to anchor) while increasing inter-class distance (push negative training samples away). Right: Normalized Softmax loss maps features and class weights onto a hypersphere, optimizing angular distances to enhance inter-class separability and intra-class compactness.

recognition, achieving near-human performance on benchmark datasets. DeepFace utilizes a large-scale dataset (4M images and 4K subjects) of facial images to train a deep neural network for face verification, while FaceNet introduces a unified embedding space for FR and clustering.
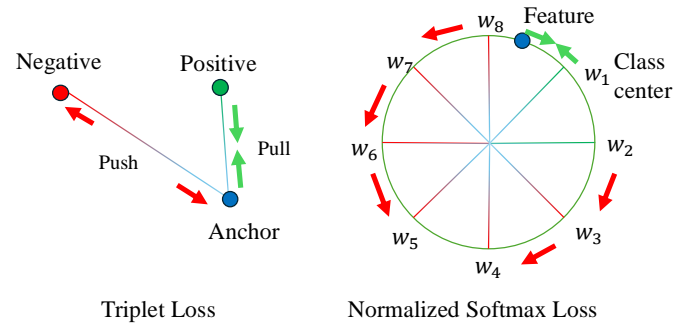
A key area of innovation in deep FR has been the development of specialized loss functions (Sec. 4.1). These loss functions are designed to improve the discriminative power of the learned features, enabling more accurate FR. Notable examples include NormFace [17], SphereFace [18], CosFace [19], ArcFace [20], CurricularFace [21] and Adaface [22] each introducing novel approaches to margin-based learning.

The performance of deep FR models is also highly dependent on the availability of large-scale training datasets (Sec. 4.2). Several large-scale face datasets have been developed to train and evaluate FR models, including CASIA-WebFace [53], VGGFace [54], MS1M [54], and WebFace260M [55]. These datasets provide a diverse range of facial images with varying pose, illumination, expression and age, enabling the training of robust FR models.

With the advent of deep learning, advanced architectures (Sec. 4.3) with billions of parameters that are trained on increasingly large face datasets bring about unprecedented improvements in accuracy and robustness. Today, face recognition stands as one of the most successful applications of deep neural networks in computer vision and AI, as shown in Fig. 7.

4 ADVANCES IN DEEP FACE RECOGNITION

Over the past decade, deep face recognition has experienced remarkable progress, driven primarily by three key factors: the development of loss functions, the availability of large-scale and diverse datasets, and advances in neural network architectures. Together, these innovations have dramatically improved the ability of models to learn highly discriminative, robust facial representations.

4.1 Loss Functions

The choice of loss function is critical in training deep face recognition models, as it directly guides the network to learn discriminative feature embeddings suitable for distinguishing between a vast number of identities. Among the key innovations that drove the advancement in FR, the highest number of publications have come from the advances in the loss function. Several distinct paradigms for loss function design have emerged.

Contrastive Learning Approaches: One major family of loss functions employs contrastive learning principles, directly shaping the embedding space by optimizing relative distances between samples. The seminal FaceNet [14] introduces the triplet loss, designed to ensure that an anchor sample's embedding is closer to its positive (same identity) counterpart than to any negative (different identity) sample by a predefined margin, typically in the Euclidean space.

While effective, triplet loss faces challenges in sampling informative triplets. Many randomly chosen triplets provide weak gradient signal, making training inefficient or necessitating complex hard-negative mining strategies. To address this, proxy-based methods are proposed. Techniques like Proxy Anchor Loss [57], originating from general deep metric learning, associate each class with learnable proxies (representative vectors), simplifying the loss computation by comparing samples to these proxies rather than exhaustively searching for pairs or triplets within a batch.

Further refining the contrastive approach, Supervised Contrastive Learning (SupCon) [69] generalizes the loss to leverage all positive samples available for an anchor within a batch, contrasting them against all negative samples. This more data-efficient approach has been successfully applied to FR, for instance, in UCFace [67]. Other works adapt contrastive ideas for specific goals: Open-Set Biometrics [70] focuses on improving open-set performance by explicitly minimizing scores between non-mated pairs, while CAFace [71] uses a contrastive-style cosine similarity loss to enforce consistency between embeddings of low-quality images and their high-quality counterparts, promoting quality invariance. Another work in this category,

TABLE 1: Summary of deep FR methods focusing on their loss functions, with their key advantages and limitations.

Name	Year	Pros	Cons
DeepID2+ [58]	2014	Joint ID + verification loss for robust features	Complex training with dual supervision losses
CenterFace [59]	2016	Center loss enhances intra-class compactness	No explicit inter-class separation; needs tuning
SphereFace [18]	2017	Angular margin enforces hyperspherical separation	Training instability from angular multiplicity
L2-Face [60]	2017	L2 norm constraint improves angular discrimination	Needs careful radius tuning; no margin enforcement
ArcFace [20]	2018	Additive angular margin boosts inter-class separation	Fixed margin may hurt low-quality samples
CosFace [19]	2018	Cosine margin improves class separability stably	Uniform margin not adaptive; needs tuning
SeqFace [61]	2018	Sequence-aware loss improves temporal supervision	Needs sequence data; dual loss increases complexity
Git Loss [62]	2018	Unified softmax + center loss boosts discrimination	Extra tuning and complexity with marginal gain
MagFace [63]	2021	Feature norm models quality for adaptive margin	Complex loss and quality-norm assumptions
AdaFace [22]	2022	Dynamic margin based on feature norm quality	Relies on norm-quality link and tuning
ElasticFace [64]	2022	Elastic margin adapts to feature variability	Stochastic margins add tuning and training cost
UniFace [65]	2023	Similarity threshold improves verification alignment	Global constraints increase optimization cost
UniTSFace [66]	2023	Sample-to-sample loss optimizes verification	Pairwise loss and threshold learning cause overhead
UCFace [67]	2024	Uncertainty and probability density aware contrastive learning	Cannot be used by itself, must be accompanied by margin loss
LAFS [68]	2024	Landmark based SSL pretraining helps face recognition	Loss depends on pretrained model and the landmark quality.

Related efforts also explore optimizing embedding spaces to better align with recognition objectives, such as in [72], where features are learned to directly improve performance metrics for face identification and verification separately.

Margin-based Softmax Losses: A dominant and highly successful approach in deep FR involves modifying the standard softmax cross-entropy loss to directly enhance feature discriminability. The core motivation is to learn embeddings that exhibit smaller intra-class variations (features of the same person are close together) while simultaneously maximizing inter-class separation (features of different people are far apart).

The standard softmax loss, often used as a baseline in classification tasks, is formulated for a sample $\mathbf{x}_i$ with feature embedding $\mathbf{z}_i \in \mathbb{R}^d$ belonging to the y_i-th class as:

$$\mathcal{L}_{CE}(\mathbf{x}_i) = -\log \frac{\exp(\mathbf{W}_{y_i}^T \mathbf{z}_i + b_{y_i})}{\sum_{j=1}^{C} \exp(\mathbf{W}_j^T \mathbf{z}_j + b_j)}, \quad (1)$$

where $\mathbf{W}_j$ is the weight vector for the j-th class, b_j is the bias term, and C is the total number of classes or identities in the training set. While effective for classification, this formulation doesn't explicitly enforce the metric learning objective crucial for face recognition where we encounter identities not seen during training.

An early work moving in this direction is Center Loss [59], which adds an auxiliary loss term to the standard softmax. This term penalizes the Euclidean distances between the deep features and their corresponding learned class centers, directly encouraging intra-class compactness. A significant breakthrough comes with the normalization of both feature embeddings ($\|\mathbf{z}_i\| = 1$) and classification weights ($\|\mathbf{W}_j\| = 1$, and setting $b_j = 0$). This reformulation, pioneered by SphereFace [18], maps the optimization problem onto a hypersphere where the dot product $\mathbf{W}_j^T \mathbf{z}_i$ becomes equivalent to $\cos \theta_j$, the cosine of the angle between the feature vector $\mathbf{z}_i$ and the weight vector $\mathbf{W}_j$. A scaling factor s is typically introduced to control the radius of the hyperspherical feature space. The loss then becomes:

$$\mathcal{L}_{cos}(\mathbf{x}_i) = -\log \frac{\exp(s \cdot \cos \theta_{y_i})}{\sum_{j=1}^{C} \exp(s \cdot \cos \theta_j)}. \quad (2)$$

Building on this normalized angular space, the key innovation is the introduction of explicit margins to make the learning objective more stringent. CosFace [19] introduces an additive cosine margin (m) by modifying the target logit to $s \cdot (\cos \theta_{y_i} - m)$. ArcFace [20] proposes an additive angular margin (m) by modifying the target angle itself, resulting in a target logit of $s \cdot \cos(\theta_{y_i} + m)$. Both approaches effectively create a decision boundary gap, forcing learned features for the same identity to cluster more tightly in the angular space, thereby significantly improving discriminative power. A visual comparison of contrastive triplet loss and margin-based normalized softmax loss is illustrated in Fig. 8, highlighting how each paradigm optimizes the embedding space to enhance FR performance.

Subsequent research further refines these margin-based concepts. MagFace [63] proposes leveraging the magnitude of the feature vector (before normalization) as an indicator of face image quality, incorporating an auxiliary loss to promote larger magnitudes for higher-quality samples. AdaFace [22] addresses the challenge posed by low-quality or difficult samples by introducing an adaptive margin function. It dynamically adjusts the margin stringency based on image quality indicators, reducing the negative impact of potentially unrecognizable faces in the training process.

These advancements in margin-based softmax losses lead to remarkable performance gains, pushing verification accuracy on high-quality benchmarks like LFW [26] and CFP-FP [27] towards saturation (often exceeding 99%). This success shifts the community's focus towards improving performance in more challenging, unconstrained scenarios, particularly those involving low-quality images, as represented by benchmarks like IJB-S [25], [73].

Further refinements continued to explore margin dynamics; for example, ElasticFace [64] introduces randomized margins for greater flexibility during training, while UniFace [65] proposes the Unified Cross-Entropy (UCE) loss specifically aiming to guarantee a clear separation threshold between positive and negative pairs.

Margin-based softmax variants [20], [22], [65] currently dominate SoTA results. The specific choice among these often depends on the dataset characteristics (*e.g.*, presence of low-quality images) and desired properties. Contrastive
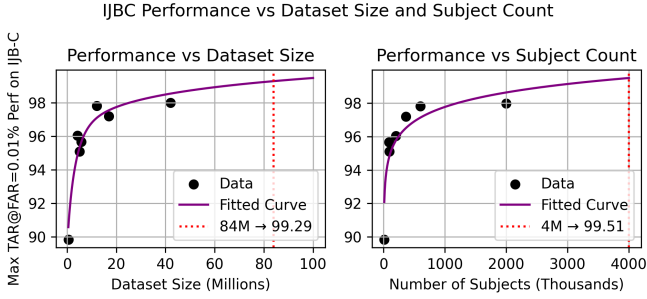
Fig. 9: Recognition performance on IJB-C dataset as a function of training dataset size (left) and training number of subjects (right). The dots show the the best publically available algorithms' performance for the given training dataset. Curves are fitted using the logarithmic function. Both increasing the number of images and expanding subject diversity significantly improve performance. However, the performance begins to saturate around 42M images and 2M subjects, suggesting diminishing returns at larger scales. While further gains are still possible, it may require novel embeddings.

Fig. 10: Plots showing the growth of face recognition datasets over time. The left plot illustrates the number of images (in millions), and the right plot shows the number of subjects (in thousands) for each dataset.

methods remain a helpful auxiliary loss, on top of margin-based softmax losses. A summary of various loss functions are shown in Tab. 1.

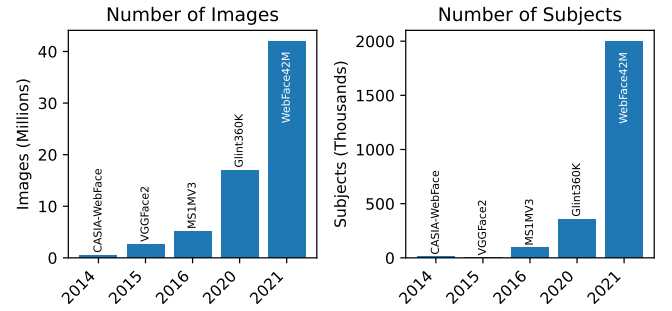**Auxiliary Losses for Interpretability and Distillation:** Beyond optimizing the core embedding space based purely on identity labels, another category of loss functions incorporates auxiliary objectives to achieve specific goals, such as enhancing model interpretability or improving performance in challenging conditions like low resolution. These losses often supplement the primary identity discrimination loss.

A significant effort has focused on improving model interpretability, *e.g.*, understand *how* the network makes decisions. Towards this, Yin *et al.* [74] propose spatial and feature activation diversity losses. These encourage the network to learn more structured representations where different spatial activations may correspond to different facial aspects, while also making these interpretable features discriminative and robust to occlusions. Similarly, the Explainable Channel Loss (ECLoss), also framed as Activation Template Matching Loss [75], encourages specific channels within convolutional layers to specialize in detecting distinct facial parts (*e.g.*, eyes) without explicit part annotations, thereby providing a direct interpretation of channel function.

Knowledge distillation (KD) offers another avenue, often targeting specific challenges. For low-resolution FR, Attention Similarity KD (A-SKD) [76] transfers teacher attention maps to guide the student's focus. For efficient large-scale training, Li *et al.* [77] develop feature-based KD techniques, like reverse distillation, that importantly remove the need for identity supervision for the student, saving memory while addressing the teacher-student 'intrinsic gap'. These examples highlight the use of specialized loss functions and KD strategies to imbue models with desirable properties like explainability, robustness, or training efficiency, complementing the primary task of recognition.

4.2 Datasets

The availability and scale of training data have been pivotal factors driving the remarkable progress in deep learn-

ing [15]. Publicly available large-scale face datasets (*e.g.*, MS-Celeb-1M, VGGFace2) spurred rapid advancements in the mid-2010s. Tab. 2 provides an overview of several influential datasets commonly used in the field, detailing the number of images and unique identities they contain. A clear trend emerges from this summary: a dramatic increase in dataset size over time. Early benchmark datasets like CASIA-WebFace [53] offer around half a million images from ten thousand subjects. In contrast, subsequent collections such as MS1MV2/V3 [78], Glint360K [79], and particularly the WebFace series [55], have pushed these numbers significantly higher, culminating in WebFace42M with over 42 million face images spanning 2 million identities. This growth reflects the community's understanding that larger and more diverse datasets are crucial for training accurate and generalizable FR models. It is important to note that MS-Celeb-1M and VGGFace2 have been redacted from the dataset creators due to lack of user consent.

In Fig. 9, we show the FR performance on IJB-C at TAR@FAR=0.01% with varied dataset size and number of subjects. The performance is taken as the maximum of FR algorithms that were trained on the particular dataset. And we fit a curve to see the trend. We observe that both increasing dataset size and subject number lead to substantial improvements in performance. However, the trend indicates a saturation point around 42M images or 2M subjects, beyond which additional data yields diminishing returns.

It is important to note the origin and labeling methodology of many of these large-scale datasets. A significant portion, including prominent datasets like MS-Celeb-1M and the WebFace series, are curated by collecting images from publicly accessible sources on the Internet, often leveraging search engines or social media platforms. Consequently, the identity labels associated with these images are frequently "pseudo-labels," because web searches of celebrities may return different subject images. Due to the volumn of dataset, the labels are generated through automated clustering algorithms or matching, rather than manual verification. While efforts are made to clean and refine these labels, noise and inaccuracies can persist. Some approaches, like that used for WebFace260M [55], employ iterative self-labeling and retraining of specialized labeler models to improve the quality of these pseudo-labels over multiple cycles. Since benchmark datasets [23], [24], [26], [27], [29] are also curated

TABLE 2: Overview of face recognition datasets, used primarily for training deep networks, with corresponding image and subject counts, year of release, and authentication performance reported as TAR@FAR=0.01%.

Dataset	Year	# Images	# Subjects	Max IJB-C TAR@FAR=0.01%
CASIA-WebFace [53]	2014	0.5M	10K	89.84 (UniTSFace)
VGGFace2 [54]	2018	2.6M	2.6K	—
MS1MV2 [78]	2019	5.8M	85K	95.67 (AdaFace)
MS1MV3 [78]	2019	5.1M	93K	95.10 (ArcFace)
Glint360K [79]	2021	17M	360K	97.20 (PartialFC)
WebFace4M [55]	2021	4.2M	200K	96.03 (AdaFace)
WebFace12M [55]	2021	12M	600K	97.82 (ViT KP-RPE)
WebFace42M [55]	2021	42M	2M	**97.99 (UniTSFace)**

from public web, training datsets need to ensure that the identities in training and test sets do not overlap.

Also, the practice of web-scraping facial images raises significant privacy and ethical concerns within the research community and society at large [80], as individuals may not have provided explicit consent for their images to be used in this manner for developing and training recognition systems. This issue remains an active area of discussion and necessitates careful consideration of data governance and ethical guidelines moving forward.

In addition to the large-scale 2D image datasets collected primarily from the web, the field also utilizes 3D face datasets [81]–[92]. These datasets capture the geometric structure of the face, often along with texture information, using specialized acquisition techniques like 3D scanners, structured light, or multi-camera stereo systems. A key distinction is that 3D datasets are typically collected under controlled laboratory settings with the explicit consent of the participants. This controlled acquisition allows for high-quality, precise capture of facial shape in RGBd (depth), which can offer inherent robustness advantages against variations in pose and illumination compared to 2D images.

3D face recognition has also emerged as a parallel research track, with dedicated benchmarks and evaluation campaigns [93]. Notably, datasets such as FRGC [93] and Lock3DFace [94], which leverages low-cost RGB-D sensors like Kinect, have been widely used in the community to advance recognition algorithms under realistic conditions.

However, the process of 3D data acquisition is significantly more complex, time-consuming, and expensive. Consequently, the volume of available 3D face data, both in terms of the number of scans and the number of unique subjects, is substantially smaller compared to the massive scale achieved by web-scraped 2D datasets. This difference in scale limits the use of 3D face datasets for training the deep models that dominate current FR research, although they remain valuable for specific research tasks, evaluation, and applications where 3D information is critical. For general face recognition applications, such as in law enforcement, immigration, or airport screening, RGB cameras offer a more practical solution considering legacy databases and return on investment.

4.3 Neural Network Architectures

CNN Architectures in FR: The backbone neural network architecture plays a crucial role in extracting discriminative features from face images. The revolution brought by deep learning in computer vision, largely initiated by AlexNet [15] on the ImageNet challenge, quickly permeates the field of FR. Early deep FR models adapt existing
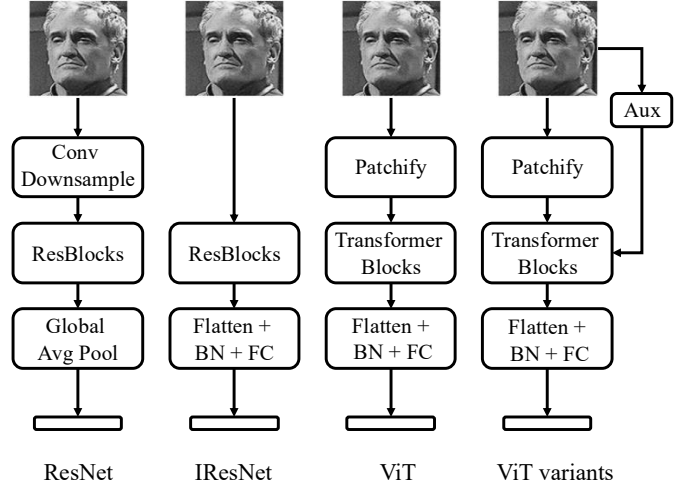
Fig. 11: Comparison of architectures used in FR. From left to right: ResNet [16] is an architecture used for classification [15]. IResNet [20] modifies this by removing downsampling and using feature flattening, batch normalization (BN), and fully connected (FC) layers that are helpful for metric learning [101]. Vision Transformer [102] (ViT) replaces convolutions with a patchify operation and transformer blocks; ViT variants further extend this by incorporating auxiliary information such as facial keypoints [39], [99] to improve learning.

successful CNN architectures designed for general object recognition.

Architectures like GoogLeNet [95] demonstrates the power of increased network depth and led to its adoption in FaceNet [14]. The introduction of Residual Networks (ResNets) [16] which addresses the vanishing gradient problem in very deep networks through the use of residual connections (shortcuts) leads to training of much deeper models (*e.g.*, ResNet-50, ResNet101, ResNet-152). Variants of ResNet (*e.g.*, SE blocks [96]), become the popular backbone for many SoTA face recognition systems developed in the late 2010s [19], [20]. ArcFace [20]'s adoption of input size 112×112 leads to the widely used IR-ResNet face recognition backbones which removed first downsampling blocks to compensate for the small resolution. Fig. 12 shows the progression of facial image sizes in the FR datasets.

Facial alignment is a crucial preprocessing step in FR systems, ensuring that key facial features (*i.e.* eyes, nose, and mouth) are consistently positioned across different images. Earlier FR datasets [53] utilize Multi-task Cascaded Convolutional Network (MTCNN) [97], which jointly performs face detection and landmark localization via a series of cascaded networks. With the advent of single-stage detectors like SSD [98], more efficient and accurate methods emerge. Notably, RetinaFace [99] has become a popular solution, offering precise face detection and alignment. When trained on strong datasets such as WiderFace [100] and paired with an improved backbone, RetinaFace is a robust choice for preprocessing large-scale face datasets [55]. Some example alignments are shown in the last row of Fig. 12.

Vision Transformers in FR: Mirroring trends in natural language processing and broader computer vision, Vision Transformers (ViTs) [102] have emerged as a powerful alternative to CNNs. ViTs process images by dividing them into
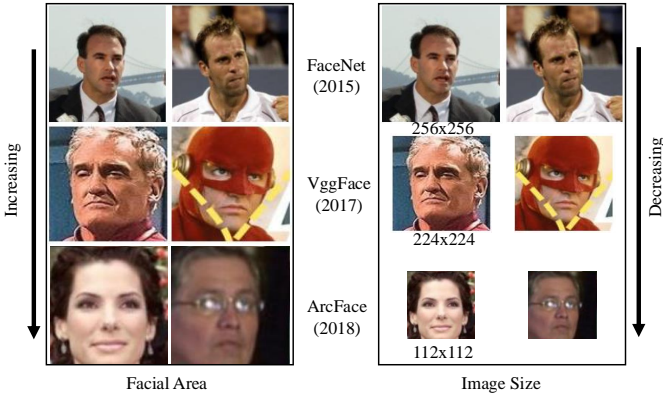
Fig. 12: Comparison of face recognition methods in terms of facial contextual area and image resolution. Models have evolved to focus on more tightly cropped facial regions while reducing resized input image size (256×256 to 112×112), enabling more efficient feature extraction.

Fig. 13: Examples of real and synthetic datasets. Real datasets evolve from the primarily frontal CASIA-WebFace to the large-scale, diverse WebFace260M, including wide variations in pose and expression. Synthetic datasets also advance, focusing on improving diversity and maintaining identity consistency.

patches, and feeding the resulting sequence into a Transformer encoder [103]. The self-attention mechanism within Transformers allows the model to weigh the importance of different image patches globally, potentially capturing long-range dependencies that might be missed by the local receptive fields of CNNs. ViTs have also shown great performance in face recognition domains [39], [104], [105]. And adoption of ViT in face recognition implies adoption of advances around ViT. SwinFace [105] is an application of Swin Transformer [106]. KP-RPE [39] integrates facial landmarks into relative position encodings in ViT, improving robustness to pose variations.

Empirically, compared to ResNets, training ViT on FR models entails more augmentations and requires larger-scale training set [39]. Fig. 11 shows how FR models have changed over time, from ResNets to newer ViT models that use attention and extra information to recognize faces better.

While both CNNs, particularly ResNet variants like IR-SE models, and Vision Transformers (ViTs) have demonstrated SoTA performance, there is no single definitively best architecture for all FR tasks. ViTs have shown potential for marginally higher accuracy on some benchmarks, especially when trained on extremely large datasets, leveraging their capability to capture global image dependencies. However, they often necessitate more extensive training data and sophisticated augmentation strategies. ResNets remain highly competitive, often providing a more efficient balance between performance and training/inference cost, particularly with moderate-sized datasets. The optimal choice often depends on the specific application's constraints, including dataset size, available compute resources, and deployment environment.

Face recognition model deployment depends heavily on computational demands, mainly measured by FLOPs and model size. For example, IResNet50 needs 12.62 GFLOPs and has 43.59M parameters, while IResNet101 uses 24.19 GFLOPs and 65.15M parameters. ViT models are more demanding—ViT Small has 17.42 GFLOPs and 95.95M parameters, and ViT Base requires 24.83 GFLOPs with 114.87M parameters. On consumer GPUs such as Nvidia 3090, IRes-
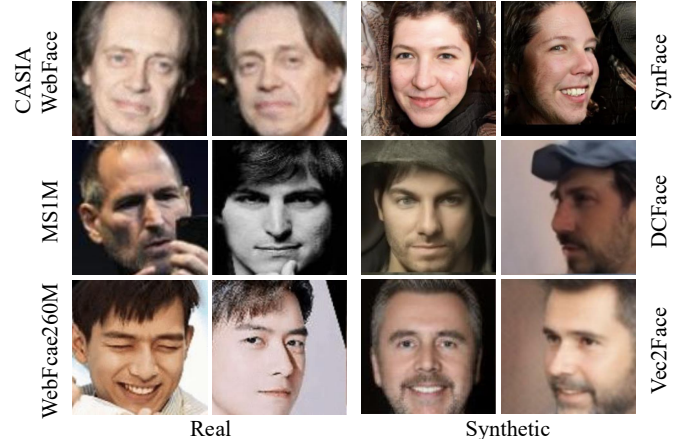
Net50 can process over 1400 images/second, while ViT Base handles about 640 images/second. Unlike academic research, industry models or government vendor models [107] can use model ensembles, further increasing the load. Additionally, pre-processing steps like face detection and alignment add to the overall computational cost. Note that ViT backbones can benefit from research that speed up ViT inference [108], [109].

Efficiency, Adaptation, and Compact Embeddings: Beyond achieving maximum accuracy, research has also focused on developing efficient architectures suitable for deployment on resource-constrained devices like mobile phones. MobileFaceNet [110] employs depthwise separable convolutions to significantly reduce computational cost and model size while maintaining reasonable accuracy (IResNet100 at 99.83% vs MobileFaceNet at 99.55% in LFW verification accuracy while being $60\times$ smaller). S-ViT [111] applies sparse attention to reduce computational cost without sacrificing accuracy. The continuous evolution of neural network architectures, from deeper CNNs to attention-based Transformers and efficient mobile designs, has been a key driver alongside loss function innovations and larger datasets in pushing the performance boundaries of automated FR.

Fine-tuning is crucial for adapting face recognition models to new domains, especially under quality mismatches [107] (*e.g.* low vs high quality images). Instead of full fine-tuning, which risks catastrophic forgetting, recent work like PETALface [112] uses LoRA [113], a parameter-efficient finetuning method that adds low-rank adaptation modules. By weighting LoRA blocks based on image quality, PETALface adapts effectively to low-resolution faces while preserving high-resolution performance.

Alongside the choice of network architecture, the dimensionality of the output feature vector, or template size, is another critical design decision influencing system performance and efficiency. While modern systems often employ 512-dimensional embeddings, research by Gong *et al.* [123] investigates the fundamental requirements of these representations by examining their intrinsic dimensionality. Their

TABLE 3: Comparison of synthetic face training datasets for face recognition across five standard benchmarks. "Gap to Real" shows the average performance drop relative to the use of real CASIA-WebFace dataset alone for training. Brackets in the Generator Training Dataset column denote datasets used for pretraining, which may help models learn facial priors. A more fair comparison might involve equalizing the use of pretrained models. FR model used is IR50 [114]–[116].

Methods	Venue	Generator Train Dataset	# images (# IDs× imgs/ID)	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg	Gap to Real
SynFace [114]	ICCV21	FFHQ [117]	0.5M (10K × 50)	91.93	75.03	70.43	61.63	74.73	74.75	26.04
DigiFace [115]	WACV23	511 3D Scans	1M (10K × 100)	95.40	87.40	78.87	76.97	78.62	83.45	11.34
DCFace [116]	CVPR23	CASIA-WebFace (FFHQ)	0.5M (10K × 50)	98.55	85.33	82.62	89.70	91.60	89.56	5.23
IDnet [118]	CVPR23	CASIA-WebFace [53]	0.5M (10K × 50)	92.58	75.40	74.25	63.88	79.90	79.13	15.66
ExFaceGAN [119]	IJCB23	CASIA-WebFace	0.5M (10K × 50)	93.50	73.84	71.60	78.92	82.98	80.17	14.62
SFace2 [120]	TBIS24	CASIA-WebFace	0.6M (10K × 60)	96.50	77.11	74.60	77.37	83.40	81.62	13.17
Arc2Face [121]	ECCV24	WF42M [55] (Stable Diffusion)	0.5M (10K × 50)	98.81	**91.87**	85.16	90.18	92.63	91.73	3.06
Vec2Face [122]	ICLR2025	CASIA-WebFace (WebFace4M)	0.5M (10K × 50)	**98.87**	88.97	**85.47**	**93.12**	**93.57**	**92.00**	**2.79**
CASIA-WebFace (Real)	-	NA	0.49M (Real)	99.38	96.91	89.78	94.50	93.35	94.79	0.00

work reveals that the actual degrees of freedom needed to capture essential facial variations (the intrinsic dimension) can be substantially lower than the commonly used ambient dimensions (*e.g.*, d=512). Also FaceNet [14] indicates that for a given training dataset, higher performance was achieved with d=128 compared to d=512. This finding implies that standard high-dimensional face embeddings contain significant redundancy, suggesting the potential to develop much more compact templates that could enable faster search and more efficient storage while retaining discriminative power. It is important to note, however, that in commercial systems, a template often includes more than just the feature vector itself. It may contain metadata such as quality indicators, detection attributes, or even multiple embeddings, leading to greater variation in template sizes.

4.4 Synthetic Datasets

The growing demand for large-scale, diverse, and ethically sourced training datasets has driven increasing interest in the use of *synthetic face data*. Collecting real-world facial datasets often introduces privacy and consent challenges, as well as issues of demographic imbalance and limited representation under challenging conditions (*e.g.*, extreme poses, rare ethnicities). Synthetic data offers a compelling alternative or data augmentation by enabling controlled, scalable, and bias-aware dataset creation [114]–[116].

The generation of photorealistic synthetic faces has been significantly advanced by deep generative models, particularly Generative Adversarial Networks (GANs) [124]. Variants such as StyleGAN [117], [125] are especially effective at producing high-resolution facial imagery, capable of modeling complex visual distributions and allowing fine-grained control over attributes like pose, expression, and illumination via manipulations in the latent space. The integration of 3D Morphable Models (3DMM) into GANs has further enhanced the controllability of facial attributes during generation [126]–[129]. For instance, CFSM [130] leverages GANs to synthesize faces with diverse styles, aiding in the generation of richly varied datasets.

Recently, Diffusion Models [131] emerge as a powerful generative paradigm, achieving impressive image quality and diversity. They generate images by gradually denoising a sample from pure noise, learning to reverse the diffusion process. Text-conditioned diffusion models are especially effective for controlled synthesis, enabling detailed and semantically guided generation [132], [133]. ControlNet [134]

and IP-Adapter [135] make the model adhere to input conditions such as facial landmarks, masks or other clues.

Leveraging these generative capabilities, researchers have explored creating entire synthetic datasets specifically for training FR models. The goal of face dataset generation is to create multiple images of the same subject at a large scale. The ID consistency is at an interplay with the diversity of generated images. SynFace [114] first applies GAN and latent interpolation method to generate face datasets, resulting in average face verification rate of 74.75%, marking a significant drop compared to real CASIA-Webface dataset of 94.79% as in Tab. 3. Since then, mulitple works have attempted to reduce the gap.

Following SynFace, the field has witnessed rapid growth, with numerous efforts aiming to bridge the performance gap between synthetic and real face datasets. One notable direction is the use of diffusion-based models, as exemplified by DCFace [116], which separates identity and style conditions during generation to produce identity-consistent subjects while maintaining high diversity. Arc2Face [121] takes this further by leveraging a pretrained Stable Diffusion model, highlighting the power of foundation models in capturing generalizable facial representations. On the other hand, Vec2Face [122] demonstrates that GAN-based synthesis can remain competitive when guided by a FR feature space, emphasizing the critical role of identity disentanglement. These approaches reflect the ongoing trend in synthetic dataset generation: pushing the frontier of realism and diversity while preserving identity consistency to close the gap with real-world data. Tab. 3 shows the performance of recent synthetic face datasets on FR.

There is growing interest in synthetic FR challenges, as reflected by the FRCSyn (Face Recognition Challenge in the Era of Synthetic Data) series [136], [137]. The series reveal that while synthetic-only training remains slightly behind real data in certain scenarios, it offers compelling advantages in mitigating demographic bias and improving robustness to variations in pose, age, and occlusion. Examples of the comparison between real and synthetic datsets are shown in Fig. 13. Also multiple studies show that snynthetic images can be used as an augmentation to the real dataset. For example, addition of Vec2Face [122] to CASIA-WebFace can increase about 1.00% in average verification accuracy.

Despite the promise, the effective use of purely synthetic data for training SoTA FR models faces challenges. A primary hurdle is the *domain gap* between synthetic and real images. Models trained solely on synthetic data may not

generalize very well to real-world images due to subtle differences in texture, lighting, or artifacts introduced by the generation process. Ensuring sufficient diversity and realism, particularly in capturing the nuances of identity across various conditions, remains an active research area. Furthermore, the generation process itself can be computationally expensive, and ethical concerns regarding the potential misuse of highly realistic synthetic faces (*e.g.*, deepfakes) persist. To address these risks, there is a growing emphasis on watermarking or other techniques [138] to clearly indicate that images are synthetically generated.

4.5 Feature Fusion in Face Recognition

In template-based face recognition, multiple face images of the same identity—often captured under varying conditions of pose, illumination, resolution, and occlusion—must be fused into a single, compact representation to enable efficient and accurate comparison. This fusion scenario commonly arises in gallery settings, where multiple still images (or media) per subject must be aggregated into a unified template.

A second, and increasingly popular, scenario involves video-based face recognition, where frames extracted from probe video sequences are fused into a single representation. This use case poses unique challenges, as it often requires online (on-the-fly) feature fusion to support real-time applications such as surveillance or mobile authentication. Despite differing temporal constraints, both still-image and video-based fusion share the core objective: to generate robust and compact representations that preserve discriminative identity cues.

Feature fusion is a critical step in this process, as it determines how the information from diverse images of the same person is aggregated into a unified descriptor. Naive methods like average or max pooling treat all feature embeddings equally, which can dilute discriminative cues by giving equal importance to low-quality or redundant images. Therefore, effective feature fusion must not only compress but also intelligently filter, weight, and adapt to the content of the input set. The ability to generate robust, order-invariant, and compact template representations directly impacts FR performance, especially in unconstrained or real-time scenarios.

Notably, one of the earliest contributions to video-based face recognition uses adaptive hidden Markov models to model temporal dynamics and perform recognition based on entire video sequences [72]. Over the years, feature fusion in FR has evolved from simple averaging techniques to highly adaptive and context-aware neural aggregation frameworks. Early methods such as the Neural Aggregation Network (NAN) [139] establishes the value of learning quality-aware weights through attention, enabling robust, order-invariant representations for face templates. This sets the stage for more sophisticated models like Multicolumn Networks [140] and C-FAN [141], which incorporate fine-grained quality analysis, either by separating visual and contextual importance or by estimating weights for individual feature channels. These advances are instrumental in improving FR accuracy on challenging template-based datasets such as IJB-C [24].
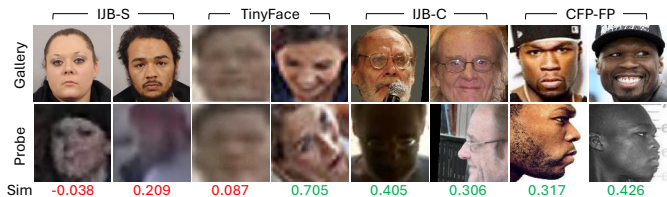
Fig. 14: Examples of faces across different datasets: IJB-S [25], TinyFace [73], IJB-C [24], and CFP-FP [27]. The top row shows gallery images, the middle row shows corresponding probe images, and the bottom row reports cosine similarity scores. Higher similarity scores (in green) indicate successful matches, while lower scores (in red) indicate mismatches. Features are extracted by KP-RPE [39] trained on WF4M [55].

Recent efforts have shifted toward scaling and generalizing feature fusion across different operating conditions. Methods such as CAFace [71], CoNAN [142] and ProxyFusion [143] focus on maintaining performance even when templates contain hundreds or thousands of diverse images. Meanwhile, practical methods like Norm Pooling [144] demonstrate that even simple heuristics can be powerful in multi-domain scenarios, especially when training data is limited. A consistent trend across these innovations is the move toward scalable and fast strategies that can tailor to long videos, achieving robustness without sacrificing scalability.

5 STATE OF THE ART IN FACE RECOGNITION

The evaluation of face recognition (FR) systems broadly falls into three categories:

Benchmark Evaluation: Self-evaluation by algorithm developers on public benchmark datasets such as LFW, IJB-C, and TinyFace. Some examples are shown in Fig. 14.

Technology Evaluation: Third-party independent testing conducted by organizations like NIST on operational datasets (*e.g.*, mugshots, visa photos, border kiosk captures) that are not publicly available. Developers submit their algorithms for unbiased evaluation.

Scenario Evaluation: Deployment of FR systems in real-world operational settings, such as airport boarding gates, where practical concerns such as device placement, lighting, usability, and capture quality become critical alongside algorithmic performance.

In this section, we present two types of evaluations: (1) Benchmark evaluations and (2) Technology evaluations conducted by NIST.

5.1 Benchmark Evaluations

Robust evaluation of modern FR systems necessitates standardized benchmark datasets that reflect various real-world challenges. Prominent evaluation datasets extensively cited in recent literature include Labeled Faces in the Wild (LFW) [26], CFP-FP [27], CP-LFW [28], AgeDB [29], YouTube Faces (YTF) [156], TinyFace [73], and several iterations of the IARPA Janus Benchmark (IJB) series such as IJB-B [23], IJB-C [24], and IJB-S [25]. Each dataset addresses specific challenges inherent in FR scenarios. Below some datasets are described in detail.

TABLE 4: Performance on CFP-FP [27] Dataset

Method Name	Backbone	Loss Function	Training Data	Verification (%)
GFace [145]	IResNet-50	GCE (LO)	Casia-WebFace	97.44
CosFace [19]	ResNet100	CosFace [19]	MS1MV2	98.13
ArcFace [20]	ResNet101	ArcFace	MS1MV2	98.27
MV-Softmax [146]	ResNet100	MV-Softmax	MS1MV2	98.28
CurricularFace [21]	ResNet101	CurricularFace	MS1MV2	98.37
TransFace-B [147]	ResNet100	ArcFace	MS1MV2	98.39
MagFace [63]	ResNet100	MagFace	MS1MV2	98.46
AdaFace [22]	ResNet101	AdaFace	MS1MV2	98.49
CQA-Face [148]	ResNet100	CQA-Face	MS1MV2	98.49
UniFace [65]	ResNet100	UniFace [65]	MS1MV2	98.63
URL [149]	ResNet101	URL	MS1MV2	98.64
LGAF [150]	ResNet100	ArcFace	MS1MV2	98.77
ArcFace [20]	ResNet50	ArcFace	Glint360K	98.77
ViT-S [102]	ViT-S	ArcFace	Glint360K	98.85
CosFace + KP-RPE	ViT	CosFace	WebFace4M	98.91
TransFace-S [147]	ViT-S	ArcFace	Glint360K	98.91
AdaFace [22]	ViT	AdaFace	WebFace4M	98.94
KP-RPE [39]	ViT	AdaFace	WebFace4M	99.01
ViT-B [102]	ViT-B	ArcFace	Glint360K	99.02
AdaFace [22]	ResNet101	AdaFace	MS1MV3	99.03
R100	ResNet100	ArcFace	Glint360K	99.04
AdaFace [22]	ViT	AdaFace	MS1MV3	99.06
ArcFace [20]	ResNet101	ArcFace	WebFace4M	99.06
KP-RPE [39]	ViT	ArcFace	WebFace4M	99.09
ViT-L	ViT-L	ArcFace	Glint360K	99.10
KP-RP [39]E	ViT	AdaFace	MS1MV3	99.11
GFace [145]	IResNet-100	GCE (LO)	MS1MV3	99.12
R200	ResNet200	ArcFace	Glint360K	99.14
AdaFace [22]	ResNet101	AdaFace	WebFace4M	99.17
TransFace-B [147]	ViT-B	ArcFace	Glint360K	99.17
AdaFace [22]	ResNet101	AdaFace	WebFace12M	99.24
KP-RPE [39]	ViT	AdaFace	WebFace12M	**99.30**
TransFace-L [147]	ViT-L	ArcFace	Glint360K	**99.32**

TABLE 5: Performance on IJB-C [24] Dataset

Method Name	Backbone	Loss Function	Training Data	TAR@FAR=1e-4
ArcFace [20]	ResNet101	ArcFace [20]	MS1MV2	96.03
MagFace [63]	ResNet101	MagFace [63]	MS1MV2	95.81
MagFace+IIC	ResNet101	MagFace	MS1MV2	95.89
ViT-S	ViT-S	ArcFace	MS1MV2	95.89
CurricularFace [21]	ResNet101	CurricularFace	MS1MV2	96.10
ViT-B	ViT-B	ArcFace	MS1MV2	96.15
ViT-L	ViT-L	ArcFace	MS1MV2	96.24
TransFace-S [147]	ViT-S	ArcFace	MS1MV2	96.45
TransFace-B [147]	ViT-B	ArcFace	MS1MV2	96.55
TransFace-L [147]	ViT-L	ArcFace	MS1MV2	96.59
ArcFace+CFSM	ResNet101	ArcFace	MS1MV2	96.60
ARoFace [151]	ResNet101	ArcFace	MS1MV2	96.66
ElasticFace [64]	ResNet101	ElasticFace	MS1MV2	96.65
TopoFR [152]	ResNet101	TopoFR	MS1MV2	96.95
GFace [145]	ResNet101	TopoFR	MS1MV2	96.96
AdaFace [22]	ResNet101	AdaFace	MS1MV2	97.09
KP-RPE [39]	ViT-B	CosFace	WebFace4M	96.98
TopoFR [152]	ResNet200	TopoFR	MS1MV2	97.08
AdaFace [22]	ViT-B	AdaFace	MS1MV3	97.10
KP-RPE [39]	ViT-B	AdaFace	WebFace4M	97.13
AdaFace [22]	ViT-B	AdaFace	WebFace4M	97.14
KP-RPE [39]	ViT-B	AdaFace	MS1MV3	97.16
KP-RPE [39]	ViT-B	ArcFace	WebFace4M	97.21
PartialFC [79]	ResNet101	ArcFace	WebFace4M	97.22
CatFace [153]	ResNet101	CatFace	MS1MV2	97.43
AdaFace [22]	ResNet101	AdaFace	WebFace4M	97.39
ARoFace [151]	ResNet101	AdaFace	WebFace4M	97.51
AdaFace [22]	ResNet101	AdaFace	WebFace12M	97.66
PartialFC [79]	ResNet101	ArcFace	WebFace12M	97.58
ARoFace [151]	ResNet101	AdaFace	WebFace12M	97.60
TopoFR [152]	ResNet101	TopoFR	Glint360K	97.60
KP-RPE [39]	ViT-B	AdaFace	WebFace12M	97.82
PartialFC [79]	ResNet101	ArcFace	WebFace42M	97.82
TopoFR [152]	ResNet200	TopoFR	Glint360K	97.84
PartialFC [79]	ViT-B	ArcFace	WebFace42M	97.90
PartialFC [79]	ResNet200	ArcFace	WebFace42M	**97.97**
UniTSFace [66]	ViT-L	UniTSFace	WebFace42M	**97.99**

TABLE 6: Performance on TinyFace [73] Dataset

Method Name	Backbone	Loss Function	Training Data	Rank1	Rank5
ArcFace+CFSM	ResNet101	ArcFace	MS1MV2	64.69	68.80
TransFace-L [147]	ViT-S	ArcFace	MS1MV2	67.52	71.00
ARoFace [151]	ResNet101	ArcFace	MS1MV3	67.54	71.05
LGAF [150]	ResNet101	ArcFace	MS1MV2	68.35	71.59
ArcFace [20]	ResNet101	ArcFace	WebFace4M	71.11	74.38
AdaFace [22]	ResNet101	AdaFace	WebFace4M	72.02	74.52
AdaFace [22]	ResNet101	AdaFace	WebFace12M	72.29	74.97
REE [154]	ResNet-50	ArcFace	Native VLR	73.06	77.22
ARoFace [151]	ResNet101	ArcFace	WebFace4M	73.80	76.53
ARoFace [151]	ResNet101	AdaFace	WebFace4M	73.98	76.47
ARoFace [151]	ResNet101	AdaFace	WebFace12M	74.00	76.87
KP-RPE [39]	ViT-B	CosFace	WebFace4M	75.48	78.30
KP-RPE [39]	ViT-B	ArcFace	WebFace4M	**75.62**	**78.57**
KP-RPE [39]	ViT-B	AdaFace	WebFace4M	**75.80**	78.49

TABLE 7: Performance on IJB-S [25] Dataset

Method Name	Backbone	Loss Function	Training Data	Rank-1	Rank-5
PFE [155]	ResNet101	PFE	MS1MV2	50.16	58.33
URL [149]	ResNet101	URL	MS1MV2	59.79	65.78
CurricularFace [21]	ResNet101	CurricularFace	MS1MV2	62.43	68.68
AdaFace [22]	ResNet101	AdaFace	MS1MV2	65.26	70.53
AdaFace [22]	ViT	AdaFace	MS1MV3	65.95	71.64
AdaFace [22]	ResNet101	AdaFace	MS1MV3	67.12	72.67
KP-RPE [39]	ViT	AdaFace	MS1MV3	67.62	73.25
ArcFace [20]	ResNet101	ArcFace	WebFace4M	69.26	74.31
AdaFace [22]	ResNet101	AdaFace	WebFace4M	70.42	75.29
ARoFace [151]	ResNet101	ArcFace	WebFace4M	70.96	75.54
AdaFace [22]	ResNet101	AdaFace	WebFace12M	71.35	76.24
AdaFace [22]	ViT	AdaFace	WebFace4M	71.90	77.09
KP-RPE [39]	ViT	CosFace	WebFace4M	72.22	77.67
ARoFace [151]	ResNet101	AdaFace	WebFace12M	72.28	77.93
KP-RPE [39]	ViT	AdaFace	WebFace4M	**72.78**	**78.20**
KP-RPE [39]	ViT	ArcFace	WebFace4M	**73.04**	**78.62**

Labeled Faces in the Wild (LFW [26]) consists of over 13,000 facial images collected from the web, annotated with identity labels. The dataset includes multiple images for approximately 1,680 individuals of high-quality images. Main usage of this dataset is for the verification task.

YouTube Faces (YTF [156]) specifically targets video-based unconstrained face recognition. The dataset comprises clips varying from 48 to 6,070 frames, with an average length of 181 frames and contains an average of 2 videos per subject, making it useful for assessing algorithms designed to handle real-world variability in videos.

CFP-FP [27] evaluates the capability of algorithms to match frontal face images with their corresponding profile ones. It is particularly challenging due to large variations in facial orientation. The dataset is widely used to benchmark algorithms designed for pose-invariant face verification.

TinyFace [73] is explicitly designed for low-resolution FR research at scale. It includes 169,403 naturally low-resolution images (average size 20x16 pixels) depicting 5,139 identities. Images in TinyFace are cropped from crowded scenes and span a diverse range of lighting, occlusion, backgrounds.

IARPA Janus Benchmark-C (IJB-C) expands upon earlier series IJB-B [23], containing imagery and videos for 3,531 subjects, including 1,661 newly added identities. It comprises approximately 138,000 images and 11,000 videos. IJB-C serves as a challenging dataset for template-based recognition tasks. It contains significant variations in pose, illumination, and image quality [24]. Performance is often reported as TAR@FAR=threshold where the FAR threshold is selected based on the target operating point.

IARPA Janus Surveillance Video Benchmark (IJB-S) targets surveillance-specific scenarios, featuring images and videos of 202 subjects collected at a Department of Defense facility. The galleries are comprised of high-quality upper torso images and the probes are videos captured by surveillance camera of varied altitude and range. It is suited for evaluating surveillance-oriented FR approaches [25].

Collectively, these datasets represent comprehensive benchmarks that drive progress in addressing the nuanced challenges of modern face recognition technologies.

5.1.1 State-of-the-Art Performance

Recent advancements in FR performance demonstrate notable progress across different benchmarks, underscoring the effectiveness of deep learning architectures and sophisticated loss functions. In Tab. 4~7 we compile the up-to-date FR performance under each evaluation datasets using relevant metrics.

For 1:1 verification tasks, determining if two images belong to the same person, Verification Accuracy is a primary metric, indicating the overall percentage of correct matches and non-matches. On more challenging datasets involving
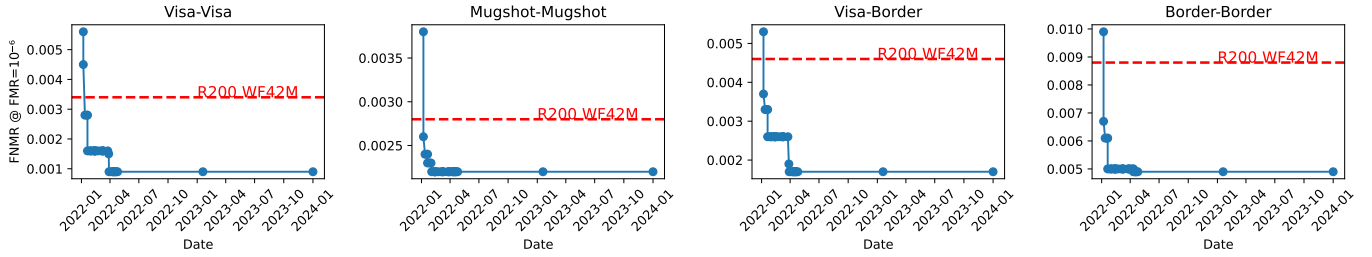
Fig. 15: Evolution of SoTA performance in NIST FRVT 1:1 verification since January 2022. Each plot shows the cumulative minimum False Non-Match Rate (FNMR, lower is better) achieved by any submitted algorithm up to the corresponding date for the Visa, Mugshot, Visa Border, and Border datasets (the plot titles indicate gallery - probe in order). Performance is evaluated at a low False Match Rate (*e.g.*, FMR=10^{-6}). The dashed red line indicates the performance level achieved by the WebFace42M entry (R200 [55] WF42M) for comparison.

template-based verification or significant variability, performance is frequently reported as the True Accept Rate (TAR) at a specific False Accept Rate (FAR), often FAR=0.01% or even lower, indicating the likelihood of correctly accepting a genuine match while maintaining a very low rate of false acceptances. In NIST's terminology, it is equivalent to 1-FNMR at a fixed FMR. For 1:N identification tasks, where an image must be matched against a gallery of known individuals, Rank-k accuracy is often used. This measures the percentage of times the correct identity is found within the top k ranked matches (Rank-1 being the most stringent). For open-set identification tasks (where the probe identity may not be in the gallery), the True Positive Identification Rate (TPIR) at a given False Positive Identification Rate (FPIR) (*e.g.*, FPIR=0.01%) is a standard metric, measuring the rate of correctly identifying individuals present in the gallery while controlling the rate of incorrectly matching unknown individuals to someone in the gallery.

CFP-FP [27]: Current methods achieve extremely high verification accuracy, often exceeding 99%. Top performance is typically seen with models utilizing ViT [102] backbones (*e.g.*, ViT-L, ViT-B, ViT-S variants) or deeper ResNet [16] architectures (*e.g.*, ResNet-101, ResNet-200). Effective loss functions like AdaFace [22] and ArcFace [20] are prevalent among the leading methods. Furthermore, training on very large datasets like Glint360K [79], WebFace [55] is crucial for reaching the highest scores, with methods like TransFace [147] reporting accuracies above 99.3%.

IJB-C [24]: IJB-C dataset presents a more challenging scenario involving template-based matching (comparing sets of images/video frames). Performance is often measured by the TAR@FAR=0.01%. SoTA methods, such as PFC [79] (utilizing ViT-L or ResNet200), KP-RPE [39] (with ViT-B), AdaFace [22], and TopoFR [152], achieve TAR values around 98% at 0.01%. Again, larger backbones (ViT-L [102], ResNet200 [16]) and extensive training data (WebFace42M [55], Glint360K [79]) are characteristic of the top-performing approaches.

TinyFace [73] (Low-Resolution Recognition): TinyFace specifically addresses the difficulty of recognizing faces from very low-resolution images. As expected, performance metrics like Rank-1 identification accuracy are considerably lower than on high-resolution datasets. Leading methods, predominantly using ViT-B backbones combined with techniques like KP-RPE that make the model robust to misalign-

ments and loss functions such as AdaFace that allow quality adaptive training achieve Rank-1 accuracies around 75-76%. Training on large datasets like WebFace12M is also helpful for performance. Methods like ARoFace [151] also show competitive results, highlighting the ongoing efforts to improve recognition under significant resolution constraints.

IJB-S [25]: Similar to TinyFace, IJB-S contains low-quality imageries and presents faces extracted from surveillance footage. We report Surveillance-to-Still (S2S) protocol. Top performance, measured by S2S Rank-1 accuracy, reaches approximately 73%. Another characteristic of this dataset is that the template size is large, making it a suitable choice to evaluate the methods for template feature fusion methods [71], [142], [143].

5.2 Technology Evaluations by NIST

The National Institute of Standards and Technology (NIST) has been conducting independent evaluations of FR technologies since 1999. Initially under the Face Recognition Vendor Test (FRVT), these activities have expanded to include the Face Recognition Technology Evaluation (FRTE) and Face Analysis Technology Evaluation (FATE) [157]. NIST evaluations are critical for assessing the technology readiness of algorithms for real-world deployment in government and operational settings.

Unlike benchmark evaluations conducted on public datasets, NIST uses sensitive operational datasets not available in the public domain, such as mugshots, visa application photos, and imagery from border kiosks. Developers submit their algorithms for third-party testing, ensuring unbiased, standardized evaluation. NIST assesses both 1:1 verification and 1:N identification scenarios, measuring metrics such as False Non-Match Rate (FNMR) at a fixed False Match Rate (FMR), and identification rates at various thresholds. NIST's reporting of 1-FNMR at a fixed FMR is equivalent to the more recently used terminology TAR@FAR metric. The number of distinct algorithms submitted to FRVT has grown over time, reflecting its increasing relevance and accessibility. Since evaluation is free and ongoing, participants can submit algorithms at any time for both 1:1 verification and 1:N identification tasks, with N now including up to 12 million enrolled identities.

Between 2014 and 2018, NIST reported that face recognition software improved by a factor of 20 in search accuracy [158], highlighting the rapid pace of advancement
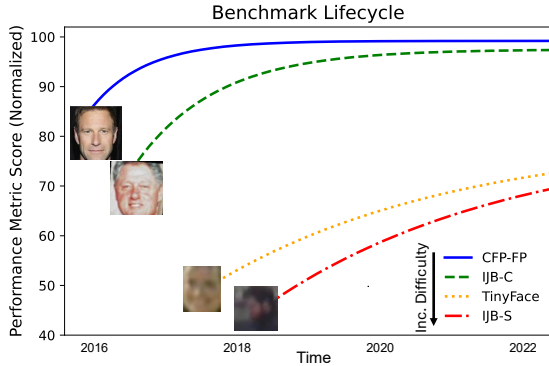
Fig. 16: Illustration of performance trends of various face recognition benchmarks over time, illustrating eventual performance saturation as datasets become extensively explored. This saturation is an indication of the progress in the field. Currently low-quality imagery benchmarks such as TinyFace and IJB-S are far from being solved.

in the field. To date, NIST has evaluated over 400 algorithms [159]. While academic benchmarks are crucial for driving research, the NIST FRVT provides an ongoing, operational evaluation of both academic and commercial algorithms under various scenarios, serving as a key indicator of the absolute SoTA deployed in real-world systems. Fig. 15 plots recent FRVT 1:1 verification performance results, including the performance trajectory of a strong academic baseline (ResNet-200 trained on the large-scale WebFace42M dataset, highlighted in red) for comparison against numerous vendor submissions. The performance results from the FRVT evaluations are publicly available on the organizer's website [159].

The results consistently show that top-performing algorithms, often developed by industry players, achieve excellent accuracy. However, many leading academic models, especially those trained on large public datasets like WebFace42M, perform competitively, demonstrating the strong impact of academic research on real-world applications. Nonetheless, the very best performing systems typically originate from industry, a difference that may stem from access to larger proprietary datasets, specialized hardware optimizations, extensive system-level engineering, or specific algorithmic refinements not yet published in academic literature. Still, the close proximity of top academic results to industrial leaders underscores the significant contribution of academic research to advancing practical FR capabilities.

6 CURRENT CHALLENGES IN FACE RECOGNITION

Despite the impressive advancements in face recognition performance, several critical challenges persist, especially when moving beyond controlled benchmark scenarios to real-world deployments. As illustrated in Fig. 16, existing benchmarks tend to reach saturation over time as methods advance and datasets become well-explored. This saturation, however, does not necessarily indicate that face recognition is a solved problem. Instead, it underscores the limitations of current evaluation methodologies in capturing the complexity of real-world conditions. Some challenging examples are shown in Fig 14. In cases involving low-quality

TABLE 8: Performance degradation on IJB-S (Survillance to Single protocol) as the gallery size increases with imposters sampled from an external dataset.

Gallery Setting	Gallery Size	Rank-1	Rank-5	TPIR @ FPIR=0.01
Baseline Gallery	202	62.0%	68.2%	46.1%
+1K External Imposters	1,202	56.1%	61.6%	43.7%
+5K External Imposters	5,202	51.1%	57.3%	40.8%
+10K External Imposters	10,202	48.4%	55.1%	38.0%

images (*e.g.,* IJB-S images), facial details are often absent, forcing models to rely on more visible soft biometric cues, such as beards or hair color, to make more informed guesses.

6.1 Recognition at Scale: Large Galleries

A significant limitation of current academic benchmarks is their inability to replicate the scale and complexity of practical biometric systems. Real-world FR deployments often involve galleries with millions to billions of identities, far exceeding the thousands typically evaluated in research benchmarks. For instance, India's Aadhaar national biometric system contains biometric data—including facial images—of over 1.4 billion individuals. At such immense scales, even minor degradations in recognition accuracy translate into substantial absolute numbers of false identifications or misses, impacting millions of users.

To highlight the disparity between benchmark evaluations and real-world scenarios, an illustrative experiment was conducted using the IJB-S dataset. The baseline IJB-S gallery (202 identities) was augmented with external imposters ranging from 1,000 to 10,000 identities. Tab. 8 presents the performance degradation observed with increasing gallery sizes. The results demonstrate a clear decline in recognition accuracy as the gallery grows, underscoring the challenge posed by real-world scenarios involving large galleries and numerous imposters.

This decline clearly highlights that conventional benchmarks do not fully represent operational scenarios, particularly when dealing with large-scale galleries containing diverse, unstructured, and noisy identities.

6.2 Multi-modal Recognition: Beyond Facial Imagery

As face recognition technologies move towards more challenging environments characterized by low resolution, extreme poses, occlusions, varying illumination conditions, and large-scale databases, reliance solely on facial imagery becomes increasingly insufficient. Real-world scenarios such as surveillance or public safety applications require robust identification techniques capable of handling severely degraded visual information.

To address these challenges, there is growing emphasis on integrating multiple biometric modalities. Incorporating additional cues such as body shape, gait, or even behavioral patterns significantly enhances recognition robustness. Traditionally outputs from multiple biometric modalites are combined using score fusion [160]. Score-level fusion combines similarity scores from multiple biometric modalities after similarity comparison. Common approaches include normalization methods like Z-score and min-max, likelihood ratio-based fusion, and simple aggregations such as mean, max, or min fusion [161]–[164]. These techniques

collectively improve robustness and accuracy in challenging recognition scenarios.

On the other hand, multi-modal biometrics can be conducted with the fusion at the input or feature level. SapiensID [41] proposes to combine face and body recognition under one model, offering particular promise in cross modality comparison, as body images offer larger visual area that can reliably distinguish individuals at lower image resolutions. Ultimately, the future of robust FR lies in embracing a multi-modal approach, harnessing complementary biometric modalities to overcome the inherent limitations of any single modality.

6.3 Capacity of Generative Models

An emerging question in synthetic dataset design is not just whether generated faces look realistic, but how many truly distinct and usable identities a generative model can produce. This is fundamentally a question of *identity capacity*: given a fixed number of real training images, how many well-separated subjects can a model generate?

DCFace [116], trained on 52k real face images, generates 20k new synthetic identities. In contrast, Vec2Face [122], trained on a much larger dataset (360k images), achieves up to 200k well-separated identities. This scaling behavior demonstrates that generative identity capacity is closely related to the diversity and richness of the real training data.

Recent work by Boddeti *et al.* [165] propose a principled statistical framework for estimating the upper bound of this capacity, framing it as a hyperspherical packing problem in the feature space of a face recognition model. They define capacity as the maximum number of identities that can be placed in this space without exceeding a predefined similarity threshold (related to a false acceptance rate). Their empirical estimates show that StyleGAN3 has a practical upper bound, approximately 1.43 million identities at a 0.1% FAR, which decreases sharply with stricter thresholds. For class-conditional models like DCFace, the capacity was significantly lower, due to its greater intra-class variation.

These results underscore an important insight: while generative models can amplify identity diversity, their capacity is not unlimited. The sampling distribution remains bounded by the identity entropy encoded during training. Thus, future research can aim to formalize these constraints, explore the theoretical upper bounds of novel identity generation, and propose methods for synthetic identities to be meaningfully distinct and diverse.

This raises a compelling question for the future: could synthetic datasets eventually surpass the utility of real datasets for training FR models? While current synthetic data often lags behind real data due to domain gaps and capacity limitations, the potential advantages of synthetic generation could be unparalleled control over attributes, scalability, and the ability to systematically generate data for rare conditions or underrepresented demographics [166].

Realizing this potential likely requires moving beyond current 2D generative paradigms. Integrating 3D modeling and rendering techniques stands out as a particularly promising direction. By leveraging explicit 3D representations, future generative pipelines could offer physically gr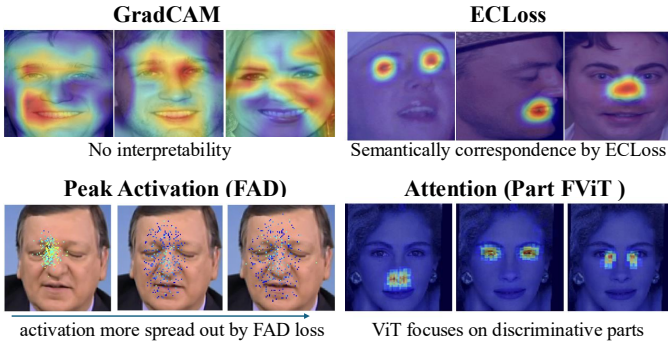ounded control over geometry, pose, illumination, and material properties (like skin texture and reflectance), potentially generating synthetic faces with greater realism, diversity, and, crucially, more distinct and well-separated identities than achievable through purely data-driven 2D synthesis alone. DigiFace [115] explores this direction and the key limitation is in the domain gap. Further research exploring these hybrid approaches, alongside developing better methods to measure and maximize the effective identity capacity, will be key to determining if and how synthetic data can ultimately overcome the limitations of, and perhaps even outperform, real-world data collection for advancing face recognition.

6.4 Role of Foundation Models in Face Recognition

Foundation models (FMs) are large-scale models pretrained on extensive image or text datasets for general-purpose tasks, rather than task-specific objectives such as face recognition. These models provide both pretrained weights and robust feature representations derived from broad visual or textual domains. Chettaoui *et al.* [167] offer a comprehensive overview of the role of foundation models in FR. Their findings indicate that, since FR models are traditionally trained on large-scale datasets, the advantages of using FMs are not clearly observed at the large scale training data.

However, fine-tuning FMs in low-data settings can significantly improve their performance [167]. Key comparative results are shown in Tab. 9. However, obtaining large scale training dataset is not difficult for face recognition, the benefit of FMs is still to be probed. Future work should focus on identifying which fine-tuning techniques, such as LoRA [113], and which foundation models, like CLIP [168] or DINOv2 [169], offer the best starting points for FR applications. Additionally, there is a need to understand why the advantages of foundation models diminish when training with large-scale FR datasets.

Recently, LAFS [68] introduces pretraining on unlabeled face data using foundation models, effectively learning critical face recognition representations and achieving strong few-shot performance. This highlights the value of specialized pretraining and motivates further exploration of domain-specific self-supervised learning (SSL) for developing specialized FR foundation models. It also raises questions about their interaction with general-purpose foundation models and potential reasons why the benefits of general models may diminish on large-scale FR datasets.

TABLE 9: Performance Comparison of Foundation Models (FMs) in Face Recognition under Different Training Regimes. Accuracies are averaged over LFW, CALFW, CPLFW, CFP-FP, and AgeDB. Rank is 16 for LoRA. CosFace [19] is used to train the models.

Model	Arch	Train Dataset	Train Setting	Avg. Acc. (%)
DINOv2	ViT-S	-	Pre-trained (Zero-shot FR)	64.70
CLIP	ViT-S	-	Pre-trained (Zero-shot FR)	82.64
ViT	ViT-S	1k IDs	Trained from Scratch	69.96
DINOv2	ViT-S	1k IDs	Fine-tuned (LoRA)	87.10
CLIP	ViT-S	1k IDs	Fine-tuned (LoRA)	90.75
ViT	ViT-S	CASIA-WebFace	Trained from Scratch	88.56
DINOv2	ViT-S	CASIA-WebFace	Fine-tuned (LoRA)	90.94
CLIP	ViT-S	CASIA-WebFace	Fine-tuned (LoRA)	92.13
ViT	ViT-L	WebFace4M	Trained from Scratch	95.65
DINOv2	ViT-L	WebFace4M	Fine-tuned (LoRA)	96.03
CLIP	ViT-L	WebFace4M	Fine-tuned (LoRA)	95.59

GradCAM

ECLoss

No interpretability

Semantically correspondence by ECLoss

Peak Activation (FAD)

Attention (Part FViT)

activation more spread out by FAD loss

ViT focuses on discriminative parts

Fig. 17: Comparison of visualization methods for face-related tasks. Grad-CAM [170] highlights broad, less interpretable regions, whereas ECLoss [171] enforces semantic correspondence with activations on meaningful facial parts. Peak Activation with FAD [74] shows that activations become more spread out across the face with the application of FAD loss. Attention maps from PartFViT [172] demonstrate that ViT models concentrate on discriminative facial parts, *e.g.*, eyes and nose.

6.5 Interpretability

Deep learning models used for face recognition are frequently viewed as black boxes. Their internal decision-making processes, involving millions of parameters, are inherently opaque, making it challenging to understand precisely why a particular decision (match vs. no-match, high vs. low quality score) was reached. This lack of transparency hinders trust, complicates debugging, and makes it difficult to assess its confidence, and impossible to be presented as evidence in the court. Several interpretability and explainability techniques are being applied or explored in the context of face recognition:
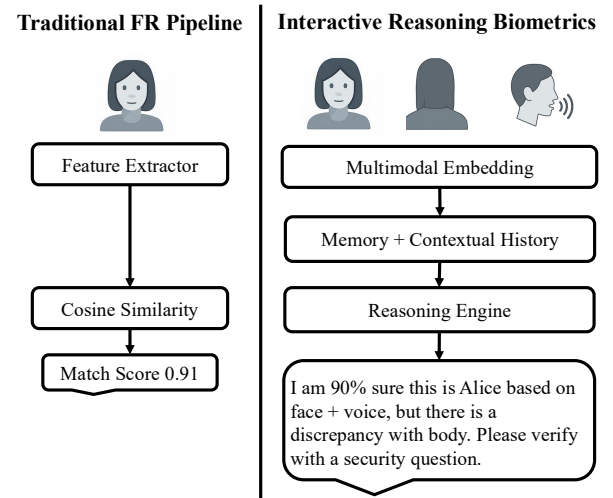
Saliency/Attribution Maps: Methods such as Grad-CAM [170] or SHAP [173] generate heatmaps highlighting the input image regions (pixels) most influential for the model's decision. For Transformer-based FR models, analyzing internal attention weights can offer insights into which parts of the input representation the model focuses on during processing [174]. FAD [74] proposes spatial activation diversity loss to learn more structured face activation. Some examples are shown in Fig. 17.

Concept-based Explanation: Moving beyond pixel importance, these approaches aim to link model decisions to higher-level, human-understandable concepts [175]. This could involve identifying the influence of specific facial attributes (*e.g.*, eye shape, nose structure) or using methods like ECLoss [171] to directly explain learned features without extra annotations. Some examples are shown in Fig. 17.

Counterfactual Explanation: These techniques explain a decision by showing minimal changes to the input that would alter the model's output [176], [177] (*e.g.*, "How would this face need to change to no longer match?").

Frequency-Domain Explanation: Another approach specifically investigated for FR involves analyzing the influence of different frequency components (*e.g.*, low vs. high frequencies representing coarse structure vs. fine details) in the input images on the matching decision [178], [179]. This provides a different perspective beyond spatial explanations.

LLM based Explanation: Recent advances demonstrate that large language models (LLMs) can significantly enhance the

Traditional FR Pipeline

Interactive Reasoning Biometrics

Feature Extractor

Cosine Similarity

Match Score 0.91

Multimodal Embedding

Memory + Contextual History

Reasoning Engine

I am 90% sure this is Alice based on face + voice, but there is a discrepancy with body. Please verify with a security question.

Fig. 18: Comparison between traditional face recognition (FR) pipelines and future paradigms integrating multimodal biometrics, reasoning, and explanations. Traditional FR systems output a simple match score based on feature similarity, whereas future systems will reason across modalities, dynamically assess uncertainty, and collaborate with humans through interpretable feedback loops.

interpretability of FR systems. Traditional interpretability tools like saliency maps or concept attributions often highlight important facial regions but fall short of conveying a full narrative understandable to users or forensic analysts. Recent approaches leverage LLMs to bridge this gap: models like XAI-CLIP [180] and vision-language alignment methods [181] can generate natural language explanations describing why two face images are matched or not, citing attributes such as "similar eyebrow curvature, matching nose bridge width, and aligned mouth corners." Concept bottleneck models [182], [183] explicitly model this process by learning interpretable features like glasses, facial hair before the recognition decision to be explained in terms of high-level, semantically meaningful traits rather than low-level pixel activations.

Beyond static descriptions, emerging research explores chain-of-thought prompting [184] to enable step-by-step reasoning over input features, providing deeper multi-stage narratives. In face recognition, a system could first note matching high-level features (*e.g.*, facial structure) before explaining divergences in fine details (*e.g.*, minor pose variation). Incorporating such language-based explanations can significantly improve transparency and accountability in FR applications. Future work should focus on grounding these explanations tightly to visual evidence, minimizing hallucination, and ensuring faithfulness even as models scale to large, foundation model-based face recognition systems.

What becomes increasingly important is not just producing a match score—but doing so with calibrated confidence and explainability. We can begin to envision systems that, in uncertain cases, simulate multiple identity hypotheses, ask follow-up questions, request additional evidence, or provide probabilistic reasoning behind their decisions. Fig. 18 illustrates this conceptual shift, contrasting traditional FR pipelines with emerging interactive, reasoning-based biometric systems designed for calibrated confidence and hu-

man collaboration. This marks a shift from fully automated systems to a more collaborative model, where humans and deep learning systems coexist in a feedback loop. What is previously missing is the model's ability to know when to intervene.

A critical aspect of enabling this loop is enhancing the model's ability to know when to initiate interaction or request intervention. While existing research in Face Image Quality Assessment (FIQA) provides valuable groundwork by estimating input quality or recognizability [154], [185], [186], the next step involves integrating these signals into broader reasoning frameworks. Moving forward, understanding when and how to involve a human—with clear reasoning—will be a critical step. To enable this, we'll also need new evaluation metrics—ones that go beyond accuracy, and reflect trust, interpretability, and decision quality in real-world scenarios.

7 SUMMARY

Over the past half-century, automated face recognition has transformed remarkably, moving from early geometric and handcrafted methods to sophisticated deep learning models that often surpass human performance on benchmarks. This paper chronicled this journey, emphasizing the deep learning revolution driven by innovations in network architectures (ResNets to ViTs), specialized loss functions (like margin-based softmax), and massive datasets (*e.g.*, Web-Face42M), which have yielded current SoTA capabilities.

Despite technical successes, evidenced by near-saturation on high-quality benchmarks (LFW [26], CFP-FP [27]) and strong results on challenging ones (IJB-C [24]) leading to widespread deployments (from mobile face unlock and payment to airport security and immigration), face recognition is not solved. Significant hurdles remain for real-world deployments, including accuracy at massive scales (billions of identities), robustness to degraded imagery (low quality, pose, occlusion) from CCTV videos, model interpretability and data privacy for public trust. The limitations of face-only data drive interest in multimodal biometrics, while synthetic data generation offers potential solutions but faces challenges like domain gaps and identity capacity.

Despite remarkable progress and widespread adoption of face recognition (FR) technologies, several key challenges remain. Distinguishing between identical twins is a significant limitation, especially for national-scale systems, given the U.S. twin birth rate of 31.2 per 1,000 live births (CDC, 2022). While FR systems claim robustness over decades, longitudinal performance can degrade, requiring re-enrollment. Current models rely on embeddings rather than human-interpretable features, which limits explainability. Additionally, there are risks where facial data is used beyond its original intent, raising privacy and ethical concerns. Ensuring template integrity, encrypted matching, and resistance to spoofing attacks are persistent challenges. Balancing recognition accuracy and latency is critical, especially for billion-scale deployments. Addressing these issues is essential for advancing FR responsibly.

Looking ahead, research can focus on overcoming these persistent challenges. Key directions include developing robust multimodal systems for person recognition (not just man collaboration. This marks a shift from fully automated multiple biomteric traits but also demographic and geolocation data), advancing explainable AI for transparency (*e.g.*, evidence in courts of law for conviction), pushing synthetic data boundaries for realism and capacity, and evaluating the role of foundation models. Addressing large-scale gallery performance, calibrated confidence, and human-AI collaboration are also crucial next steps. As face recognition technology proliferates, continued innovation must be carefully coupled with responsible development and ethical considerations. These include use consent for data capture and data sharing, maintaining data security, and using data only for the purpose for which it was collected.

REFERENCES

[1] T. Kanade, "Picture processing system by computer complex and recognition of human faces," Kyoto University, Tech. Rep., 1974.
[2] N. NZ, "Face detection vs facial recognition – what's the difference?" 2022.
[3] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, 2016.
[4] C. AI, "Why facial recognition is the best biometric," 2023.
[5] V. Technologies, "Pros and cons of facial recognition," 2023.
[6] K. Lai, L. Queiroz, V. Shmerko, K. Sundberg, and S. Yanushkevich, "Post-pandemic follow-up audit of security checkpoints," *IEEE Access*, 2023.
[7] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *T-PAMI*, 2007.
[8] L. Ding, C. Shu, C. Fang, and X. Ding, "Computers do better than experts matching faces in a large population," in *ICCI*, 2010.
[9] M. A. Turk, A. Pentland *et al.*, "Face recognition using eigenfaces." in *CVPR*, 1991.
[10] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV*, 2004.
[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.
[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.
[13] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.
[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NeurIPS*, 2012.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[17] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L2 hypersphere embedding for face verification," in *ACM MM*, 2017.
[18] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.
[19] 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 *CVPR*, 2018.
[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.
[21] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "CurricularFace: adaptive curriculum learning loss for deep face recognition," in *CVPR*, 2020.
[22] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *CVPR*, 2022.
[23] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "IARPA Janus Benchmark-B face dataset," in *CVPRW*, 2017.
[24] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark-C: Face dataset and protocol," in *ICB*, 2018.
[25] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain, "IJB–S: IARPA Janus Surveillance Video Benchmark," in *BTAS*, 2018.

[26] 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*, 2008.

[27] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *WACV*, 2016.

[28] T. Zheng and W. Deng, "Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing University of Posts and Telecommunications, Tech. Rep., 2018.

[29] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AGEDB: the first manually collected, in-the-wild age database," in *CVPRW*, 2017.

[30] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments," *arXiv*, 2017.

[31] Z. J. Wang, C. Kulkarni, L. Wilcox, M. Terry, and M. Madaio, "Farsight: Fostering responsible ai awareness during ai application prototyping," in *CHI*, 2024.

[32] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, and Y. Guo, "A survey of face recognition," *arXiv*, 2022.

[33] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, 2021.

[34] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *CVIU*, 2019.

[35] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, 2020.

[36] Y. Jing, X. Lu, and S. Gao, "3d face recognition: A comprehensive survey in 2022," *Computational Visual Media*, 2023.

[37] K. Kotwal and S. Marcel, "Review of demographic bias in face recognition," *arXiv*, 2025.

[38] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *T-PAMI*, 2022.

[39] M. Kim, Y. Su, F. Liu, A. Jain, and X. Liu, "Keypoint relative position encoding for face recognition," in *CVPR*, 2024.

[40] C. Wang, W. An, K. Jiang, X. Liu, and J. Jiang, "Llv-fsr: Exploiting large language-vision prior for face super-resolution," *arXiv preprint arXiv:2411.09293*, 2024.

[41] M. Kim, D. Ye, Y. Su, F. Liu, and X. Liu, "Sapiensid: Foundation for human recognition," in *CVPR*, 2025.

[42] Clearview AI, "Building a secure world, one face at a time," 2025.

[43] B. Klare, A. A. Paulino, and A. K. Jain, "Analysis of facial features in identical twins," in *IJCB*, 2011.

[44] T. Swearingen and A. Ross, "Lookalike disambiguation: Improving face identification performance at top ranks," in *ICPR*, 2021.

[45] E. Spearman, *Crime and Punishment in England: A Sourcebook*. Routledge, 1869.

[46] H. Faulds, "On the skin-furrows of the hand," *Nature*, 1880.

[47] W. Bledsoe, "Man-machine facial recognition," Panoramic Research, Inc., Tech. Rep., 1966.

[48] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *T-PAMI*, 1997.

[49] A. M. Martinez and A. C. Kak, "Pca versus lda," *T-PAMI*, 2001.

[50] L. Wiskott, J.-M. Fellous, N. Krüger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," in *Intelligent biometric techniques in fingerprint and face recognition*. Routledge, 2022.

[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[52] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *IJCV*, 2007.

[53] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv*, 2014.

[54] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.

[55] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *CVPR*, 2021.

[56] Fortune Business Insights, "Facial recognition market size, share & industry analysis [...] and regional forecast, 2025–2032," 2025, report ID: FBI101061.

[57] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *CVPR*, 2020.

[58] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2892–2900.

[59] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[60] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv*, 2017.

[61] W. Hu, Y. Huang, F. Zhang, R. Li, W. Li, and G. Yuan, "Seqface: make full use of sequence information for face recognition," *arXiv*, 2018.

[62] A. Calefati, M. K. Janjua, S. Nawaz, and I. Gallo, "Git loss for deep face recognition," *arXiv*, 2018.

[63] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *CVPR*, 2021.

[64] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *CVPR*, 2022.

[65] J. Zhou, X. Jia, Q. Li, L. Shen, and J. Duan, "Uniface: Unified cross-entropy loss for deep face recognition," in *ICCV*, 2023.

[66] X. Jia, J. Zhou, L. Shen, J. Duan *et al.*, "Unitsface: Unified threshold integrated sample-to-sample loss for face recognition," *NeurIPS*, 2023.

[67] K. Ahn, S. Lee, S. Han, C. Y. Low, and M. Cha, "Uncertainty-aware face embedding with contrastive learning for open-set evaluation," *T-IFS*, 2024.

[68] Z. Sun, C. Feng, I. Patras, and G. Tzimiropoulos, "Lafs: Landmark-based facial self-supervised learning for face recognition," in *CVPR*, 2024.

[69] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, 2020.

[70] Y. Su, M. Kim, F. Liu, A. Jain, and X. Liu, "Open-set biometrics: Beyond good closed-set models," in *ECCV*, 2024.

[71] M. Kim, F. Liu, A. K. Jain, and X. Liu, "Cluster and aggregate: Face recognition with large probe set," *NeurIPS*, 2022.

[72] J. Tu, X. Liu, and P. Tu, "On optimizing subspaces for face recognition," in *ICCV*, 2009.

[73] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *ACCV*, 2018.

[74] B. Yin, I. Tran, H. Li, X. Shen, and X. Liu, "Towards interpretable face recognition," in *ICCV*, 2019.

[75] H. Lin, H. Liu, Q. Li, and L. Shen, "Activation template matching loss for explainable face recognition," in *FG*, 2023.

[76] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition," in *ECCV*, 2022.

[77] J. Li, Z. Guo, H. Li, S. Han, J.-w. Baek, M. Yang, R. Yang, and S. Suh, "Rethinking feature-based knowledge distillation for face recognition," in *CVPR*, 2023.

[78] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[79] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang *et al.*, "Partial fc: Training 10 million identities on a single machine," in *ICCV*, 2021.

[80] X. Wang, Y. C. Wu, M. Zhou, and H. Fu, "Beyond surveillance: privacy, ethics, and regulations in face recognition technology," *Frontiers in Big Data*, 2024.

[81] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999.

[82] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *AVSS*, 2009.

[83] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, "Morphable face models—an open framework," in *FG*, 2018.

[84] J. Booth, A. Roussos, A. S. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *IJCV*, 2018.

[85] S. Ploumpis, S. Hu, Y. Xie, W. A. P. Smith, and S. Zafeiriou, "Combining 3d morphable models: A large scale face-and-head model," in *CVPR*, 2019.

[86] V. Abrevaya, S. Wuhrer, and E. Boyer, "Multilinear autoencoder for 3d face model learning," in *WACV*, 2018.

[87] T. Albrecht, K. Varanasi, V. Blanz, and C. Theobalt, "Statistical 3d shape models of human faces," *IJCV*, 2013.

[88] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM TOG*, 2017.

[89] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," in *ECCV*, 2018.

[90] H. Li, W. A. P. Smith, A. Tewari, H.-P. Seidel, and C. Theobalt, "Learning formation of physically-based face attributes," in *CVPR*, 2020.

[91] L. Wang, Y. Zhang, and Y. Liu, "Faceverse: A fine-grained and detail-controllable 3d face morphable model from a hybrid dataset," in *CVPR*, 2022.

[92] W. R. Schwartz and L. S. Davis, "Inface: A toolbox for illumination invariant face recognition," in *BTAS*, 2010.

[93] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *CVPR*, 2005.

[94] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3dface: A large-scale database of low-cost kinect 3d faces," in *ICB*, 2016.

[95] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[96] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[97] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.

[98] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[99] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv*, 2019.

[100] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[101] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *CVPRW*, 2019.

[102] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv*, 2020.

[103] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[104] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," *Scientific Reports*, 2024.

[105] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng, "Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation," *T-CSVT*, 2023.

[106] 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 *ICCV*, 2021.

[107] F. Liu, R. Ashbaugh, N. Chimitt, N. Hassan, A. Hassani, A. Jaiswal, M. Kim, Z. Mao, C. Perry, Z. Ren *et al.*, "Farsight: A physics-driven whole-body biometric system at large distance and altitude," in *WACV*, 2024.

[108] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *NeurIPS*, 2022.

[109] B. Lefaudeux, F. Massa, D. Liskovich, W. Xiong, V. Caggiano, S. Naren, M. Xu, J. Hu, M. Tintore, S. Zhang, P. Labatut, D. Haziza, L. Wehrstedt, J. Reizenstein, and G. Sizov, "xformers: A modular and hackable transformer modelling library," 2022.

[110] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *CCBR*, 2018.

[111] G. Kim, G. Park, S. Kang, and S. S. Woo, "S-vit: Sparse vision transformer for accurate face recognition," in *SAC*, 2023.

[112] K. Narayan, N. G. Nair, J. Xu, R. Chellappa, and V. M. Patel, "Petalface: Parameter efficient transfer learning for low-resolution face recognition," in *WACV*, 2025.

[113] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, 2022.

[114] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "Synface: Face recognition with synthetic data," in *ICCV*, 2021.

[115] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "Digiface-1m: 1 million digital face images for face recognition," in *WACV*, 2023.

[116] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *CVPR*, 2023.

[117] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[118] J. N. Kolf, T. Rieber, J. Elliesen, F. Boutros, A. Kuijper, and N. Damer, "Identity-driven three-player generative adversarial network for synthetic-based face recognition," in *CVPR*, 2023.

[119] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer, "Exface-gan: Exploring identity directions in gan's learned latent space for synthetic identity generation," in *IJCB*, 2023.

[120] F. Boutros, M. Huber, A. T. Luu, P. Siebke, and N. Damer, "Sface2: Synthetic-based face recognition with w-space identity-driven sampling," *T-BIFS*, 2024.

[121] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2face: A foundation model for id-consistent human faces," in *ECCV*, 2024.

[122] H. Wu, J. Singh, S. Tian, L. Zheng, and K. W. Bowyer, "Vec2face: Scaling face dataset generation with loosely constrained vectors," *arXiv*, 2024.

[123] S. Gong, V. N. Boddeti, and A. K. Jain, "On the intrinsic dimensionality of image representations," in *CVPR*, 2019.

[124] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.

[125] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*, 2020.

[126] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *CVPR*, 2018.

[127] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim, "Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model," in *ECCV*, 2018.

[128] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *CVPR*, 2019.

[129] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM TOG*, 2018.

[130] F. Liu, M. Kim, A. Jain, and X. Liu, "Controllable and guided face synthesis for unconstrained face recognition," in *ECCV*, 2022.

[131] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[132] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *NeurIPS*, 2021.

[133] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[134] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[135] H. Ye *et al.*, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv*, 2023.

[136] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia *et al.*, "Frcsyn challenge at wacv 2024: Face recognition challenge in the era of synthetic data," in *WACV*, 2024.

[137] I. DeAndres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fierrez, J. Ortega-Garcia *et al.*, "Frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data," in *CVPR*, 2024.

[138] V. Asnani, X. Yin, T. Hassner, S. Liu, and X. Liu, "Proactive image manipulation detection," in *CVPR*, 2022.

[139] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *CVPR*, 2017.

[140] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," *arXiv*, 2018.

[141] S. Gong, Y. Shi, N. D. Kalka, and A. K. Jain, "Video face recognition: Component-wise feature aggregation network (c-fan)," in *ICB*, 2019.

[142] B. Jawade, D. D. Mohan, D. Fedorishin, S. Setlur, and V. Govindaraju, "Conan: Conditional neural aggregation network for unconstrained face feature fusion," in *IJCB*, 2023.

[143] B. Jawade, A. Stone, D. D. Mohan, X. Wang, S. Setlur, and V. Govindaraju, "Proxyfusion: Face feature aggregation through sparse experts," *NeurIPS*, 2024.

[144] A. Nanduri and R. Chellappa, "Template-based multi-domain face recognition," in *IJCB*, 2024.

[145] W. Zhao, X. Zhu, H. Shi, X.-Y. Zhang, G. Zhao, and Z. Lei, "Global cross-entropy loss for deep face recognition," *T-IP*, 2025.

[146] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Misclassified vector guided softmax loss for face recognition," in *AAAI*, 2020.

[147] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun, "Transface: Calibrating transformer training for face recognition from a data-centric perspective," in *ICCV*, 2023.

[148] Q. Wang and G. Guo, "Cqa-face: Contrastive quality-aware attentions for face recognition," in *AAAI*, 2022.

[149] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *CVPR*, 2020.

[150] Y. Wang and W. Wei, "Local and global feature attention fusion network for face recognition," *Pattern Recognition*, 2025.

[151] M. S. E. Saadabadi, S. R. Malakshan, A. Dabouei, and N. M. Nasrabadi, "Aroface: Alignment robustness to improve low-quality face recognition," in *ECCV*, 2024.

[152] J. Dan, Y. Liu, J. Deng, H. Xie, S. Li, B. Sun, and S. Luo, "Topofr: A closer look at topology alignment on face recognition," *NeurIPS*, 2024.

[153] N. A. Talemi, H. Kashiani, and N. M. Nasrabadi, "Catface: Cross-attribute-guided transformer with self-attention distillation for low-quality face recognition," *T-BIFS*, 2024.

[154] J. C. L. Chai, T.-S. Ng, C.-Y. Low, J. Park, and A. B. J. Teoh, "Recognizability embedding enhancement for very low-resolution face recognition and quality estimation," in *CVPR*, 2023.

[155] Y. Shi and A. K. Jain, "Probabilistic face embeddings," in *ICCV*, 2019.

[156] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011.

[157] NIST, "Face technology evaluations: Frte/fate," 2024.

[158] ——, "Nist evaluation shows advance in face recognition software's capabilities," 2018.

[159] ——, "Face recognition vendor test (frvt)," 2024.

[160] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, 2019.

[161] K. Nandakumar, Y. Chen, S. C. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *T-PAMI*, 2007.

[162] N. Poh and J. Kittler, "A unified framework for biometric expert fusion incorporating quality measures," *T-PAMI*, 2011.

[163] N. Poh, J. Kittler, and T. Bourlai, "Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization," in *BTAS*, 2007.

[164] M. Vatsa, R. Singh, and A. Noore, "Integrating image quality in 2ν-svm biometric match score fusion," *International Journal of Neural Systems*, 2007.

[165] V. N. Boddeti, G. Sreekumar, and A. Ross, "On the biometric capacity of generative face models," in *IJCB*, 2023.

[166] S. Um and J. C. Ye, "Self-guided generation of minority samples using diffusion models," in *ECCV*, 2024.

[167] T. Chettaoui, N. Damer, and F. Boutros, "Froundation: Are foundation models ready for face recognition?" *IVC*, 2025.

[168] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[169] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv*, 2023.

[170] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *IJCV*, 2020.

[171] H. Lin, H. Wu, and Y. Liu, "Activation template matching loss for explainable face recognition," *arXiv*, 2024, (Note: Check for final publication venue).

[172] Z. Sun and G. Tzimiropoulos, "Part-based face recognition with vision transformers," *arXiv*, 2022.

[173] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.

[174] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos, "Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions," *IEEE Access*, 2024.

[175] G. Mikriukov, J. H. Lee, G. Schwalbe, and S. Wermter, "Explainable concept generation through vision-language preference learning," in *NeurIPS Workshop on Interpretable AI*, 2024.

[176] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, 2021.

[177] B. Sobieski and P. Biecek, "Global counterfactual directions," in *ECCV Workshop on Explainable Computer Vision*, 2024.

[178] M. Huber and N. Damer, "Beyond spatial explanations: Explainable face recognition in the frequency domain," *arXiv*, 2024.

[179] ——, "Frequency matters: Explaining biases of face recognition in the frequency domain," *arXiv*, 2025.

[180] A. Yao *et al.*, "Xai-clip: Explainable vision-language pretraining," in *CVPR*, 2024.

[181] M. Liu *et al.*, "Interpretable predictions via vision-language alignment," in *ICLR*, 2024.

[182] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *ICML*, 2020.

[183] A. Dombrowski *et al.*, "Faithful vision-language concept bottlenecks," in *ICLR*, 2024.

[184] J.-B. Alayrac *et al.*, "Chain-of-thought prompting for visual reasoning with flamingo," in *NeurIPS*, 2024.

[185] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "Crfiqa: face image quality assessment by learning sample relative classifiability," in *CVPR*, 2023.

[186] F.-Z. Ou, C. Li, S. Wang, and S. Kwong, "Clib-fiqa: face image quality assessment with confidence calibration," in *CVPR*, 2024.