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

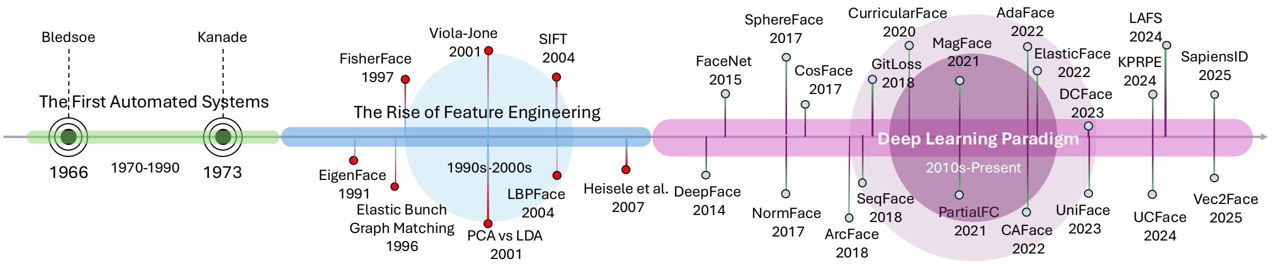


Fig. 1: Historical evolution of automated face-recognition research over the past five decades. The timeline illustrates key milestones—from early geometric models (1960s–1980s), through the feature-engineering era (1990s–2000s), to modern deep-learning systems (2010s–present). Colors indicate each era; transparency is for visual purposes only.

Abstract—Over the past five decades, automated face recognition (FR) has progressed from handcrafted geometric and statistical approaches to advanced deep learning architectures that now approach, and in many cases exceed, human performance. This paper traces the historical and technological evolution of FR, encompassing early algorithmic paradigms through to contemporary neural systems trained on extensive real and synthetically generated datasets. We examine pivotal innovations that have driven this progression, including advances in dataset construction, loss function formulation, network architecture design, and feature fusion strategies. Furthermore, we analyze the relationship between data scale, diversity, and model generalization, highlighting how dataset expansion correlates with benchmark performance gains. Recent systems have achieved near-perfect large-scale identification accuracy, with the leading algorithm in the latest NIST FRTE 1:N benchmark reporting a FNIR of 0.15 percent at FPIR of 0.001 on a gallery of over 10 million identities . We delineate key open problems and emerging directions, including scalable training, multi-modal fusion, synthetic data, and interpretable recognition frameworks.

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). Fig.1 presents a timeline summarizing major milestones across five decades of automated FR research.

FR has become one of the most prevalent biometric modalities employed today [2]. 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 [3]. Face acquisition can be achieved using low-cost cameras, making it accessible and scalable across diverse applications [2], [3]. The non-contact nature of FR offers hygienic advantages, especially salient in a post-pandemic era [4], [5]. Furthermore, FR 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 [2].

Notably, even prior to deep mode, automated FR 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 [6]. Further research in 2010 revealed that a specific algorithm exceeded the accuracy of thousands of customs inspectors when dealing with straightforward facial comparisons in operational settings [7]. Some examples of challenging pairs are given in Fig. 3. While these early

a) Phone b) Airport c) Boarding d) Surveillance e) Door bell

Fig. 2: Examples of real-world 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.

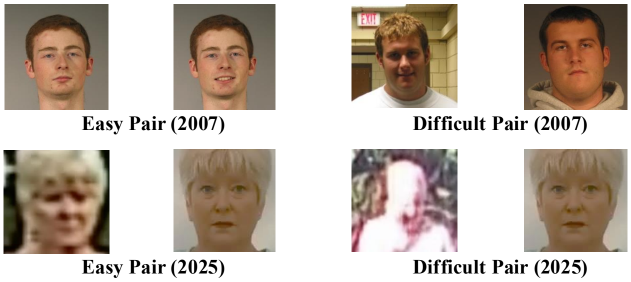
successes are significant, the field has undergone a dramatic transformation since then, fueled by advancements in deep networks and the availability of large-scale datasets. This paper will revisit this critical juncture, exploring how the field has evolved and the extent to which current FR technology has surpassed human abilities across a wider range of challenging conditions.

The progress in FR has been driven by advances in computing power, the availability of large-scale datasets, and a shift in understanding facial image representation and comparison. Early methods relied on handcrafted features capturing facial structure, including appearance-based approaches like Eigenfaces [8] and texture-based ones like Local Binary Patterns (LBP) [9]. Machine learning methods such as statistical models and cascaded classifiers [10] improved detection accuracy, enabling more reliable FR pipelines. The advent of deep learning introduced a paradigm shift, allowing algorithms to learn discriminative features directly from large training data, as shown by DeepFace [11], DeepID [12], and FaceNet [13]. These developments greatly enhanced FR performance in accuracy and efficiency but also introduced challenges related to data bias and robustness under pose and occlusion variations.

This paper will trace the historical development of FR, 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 [14], [15], loss function design [16]–[21], and the utilization of increasingly large and diverse datasets [22]–[29]. We will address the emerging role of synthetic data 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 field. Unless otherwise specified, this survey focuses primarily on 2D visible-light FR, with brief discussions of 3D and other sensing modalities where relevant.

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 State-of-the-Art (SoTA) within its broader trajectory. Recent surveys have also addressed specialized topics within the broader face analysis domain, such as 3D FR [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 FR itself, with slight mention of important related tasks.


Easy Pair (2007) **Difficult Pair (2007)**

Easy Pair (2025) **Difficult Pair (2025)**

Fig. 3: Visualization of easy and difficult face pairs for 2007 (top) and 2025 (bottom) FR, where difficulty is defined by the pairs that (State-of-the-Art) SoTA models of the time struggle to correctly identify [6], [30]. 2007 subjects and images are from [6]. 2025 subject and images are from BRIAR dataset [31]

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 FR system operates in two phases: *enrollment* and *recognition*[1]. Enrollment captures and stores facial data as a baseline, while recognition uses it to confirm or determine identity. Recognition involves two tasks: **Verification**, a one to one comparison verifying if a face matches a claimed identity; and **Identification**, a one to many search identifying a face from a database of enrolled faces.

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 facial images [39] .

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

[1] *Comparison* and *recognition* are used interchangeably, with recognition applying a threshold to a similarity score.

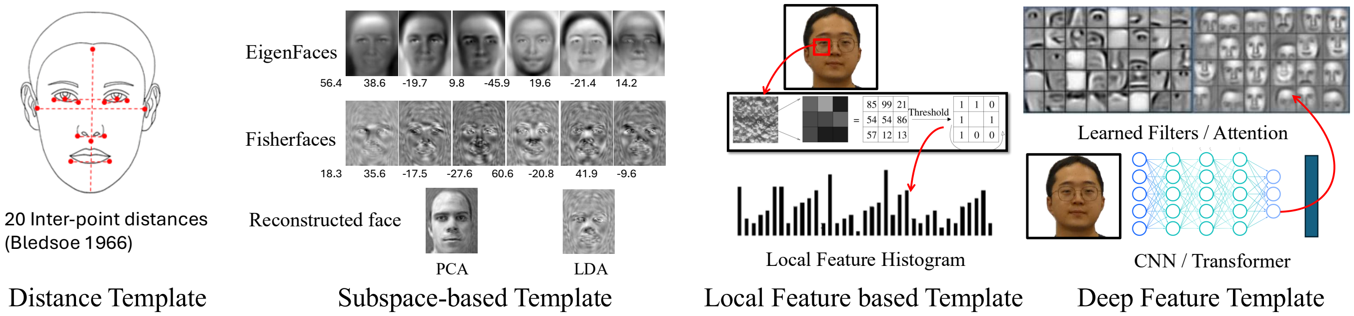


Fig. 4: Evolution of face template representations. Early methods used geometric distances between facial landmarks (red dots and dashed lines), then subspace projections (*e.g.*, PCA, LDA), and local texture histograms from small patches (red square). Modern approaches employ CNNs and Transformers that learn deep feature embeddings directly from data.

characteristics are distilled from the facial image, creating a compact and informative 'template'. This template is stored in a database and the original image can optionally be discarded 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, semantic attributes 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 FR 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 [25], [27]–[29]. Variations in image quality (*e.g.*, resolution, blur, and noise) further exacerbate the problem [22]–[24]. Early FR systems often struggled to address these challenges, necessitating carefully controlled imaging environments with constrained pose and illumination conditions [8], [9].

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 [40]. Moreover, non-related individuals may coincidentally look alike (so-called doppelgangers) further complicating the discrimination task [41]. 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 [12], [13], [16], [18]–[21]. 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.

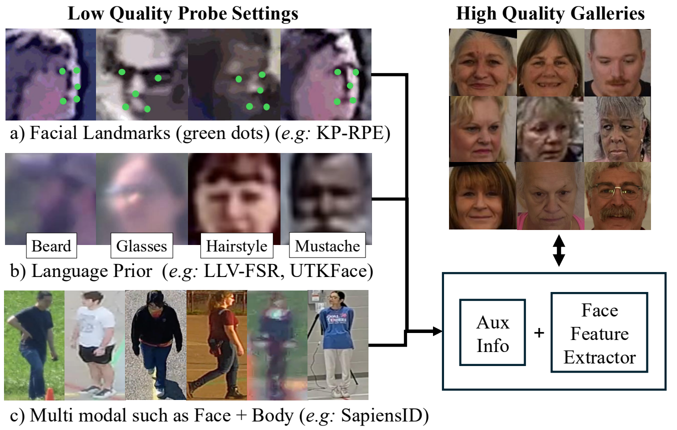
The human face encodes a wide spectrum of information. As illustrated in Fig. 6, a single image can reveal identity, demographic traits, physical attributes, and social cues. However, deep learning-based FR systems typically do not treat these aspects independently. Instead, they amalgamate all visible 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 FR 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 [42]. 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 [43]. These pre-digital efforts established the conceptual foundation for biometric identification, which would later evolve into automated computer vision systems.

The First Automated Systems (1960s–1980s). Early computer-based FR systems emerged in the 1960s, notably with the work of Bledsoe, who used manually annotated facial landmarks and feature distances to identify individuals [44]. Kelly [45] attempted to automate facial identification by using computer vision to measure distances between key facial landmarks, such as the eyes and nose. A major milestone followed with Kanade's 1973 dissertation, which presented the first fully automated FR system [1].

The Rise of Feature Engineering (1990s–2000s). The 1990s mark a paradigm shift in FR, moving from handcrafted geometric features to holistic, appearance-based representations. A seminal contribution is the Eigenfaces by Turk and Pentland [8], which leverages PCA to represent faces as linear combinations of orthogonal basis images. This



a) Facial Landmarks (green dots) (*e.g:* KP-RPE)

b) Language Prior (*e.g:* LLV-FSR, UTKFace)

c) Multi modal such as Face + Body (*e.g:* SapiensID)

Fig. 5: Illustration of template enhancement by incorporating auxiliary information such as facial landmarks (*e.g.*, KP-RPE [50]), language priors (*e.g.*, LLV-FSR [51]), and multi-modal cues (*e.g.*, SapiensID [52]).

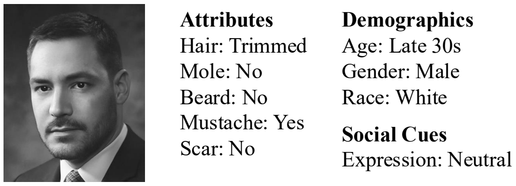
approach enables more compact and discriminative facial representations. However, Eigenfaces exhibit limitations in handling variations in lighting and facial expression. To address this, Belhumeur *et al.* [46] introduce Fisherfaces, which apply linear discriminant analysis (LDA) to better separate individuals, improving robustness under varying illumination. This PCA-versus-LDA debate is further explored by Martinez and Kak [47], who highlight the strengths and weaknesses of both in practical scenarios.

Building on holistic approaches, researchers aimed to model both facial appearance and shape variability. Lanitis, Taylor, and Cootes (1995) [48] proposed a PCA-based framework combining geometry and grey-level appearance for automatic face analysis. Moghaddam and Pentland (1997) [49] introduced a probabilistic eigenspace, modeling object classes as Gaussian or Mixture of Gaussians and framing recognition as maximum likelihood estimation. These works established a unified, data-driven foundation for appearance-based recognition.

Model-based techniques such as Elastic Bunch Graph Matching [53] 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) [9] and Scale-Invariant Feature Transform (SIFT) [54], 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 [10], 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.* [55] proposed 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.

The FERET program [56] advanced evaluation protocols by introducing standardized datasets and methods for comparing FR algorithms. Its gallery probe design and reporting metrics became key benchmarks in the 1990s and early 2000s, forming the foundation for empirical progress.



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

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

Social Cues
Expression: Neutral

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

Deep Learning Paradigm (2010s–Present). The advent of deep learning [14], [15] in the 2010s revolutionized 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 [14] and ResNet [15] demonstrate the power of convolutional networks for image recognition, paving the way for their adoption in FR. The ImageNet dataset [14] provides a crucial resource for pre-training these large models, which were then fine-tuned for FR tasks.

Early pioneering works like DeepFace [11] and FaceNet [13] demonstrate the potential of deep learning for FR, 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 face grouping.

A key area of innovation in deep FR has been the development of 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 [16], SphereFace [17], CosFace [18], ArcFace [19], CurricularFace [20] and Adaface [21], 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 [57], VGGFace [58], MS1M [58], and WebFace260M [59]. 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, FR 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

Deep face recognition has advanced through improved loss functions, large diverse datasets, and better neural architectures, enabling more discriminative representations.

4.1 Loss Functions

Before deep learning, FR relied on analytical methods lacking desired discriminative power. Eigenfaces [8] used Principal Component Analysis to maximize variance, and Fisherfaces [46] applied Linear Discriminant Analysis to

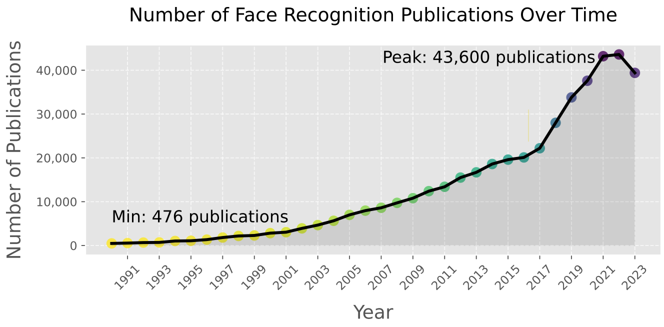


Fig. 7: Number of FR publications over time. Research activity in the field grew steadily until the early 2010s, followed by an explosive increase coinciding with the rise of deep learning. The peak in 2022 reflects the technology's mainstream adoption, though recent years suggest a slight cooling-off 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 [60].

enhance class separability, both requiring costly covariance matrix diagonalization [47]. Deep learning replaced these with gradient based optimization and specialized loss functions [16]–[21] that directly enforce inter-class separability and intra-class compactness.

The choice of loss function is critical in training deep FR 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 [13] 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 [61], 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) [73] 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 [71]. Other works adapt contrastive ideas for specific goals: Open-Set Biometrics [74] focuses on improving open-set performance by explicitly minimizing scores between non-mated pairs, while CAFace [75] uses a contrastive-style cosine similarity loss to enforce consistency between embeddings of low-quality images and their high-quality counterparts, promoting quality invariance. Related efforts also explore optimizing embedding spaces to better align with recognition objectives, such as in [76], where features are learned to directly improve performance metrics for identification and verification separately.

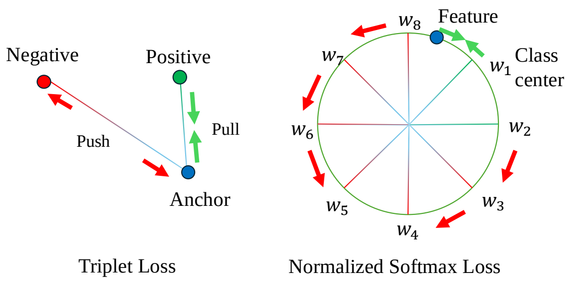


Fig. 8: Comparison of loss function paradigms in deep FR. Left: Triplet loss in contrastive learning reduces intra-class distance (pull positive training samples closer to anchor) while increasing inter-class distance (push negative training samples away). Right: Normalized Softmax loss maps features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

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 (same person are close together) while simultaneously maximizing inter-class separation (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:

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

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 FR where we encounter identities not seen during training.

An early work moving in this direction is Center Loss [63], 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 [17], 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:

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

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

Name	Year	Pros	Cons
DeepID2+ [62]	2014	Joint ID + verification loss for robust features	Complex training with dual supervision losses
CenterFace [63]	2016	Center loss enhances intra-class compactness	No explicit inter-class separation; needs tuning
SphereFace [17]	2017	Angular margin enforces hyperspherical separation	Training instability from angular multiplicity
L2-Face [64]	2017	L2 norm constraint improves angular discrimination	Needs careful radius tuning; no margin enforcement
ArcFace [19]	2018	Additive angular margin boosts inter-class separation	Fixed margin may hurt low-quality samples
CosFace [18]	2018	Cosine margin improves class separability stably	Uniform margin not adaptive; needs tuning
SeqFace [65]	2018	Sequence-aware loss improves temporal supervision	Needs sequence data; dual loss increases complexity
Git Loss [66]	2018	Unified softmax + center loss boosts discrimination	Extra tuning and complexity with marginal gain
MagFace [67]	2021	Feature norm models quality for adaptive margin	Complex loss and quality-norm assumptions
AdaFace [21]	2022	Dynamic margin based on feature norm quality	Relies on norm-quality link and tuning
ElasticFace [68]	2022	Elastic margin adapts to feature variability	Stochastic margins add tuning and training cost
UniFace [69]	2023	Similarity threshold improves verification alignment	Global constraints increase optimization cost
UniTSFace [70]	2023	Sample-to-sample loss optimizes verification	Pairwise loss and threshold learning cause overhead
UCFace [71]	2024	Uncertainty and probability density aware contrastive learning	Cannot be used by itself, must be accompanied by margin loss
LAFS [72]	2024	Landmark based SSL pretraining helps FR	Loss depends on pretrained model and the landmark quality.

Building on this normalized angular space, the key innovation is the introduction of explicit margins to make the learning objective more stringent. CosFace [18] introduces an additive cosine margin (m) by modifying the target logit to $s \cdot (\cos\theta_{y_i} - m)$. ArcFace [19] 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 [67] 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 [21] 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 [25] and CFP-FP [26] 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,TinyFace or BRIAR [24], [77], [78]. The IJB-S and BRIAR datasets are limited-access evaluation sets developed under U.S. Government research programs and are distributed selectively for research use.

Further refinements continued to explore margin dynamics; for example, ElasticFace [68] introduces randomized margins for greater flexibility during training, while UniFace [69] proposes the Unified Cross-Entropy (UCE) loss specifically aiming to guarantee a clear separation threshold between positive and negative pairs.

Margin-based softmax variants [19], [21], [69] currently dominate SoTA results. Contrastive 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.* [79] 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 [80], 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) [81] transfers teacher attention maps to guide the student's focus. For efficient large-scale training, Li *et al.* [82] 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

Before deep learning, FR was limited by small datasets. Prior to the 1996 FERET database, which had 1,199 individuals and became a major benchmark [56], [83], most studies used fewer than 50 subjects, restricting generalization under unconstrained conditions. These small datasets sufficed for classical approaches like PCA based Eigenfaces [8], which relied on analytical solutions. The advent of deep learning [11], [13], [14] created a vast need for data, as neural networks require far more samples to learn model parameters [58], [59], [84].

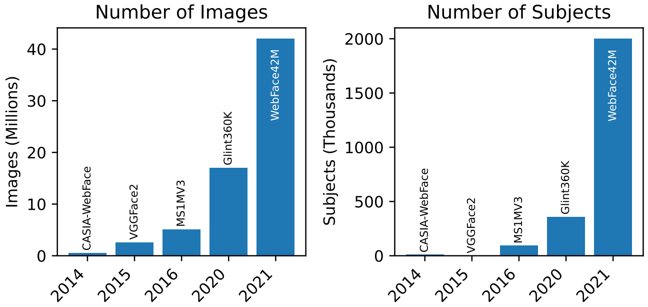


Fig. 9: Plots showing the growth of FR datasets over time. The left illustrates the number of images (in millions), and the right shows the number of subjects (in thousands) for each dataset.

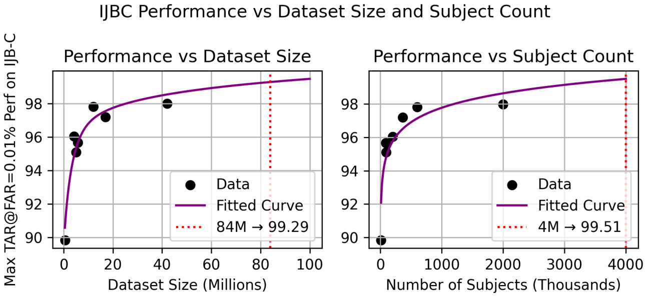


Fig. 10: Recognition performance on IJB-C dataset as a function of training dataset size (left) and training number of subjects (right). The dots show the the best publically available algorithms' performance for the given training dataset. Curves are fitted using the logarithmic function. Both increasing 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.

The availability and scale of training data have been pivotal factors driving the remarkable progress in deep learning [14]. Publicly available large-scale face datasets (*e.g.*, MS-Celeb-1M, VGGFace2) spurred rapid advancements in the mid-2010s. Fig. 9 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 [57] offer around half a million images from ten thousand subjects. In contrast, subsequent collections such as MS1MV2/V3 [84], Glint360K [85], and particularly the WebFace series [59], 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 discontinued by the dataset creators. Some popular public datasets and pretrained model checkpoints are available in this link.

In Fig. 10, 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 these datasets, 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 [59], employ iterative self-labeling and retraining of specialized labeler models to improve the quality of these pseudo-labels over multiple cycles. Since benchmark datasets [22], [23], [25], [26], [28] are also curated 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 [86], 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 [87]–[98]. 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 [99]. Notably, datasets such as FRGC [99] and Lock3DFace [100], 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 FR applications, such as in law enforcement, immigration, or airport screening, RGB cameras offer a more practical solu-

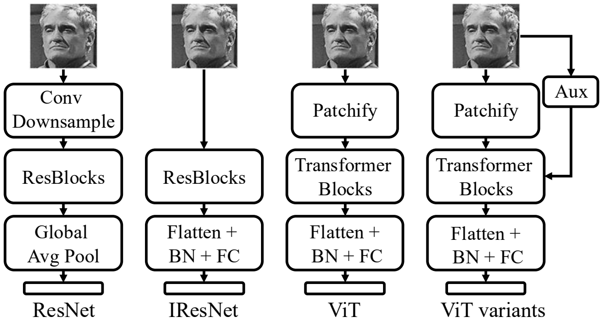


Fig. 11: Comparison of architectures used in FR. From left to right: ResNet [15] is an architecture used for classification [14]. IResNet [19] modifies this by removing downsampling and using feature flattening, batch normalization (BN), and fully connected (FC) layers that are helpful for metric learning [107]. Vision Transformer [108] (ViT) replaces convolutions with a patchify operation and transformer blocks; ViT variants further extend this by incorporating auxiliary information such as facial keypoints [50], [105] to improve learning.

tion 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 [14] on the ImageNet challenge, quickly permeates the field of FR. Early deep FR models adapt existing CNN architectures designed for general object recognition.

Architectures like GoogLeNet [101] demonstrate the power of increased network depth and led to its adoption in FaceNet [13]. The introduction of Residual Networks (ResNets) [15] 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 [102]), become the popular backbone for many SoTA FR systems developed in the late 2010s [18], [19]. ArcFace [19]'s adoption of input size 112 × 112 leads to the widely used IR-ResNet 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 [57] utilize Multi-task Cascaded Convolutional Network (MTCNN) [103], which jointly performs face detection and landmark localization via a series of cascaded networks. With the advent of single-stage detectors like SSD [104], more efficient and accurate methods emerge. Notably, RetinaFace [105] has become a popular solution, offering precise face detection and alignment. When trained on strong datasets such as WiderFace [106] and paired with an improved backbone, RetinaFace is a robust choice for preprocessing large-scale face datasets [59]. 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) [108] have emerged as a powerful alternative to CNNs. ViTs process images by dividing them into patches, and feeding the resulting sequence into a Transformer encoder [109]. 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 FR domains [50], [110], [111]. And adoption of ViT in FR implies adoption of advances around ViT. SwinFace [111] is an application of Swin Transformer [112]. KP-RPE [50] 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 [50]. 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, compute resources, and deployment setting.

FR 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, IResNet50 can process over 1400 images/second, while ViT Base handles about 640 images/second. Unlike academic research, industry models or government vendor models [113] can use model ensembles, further increasing the load. Preprocessing steps like face detection and alignment add to the computational cost. Note that ViT backbones can benefit from research that speed up ViT inference [114], [115].

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 [116] 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× smaller). S-ViT [117] 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

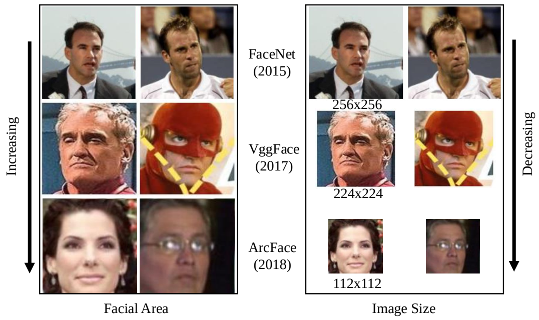


Fig. 12: Comparison of FR in terms of facial contextual area and image resolution. Models have evolved to focus on more tightly cropped regions while reducing resized input image size (256×256 to 112×112), enabling more efficient feature extraction.

pushing the performance boundaries of automated FR.

Fine-tuning is crucial for adapting FR models to new domains, especially under quality mismatches [113] (*e.g.* low vs high quality images). Instead of full fine-tuning, which risks catastrophic forgetting, recent work like PETALface [118] uses LoRA [119], 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.

Also, FaceNet [13] 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.

The optimal embedding dimensionality varies with the scale and diversity of the training data. The embedding space must fit all subjects within a hypersphere while preserving sufficient inter-class distances. As the number of identities increases, a higher-dimensional feature space may be needed to maintain discriminability. Hence, smaller dimensions (e.g., 128) may suffice for moderate datasets, whereas larger datasets benefit from higher-dimensional representations to preserve subject separation.

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 [120]–[122].

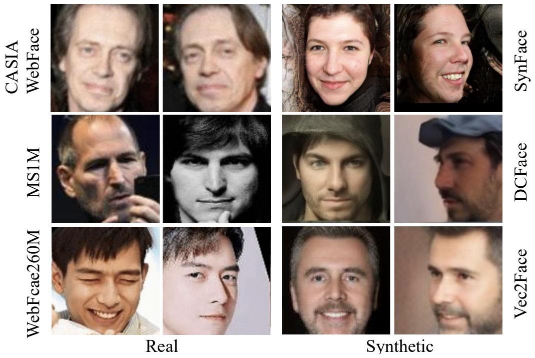


Fig. 13: Examples of real and synthetic datasets. Real datasets evolve from the primarily frontal CASIA-WebFace to the large-scale, diverse WebFace260M, including wide variations in pose and expression. Synthetic datasets also advance, focusing on improving diversity and maintaining identity consistency.

The generation of photorealistic synthetic faces has been significantly advanced by deep generative models, particularly Generative Adversarial Networks (GANs) [129]. Variants such as StyleGAN [123], [130] 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 [131]–[134]. For instance, CFSM [135] leverages GANs to synthesize faces with diverse styles, aiding in the generation of richly varied datasets.

Recently, Diffusion Models [136] 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 [137], [138]. ControlNet [139] and IP-Adapter [140] 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 [120] 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. 2. Since then, mulitple works have attempted to reduce the gap.

Following SynFace, rapid advances have sought to bridge the gap between synthetic and real face datasets. Diffusion-based models such as DCFace [122] separate identity and style conditions to produce identity-consistent and diverse subjects, while Arc2Face [127] builds on pretrained Stable Diffusion to exploit the generalization power of foundation models. Vec2Face [128] shows that GAN-based synthesis remains competitive when guided by a FR feature space, underscoring the importance of identity disentanglement. These methods exemplify the trend toward improving realism, diversity, and identity consistency to narrow the gap with real data. Tab. 2 presents the performance of recent synthetic face datasets on FR.

There is growing interest in synthetic FR challenges. FRCSyn series [141], [142], show that while synthetic-only training trails real data slightly, it reduces demographic bias and improves robustness to pose, age, and occlusion. Figure 13 compares real and synthetic datasets. Adding

TABLE 2: Comparison of synthetic face training datasets for FR 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 [120]–[122].

Methods	**Venue**	**Generator Train Dataset**	**# images (# IDs× imgs/ID)**	**LFW**	**CFP-FP**	**CPLFW**	**AgeDB**	**CALFW**	**Avg**	**Gap to Real**
SynFace [120]	ICCV21	FFHQ [123]	0.5M (10K × 50)	91.93	75.03	70.43	61.63	74.73	74.75	26.04
DigiFace [121]	WACV23	511 3D Scans	1M (10K × 100)	95.40	87.40	78.87	76.97	78.62	83.45	11.34
DCFace [122]	CVPR23	CASIA-WebFace (FFHQ)	0.5M (10K × 50)	98.55	85.33	82.62	89.70	91.60	89.56	5.23
IDnet [124]	CVPR23	CASIA-WebFace [57]	0.5M (10K × 50)	92.58	75.40	74.25	63.88	79.90	79.13	15.66
ExFaceGAN [125]	IJCB23	CASIA-WebFace	0.5M (10K × 50)	93.50	73.84	71.60	78.92	82.98	80.17	14.62
SFace2 [126]	TBIS24	CASIA-WebFace	0.6M (10K × 60)	96.50	77.11	74.60	77.37	83.40	81.62	13.17
Arc2Face [127]	ECCV24	WF42M [59] (Stable Diffusion)	0.5M (10K × 50)	98.81	**91.87**	85.16	90.18	92.63	91.73	3.06
Vec2Face [128]	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

synthetic data, such as Vec2Face [128] to CASIA-WebFace, can raise average verification accuracy by about 1.00%.

Despite its promise, using purely synthetic data to train SoTA FR models faces key challenges, primarily the *domain gap* between synthetic and real images. Models trained only on synthetic data often struggle to generalize to real-world due to subtle differences in texture, lighting, or artifacts from the generation process. Achieving sufficient diversity and realism, especially in capturing identity nuances under varying conditions, remains an active research focus. Generative models can mitigate privacy and consent concerns, yet their training on real web-sourced images raises doubts about whether generated content removes or only obscures data ownership and consent issues [143]. To mitigate these risks, growing attention is given to watermarking and related methods [144] that clearly mark images as synthetic.

4.5 Feature Fusion in Face Recognition

In template-based FR, 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 FR, 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. 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 order-invariant, and compact template representations directly impacts FR performance, especially in unconstrained or real-time scenarios.

Early video-based FR approaches employed adaptive hidden Markov models to capture temporal dynamics and recognize entire video sequences [76]. Over time, feature fusion in FR advanced from simple averaging to adaptive and context-aware neural aggregation. The Neural Aggregation Network (NAN) [145] demonstrated the effectiveness of learning quality-aware attention weights for robust, order invariant face templates. Building on this, Multicolumn Networks [146] and C-FAN [147] introduced fine-grained quality analysis by modeling visual and contextual importance or by weighting individual feature channels. These developments significantly improved FR performance on challenging template-based benchmarks such as IJB-C [23].

Recent work emphasizes scalable and generalized feature fusion across diverse conditions. Methods like CAFace [75], CoNAN [148], and ProxyFusion [149] sustain performance even with templates containing many varied images. Practical approaches such as Norm Pooling [150] show that simple heuristics can be effective in multi domain settings, especially with limited training data. Overall, these innovations reflect a trend toward scalable, efficient strategies that handle long videos while maintaining robustness.

5 State of the art in Face Recognition

FR evaluation progressed from lab specific testing to standardized protocols [56], [83], [99]. Before deep learning, small proprietary datasets with fewer than 100 subjects and over 95% reported accuracy lacked generalizability [8], [46], [47]. The 1996 FERET program established large scale standardized evaluation, enabling systematic benchmarking and transparent comparison [56], [83]. This shift to structured evaluation allowed consistent progress tracking and set stage for deep learning's transformative impact on FR [11], [13], [14], [32]–[34].

The FERET program [56] in the 1990s standardized FR benchmarks with 14,126 images of 1,199 individuals. A decade later, the FRGC [99] expanded this with 50,000 high resolution images and 3D scans, defining protocols that guided later NIST FRVT [151] series. Together, they formed the basis of modern FR benchmarks.

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) [25], CFP-FP [26], CP-LFW [27], AgeDB [28], YouTube Faces (YTF) [163], TinyFace [77], and several iterations of

TABLE 3: Performance on CFP-FP [26] Dataset

Method Name	Backbone	Loss Function	Training Data	Verification (%)
GFace [152]	IResNet-50	GCE (LO)	Casia-WebFace	97.44
CosFace [18]	ResNet100	CosFace [18]	MS1MV2	98.13
ArcFace [19]	ResNet101	ArcFace	MS1MV2	98.27
MV-Softmax [153]	ResNet100	MV-Softmax	MS1MV2	98.28
CurricularFace [20]	ResNet101	CurricularFace	MS1MV2	98.37
TransFace-B [154]	ResNet100	ArcFace	MS1MV2	98.39
MagFace [67]	ResNet100	MagFace	MS1MV2	98.46
AdaFace [21]	ResNet101	AdaFace	MS1MV2	98.49
CQA-Face [155]	ResNet100	CQA-Face	MS1MV2	98.49
UniFace [69]	ResNet100	UniFace [69]	MS1MV2	98.63
URL [156]	ResNet101	URL	MS1MV2	98.64
LGAF [157]	ResNet100	ArcFace	MS1MV2	98.77
ArcFace [19]	ResNet50	ArcFace	Glint360K	98.77
ViT-S [108]	ViT-S	ArcFace	Glint360K	98.85
CosFace + KP-RPE	ViT	CosFace	WebFace4M	98.91
TransFace-S [154]	ViT-S	ArcFace	Glint360K	98.91
AdaFace [21]	ViT	AdaFace	WebFace4M	98.94
KP-RPE [50]	ViT	AdaFace	WebFace4M	99.01
ViT-B [108]	ViT-B	ArcFace	Glint360K	99.02
AdaFace [21]	ResNet101	AdaFace	MS1MV3	99.03
R100	ResNet100	ArcFace	Glint360K	99.04
AdaFace [21]	ViT	AdaFace	MS1MV3	99.06
ArcFace [19]	ResNet101	ArcFace	WebFace4M	99.06
KP-RPE [50]	ViT	ArcFace	WebFace4M	99.09
ViT-L	ViT-L	ArcFace	Glint360K	99.10
KP-RP [50]E	ViT	AdaFace	MS1MV3	99.11
GFace [152]	IResNet-100	GCE (LO)	MS1MV3	99.12
R200	ResNet200	ArcFace	Glint360K	99.14
AdaFace [21]	ResNet101	AdaFace	WebFace4M	99.17
TransFace-B [154]	ViT-B	ArcFace	Glint360K	99.17
AdaFace [21]	ResNet101	AdaFace	WebFace12M	99.24
KP-RPE [50]	ViT	AdaFace	WebFace12M	**99.30**
TransFace-L [154]	ViT-L	ArcFace	Glint360K	**99.32**

TABLE 4: Performance on IJB-C [23] Dataset

Method Name	Backbone	Loss Function	Training Data	TAR@FAR=1e-4
ArcFace [19]	ResNet101	ArcFace [19]	MS1MV2	96.03
MagFace [67]	ResNet101	MagFace [67]	MS1MV2	95.81
MagFace+IIC	ResNet101	MagFace	MS1MV2	95.89
ViT-S	ViT-S	ArcFace	MS1MV2	95.89
CurricularFace [20]	ResNet101	CurricularFace	MS1MV2	96.10
ViT-B	ViT-B	ArcFace	MS1MV2	96.15
ViT-L	ViT-L	ArcFace	MS1MV2	96.24
TransFace-S [154]	ViT-S	ArcFace	MS1MV2	96.45
TransFace-B [154]	ViT-B	ArcFace	MS1MV2	96.55
TransFace-L [154]	ViT-L	ArcFace	MS1MV2	96.59
ArcFace+CFSM	ResNet101	ArcFace	MS1MV2	96.60
ARoFace [158]	ResNet101	ArcFace	MS1MV2	96.66
ElasticFace [68]	ResNet101	ElasticFace	MS1MV2	96.65
TopoFR [159]	ResNet101	TopoFR	MS1MV2	96.95
GFace [152]	ResNet101	TopoFR	MS1MV2	96.96
AdaFace [21]	ResNet101	AdaFace	MS1MV2	97.09
KP-RPE [50]	ViT-B	CosFace	WebFace4M	96.98
TopoFR [159]	ResNet200	TopoFR	MS1MV2	97.08
AdaFace [21]	ViT-B	AdaFace	MS1MV3	97.10
KP-RPE [50]	ViT-B	AdaFace	WebFace4M	97.13
AdaFace [21]	ViT-B	AdaFace	WebFace4M	97.14
KP-RPE [50]	ViT-B	AdaFace	MS1MV3	97.16
KP-RPE [50]	ViT-B	ArcFace	WebFace4M	97.21
PartialFC [85]	ResNet101	ArcFace	WebFace4M	97.22
CatFace [160]	ResNet101	CatFace	MS1MV2	97.43
AdaFace [21]	ResNet101	AdaFace	WebFace4M	97.39
ARoFace [158]	ResNet101	AdaFace	WebFace4M	97.51
AdaFace [21]	ResNet101	AdaFace	WebFace12M	97.66
PartialFC [85]	ResNet101	ArcFace	WebFace12M	97.58
ARoFace [158]	ResNet101	AdaFace	WebFace12M	97.60
TopoFR [159]	ResNet101	TopoFR	Glint360K	97.60
KP-RPE [50]	ViT-B	AdaFace	WebFace12M	97.82
PartialFC [85]	ResNet101	ArcFace	WebFace42M	97.82
TopoFR [159]	ResNet200	TopoFR	Glint360K	97.84
PartialFC [85]	ViT-B	ArcFace	WebFace42M	97.90
PartialFC [85]	ResNet200	ArcFace	WebFace42M	**97.97**
UniTSFace [70]	ViT-L	UniTSFace	WebFace42M	**97.99**

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

Method Name	Backbone	Loss Function	Training Data	Rank-1	Rank-5
PFE [161]	ResNet101	PFE	MS1MV2	50.16	58.33
URL [156]	ResNet101	URL	MS1MV2	59.79	65.78
CurricularFace [20]	ResNet101	CurricularFace	MS1MV2	62.43	68.68
AdaFace [21]	ResNet101	AdaFace	MS1MV2	65.26	70.53
AdaFace [21]	ViT	AdaFace	MS1MV3	65.95	71.64
AdaFace [21]	ResNet101	AdaFace	MS1MV3	67.12	72.67
KP-RPE [50]	ViT	AdaFace	MS1MV3	67.62	73.25
ArcFace [19]	ResNet101	ArcFace	WebFace4M	69.26	74.31
AdaFace [21]	ResNet101	AdaFace	WebFace4M	70.42	75.29
ARoFace [158]	ResNet101	ArcFace	WebFace4M	70.96	75.54
AdaFace [21]	ResNet101	AdaFace	WebFace12M	71.35	76.24
AdaFace [21]	ViT	AdaFace	WebFace4M	71.90	77.09
KP-RPE [50]	ViT	CosFace	WebFace4M	72.22	77.67
ARoFace [158]	ResNet101	AdaFace	WebFace12M	72.28	77.93
KP-RPE [50]	ViT	AdaFace	WebFace4M	**72.78**	**78.20**
KP-RPE [50]	ViT	ArcFace	WebFace4M	**73.04**	**78.62**

TABLE 6: Performance on TinyFace [77] Dataset

Method Name	Backbone	Loss Function	Training Data	Rank1	Rank5
ArcFace+CFSM	ResNet101	ArcFace	MS1MV2	64.69	68.80
TransFace-L [154]	ViT-S	ArcFace	MS1MV2	67.52	71.00
ARoFace [158]	ResNet101	ArcFace	MS1MV3	67.54	71.05
LGAF [157]	ResNet101	ArcFace	MS1MV2	68.35	71.59
ArcFace [19]	ResNet101	ArcFace	WebFace4M	71.11	74.38
AdaFace [21]	ResNet101	AdaFace	WebFace4M	72.02	74.52
AdaFace [21]	ResNet101	AdaFace	WebFace12M	72.29	74.97
REE [162]	ResNet-50	ArcFace	Native VLR	73.06	77.22
ARoFace [158]	ResNet101	ArcFace	WebFace4M	73.80	76.53
ARoFace [158]	ResNet101	AdaFace	WebFace4M	73.98	76.47
ARoFace [158]	ResNet101	AdaFace	WebFace12M	74.00	76.87
KP-RPE [50]	ViT-B	CosFace	WebFace4M	75.48	78.30
KP-RPE [50]	ViT-B	ArcFace	WebFace4M	**75.62**	**78.57**
KP-RPE [50]	ViT-B	AdaFace	WebFace4M	**75.80**	**78.49**

the IARPA Janus Benchmark (IJB) series such as IJB-B [22], IJB-C [23], and IJB-S [24]. Each dataset addresses specific challenges inherent in FR scenarios. Below some datasets are described in detail.

Labeled Faces in the Wild (LFW [25]) 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 [163]) specifically targets video-based unconstrained FR. 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 [26] 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 [77] 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 [22], 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 [23]. 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 [24].

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

5.1.1 *State-of-the-Art Performance*

FR performance shows progress across different benchmarks, underscoring the effectiveness of deep learning architectures and sophisticated loss functions. In Tab. 3∼5 we compile the up-to-date FR performance under each evaluation datasets using relevant metrics.

In 1:1 verification, Verification Accuracy is the proportion of correct matches and nonmatches. On difficult datasets, performance is given by the True Accept

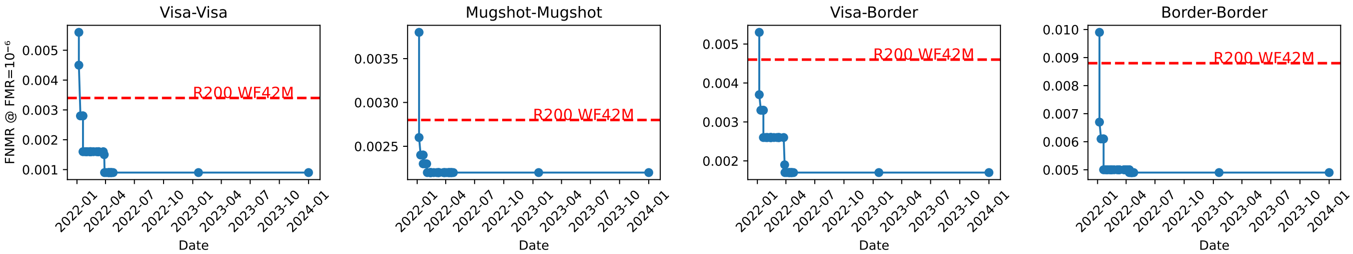


Fig. 14: SoTA performance in NIST FRVT 1:1 verification since January 2022. Plots show the cumulative minimum False Non-Match Rate (FNMR) achieved by any submitted algorithm up to the corresponding date for the Visa, Mugshot, Visa Border, and Border datasets (the plot titles indicate gallery - probe in order). Performance shows a low False Match Rate (*e.g.*, FMR=10^{-6}). The dashed red line indicates the performance level achieved by the WebFace42M entry (R200 [59] WF42M) for comparison.

Rate (TAR) at a fixed False Accept Rate (FAR), usually FAR=0.01%, representing $1-\text{FNMR}$ at a fixed FMR. For 1:N identification, Rank-k accuracy shows how often the correct identity is within the top k results (Rank-1 is strictest). In open set identification, the True Positive Identification Rate (TPIR) at a given False Positive Identification Rate (FPIR), such as FPIR=0.01%, measures correct identifications while limiting false matches of unknowns.

CFP-FP [26]: Current methods achieve extremely high verification accuracy, often exceeding 99%. Top performance is typically seen with models utilizing ViT [108] backbones (*e.g.*, ViT-L, ViT-B, ViT-S variants) or deeper ResNet [15] architectures (*e.g.*, ResNet-101, ResNet-200). Effective loss functions like AdaFace [21] and ArcFace [19] are prevalent among the leading methods. Furthermore, training on very large datasets like Glint360K [85], WebFace [59] is crucial for reaching the highest scores, with methods like TransFace [154] reporting accuracies above 99.3%.

IJB-C [23]: 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 [85] (utilizing ViT-L or ResNet200), KP-RPE [50] (with ViT-B), AdaFace [21], and TopoFR [159], achieve TAR values around 98% at 0.01%. Again, larger backbones (ViT-L [108], ResNet200 [15]) and extensive training data (WebFace42M [59], Glint360K [85]) are characteristic of the top-performing approaches.

TinyFace [77] (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 misalignments 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 [158] also show competitive results, highlighting the ongoing efforts to improve recognition under significant resolution constraints.

IJB-S [24]: 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 [75], [148], [149].

5.2 Technology Evaluations by NIST

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) [164]. NIST evaluations are critical for assessing the readiness of algorithms for real-world deployment.

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 FR software improved by a factor of 20 in search accuracy [165], highlighting the rapid pace of advancement in the field. To date, NIST has evaluated over 400 algorithms [151]. 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. 14 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 [151].

The results consistently show that top-performing algorithms, often developed by industry players, achieve excellent accuracy. However, many leading academic models,

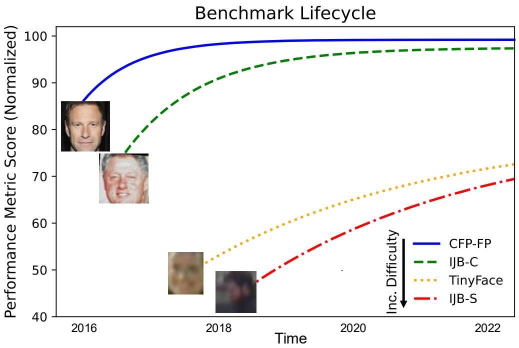


Fig. 15: Illustration of performance trends of various FR benchmarks over time, illustrating eventual performance saturation as datasets become extensively explored. This saturation is an indication of the progress in the field.

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

Although FR performance has advanced greatly, challenges persist in real-world use. As Fig. 15 shows, benchmark saturation reflects limits in current evaluations rather than problem resolution. Fig. 16 illustrates that for low-quality images (*e.g.*, IJB-S), missing facial details force models to depend on soft biometric cues like beards or hair color.

6.1 Recognition at Scale: Large Galleries

A major limitation of current academic benchmarks is their failure to match the scale and complexity of real biometric systems. Practical FR deployments often manage galleries with millions to billions of identities, far beyond the thousands used in research. For example, India's Aadhaar system stores biometric data, including facial images, for over 1.4 billion people. At such scales, even small drops in recognition accuracy lead to large numbers of false matches or misses, affecting millions of users.

An experiment on the IJB S dataset was conducted to illustrate the gap between benchmark evaluations and real world scenarios. The baseline IJB S gallery (202 identities) was expanded with external imposters ranging from 1,000 to 10,000 identities. Tab. 7 shows a performance drop as gallery size increases, revealing reduced recognition accuracy with larger galleries. This demonstrates that benchmarks can fail to represent real world conditions involving extensive galleries with diverse, unstructured, and noisy identities.

6.2 Practical Applications

While benchmark performance provides a useful measure of algorithmic progress, the ultimate test of FR lies in its deployment across real-world applications. Each application domain introduces unique constraints, ethical considerations and requirements that continue to shape FR research.

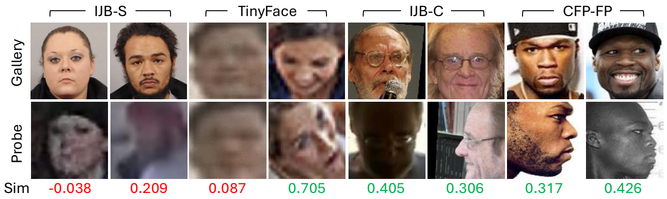


Fig. 16: Examples of faces across different datasets: IJB-S [24], TinyFace [77], IJB-C [23], and CFP-FP [26]. The top row shows gallery images, the middle row shows corresponding probe images, and the bottom row reports cosine similarity scores. Higher similarity scores (in green) indicate successful matches, while lower scores (in red) indicate mismatches. Features are extracted by KP-RPE [50] trained on WF4M [59].

Surveillance and Security. Surveillance applications revealed limits of low-quality, cross-resolution FR, prompting benchmarks like IJB-S [24] and TinyFace [77] and quality-aware models [21], [67]. More recent BRIAR project [78] extends evaluation up to 1,000 m standoff distances, integrating body and gait cues for robust recognition [31]. Under such setting, face-only systems achieve under 80% TAR@1% FAR and about 84% Rank-20 accuracy, while multimodal fusion exceeds 90% TAR, underscoring the need for cues beyond the face for long-range recognition [113].

Cross Age and Child Recognition. Age progression causes strong intra-class variation. NIST studies [151] show low errors for middle aged adults but higher rates for the youngest and oldest. YFA dataset [166] shows similarity degradation with even short age gaps, for ages under 36 months. Child recognition is reliable only over short term, Systems should use age aware thresholds, periodic re-enrollment, and multimodal fusion.

Other Specialized Domains. In *mobile authentication and access control*, FR achieves high reliability with emphasis on real time performance and privacy on edge devices [167]. In *forensic and post mortem identification*, robustness to extreme degradations, aging, and cross spectral imagery remains active, amid debates on accuracy, bias, and evidentiary standards [168]. These applications illustrate the breadth of FR but merit separate detailed study.

6.3 Multi-modal Recognition: Beyond Facial Imagery

As FR 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 [169]. Score-level fusion combines similarity scores from multiple biometric modalities

TABLE 7: 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%

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 [170]–[173]. 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 [52] 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 distinguish individuals at lower image resolutions. The future of robust FR lies in embracing a multi-modal approach, harnessing complementary biometric modalities to overcome the limitations of any single modality.

6.4 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 [122], trained on 52k real face images, generates 20k new synthetic identities. In contrast, Vec2Face [128], 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.* [174] 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 FR 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 [175].

TABLE 8: Performance Comparison of Foundation Models (FMs) in FR under Different Training Regimes. Accuracies are averaged over LFW, CALFW, CPLFW, CFP-FP, and AgeDB. Rank is 16 for LoRA. CosFace [18] 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

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 grounded 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 [121] 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 overcome the limitations of, and perhaps even outperform, real-world data collection for advancing FR.

6.5 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 FR. These models provide both pretrained weights and robust feature representations derived from broad visual or textual domains. Chettaoui *et al.* [176] 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 [176]. Key comparative results are shown in Tab. 8. However, obtaining large scale training dataset is not difficult for FR, the benefit of FMs is still to be probed. Future work should focus on identifying which fine-tuning techniques, such as LoRA [119], and which foundation models, like CLIP [177] or DINOv2 [178], 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 [72] introduces pretraining on unlabeled face data using foundation models, effectively learning crit-

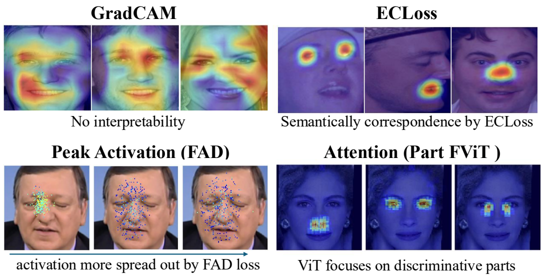

GradCAM
ECLoss
No interpretability
Semantically correspondence by ECLoss
Peak Activation (FAD)
Attention (Part FViT)
activation more spread out by FAD loss
ViT focuses on discriminative parts

Fig. 17: Comparison of visualization methods for face-related tasks. Grad-CAM [179] highlights broad, less interpretable regions, whereas ECLoss [80] enforces semantic correspondence with activations on meaningful facial parts. Peak Activation with FAD [79] shows that activations become more spread out across the face with the application of FAD loss. Attention maps from PartFViT [180] demonstrate that ViT models concentrate on discriminative facial parts, *e.g.*, eyes and nose.

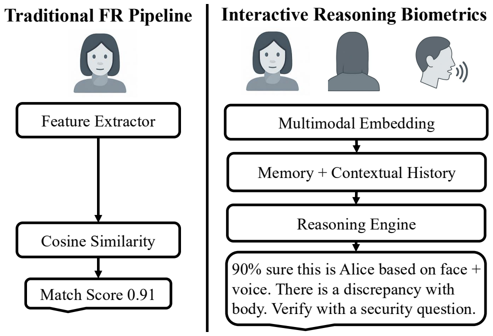

Traditional FR Pipeline
Interactive Reasoning Biometrics
Feature Extractor
Cosine Similarity
Match Score 0.91
Multimodal Embedding
Memory + Contextual History
Reasoning Engine
90% sure this is Alice based on face + voice. There is a discrepancy with body. Verify with a security question.

Fig. 18: Comparison between traditional 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 can reason across modalities, dynamically assess uncertainty, and collaborate with humans through feedback loops.

ical FR 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.

6.6 Interpretability

Deep learning models used for FR 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 FR:

Saliency/Attribution Maps: Methods such as Grad-CAM [179] or SHAP [181] 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 [182]. FAD [79] 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 [183]. This could involve identifying the influence of specific facial attributes (*e.g.*, eye shape, nose structure) or using methods like ECLoss [80] 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 [184], [185] (*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 [186], [187]. This provides a different perspective beyond spatial explanations.

LLM-based Explanation: Recent advances show that large language models (LLMs) are improving the interpretability of FR systems. Traditional tools such as saliency maps or concept attributions highlight key facial regions but fail to provide coherent, human-readable rationales. New approaches like XAI-CLIP [188] and interpretable vision–language alignment methods [189] use LLMs to generate natural language explanations of why two faces match or differ, citing traits such as "similar eyebrow curvature, matching nose bridge width, and aligned mouth corners." Complementary work on concept bottleneck models [190], [191] enhances interpretability by learning high-level concepts (e.g., glasses, facial hair), allowing explanations grounded in semantic features rather than pixel activations. Such language-based reasoning enables FR systems to describe shared and divergent traits transparently, emphasizing the need for tighter grounding between textual justifications and visual evidence.

As interpretability advances, the next challenge is to pair match scores with calibrated confidence and clear reasoning. Future FR systems may, in uncertain cases, simulate multiple identity hypotheses, request additional evidence, or express probabilistic justifications. Fig. 18 illustrates this shift from static pipelines to interactive, reasoning-based systems that foster human collaboration. A key step toward such feedback loops is enabling models to determine when to involve humans. Building on Face Image Quality Assessment (FIQA) research that estimates input quality or recognizability [162], [192], [193], future work should integrate these signals into broader reasoning frameworks and develop metrics that go beyond accuracy to capture trust, interpretability, and decision quality.

6.7 Fairness and Bias in Face Recognition

As FR moved from research to practical deployment, ensuring equitable performance across demographic groups became critical. Fairness concerns intensified in the late 2010s following incidents of wrongful arrests [194], biased airport screening [195], and misidentification of public figures [196]. Since 2019, NIST's FRVT includes demographic analyses [151]. For an extensive review, see Kotwal and Marcel [37]; this section summarizes key sources of bias, datasets, metrics, and mitigation strategies.

Sources of Disparities. Bias arises from multiple factors. Training data imbalance has long been cited [197], [198], though disparities persist even in balanced sets [199]; Skin reflectance affects accuracy under poor lighting [200], while image quality and illumination improvements reduce group gaps [201]. Appearance factors such as hairstyle, facial hair, makeup, and occlusion drive gender differences more than gender itself [202].

Datasets and Evaluation. Fairness evaluation requires demographically annotated datasets. Public options include RFW [203], BFW [204], BUPT and BalancedFace [205].

Metrics and Mitigation. Fairness metrics captures shifts in similarity score distributions. NIST FRVT [151] contains large-scale demographic performance variations. Fairness Discrepancy Rate (FDR) quantifies these disparities, with values near one indicating greater parity [206]. Mitigation occurs through (1) preprocessing via demographic augmentation or synthetic data generation [207], (2) using adaptive architecture, losses or adversarial debiasing [205], [208], [209], and (3) *postprocessing* with subgroup calibration or score normalization [204].

6.8 Interconnected Areas with FR

Face Image Quality Assessment (FIQA). Recognition reliability depends strongly on image quality, influenced by pose, blur, and illumination. FIQA models estimate a recognizability score to guide quality aware training and inference. FIQA methods are either unsupervised, predicting quality from recognition feature certainty [67], [210]–[212], or supervised, deriving pseudo quality labels through characterization and training an independent model [192], [213]–[215]. FIQA now underpins FR by quantifying uncertainty and improving reliability with low quality inputs.

Presentation Attacks and Spoof Detection (PAD). As FR systems expand to consumer and border uses, they face attacks like printed photos, replayed videos, and three dimensional masks. PAD defends using texture cues, depth, or multi sensor fusion [216]. Progress in image forensics and deepfake detection adds augmentation [217], [218], frequency analysis [219], [220], and disentanglement learning [221]–[223]. Interpretable detectors such as DDVQA BLIP [224] or M2F2Det [225] employ vision language models for explanations, enhancing transparency.

Privacy Preserving Face Recognition (PFR). With FR in sensitive contexts, privacy protection is crucial. Cryptographic techniques use homomorphic encryption or secure multiparty computation for similarity computation without exposing raw faces [226], [227], but computationally heavy. Transform based methods conceal by modifying representations. Early works use obfuscation [228], while modern ones use adversarial or diffusion methods [229], [230].

6.9 Regulatory and Policy Perspectives.

Global scrutiny has turned fairness and accountability into regulatory imperatives. The *GDPR* [231], [232] classifies facial imagery and other biometric data as special personal data, requiring explicit consent or anonymization of major FR datasets. The *EU AI Act* [233] designates biometric identification as high risk, mandating transparency, bias evaluation, and human oversight. Meanwhile, cities like San Francisco and Boston restrict public FR use, while others expand it for security purposes. These measures embed fairness as both scientific and regulatory.

7 Summary

Over the past fifty years, FR has advanced from geometric and handcrafted methods to deep learning models surpassing human accuracy. Progress in architectures (ResNets, ViTs), loss functions (margin based softmax), and large datasets (*e.g.*, WebFace42M) has driven state of the art performance. Despite success on benchmarks (LFW [25], CFP FP [26], IJB-C [23]) and deployment in authentication and security, challenges persist: scalability to billions, degraded imagery, limited interpretability, data privacy and fairness.

Future work should pursue multimodal and explainable systems, realistic synthetic data, and foundation models, emphasizing large scale robustness, confidence calibration, and ethical use through consent, security, and purpose limitation. Recent developments of FR datasets in the community have highlighted growing attention to the ethical and legal implications of datasets that lack clear and informed user consent, with increasing efforts to promote transparency and responsible data sourcing.

Acknowledgments This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] T. Kanade, "Picture processing system by computer complex and recognition of human faces," Kyoto University, Tech. Rep., 1974.

[2] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, 2016.

[3] C. AI, "Why facial recognition is the best biometric," 2023.

[4] V. Technologies, "Pros and cons of facial recognition," 2023.

[5] K. Lai, L. Queiroz, V. Shmerko, K. Sundberg, and S. Yanushkevich, "Post-pandemic follow-up audit of security checkpoints," *IEEE Access*, 2023.

[6] 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.

[7] L. Ding, C. Shu, C. Fang, and X. Ding, "Computers do better than experts matching faces in a large population," in *ICCI*, 2010.

[8] M. A. Turk, A. Pentland *et al.*, "Face recognition using eigenfaces." in *CVPR*, 1991.

[9] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV*, 2004.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.

[11] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[12] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NeurIPS*, 2012.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[16] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L2 hypersphere embedding for face verification," in *ACM MM*, 2017.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.

[18] 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.

[19] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[20] 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.

[21] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *CVPR*, 2022.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] G. Jager, D. Cornett, G. Glenn, D. Aykac, C. Johnson, R. Zhang, R. Shivers, D. Bolme, L. Davies, S. Dolvin *et al.*, "Expanding on the briar dataset: A comprehensive whole body biometric recognition resource at extreme distances and real-world scenarios," in *FG*, 2025.

[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 fairness in face recognition," *T-BIOM*, 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] Clearview AI, "Building a secure world, one face at a time," 2025.

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

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

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

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

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

[45] M. D. Kelly, "Visual identification of people by computer," Ph.D. dissertation, Stanford University, 1970.

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

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

[48] A. Lanitis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *ICCV*, 1995.

[49] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object detection," *TPAMI*, 1997.

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

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

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

[53] 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*, 2022.

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

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

[56] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, 1998.

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

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

[59] 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.

[60] Fortune Business Insights, "Facial recognition market size, share & industry analysis and regional forecast," 2025.

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

[62] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015.

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

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

[65] 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.

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

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

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

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

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

[71] 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.

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

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

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

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

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

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

[78] Intelligence Advanced Research Projects Activity, “Biometric recognition and identification at altitude and range (briar),” https://www.iarpa.gov/research-programs/briar, 2021.

[79] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu, “Towards interpretable face recognition,” in *ICCV*, 2019.

[80] H. Lin, H. Wu, and Y. Liu, “Activation template matching loss for explainable face recognition,” *arXiv*, 2024.

[81] 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.

[82] 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.

[83] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The feret evaluation methodology for face-recognition algorithms,” *TPAMI*, 2000.

[84] 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.

[85] 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.

[86] 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.

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

[88] 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.

[89] 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.

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

[91] 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.

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

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

[94] 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.

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

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

[97] 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.

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

[99] 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.

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

[101] 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.

[102] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018.

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

[104] 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.

[105] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *CVPR*, 2020.

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

[107] 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.

[108] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *ICLR*, 2021.

[109] 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.

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

[111] 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.

[112] 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.

[113] 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.

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

[115] 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.

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

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

[118] 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.

[119] 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.

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

[121] 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.

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

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

[124] 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.

[125] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer, “Exfacegan: Exploring identity directions in gan’s learned latent space for synthetic identity generation,” in *IJCB*, 2023.

[126] 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.

[127] 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.

[128] H. Wu, J. Singh, S. Tian, L. Zheng, and K. W. Bowyer, “Vec2face: Scaling face dataset generation with loosely constrained vectors,” in *ICLR*, 2025.

[129] 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.

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

[131] 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.

[132] 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.

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

[134] 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.

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

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

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

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

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

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

[141] 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.

[142] 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.

[143] A. Silberling, "Studio ghibli hasn't commented on openai's onslaught of ai copies, but the fan subreddit has," *TechCrunch*.

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

[145] 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.

[146] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," in *BMVC*, 2018.

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

[148] 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.

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

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

[151] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (FRVT) part 3: Demographic effects," NIST, Tech. Rep., 2019.

[152] 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.

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

[154] 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.

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

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

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

[158] 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.

[159] 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.

[160] 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.

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

[162] 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.

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

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

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

[166] K. Bahmani and S. Schuckers, "Face recognition in children: A longitudinal study," in *IWBF*. IEEE, 2022.

[167] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello, "Continuous user authentication on mobile devices: Recent progress and remaining challenges," *IEEE Signal Processing Magazine*, 2016.

[168] C. G. Zeinstra, D. Meuwly, A. Ruifrok, R. N. Veldhuis, and L. J. Spreeuwers, "Forensic face recognition as a means to determine strength of evidence: a survey," *Forensic science review*, 2018.

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

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

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

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

[173] M. Vatsa, R. Singh, and A. Noore, "Integrating image quality in 2ν-svm biometric match score fusion," *IJNS*, 2007.

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

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

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

[177] 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.

[178] 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.

[179] 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.

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

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

[182] 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.

[183] 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.

[184] 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.

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

[186] M. Huber and N. Damer, "Beyond spatial explanations: Explainable face recognition in the frequency domain," in *WACV*, 2025.

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

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

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

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

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

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

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

[194] K. Hill, "Wrongfully accused by an algorithm," *The New York Times*, June 2020.

[195] C. Garvie, "America under watch: Face surveillance in the united states," Georgetown Law Center on Privacy & Technology, Tech. Rep., 2019.

[196] J. Snow, "ACLU claims amazon's face recognition falsely matched 28 members of congress with mugshots," *MIT Technology Review*, 2018.

[197] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *TIFS*, 2012.

[198] K. S. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K. W. Bowyer, "Characterizing the variability in face recognition accuracy relative to race," in *CVPRW*, 2019.

[199] M. Gwilliam, S. Hegde, L. Tinsley, and A. Bouamra, "Rethinking common assumptions to mitigate racial bias in face recognition datasets," in *ICCVW*, 2021.

[200] C. M. Cook, J. J. Howard, Y. B. Sirotin *et al.*, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," Maryland Test Facility, Tech. Rep., 2019.

[201] K. S. Krishnapriya, V. Albiero, K. Vangara *et al.*, "Issues related to face recognition accuracy varying based on race and skin tone," in *TTS*, 2020.

[202] N. Kurz, H. Fang, and J. P. Robinson, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation," in *CVPRW*, 2022.

[203] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *ICCV*, 2019.

[204] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?" in *CVPRW*, 2020.

[205] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *CVPR*, 2020.

[206] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," in *Pattern Recognition Letters*, 2020.

[207] A. Kortylewski, B. Egger, A. Schneider *et al.*, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *CVPRW*, 2019.

[208] S. Gong, X. Liu, and A. Jain, "Mitigating face recognition bias via group adaptive classifier," in *CVPR*, 2021.

[209] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *ECCV*, 2020.

[210] Z. Babnik, P. Peer, and V. Štruc, "Faceqan: Face image quality assessment through adversarial noise exploration," in *ICPR*, 2022.

[211] ——, "Diffiqa: Face image quality assessment using denoising diffusion probabilistic models," in *IJCB*, 2023.

[212] N. Ozay, Y. Tong, F. W. Wheeler, and X. Liu, "Improving face recognition with a quality-based probabilistic framework," in *CVPRW*, 2009.

[213] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality assessment for face recognition based on deep learning," in *ICB*, 2019.

[214] F. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y. Wang, "Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance," in *CVPR*, 2021.

[215] J. Zhu, Y. Su, M. Kim, A. Jain, and X. Liu, "A quality-guided mixture of score-fusion experts framework for human recognition," in *ICCV*, 2025.

[216] Y. Liu and X. Liu, "Spoof trace disentanglement for generic face anti-spoofing," *T-PAMI*, 2022.

[217] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *CVPR*, 2020.

[218] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts."

[219] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *CVPR*, 2021.

[220] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*, 2021.

[221] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *ICCV*, 2023.

[222] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP*, 2019.

[223] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *CVPR*, 2023.

[224] Y. Zhang, B. Colman, X. Guo, A. Shahriyari, and G. Bharaj, "Common sense reasoning for deepfake detection," in *ECCV*, 2024.

[225] X. Guo, X. Song, Y. Zhang, X. Liu, and X. Liu, "Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector," in *CVPR*, 2025.

[226] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *PETS*, 2009.

[227] W. Yang, S. Wang, H. Cui, Z. Tang, and Y. Li, "A review of homomorphic encryption for privacy-preserving biometrics," *Sensors*, 2023.

[228] T. A. M. Kevenaar, G. J. Schrijen, M. van der Veen, A. H. M. Akkermans, and F. Zuo, "Face recognition with renewable and privacy preserving binary templates," in *AutoID*, 2005.

[229] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer, "idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model," in *ICCV*, 2023.

[230] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *CVPR*, 2023.

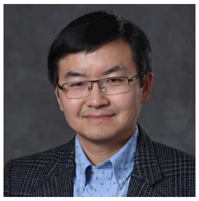
[231] European Parliament & Council of the European Union, "Regulation (eu) 2016/679 of the european parliament," 2016.

[232] S. Zaborska, "Legal regulation of the protection of biometric data under the GDPR," *Studia Iuridica Lublinensia*, 2019.

[233] European Parliament & Council of the European Union, "Regulation (eu) 2024/1689 of the european parliament," 2024.

Minchul Kim received his Ph.D. degree in Computer Science and Engineering from Michigan State University, East Lansing, MI, USA. He is currently a software engineer at Google, where he works on machine learning and computer vision applications. His research interests include face recognition, biometrics, and deep learning. During his Ph.D., he published in top-tier conferences and journals in the field of computer vision and biometrics.

Anil K. Jain (*Life Fellow, IEEE*) is a University Distinguished Professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition, computer vision, and biometric authentication. Jain is a member of the U.S. National Academy of Engineering, the Indian National Academy of Engineering, the World Academy of Sciences, and the Chinese Academy of Sciences.



Xiaoming Liu (*Fellow, IEEE*) is a MSU Foundation Professor, and Anil and Nandita Jain Endowed Professor in the Department of Computer Science and Engineering at Michigan State University. He received his Ph.D. from Carnegie Mellon University in 2004. His research interests span computer vision, machine learning, and biometrics. He is an Associate Editor for IEEE Transactions on Pattern Analysis and Machine Intelligence. He is a fellow of IEEE and IAPR.