

An Identity Level Protocol for Calibrated Open Set Face Recognition in Identity Scarce Scenarios

Divya Saleela¹(Corresponding Author), Rani Saritha R², Rahila N A³, Reji R⁴, Teena Joseph⁵, Sujith Kumar P S⁶, Gunjan Mittal Roy⁷, Syed Arshad Ali⁸, Chinchu M S⁹

¹Assistant Professor, College of Engineering Aranmula, Kerala, India

²Associate Professor, Saintgits College of Engineering (Autonomous), Kerala, India

³Research Scholar, Thangal Kunju Musaliar College of Engineering, Kerala, India

⁴Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

⁵Professor, St. Thomas College of Engineering & Technology, Kerala, India

⁶Professor, Sree Buddha College of Engineering, Kerala, India

⁷Associate Professor, IILM University, Greater Noida, India

⁸Assistant Professor, Sharda University, Greater Noida, India

⁹Assistant Professor, Sree Buddha College of Engineering, Kerala, India

divyasaleela009@cearanmula.ac.in¹, ranisaritha3090@gmail.com², rahilafiroz12@gmail.com³, rejir@tkmit.ac.in⁴, tinaajo85@gmail.com⁵, kcpsulickal@gmail.com⁶, gunjanmittal2013@gmail.com⁷, syedarshadali1987@gmail.com⁸, chinchums18@gmail.com⁹

Abstract: Open-set face recognition in unconstrained environments is hindered by identity scarcity, unreliable rejection of unseen individuals, and poorly calibrated decision confidence. This paper introduces an identity-level open-set evaluation protocol on the Labeled Faces in the Wild (LFW) dataset that explicitly models unknown identities during inference and enforces strict separation between training, gallery, and probe identities. The protocol enables joint assessment of open-set verification, identification accuracy, and decision calibration under both balanced and full-open operating conditions. Using this framework, a face-specific embedding model based on FaceNet with VGGFace2 pretraining is evaluated against strong generic and foundation-model baselines. Experimental results show that, using the proposed evaluation protocol, the best-performing face-specialized baseline (FaceNet trained on VGGFace2) achieves robust open-set discrimination and identification performance, with a ROC-AUC of 0.964, an equal error rate of 0.059, a Top-1 identification accuracy of 0.933, and an accepted-known identification accuracy of 0.982 under full-open evaluation. Furthermore, post-hoc calibration significantly improves confidence reliability without affecting discriminative performance. The proposed protocol and results demonstrate the importance of identity-aware evaluation and calibrated decision mechanisms for practical open-set face recognition.

Keywords

Open-set face recognition, identity scarcity, identity-level evaluation, face embeddings, decision calibration, LFW dataset, unknown identity rejection, reliability-aware recognition.

1. Introduction

Deep learning based face recognition systems have achieved strong performance in closed-set scenarios through discriminative embedding learning and large-scale curated datasets such as VGGFace2 and MS-Celeb-1M [1]. In these settings, inference assumes that every probe identity belongs to the same identity universe observed during training. Under this assumption, similarity-based matching can be optimized exclusively for inter-class separation and intra-class compactness. However, this formulation does not reflect

operational environments, where systems must encounter and correctly handle identities that were not present during training.

Open-set face recognition extends the classical formulation by requiring both accurate identification of known identities and reliable rejection of previously unseen individuals [2]. Unlike closed-set identification or pairwise verification, open-set recognition introduces asymmetric error costs and requires explicit decision thresholds to balance false acceptance and false rejection. Consequently, evaluation must incorporate operating-point-dependent metrics such as false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER), rather than relying solely on closed-set accuracy.

A fundamental challenge in open-set face recognition is identity scarcity. In unconstrained datasets such as LFW, a substantial proportion of identities are represented by only one or two images [3]. This scarcity restricts the modeling of intra-identity variation and increases overlap between similarity distributions of known and unknown identities. As a result, even highly discriminative embeddings may assign elevated similarity scores to unseen individuals, leading to systematic failures in unknown identity rejection.

Despite this challenge, many existing evaluation protocols do not explicitly account for identity scarcity. Standard verification benchmarks focus on pairwise similarity scores without distinguishing between known and unknown identity distributions [4] [5]. Closed-set identification protocols further assume that the correct identity is always present in the gallery. Even when open-set performance is reported, identity-level separation between training, gallery, and probe sets is often relaxed, allowing implicit identity leakage and optimistic performance estimates.

Recent advances in representation learning have introduced powerful generic and foundation vision models trained on large-scale object-centric or multimodal datasets [6]. While these representations exhibit strong transferability and retrieval performance, they are not optimized for identity discrimination. In open-set face recognition, such embeddings may rank identities effectively while still failing to provide sufficient separation between known and unknown score distributions, particularly under identity-scarce conditions.

Beyond discriminative performance, decision calibration plays a critical role in open-set recognition systems. Similarity scores produced by embedding models are frequently treated as confidence measures despite lacking probabilistic interpretation [7]. Poor calibration leads to overconfident acceptance of unknown identities and unstable operating points across datasets. In threshold-based open-set systems, even minor calibration errors can result in substantial degradation of FAR or FRR, yet calibration is rarely evaluated alongside open-set recognition metrics.

This research addresses these limitations by introducing an identity-level open-set evaluation protocol that explicitly models identity scarcity and unknown identity rejection. The protocol enforces strict separation between training, gallery, and probe identities and supports evaluation under both balanced and full-open operating conditions. Balanced validation sets are used for principled threshold selection, while full-open test sets provide stable estimates of operating-point behavior. In addition, post-hoc calibration is incorporated to assess and improve decision reliability without retraining embedding models.

Using the proposed protocol, a face-specific embedding model based on FaceNet with VGGFace2 pretraining is evaluated against generic ImageNet-trained convolutional networks and a modern foundation vision model on the Labeled Faces in the Wild dataset. Experimental results demonstrate that face-trained embeddings substantially outperform generic representations in open-set discrimination, identification accuracy, and calibrated decision behavior under identity-scarce conditions. These findings establish identity-aware evaluation and calibration as essential components for deploying robust open-set face recognition systems in real-world scenarios. This work does not propose a new face recognition model; instead, it introduces a rigorously designed identity-level evaluation and calibration protocol, addressing critical gaps in how open-set face recognition systems are assessed under identity-scarce conditions

The remainder of this paper is organized as follows. Section II reviews related work in open-set face recognition, identity-scarce evaluation protocols, and calibration methods. Section III describes the proposed identity-level evaluation protocol and dataset construction. Section IV details the experimental setup and calibration procedures. Section V presents quantitative results and comparative analysis, and Section VI concludes the paper with future work.

II. Literature Review

Research on open-set face recognition spans multiple interconnected directions that influence system performance under realistic deployment conditions. Existing work can be broadly grouped into five categories: closed-set and verification-based face recognition, open-set and unknown identity rejection methods, identity scarcity and evaluation protocols, generic and foundation-model representations for face analysis, and calibration and confidence estimation in recognition systems. This section reviews these areas with a focus on their limitations in identity-scarce open-set scenarios.

Early and contemporary face recognition methods have largely been developed under closed-set or verification-based assumptions, where all identities encountered during testing are presumed to be known [8] [9]. Discriminative embedding learning using metric-based objectives, including contrastive loss, triplet loss, and angular-margin losses, has been highly successful in improving inter-class separability and intra-class compactness [10] [11]. While these approaches yield strong performance on standard benchmarks, they do not explicitly address the rejection of unseen identities. When deployed in open environments, such systems tend to assign confident predictions to unknown individuals, revealing a fundamental mismatch between training assumptions and operational requirements.

Open-set recognition has been studied as a general pattern recognition problem, with approaches based on score thresholding, probabilistic modeling, and extreme value theory proposed to detect unknown classes [12] [13] [14]. In face recognition, extensions such as OpenMax and score normalization techniques have been explored to mitigate false acceptance of unseen identities [15]. However, many of these methods rely on assumptions about score distributions or require sufficient samples per class, which are often violated in unconstrained face datasets. Furthermore, open-set evaluations are frequently conducted

without strict identity-level separation, limiting their relevance to identity-centric recognition tasks.

Identity scarcity is a defining characteristic of widely used unconstrained face datasets, where the number of identities is large but the number of samples per identity is extremely limited [16] [17]. In datasets such as LFW, a substantial fraction of identities appear only once, preventing robust modeling of intra-identity variation. Despite this, many evaluation protocols reuse identities across training, gallery, and probe sets or rely on pairwise sampling strategies that obscure identity-level leakage [18]. These practices can lead to optimistic performance estimates and do not accurately reflect system behavior when encountering genuinely unseen identities during inference.

Recent advances in representation learning have introduced generic and foundation vision models trained on large-scale object-centric or multimodal datasets [19] [20]. Such models exhibit strong transfer learning capabilities and have been applied to face recognition tasks without domain-specific optimization. While these representations often achieve competitive retrieval and ranking performance, they are not explicitly trained for identity discrimination or open-set rejection. Under identity-scarce conditions, generic embeddings may produce overlapping similarity distributions for known and unknown identities, resulting in unreliable threshold-based decisions despite high ranking accuracy.

Decision calibration has been widely studied in classification systems as a means of aligning predicted confidence with empirical accuracy [21] [22]. Post-hoc techniques such as temperature scaling have proven effective in improving calibration without retraining models [23]. In face recognition, however, similarity scores are commonly interpreted as confidence measures without probabilistic grounding, particularly in open-set settings. Calibration metrics such as expected calibration error (ECE) and Brier score are rarely reported alongside open-set recognition results, leaving the reliability of decision confidence largely unexamined.

Thus, existing literature reveals several critical gaps. Closed-set training and evaluation paradigms dominate face recognition research, despite limited applicability to open environments. Open-set methods often neglect identity scarcity and lack rigorous identity-level evaluation protocols. Generic and foundation-model representations remain insufficiently assessed under identity-scarce open-set conditions, and decision calibration is seldom integrated into open-set face recognition evaluation. To address these limitations, this work introduces an identity-level open-set evaluation protocol that explicitly models unknown identities and identity scarcity while incorporating calibration analysis. The proposed methodology and experimental framework are detailed in the following section.

III. Methodology

This section describes the experimental methodology used to evaluate open-set face recognition under identity-scarce conditions. The objective of the methodology is to rigorously analyze the behavior of different face embedding paradigms when both known and unknown identities are present at inference time. The methodology is evaluation-driven and focuses on protocol design rather than architectural modification. Specifically, this

section details (i) identity-level dataset partitioning, (ii) single-template gallery and probe construction under identity scarcity, (iii) face embedding extraction using multiple backbone networks, (iv) similarity computation and open-set decision modeling, (v) validation-driven threshold selection, (vi) open-set and identification performance metrics, and (vii) post-hoc calibration and reliability analysis.

A. Identity-Level Dataset Partitioning

Experiments are conducted on the Labeled Faces in the Wild (LFW) dataset [24], a widely used benchmark for unconstrained face recognition. LFW contains 13,233 face images from 5,749 identities, exhibiting a highly imbalanced identity distribution. In particular, 4,069 identities are represented by a single image, while only 1,680 identities have two or more images, making the dataset inherently identity-scarce. This characteristic makes LFW especially suitable for evaluating open-set face recognition protocols that operate under limited enrollment conditions. Figure 1 illustrates intra-class variability in LFW by showing multiple images of the same identity captured under different poses, expressions, and illumination conditions.

The dataset is partitioned strictly at the identity level to ensure that known and unknown identities are fully disjoint. Let $\mathcal{I}$ denote the set of all identities in the dataset. This set is divided into known identities $\mathcal{I}_K$ and unknown identities $\mathcal{I}_U$, such that no identity appears in both subsets. Identities assigned to $\mathcal{I}_U$ are completely excluded from training, validation, and threshold selection and are introduced only during inference, ensuring that all unknown probes correspond to genuinely unseen identities.

Known identities $\mathcal{I}_K$ are further partitioned into training, validation, and test subsets. This partitioning is performed at the identity level rather than the image level, preventing any identity overlap across splits. The validation subset is reserved exclusively for threshold selection and calibration parameter estimation, while the test subset is used only for final performance reporting. This separation ensures that no test information influences decision thresholds or calibration parameters.

Identity-level partitioning is particularly critical for LFW due to its severe identity imbalance. Image-level splits would introduce implicit identity leakage, allowing the same identity to appear across different evaluation stages and artificially inflating performance. By enforcing identity-level separation, the methodology ensures that evaluation reflects genuine generalization to unseen identities rather than memorization of identity-specific characteristics. No model parameters are trained or fine-tuned on LFW; all identity splits are used exclusively for protocol construction, threshold selection, and evaluation.

Three sample images of the same identity (Theresa_May) from LFW



Fig. 1. Three representative images of the same identity (Theresa_May) from the Labeled Faces in the Wild (LFW) dataset, illustrating intra-class appearance variation under unconstrained imaging conditions.

B. Gallery-Probe Construction under Identity Scarcity

To model identity scarcity during inference, a single-template enrollment strategy is adopted for all known identities. For each identity in $\mathcal{I}_K$, exactly one image is selected to form the gallery set $\mathcal{G}$. All remaining images belonging to that identity are assigned to the known probe set $\mathcal{P}_K$. This design simulates realistic deployment scenarios where only one reference image per identity is available.

The single-template constraint increases the difficulty of the identification task by limiting intra-identity representation in the gallery. This is particularly important in identity-scarce datasets, where additional enrollment images are rarely available. The methodology therefore avoids multi-template enrollment, which would artificially simplify the recognition problem. The resulting gallery contains exactly one embedding per known identity.

Unknown probes $\mathcal{P}_U$ are sampled from identities in $\mathcal{I}_U$. These probes are visually unconstrained and may share demographic or appearance characteristics with known identities. Known and unknown probes are evaluated jointly, ensuring that the system must simultaneously perform identity assignment and unknown rejection, reflecting realistic open-set operation. Identities represented by a single image are used either exclusively for gallery enrollment or exclusively as unknown probes, but are never allowed to contribute simultaneously to known enrollment and known probe sets, thereby preventing implicit identity leakage.

Two probe configurations are used. In the balanced open-set configuration, the number of known and unknown probes is equal and this setting is used for validation and threshold selection. In the full-open configuration, all available unknown probes are included, resulting in a strong imbalance where unknown probes significantly outnumber known probes. This full-open setting enables stable estimation of false acceptance behavior under realistic

operating conditions. Figure 2 provides an overview of the proposed open-set evaluation pipeline, illustrating gallery construction, probe processing, embedding extraction, similarity computation, and threshold-based decision making.

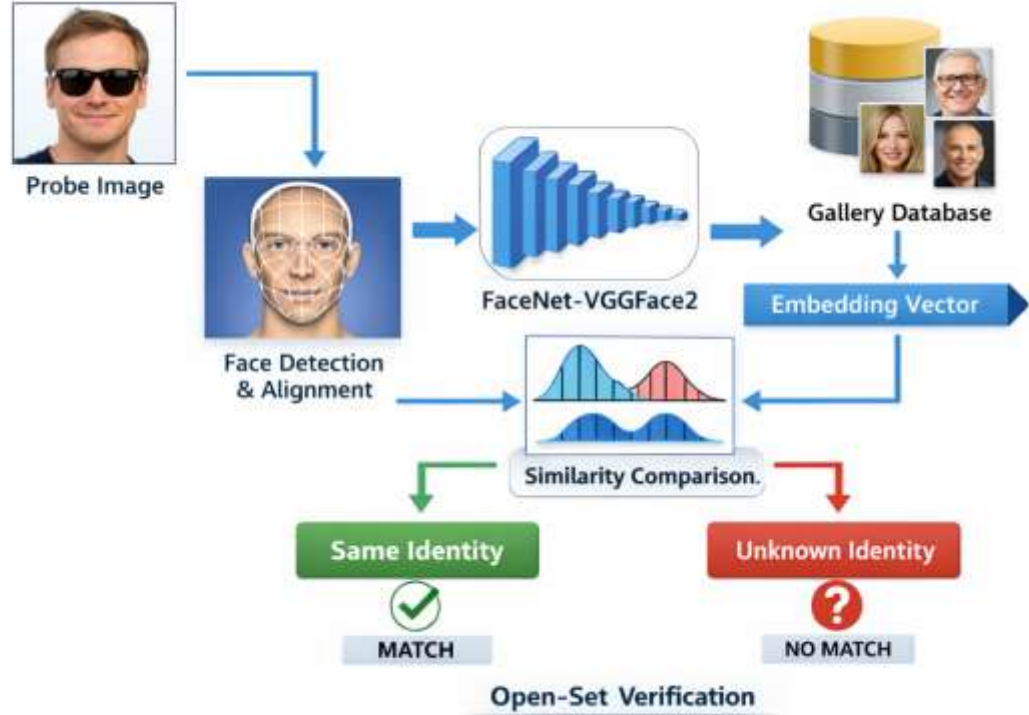


Fig. 2. Overview of the proposed open-set face recognition evaluation framework. A probe image undergoes face detection and alignment, embedding extraction using a pre-trained backbone (FaceNet-VGGFace2), similarity comparison against a single-template gallery, and threshold-based acceptance or rejection.

C. Face Embedding Extraction

Each face image is mapped to a fixed-dimensional embedding using a pre-trained backbone network. All backbone models are used in inference mode without task-specific fine-tuning. Input images are resized and normalized according to the preprocessing requirements of each backbone. This ensures that differences in performance arise from the embedding representations rather than preprocessing inconsistencies.

Let $f(\cdot)$ denote the embedding function implemented by a given backbone. For an input image x , the embedding vector is computed as $f(x)$ and subsequently ℓ_2 -normalized. Normalization constrains embeddings to lie on a unit hypersphere, enabling consistent cosine similarity computation across models. This normalization step is applied uniformly to all embeddings. The evaluation framework is backbone-agnostic. Both face-specific embedding networks and generic or foundation vision models are integrated without modifying downstream evaluation logic. This allows fair comparison across models under identical open-set

conditions. The dimensionality of embeddings may vary across models, but normalization ensures compatibility in similarity computation.

Embedding extraction is performed in batches to ensure computational efficiency. All embeddings are computed once per split and cached for subsequent similarity computation. This design choice avoids redundant computation and ensures that evaluation results are deterministic and reproducible.

D. Similarity Computation and Open-Set Decision Modeling

For each probe embedding $\mathbf{z}_p$, cosine similarity is computed with respect to all gallery embeddings $\{\mathbf{z}_g\}$. Cosine similarity is selected due to its widespread use in face recognition and its compatibility with normalized embeddings. The similarity scores represent the angular closeness between probe and gallery embeddings.

The maximum similarity score is retained as the open-set decision score:

$$s(\mathbf{z}_p) = \max_{\mathbf{z}_g \in \mathcal{G}} \mathbf{z}_p^\top \mathbf{z}_g \quad (1)$$

The gallery identity associated with this maximum similarity is selected as the predicted identity. This formulation corresponds directly to the implemented matching logic.

Open-set rejection is performed using a fixed decision threshold τ . If $s(\mathbf{z}_p) \geq \tau$, the probe is accepted as a known identity and assigned the predicted label. If $s(\mathbf{z}_p) < \tau$, the probe is rejected as unknown. This single-score formulation jointly handles identification and rejection without additional classifiers or probability models.

Using a single similarity score simplifies analysis and avoids introducing additional model assumptions. It also enables direct interpretation of score distributions for known and unknown probes. This design aligns with practical deployment scenarios where similarity thresholds are commonly used for access control and identity verification.

E. Threshold Selection Strategy

Decision thresholds are selected exclusively using validation data under balanced open-set conditions. Similarity scores from validation probes are used to compute the receiver operating characteristic (ROC) curve. The ROC curve captures the trade-off between false acceptance rate and true acceptance rate across all possible thresholds.

The equal error rate (EER) operating point is identified from the validation ROC curve. This operating point corresponds to the threshold at which false acceptance rate equals false rejection rate. The similarity value at this point is selected as the decision threshold τ .

Once selected, the threshold is fixed and applied unchanged during all test evaluations. No threshold tuning is performed on test data, including full-open testing. This strict separation between threshold selection and evaluation prevents optimistic bias and ensures that reported results reflect true generalization.

This threshold selection strategy mirrors real-world deployment practices, where thresholds are tuned offline using representative validation data and then deployed without modification.

F. Open-Set and Identification Performance Metrics

Open-set discrimination performance is evaluated using receiver operating characteristic curves, area under the ROC curve (AUC), and equal error rate. These metrics quantify the system's ability to separate known and unknown probes based on similarity scores. Both balanced and full-open test conditions are evaluated to capture different operating regimes. Operating-point behavior is further quantified using false acceptance rate and false rejection rate at the selected threshold. In addition, false rejection rate at fixed false acceptance levels is reported to assess behavior under security-sensitive constraints. These metrics are particularly important in full-open settings with high unknown prevalence.

Identification performance is evaluated only for probes that are correctly accepted as known. Metrics include Top-1 and Top-5 identification accuracy, which measure the correctness of identity assignment. Cumulative match characteristic curves are also reported to analyze rank-based identification performance.

Accepted-known identification accuracy is additionally reported. This metric measures identification accuracy conditioned on correct acceptance decisions and isolates identity classification performance from rejection behavior. Together, these metrics provide a comprehensive assessment of open-set recognition performance.

G. Decision Calibration

Raw similarity scores are treated as uncalibrated confidence estimates. Calibration quality is evaluated using expected calibration error and Brier score, which quantify the alignment between predicted confidence and empirical correctness. These metrics are computed under full-open test conditions.

To improve calibration, post-hoc temperature scaling is applied. The temperature parameter is learned using validation similarity scores without modifying the embedding networks. This approach preserves discriminative performance while adjusting confidence distributions.

The learned calibration parameters are applied to test similarity scores. Both pre- and post-calibration metrics are reported to quantify improvements in reliability. This analysis highlights the importance of calibration in threshold-based open-set face recognition systems.

IV. Experimental Setup and Calibration Procedures

This section describes the experimental setup used to instantiate the evaluation protocol defined in Section III. It details the backbone networks evaluated, implementation settings, validation and test configurations, and the procedure used for post-hoc score calibration. All experiments are conducted under a strictly inference-only regime without model fine-tuning.

A. Backbone Networks

Four representative face embedding paradigms are evaluated to assess open-set recognition behavior under identity-scarce conditions. These include both face-specialized and general-purpose visual encoders:

1. ResNet-18 (ImageNet): A lightweight convolutional network pre-trained on ImageNet and adapted for embedding extraction by removing the classification head.
2. ResNet-50 (ImageNet): A deeper convolutional architecture pre-trained on ImageNet, representing higher-capacity generic visual embeddings.
3. OpenCLIP ViT-B/32: A vision-language foundation model pre-trained using contrastive learning on large-scale image-text pairs. Only the image encoder is used.
4. FaceNet (VGGFace2): A face-specific embedding network trained on the VGGFace2 dataset and optimized for identity discrimination. This model serves as the primary face-specialized baseline.

Although more recent margin-based face recognition models exist, FaceNet pre-trained on VGGFace2 is deliberately used as a representative face-specialized baseline to isolate the impact of the proposed evaluation protocol from architectural or loss-function complexity. All models are evaluated in inference mode using fixed pre-trained weights, and no domain adaptation or task-specific training is performed. This design ensures that observed performance differences arise solely from the embedding representations rather than training or adaptation effects

B. Face Detection and Preprocessing

For face-specific embedding extraction, face detection and alignment are performed prior to embedding computation. A multi-task cascaded convolutional neural network (MTCNN) is used to detect facial regions and perform geometric alignment. Detected faces are resized to a fixed spatial resolution and normalized according to the requirements of the corresponding backbone.

For generic vision models that do not assume face input, images are resized and normalized using standard ImageNet or CLIP preprocessing pipelines. Preprocessing is applied consistently across all splits to avoid evaluation bias.

C. Validation and Test Configurations

Two distinct evaluation configurations are employed:

- **Balanced open-set configuration:** The number of known and unknown probes is equal. This configuration is used exclusively for validation, threshold selection, and calibration parameter estimation.
- **Full-open configuration:** All available unknown probes are included, resulting in a strong imbalance between unknown and known identities. This configuration is used only for final test evaluation.

Validation identities are never used for performance reporting, and test identities are never used for threshold selection or calibration. This strict separation prevents optimistic bias and ensures reproducibility.

D. Similarity Score Generation

For each probe image, cosine similarity is computed between its embedding and all gallery embeddings. The maximum similarity score is retained as the decision score, and the corresponding gallery identity is selected as the predicted identity label.

This matching strategy is applied uniformly across all backbone networks. No score normalization, cohort modeling, or adaptive decision mechanisms are introduced beyond the global threshold defined in Section III-E.

E. Calibration Procedure

Similarity scores are treated as uncalibrated confidence estimates. To assess reliability, expected calibration error (ECE) and Brier score are computed under full-open test conditions. Post-hoc temperature scaling is applied to improve calibration without modifying embedding representations. A single temperature parameter is learned using validation similarity scores by minimizing negative log-likelihood under a logistic link function. The learned temperature is then applied unchanged to test similarity scores. Both pre- and post-calibration confidence estimates are evaluated to quantify the impact of calibration on reliability and decision consistency.

F. Implementation Details and Reproducibility

All experiments are implemented using PyTorch and executed with fixed random seeds to ensure reproducibility. Embedding extraction is performed in batches and cached per split to avoid redundant computation. Evaluation metrics are computed using standard implementations, and all reported results correspond to a single deterministic evaluation run per model.

V. Results and Comparative Analysis

This section presents quantitative results obtained under the evaluation protocol described in Sections III and IV. Results are reported for both balanced and full-open open-set configurations. Performance is analyzed from three complementary perspectives: open-set discrimination, identification accuracy, and decision calibration. Comparisons are conducted across all evaluated backbone networks under identical conditions.

A. Open-Set Discrimination Performance

Open-set discrimination performance is evaluated using receiver operating characteristic (ROC) curves, area under the ROC curve (AUC), and equal error rate (EER). Figure 3 shows ROC curves obtained under the balanced open-set test configuration, where the number of known and unknown probes is equal. This setting provides a clear visualization of discriminative capability without class imbalance effects.

Across all models, face-specialized embeddings demonstrate stronger separation between known and unknown probes than generic visual encoders. The FaceNet-VGGFace2 model consistently achieves higher true acceptance rates at low false acceptance rates, indicating superior open-set discrimination under identity scarcity.

Figure 4 presents ROC curves under the full-open configuration, where unknown probes significantly outnumber known probes. Although absolute performance decreases for all methods due to the increased difficulty of this setting, relative trends remain consistent.

FaceNet-VGGFace2 maintains the highest AUC and lowest EER, highlighting its robustness to unknown identity prevalence.

To further analyze operating behavior under security-sensitive regimes, Detection Error Tradeoff (DET) curves are shown in Figure 5. The DET plots emphasize performance differences at low false acceptance rates. FaceNet-VGGFace2 exhibits a more favorable tradeoff between false acceptance and false rejection, particularly in regions relevant to access-control and authentication scenarios.

B. Operating-Point Analysis

Decision thresholds are selected using validation data at the equal error rate (EER) operating point and then applied unchanged during testing. Table I reports false acceptance rate (FAR) and false rejection rate (FRR) at the selected threshold under full-open test conditions.

In addition to threshold-based metrics, false acceptance rates at fixed false rejection levels (1% and 5%) are reported. These metrics provide insight into system behavior under constrained security requirements. FaceNet-VGGFace2 consistently achieves lower FAR at fixed FRR levels compared to generic encoders, indicating improved resistance to unknown identity acceptance.

C. Score Distribution Analysis

Figure 6 illustrates the distribution of maximum similarity scores for known and unknown probes under the full-open test configuration. Clear separation between the two score distributions is observed for face-specialized embeddings, whereas generic encoders exhibit substantial overlap.

The vertical dashed line in each plot denotes the decision threshold selected on validation data. For FaceNet-VGGFace2, the threshold lies in a region of minimal overlap between known and unknown distributions, explaining its superior operating-point performance. In contrast, generic encoders show greater distributional overlap, leading to higher false acceptance rates.

D. Identification Performance

Identification performance is evaluated only for probes that are correctly accepted as known. Top-1 and Top-5 identification accuracies are reported in Table I, along with cumulative match characteristic (CMC) curves shown in Figure 7.

CMC curves demonstrate that FaceNet-VGGFace2 achieves the highest rank-based identification accuracy across all ranks. Performance gaps are most pronounced at Rank-1, reflecting improved identity discrimination when only a single gallery template per identity is available.

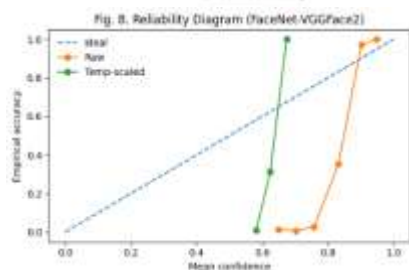
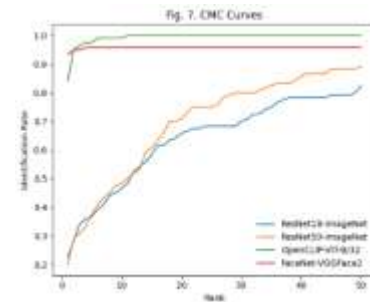
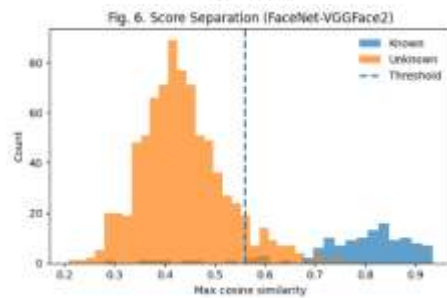
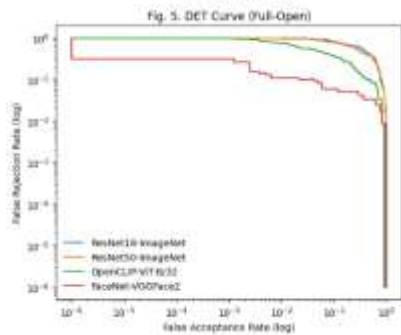
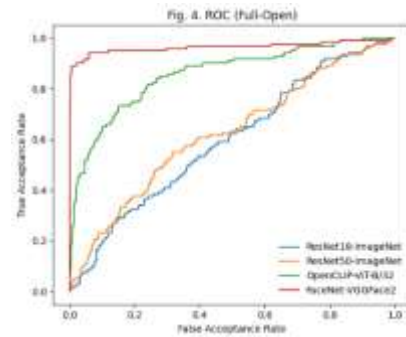
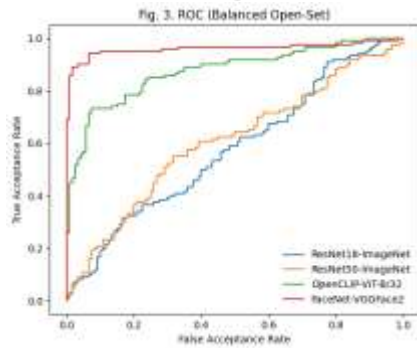
Accepted-known identification accuracy is also reported to decouple identification performance from rejection behavior. Results indicate that face-specialized embeddings not only reject unknown identities more effectively but also assign correct identities more reliably once acceptance occurs.

E. Decision Calibration and Reliability Analysis

Raw similarity scores are treated as uncalibrated confidence estimates. Calibration quality is evaluated using expected calibration error (ECE) and Brier score under full-open test conditions. Reliability diagrams before and after calibration are shown in Figure 8.

Prior to calibration, all models exhibit varying degrees of miscalibration, with confidence estimates either overconfident or underconfident relative to empirical correctness. FaceNet-VGGFace2 shows better initial calibration than generic encoders, but still exhibits measurable deviation from ideal reliability.

Post-hoc temperature scaling substantially improves calibration across all models without affecting discriminative performance. As shown in Figure 8, calibrated confidence curves more closely follow the ideal diagonal, and both ECE and Brier scores are consistently reduced. These results highlight the importance of explicit calibration when deploying threshold-based open-set face recognition systems.



Overall performance across all metrics is summarized in Table I. FaceNet-VGGFace2, as a representative face-specialized baseline, consistently outperforms generic visual encoders across open-set discrimination, operating-point robustness, identification accuracy, and calibration quality. Generic encoders, while competitive in balanced settings, exhibit notable degradation under full-open conditions and demonstrate weaker calibration behavior.

Table I. Comparative performance of face-specialized and generic embedding models under the proposed identity level open set evaluation protocol. Results are reported for balanced and full-open configurations, including open set discrimination metrics (AUC, AP, EER), operating-point behavior, identification accuracy, and calibration quality before and after post-hoc temperature scaling.

Model	AUC bal	AP bal	EER bal	AUC full	APf ull	EER full	thr	FAR @thr	FR R@thr	FA R@FR R1%	FAR @F RR5 %	FRR @FA R1%	To p-1	T o p- 5	ID ac c T	EC Er aw	ECE cal	Brier Raw	Brier Cal	
ResNe t18- Image Net	0.57 7	0.5 84	0.46 3	0.59 5	0.1 82	0.43 4	0.8 27	0.36 2	0.0 61	0.9 73	0.90 1	0.99 2	0.2 25	0. 3 6 7	0. 2 3 4	4.0 93	0.78 0	0.42 0	0.720	0.28 9
ResNe t50- Image Net	0.60 5	0.6 08	0.39 2	0.62 2	0.2 14	0.39 3	0.7 11	0.40 1	0.0 50	0.9 84	0.90 1	0.96 7	0.2 00	0. 3 7 5	0. 2 5 7	4.0 21	0.71 9	0.41 3	0.629	0.28 3
OpenC LIP- ViT- B/32	0.87 5	0.8 96	0.21 7	0.85 7	0.5 86	0.21 8	0.7 76	0.18 0	0.0 33	0.8 31	0.68 0	0.78 3	0.8 42	0. 9 7 5	0. 8 8 9	3.7 84	0.74 4	0.41 9	0.659	0.28 8
FaceN et- VGGFa ce2	0.96 1	0.9 74	0.06 2	0.96 4	0.9 39	0.05 9	0.5 60	0.06 7	0.0 08	0.8 33	0.12 0	0.13 3	0.9 33	0. 9 5 8	0. 9 8 2	1.2 17	0.62 3	0.49 4	0.452	0.31 7

These findings emphasize the importance of face-specific embedding representations for open-set recognition under identity scarcity and demonstrate that evaluation protocols must explicitly account for unknown identity prevalence and decision reliability.

VI. Conclusion and Future Work

This work proposed an identity-level evaluation protocol for open-set face recognition under identity-scarce conditions, emphasizing protocol design rather than model development. Experiments on the LFW dataset show that, under the proposed identity-level evaluation protocol, face-specialized embeddings (represented by FaceNet trained on VGGFace2) outperform generic visual encoders in open-set discrimination, operating-point stability, and identification accuracy, particularly under full-open conditions. The results also highlight the importance of calibration, with post-hoc temperature scaling improving decision reliability without affecting discriminative performance. Future work will extend the protocol to larger and more diverse datasets and investigate multi-template enrollment and adaptive thresholding strategies for robust open-set face recognition.

References

- [1] A. Zhalgas, B. Amirgaliyev, and A. Sovet, 'Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges', *Appl. Sci.*, vol. 15, no. 17, p. 9390, Aug. 2025, doi: 10.3390/app15179390.
- [2] R. H. Vareto, Y. Linghu, T. E. Boulton, W. R. Schwartz, and M. Günther, 'Open-set face recognition with maximal entropy and Objectosphere loss', *Image Vis. Comput.*, vol. 141, p. 104862, Jan. 2024, doi: 10.1016/j.imavis.2023.104862.
- [3] M. H. Fahim, A. A. Donkol, A. Bedair, and M. Abdel-Nasser, 'Image Processing A Decade-by-Decade Review with a Focus on Face Recognition', *SVU-Int. J. Eng. Sci. Appl.*, vol. 6, no. 2, pp. 29–46, Dec. 2025, doi: 10.21608/svusrc.2025.339279.1251.
- [4] M. Kim, A. Jain, and X. Liu, '50 Years of Automated Face Recognition', 2025, arXiv. doi: 10.48550/ARXIV.2505.24247.
- [5] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, 'Past, Present, and Future of Face Recognition: A Review', *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020, doi: 10.3390/electronics9081188.
- [6] A. S. Khan, M. J. Abbass, and A. H. Khan, 'Towards Fault-Aware Image Captioning: A Review on Integrating Facial Expression Recognition (FER) and Object Detection', *Sensors*, vol. 25, no. 19, p. 5992, Sep. 2025, doi: 10.3390/s25195992.
- [7] A. Chouldechova, S. Deng, Y. Wang, W. Xia, and P. Perona, 'Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition', in *Computer Vision – ECCV 2022*, vol. 13673, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds, in *Lecture Notes in Computer Science*, vol. 13673, Cham: Springer Nature Switzerland, 2022, pp. 289–306. doi: 10.1007/978-3-031-19778-9_17.
- [8] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, 'Face Recognition Systems: A Survey', *Sensors*, vol. 20, no. 2, p. 342, Jan. 2020, doi: 10.3390/s20020342.
- [9] A. K. Bahjat, 'A survey of facial recognition techniques', *Int. J. Commun. Inf. Technol.*, vol. 6, no. 2, pp. 214–225, Jul. 2025, doi: 10.33545/2707661X.2025.v6.i2c.167.
- [10] A. A. A. Khalifa, 'Efficient and robust face recognition in the wild', *Universitäts- und Landesbibliothek Sachsen-Anhalt*, 2025. doi: 10.25673/118968.
- [11] A. P. Fard and M. H. Mahoor, 'Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild', *IEEE Access*, vol. 10, pp. 26756–26768, 2022, doi: 10.1109/ACCESS.2022.3156598.

- [12] Fayin Li and H. Wechsler, 'Open set face recognition using transduction', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1686–1697, Nov. 2005, doi: 10.1109/TPAMI.2005.224.
- [13] W. J. Scheirer, L. P. Jain, and T. E. Boult, 'Probability Models for Open Set Recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014, doi: 10.1109/TPAMI.2014.2321392.
- [14] C. Geng, S.-J. Huang, and S. Chen, 'Recent Advances in Open Set Recognition: A Survey', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021, doi: 10.1109/TPAMI.2020.2981604.
- [15] E. Lopez-Lopez, X. M. Pardo, and C. V. Regueiro, 'Incremental Learning from Low-labelled Stream Data in Open-Set Video Face Recognition', *Pattern Recognit.*, vol. 131, p. 108885, Nov. 2022, doi: 10.1016/j.patcog.2022.108885.
- [16] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, 'Unconstrained Face Recognition: Identifying a Person of Interest From a Media Collection', *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014, doi: 10.1109/TIFS.2014.2359577.
- [17] L. Boudry, J. D. Nador, and M. Ramon, 'Determinants of Face Recognition: The Role of Target Prevalence and Similarity', *J. Cogn.*, vol. 7, no. 1, p. 27, Feb. 2024, doi: 10.5334/joc.339.
- [18] L. Laishram, M. Shaheryar, J. T. Lee, and S. K. Jung, 'Toward a Privacy-Preserving Face Recognition System: A Survey of Leakages and Solutions', *ACM Comput. Surv.*, vol. 57, no. 6, pp. 1–38, Jun. 2025, doi: 10.1145/3673224.
- [19] Y. Yang et al., 'Recent Advances of Foundation Language Models-based Continual Learning: A Survey', *ACM Comput. Surv.*, vol. 57, no. 5, pp. 1–38, May 2025, doi: 10.1145/3705725.
- [20] S. M. H. Abidi, S. A. Hassan, S. M. Raza, and M. J. Beliatas, 'Advances in Face Recognition: A Comprehensive Review of Feature Extraction and Dataset Evaluation', *Electronics*, vol. 15, no. 2, p. 338, Jan. 2026, doi: 10.3390/electronics15020338.
- [21] N. Weber and N. Brewer, 'The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition.', *J. Appl. Psychol.*, vol. 88, no. 3, pp. 490–499, 2003, doi: 10.1037/0021-9010.88.3.490.
- [22] C. Eliades, L. Lenc, P. Král, and H. Papadopoulos, 'Automatic face recognition with well-calibrated confidence measures', *Mach. Learn.*, vol. 108, no. 3, pp. 511–534, Mar. 2019, doi: 10.1007/s10994-018-5756-7.
- [23] C. Tomani, D. Cremers, and F. Buettner, 'Parameterized Temperature Scaling for Boosting the Expressive Power in Post-Hoc Uncertainty Calibration', in *Computer Vision – ECCV 2022*, vol. 13673, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds, in *Lecture Notes in Computer Science*, vol. 13673, Cham: Springer Nature Switzerland, 2022, pp. 555–569. doi: 10.1007/978-3-031-19778-9_32.
- [24] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, 'Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments', *University of Massachusetts, Amherst, MA, USA*, 07–49, Oct. 2007.