F.R.A.M.E.
Facial Representation Averaging for Matching Efficiency

Paolo Cursi (2155622) *Computer science*
Michele Palma (1849661) *Computer science*
cursi.2155622@studenti.uniroma1.it
palma.1849661@studenti.uniroma1.it

Biometric Project Fall 2024

Università degli Studi di Roma La Sapienza

Abstract

In this project, we aim to design and develop a lightweight model capable of performing in real time without losing performance. To do this we generate the embeddings by averaging multiple embeddings of the same person, which can also be reproduced in real time using frames.

1 Introduction

Face identification systems play a critical role in a wide range of biometric and security applications, from surveillance to access control. Traditional identification models typically operate under the *closed set* assumption, where every probe is known to belong to one of the enrolled identities in the gallery. However, real-world scenarios often require handling inputs from unknown individuals which are not present in the gallery, necessitating a shift towards *open set identification*.

Open set face identification presents a more challenging problem: the system must not only *match* probes to the correct identity when present in the gallery, but also reliably *reject* probes belonging to unknown subjects. This dual objective requires models that are both highly discriminative and capable of uncertainty estimation or threshold-based rejection.

In this work, we develop and evaluate a lightweight, real-time face identification system tailored for open set conditions. The system employs a compact embedding extractor trained to produce robust facial representations and aggregates information from multiple frames to enhance recognition accuracy. We benchmark the model on open set identification metrics such as Receiver Operating Characteristic (ROC) curves and Cumulative Match Characteristic with rejection (CMC-K), in order to analyze both its identification performance and rejection capability.

Our goal is to demonstrate that high precision, efficient face identification is feasible even under the open set assumption where rejecting impostors is as important as correctly recognizing known identities.

2 Dataset

The dataset we chose is VGGFace2, a public research face recognition dataset released by Oxford's Visual Geometry Group. It contains millions of images covering thousands of different people and each person has hundreds of pictures shot "in the wild," so we get tons of variety in pose, lighting, age, expression and even partial occlusions like hats or sunglasses. Every image already comes with a tight face bounding box and a verified label and the dataset is split into train, validation and test folders. Because the photos were scraped from real-world sources (Google Images and YouTube thumbnails) and then manually cleaned, the collection feels much closer to what a real app would see than studio shots.

3 The F.R.A.M.E network

3.1 The network structure

3.2 High-level topology

Our model is a two-stage hierarchical Vision Transformer (ViT) inspired by TinyViT [2]. Starting from an RGB input of 224 × 224:

- We first apply a shallow **3x3 convolution** that halves the spatial size
- We embed the image using a **4x4 convolution** using a stride of 4 (non-overlapping patches), it returns the tokens.
- We flatten the tokens, add the CLS token and the relative position biases (to ensure the learning of the relative position of the features).
- We give as input the tokens to a TinyViT (**Vision Transformer**) using Windowed self-attention in order to make it faster [Hierarchical stage 1]
- We merge the patches
- We put another **transformer block**
- We project the features in a 512-dimensional space using a **linear layer** (those will be the computed embeddings)
- Using an **ArcFace classification** head (state of the art for the face identification task) we classify through 100 classes.

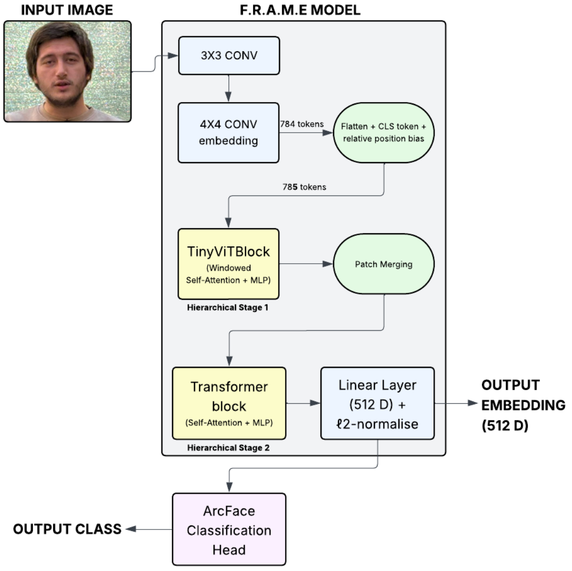
INPUT IMAGE
F.R.A.M.E MODEL
3X3 CONV
4X4 CONV embedding
784 tokens
Flatten + CLS token + relative position bias
785 tokens
TinyViTBlock (Windowed Self-Attention + MLP)
Hierarchical Stage 1
Patch Merging
Transformer block (Self-Attention + MLP)
Hierarchical Stage 2
Linear Layer (512 D) + ℓ2-normalise
OUTPUT EMBEDDING (512 D)
ArcFace Classification Head
OUTPUT CLASS

Figure 1: Training Architecture

3.3 ArcFace classifier

For supervised training we adopt the ArcFace margin-based cosine loss [1]. Given a label y and embedding $\mathbf{e}$, logits are computed as

$$\ell_k = s\,\cos\bigl(\theta_k + \mathbf{1}_{k=y} m\bigr), \quad \theta_k = \arccos\bigl(\langle \mathbf{e}, \mathbf{w}_k \rangle\bigr),$$

with scale $s = 64$ and margin $m = 0.5$. This forces intra-class embeddings to cluster tightly on the hypersphere while pushing different identities apart.

3.4 Inference

At inference time we get rid of the classification head and we just use the output embeddings to classify a subject based on the cosine similarity with other embeddings.

3.5 Real time inference

We apply a pre-processing phase to the frames by cropping them and then calculating the embeddings. To support continuous recognition in streaming scenarios, the system maintains a sliding window of the most recent k frames. Initially the embeddings are computed for the first k frames to establish a reference. For each subsequent frame, the system computes the new embedding, discards the oldest frame embedding in the window and updates the average embedding accordingly. This mechanism enables efficient, low-latency identity representation over time while maintaining robustness.

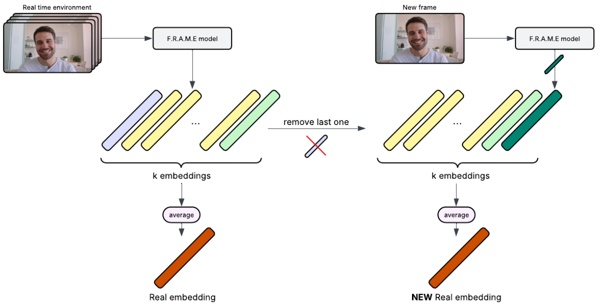

Figure 2: Real time inference

3.6 Model size and computation

The F.R.A.M.E model has **808,032 parameters**, a very small number compared to state-of-the-art models for this task, making it very suitable for real-time applications.

3.7 Embedding aggregation (the F.R.A.M.E. core)

At inference time we observe T frames of the same subject and compute their embeddings $\{\mathbf{e}_1, \ldots, \mathbf{e}_T\}$. The embedding of a single subject is computed by averaging and then normalizing k embeddings of the same person:

$$\bar{\mathbf{e}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{e}_t, \qquad \|\bar{\mathbf{e}}\|_2 = 1.$$

As we'll see later, increasing the value of k makes the model significantly more accurate.

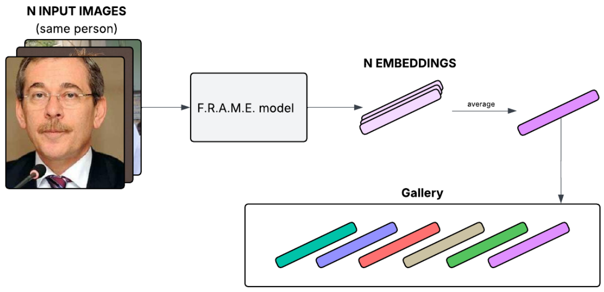

Figure 3: Gallery generation

4 Network Evaluation

4.1 Evaluation Metrics

To assess the performance of the proposed open set face identification system we rely on standard biometric metrics suitable for open set scenarios: *Receiver Operating Characteristic* (ROC), *Detection and Identification Rate* (DIR), *False Positive Identification Rate* (FPIR), *True Positive Labeling Rate* (TPLR), *True Accept Rate at Fixed False Accept Rate* (TAR@FAR) and the *Equal Error Rate* (EER).

4.2 Receiver Operating Characteristic (ROC)

The ROC curve characterizes the trade-off between the **True Positive Rate (TPR)** and the **False Positive Rate (FPR)** at varying thresholds τ:

$$\mathrm{TPR}(\tau) = \frac{\#\text{ known probes accepted}}{\#\text{ total known probes}}$$

$$\mathrm{FPR}(\tau) = \frac{\#\text{ unk. probes incorrectly accepted}}{\#\text{ total unknown probes}}$$

Each ROC point corresponds to a threshold τ, with the Area Under the Curve (AUC) summarizing overall separability between known and unknown identities.

4.3 Detection & Identification Rate (DIR)

The Detection & Identification Rate (DIR) is a comprehensive metric that measures the system's ability to both detect the presence of a known identity and correctly identify it. Unlike traditional identification metrics, DIR accounts for the open-set nature of the problem by considering both detection and identification performance simultaneously.

$$\mathrm{DIR}(\tau) = \frac{\#\text{ correctly probes at threshold } \tau}{\#\text{ total known probes}}$$

This metric provides a holistic view of system performance by combining the detection of known identities with their correct identification, making it particularly suitable for open-set face identification scenarios.

4.4 False Positive Identification Rate (FPIR)

In open set scenarios FPIR measures the rate at which unknown identities are incorrectly accepted and matched to a gallery identity. It is defined as:

$$\mathrm{FPIR}(\tau) = \frac{\#\text{ accepted unknown probes}}{\#\text{ total unknown probes}}$$

A lower FPIR indicates better ability to reject identities not present in the gallery.

4.5 True Positive Labeling Rate (TPLR)

TPLR evaluates the proportion of known probes that are correctly accepted and correctly identified (i.e. their true identity is ranked at the top position). It is defined as:

$$\mathrm{TPLR}(\tau) = \frac{\#\text{ accepted known probes (true ID)}}{\#\text{ total known probes}}$$

The accepted known probes are calculated with rank 1. This metric captures the overall ability of the system to both accept and correctly recognize known identities.

4.6 True Accept Rate at Fixed False Accept Rate (TAR@FAR)

TAR@FAR is a standard metric in face verification that quantifies how well a system correctly verifies genuine identity pairs while maintaining a strict limit on false acceptances. It is defined as:

$$\mathrm{TAR}(\tau) = \frac{\#\text{ accepted genuine pairs at threshold } \tau}{\#\text{ total genuine pairs}}$$

The threshold τ is selected such that the False Accept Rate (FAR) satisfies:

$$\mathrm{FAR}(\tau) = \frac{\#\text{ accepted impostor pairs at } \tau}{\#\text{ total impostor pairs}} = \alpha$$

Thus TAR@FAR evaluates the system's ability to correctly verify genuine users, given that only a small fraction α (e.g., 10^{-4}) of impostor pairs are incorrectly accepted.

4.7 Equal Error Rate (EER)

Equal Error Rate is a widely used performance metric in biometric verification and open set face recognition. It represents the operating point where the rate of false acceptances equals the rate of false rejections. This threshold provides a single scalar value summarizing the trade-off between security (low FAR) and usability (low FRR).

Formally the EER is defined by the threshold τ^* at which the False Accept Rate (FAR) and False Reject Rate (FRR) are equal:

$$\tau^* = \arg\min_{\tau} |\text{FAR}(\tau) - \text{FRR}(\tau)|$$

$$\text{EER} = \text{FAR}(\tau^*) = \text{FRR}(\tau^*)$$

Where:

- $\text{FAR}(\tau) = \frac{\#\text{ accepted impostor pairs at } \tau}{\#\text{ total impostor pairs}}$
- $\text{FRR}(\tau) = \frac{\#\text{ rejected genuine pairs at } \tau}{\#\text{ total genuine pairs}}$

The EER is particularly useful when no prior preference is given to minimizing either FAR or FRR. A lower EER indicates a more balanced and generally more accurate system.

4.8 Score and Label Preparation

To compute the above metrics the following processing steps are required:

- For each probe:
  - Compute similarity scores to all gallery identities
  - Identify the maximum similarity score
  - Determine the rank of the correct identity (e.g. for CMC or DIR)
- Apply threshold τ:
  - Accept if $\max(\text{similarity}) \geq \tau$
  - Reject otherwise
- Label probes as *known* or *unknown* accordingly

4.9 Evaluation

Evaluation involves sweeping over thresholds τ to generate ROC curves. Moreover it has been used to computer DIR, FPIR, TPLR, FAR and TAR. Together, these metrics provide a comprehensive picture of the system's open set identification performance. Moreover we'll also mention a previously trained model with 600000 parameters that had some issues regarding meaningful embedding extraction.

In the preliminary phase the first evaluation metric used was classification accuracy. Although not strictly aligned with the task, this metric was essential to monitor the performance of the classification head within FRAME. Achieving high accuracy, particularly on the validation set, indicated that the model was learning to extract semantically meaningful embeddings from input images. Once satisfactory accuracy was reached, the classification head was removed, allowing the model to function purely as an embedding extractor for the probe images.

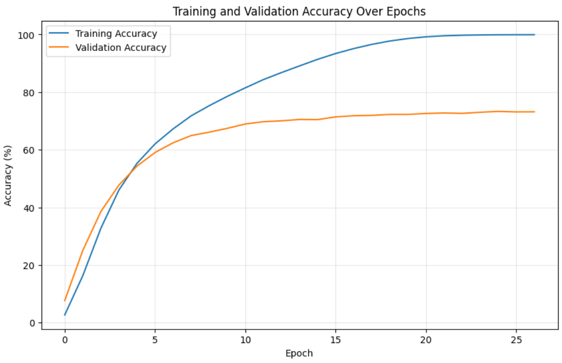

Figure 4: Train and Validation Accuracy of FRAME

We tested the FRAME model also in a closed set scenario in which we extracted an embedding from a subject present in the gallery and then verified the accuracy of the model. The trend was a logarithmic increase of the accuracy as a function of the amount of frames utilized to calculate the subject's embedding. The precision with a number of frames ≥ 9 was 100%.

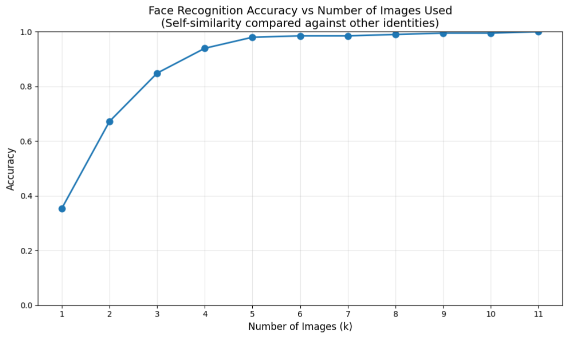

Figure 5: Rank 1 accuracy with no rejection with FRAME (closed set evaluation)

The previous model performance on the same task was quite different, in fact even at more than double the amount of frames needed by FRAME to obtain 100% accuracy, its accuracy was as good as random guessing. This indicated that using only 600 thousand parameters wasn't enough in order to be able to extract meaningful embeddings.

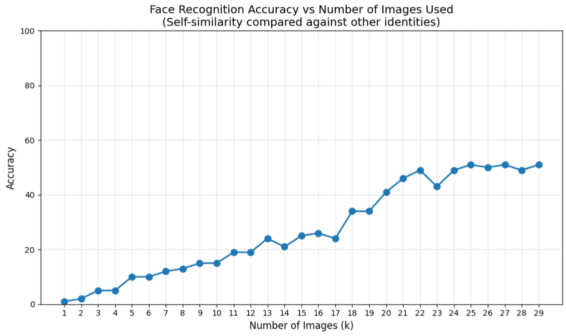

Figure 6: 600k Model rank 1 accuracy with no rejection

Figure 7 shows the True Accept Rate (TAR) at various False Accept Rate (FAR) thresholds as a function of the number of images per identity (k) in the gallery. This evaluation is performed in the open-set face re-identification setting, where not all probe identities are present in the gallery.

- **Impact of k:** As the number of images per identity increases, TAR consistently improves across all FAR levels. This demonstrates that having more images per person in the gallery significantly enhances recognition robustness, likely due to better representation of intra-class variations.

- **Effect of FAR threshold:** At high FAR thresholds (e.g., 0.01 and 0.001), the model achieves near-perfect TAR even with relatively small k. However at more stringent FARs (e.g., 10^{-5} and 10^{-6}) the TAR improves more gradually and requires a larger k to approach saturation. This indicates the system's challenge in maintaining high precision under stricter acceptance criteria.

- **Model robustness:** The steep increase in TAR with k at all FAR levels demonstrates the model's ability to generalize well with more data per identity.

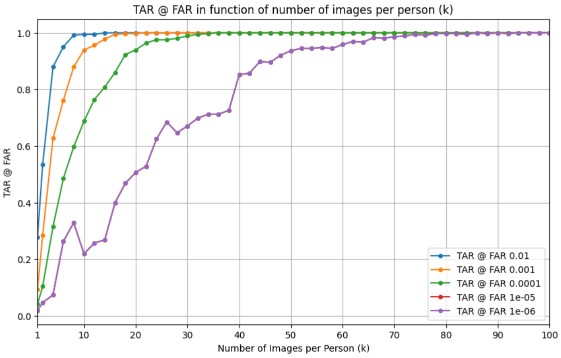

Figure 7: TAR@FAR as a function of the number of frames used by FRAME

Figure 8 illustrates Receiver Operating Characteristic (ROC) curves for different values of k (images per identity).

- **Effect of k on discriminability:** As k increases, the model's ability to distinguish between known and unknown identities improves substantially. This is reflected in the increasing Area Under the Curve (AUC) values from 0.8166 for $k = 1$ to nearly perfect AUC values above 0.9999 for $k = 20$ and beyond

- **High-FAR vs. Low-FAR Performance:** When only one image per identity is available ($k = 1$) the ROC curve shows weaker performance, especially in the low-FAR region which is critical in high-security applications. As k increases the curve shifts closer to the top left corner, indicating stronger performance even under strict false acceptance constraints

- **Saturation Point:** The model's performance saturates quickly with $k \geq 10$, where the ROC curves become nearly indistinguishable. This suggests that the model can achieve near-optimal discrimination with a moderate number of images per person

- **Random Baseline:** The dotted line represents the performance of a random classifier (AUC = 0.5) included for reference

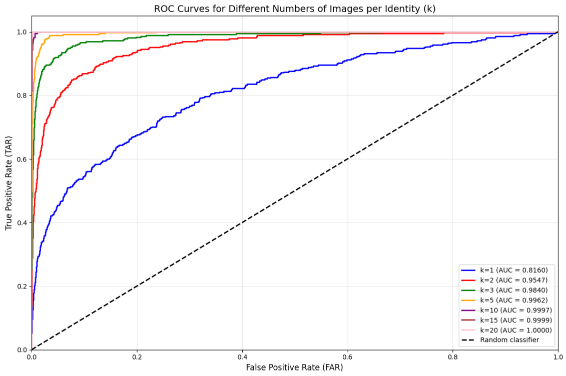

Figure 8: ROC as a function of the number of frames used by FRAME

Figure 9 illustrates the values of FPIR and TPLR for different values of k (images per identity) at different τ thresholds.

- **Effect of τ threshold on discriminability:** as τ increases the False Positive Identification Rate significantly drops independently of the number of frames used. In the case of τ equal to 0.9 the amount of frames needed to correctly identify a subject is more than double with respect to τ equal to 0.8

- **Effect of k on discriminability:** The general trend of the model is to become increasingly more accurate with more frames. On average with a number of $k \geq 20$ the model reaches a very high TPLR

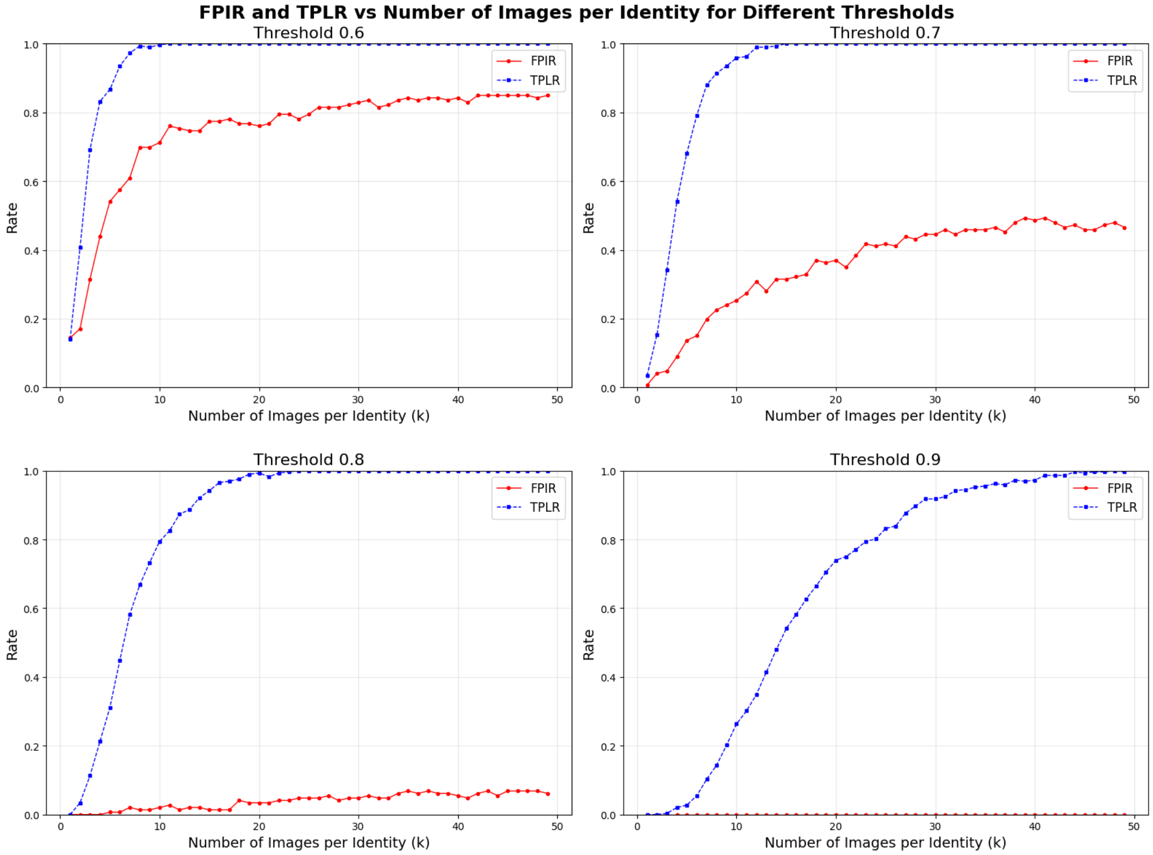

Figure 9: FPIR and TPLR at different Frames and Thresholds

Figure 10 compares the performance of the face recognition model under two settings: $k = 1$ and $k = 50$ images per identity. The left column shows the relationship between the similarity threshold and the FAR (False Acceptance Rate) and TAR (True Acceptance Rate), while the right column presents histograms of maximum cosine similarities for impostor and legitimate comparisons.

- **Performance with limited data ($k = 1$):** With only one image per identity, the overlap between impostor and legitimate similarity distributions is substantial (top-right), leading to high FAR even at moderate thresholds. The TAR is modest across thresholds, reflecting the model's limited capacity to generalize with sparse data (top-left)

- **Improvement with more data ($k = 50$):** With richer identity representation ($k = 50$), the legitimate similarities shift sharply toward higher cosine values while impostor similarities remain low (bottom-right). This clear separation allows the model to achieve near-perfect TAR and very low FAR at reasonable thresholds (bottom-left)

- **Threshold robustness:** For $k = 50$, the model maintains a stable TAR ≈ 1.0 across a wide range of thresholds, demonstrating robustness and reliability

- **Discriminative margin:** The visual contrast in the histogram plots underscores the increased discriminative margin between impostor and legitimate similarities as k increases. This margin is key to achieving high verification performance with fewer false matches

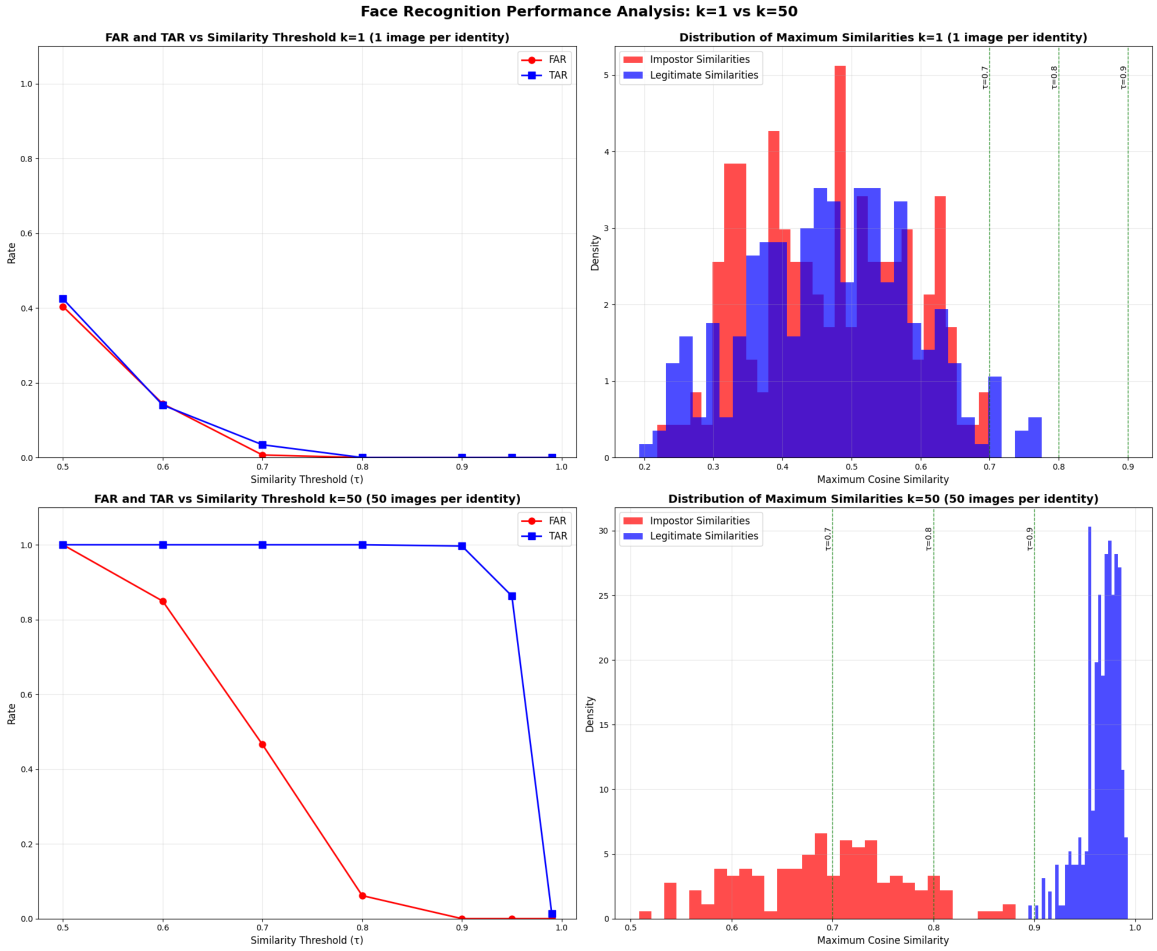

Figure 10: General performance of FRAME with 1 frame vs 50 frames used for the embeddings

Figure 11 illustrates the equal error rate by varying the number of utilized frames by the FRAME model.

- **Impact of k on model accuracy:** As the number of images per identity increases from $k = 1$ to $k = 50$, the EER significantly decreases. This indicates that the model benefits from a more comprehensive representation of each identity, allowing for more accurate similarity comparisons
- **EER trend:** For $k = 1$, the EER is relatively high due to limited identity information, making the model more prone to both false acceptances and rejections. At $k = 15$, there is a noticeable reduction in EER, and by $k = 50$, the EER is minimal, reflecting highly reliable verification performance
- **Threshold sensitivity:** The FAR and FRR curves for low k values intersect closer to the center of the threshold range, indicating higher uncertainty. In contrast, as k increases, the intersection shifts toward higher similarity thresholds, implying greater model confidence in distinguishing between identities

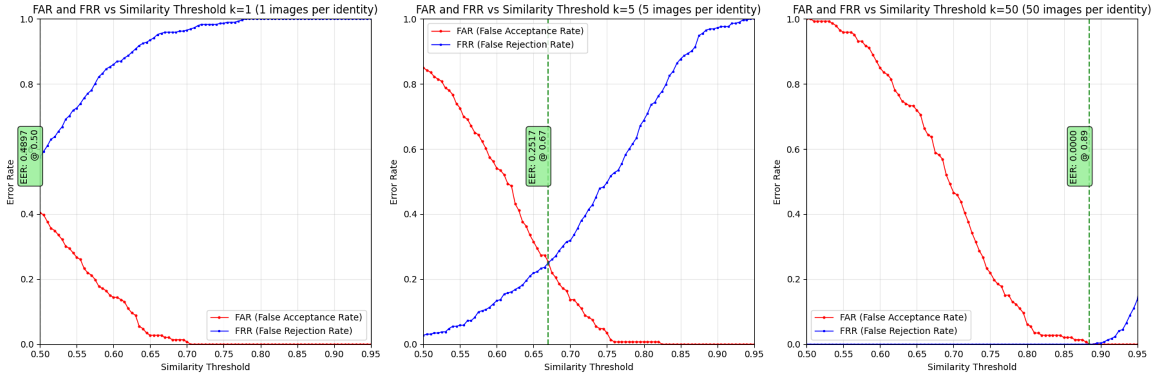

Figure 11: Equal Error Rate for FRAME

4.10 Detection & Identification Rate (DIR)

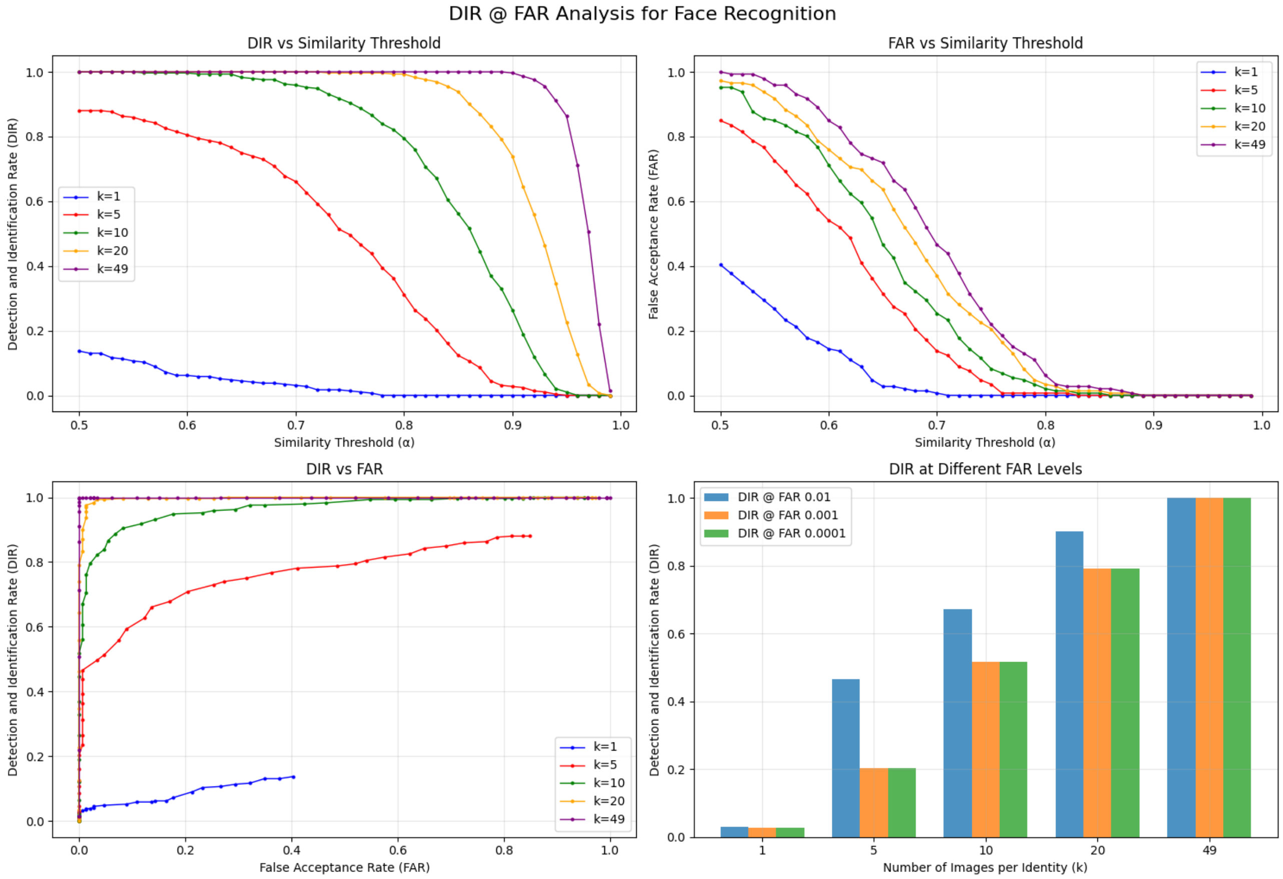

Figure 12: DIR–FAR analysis for different gallery depths k.

- **DIR vs. threshold (top-left)** – One-shot enrollment (k=1) tops out near 15 % DIR; from k=10 upward the curve stays at ≈1.0 until $\alpha \sim 0.8$.
- **FAR vs. threshold (top-right)** – FAR plummets below 10^{-3} once $\alpha > 0.8$ for any k; larger k slightly raises FAR only at very low thresholds.
- **DIR vs. FAR (bottom-left)** – With $k \geq 10$ the system achieves DIR≈1.0 even when FAR is forced under 10^{-3}; k=1 remains far from the top-left corner.
- **DIR at fixed FAR (bottom-right)** – At FAR = 10^{-4}, DIR climbs from 1 % (k=1) to 100 % once $k \geq 20$, showing clear gains from embedding averaging.

Practical takeaway: averaging 10–20 frames per identity delivers near-perfect identification (DIR ≥ 99 %) while keeping impostor acceptance below one in a thousand.

5 Conclusions

The proposed model demonstrates high stability and precision when provided with multiple frames. While its performance exhibits pseudo-random behavior with a single frame input, it becomes significantly more consistent and accurate with 10 frames. Beyond 20 frames the model reliably distinguishes between genuine identities and impostor probes, indicating a strong discriminative capability under multi-frame settings.

Despite its lightweight architecture (≤1 million parameters) the model achieves competitive performance compared to larger architectures with ≥14 million parameters, which are typically capable of achieving similar accuracy using only a single frame. The model's compact size is a deliberate design choice to support real-time inference.

References

[1] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2019, pp. 4690–4699.

[2] Kan Wu et al. “TinyViT: Fast Pretraining Distillation for Small Vision Transformers”. In: *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 68–85.