Efficient Face Recognition Using Embedding-Based Distillation

Hana Remma, Chaimaa Ouarezki, Youcef Ouadjer, Mourad Adnane, and Sid-Ahmed Berrani

Abstract—Knowledge distillation (KD) facilitates the compression of large, high-performing neural networks into efficient student models, enabling deployment on resource-limited devices like mobile phones and IoT systems. This paper introduces a KD methodology, which involves training a student to capture a teacher's soft labels, intermediate feature representations, and ground truth labels, ensuring both compactness and accuracy. Applied to face recognition, our hybrid KD framework trains a MobileFaceNet student under an InceptionResNetV1 teacher, achieving 90.40% accuracy and a 96.25% AUC, outperforming lightweight models while remaining suitable for edge devices. These results highlight the potential of KD to enable robust, scalable face recognition solutions for real-world, resource-constrained environments.

***Keywords*—Face Recognition, Knowledge Distillation, Mobile devices, Efficient Deep learning.**

NOMENCLATURE

KD	Knowledge Distillation.
FAR	False Acceptance Rate.
FRR	False Rejection Rate.
EER	Equal Error Rate.
AUC	Area Under the Curve
ROC	Receiver Operating Characteristics

I. INTRODUCTION

Face recognition is a technology with numerous real-world applications, such as biometric authentication, video surveillance, and human-computer interaction. Recent advances in deep learning have led to high-performing models like ArcFace [1], CosFace [2], and MagFace [3], which achieve remarkable accuracy on large-scale benchmarks. However, these state-of-the-art (SOTA) models typically rely on deep architectures with hundreds of millions (100) of parameters and high computational complexity, making them unsuitable for deployment on resource-constrained devices such as smartphones, IoT nodes, and embedded systems.

To address this limitation, several strategies have been explored to reduce model size while preserving accuracy. Model compression techniques, including pruning [4] and quantization [5], aim to eliminate redundant weights or reduce numerical precision. While these methods significantly reduce storage and inference costs, they often require meticulous fine-tuning and may result in performance degradation, especially on challenging face recognition tasks. For example, the lightweight architecture of ShuffleFaceNet [6], constructs compact models from the ground up using depthwise separable convolutions or channel shuffling. These models offer faster inference and lower memory footprints but still fall short in recognition accuracy compared to their larger counterparts.

An alternative and increasingly popular direction is knowledge distillation (KD), initially introduced by Hinton et al. [7, 8], which transfers knowledge from a large teacher model to a smaller student. In the context of face recognition, KD has been applied to improve embedding quality by aligning student features with those of a pretrained teacher network.

In this paper, we propose a lightweight and effective face recognition framework that leverages a hybrid knowledge distillation approach. Our method enables a compact student model to be trained under the supervision of a pre-trained teacher network using both classification loss and embedding-based distillation loss. Unlike existing distillation methods that often require complex formulations or additional modules, our approach adopts a simple yet powerful dual-loss strategy that encourages the student to simultaneously learn identity labels and replicate the teacher's rich embedding space. This ensures that the student inherits both the discriminative capability and relational structure of the teacher, achieving a balance between compactness and accuracy suitable for on-device deployment. This paper is organized as follows: Section II. dives into the methodology of the proposed knowledge distillation framework. Section III. provides interpretation and discussion of the obtained results. Finally, section IV. concludes the paper with future directions.

II. METHODOLOGY

Knowledge distillation (KD) offers a robust set of benefits, making it a pivotal technique for model compression and deployment. It enables a compact model, often termed the student, with significantly reduced computational cost, to emulate the

Manuscript received July, 2025; revised December 28, 2025.

H. Remma, C. Ouarezki, and Y. Ouadjer are with the Electronics Department, Ecole Nationale Polytechnique, 10 Rue des Frères OUDEK, El Harrach 16200, Algiers, Algeria (e-mail: hana.remma@g.enp.edu.dz; chaimaa.ouarezki@g.enp.edu.dz; youcef.ouadjer@g.enp.edu.dz).

M. Adnane is with National Higher School of Autonomous Systems Technology, and LDCCP LAB, Ecole Nationale Polytechnique, 10 Rue des Frères OUDEK, El Harrach 16200, Algiers, Algeria (e-mail: mourad.adnane@g.enp.edu.dz).

S-A. Berrani is with National School of Artificial Intelligence, Route de Mahelma, 16201 Sidi Abdellah, Algiers, Algeria (email: sidahmed.berrani@ensia.edu.dz).

Digital Object Identifier (DOI): 10.53907/enpesj.v5i2.335

2716-912X © 2025 Ecole Nationale Polytechnique

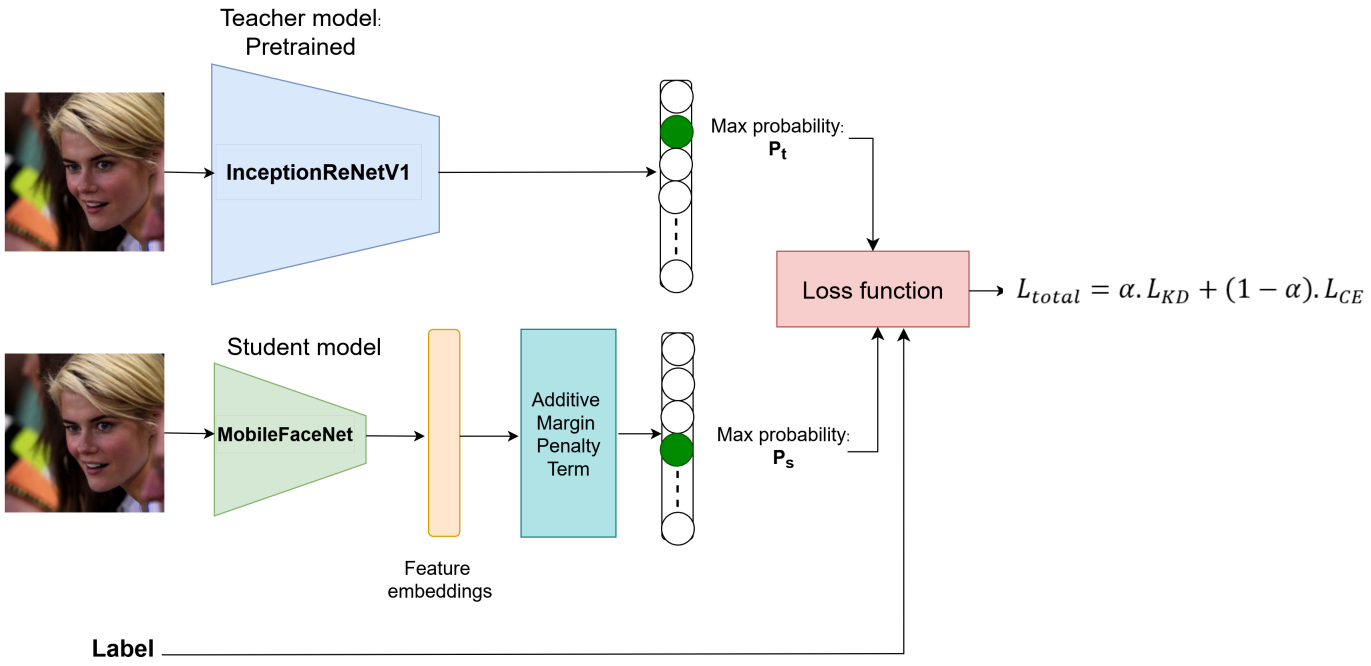

Fig. 1: Conceptual diagram of knowledge distillation, illustrating the teacher model (top) transferring soft labels to the student model (bottom), alongside ground truth labels shaping the total loss.

capabilities of a larger, high-performing teacher model, facilitating deployment on resource-constrained platforms such as mobile devices or IoT systems.

As illustrated in Figure 1, the InceptionResNetV1 [9] and MobileFaceNet [10] are used as teacher and student models respectively. The main advantage of using InceptionResnNetV1 is that it is trained on large publicly available datasets, and it encompasses general face feature representations. The MobileFaceNet on the othe hand is a compact deep neural network dedicated for facial recognition applications.

It is worth noting that the structure of InceptionResNetV1 and MobileFaceNet are different, the motivation for this difference is to enable knowledge transfer across diverse architectures, as evidenced by [11] and [12]. This flexibility allows tailoring student models to specific hardware or latency requirements while leveraging the teacher's expertise.

The proposed KD framework relies on three critical components: the teacher's soft labels, intermediate feature representations, and ground truth labels from the dataset. Drawing on foundational contributions by Hinton et al. [7] and Deng et al. [1], the following subsections detail the learning mechanisms, weaving mathematical precision with practical considerations.

A. Soft Labels: Capturing Class Relationships

The teacher, a deep neural network with significant capacity, produces soft labels-probability distributions over classes for each input that encode inter-class similarities. The distilled model often matches or surpasses the teacher's performance on specific tasks despite its smaller size. By learning softened probability distributions, the student can replicate nuanced patterns, offering more information than binary hard labels. The teacher's logits, $z_t = [z_{t,1}, \ldots, z_{t,C}]$, where C denotes the number of classes, are softened using a temperature parameter T:

$$P_{t,i} = \frac{\exp(z_{t,i}/T)}{\sum_{j=1}^{C} \exp(z_{t,j}/T)} \tag{1}$$

where $P_{t,i}$ is the softened probability for class i. Similarly to the teacher model, the student model produces soft-labels probability $P_{s,i}$ using its logits $z_s = [z_{s,1}, \ldots, z_{s,C}]$ to mimic teacher's output:

$$P_{s,i} = \frac{\exp(z_{s,i}/T)}{\sum_{j=1}^{C} \exp(z_{s,j}/T)} \tag{2}$$

To align these distributions, the student minimizes the Kullback-Leibler (KL) divergence, defining the distillation loss:

$$\mathcal{L}_{\text{KD}} = T^2 \sum_{i=1}^{C} P_{t,i} \log \frac{P_{t,i}}{P_{s,i}} \tag{3}$$

The T^2 scaling ensures the loss remains balanced for large T, enabling the student to internalize the teacher's generalizations, particularly for complex class boundaries [7].

By leveraging soft labels, KD enhances generalization capability of the student, and provides richer information than hard labels. These labels reveal class similarities, enabling the student to handle ambiguous inputs effectively. For example, a teacher assigning probabilities to "dog" and "wolf" guides the student toward nuanced patterns. Hinton et al. [7] showed KD-trained students outperform hard-label-trained models on complex datasets. Peng et al. [13] found that KD promotes smoother class distributions in remote sensing tasks, improving robustness to input variations.

B. Intermediate Feature Representations

Building on soft labels, and feature embeddings of the student, it is possible to learn robust intermediate representations. A method advanced by Deng et al. [1] in their ArcFace framework, proposed a highly discriminative face feature Alorithm, based on an additive angular margin penalty term m.

In this work we use the ArcFace additive angular margin penalty to enhance the discriminative power of deep features for face recognition. Given an input feature embedding vector $\mathbf{x_s} \in \mathbb{R}^d$, produced by the student model, and a weight vector $\mathbf{W}_i \in \mathbb{R}^d$ for class i, both are first normalized to compute the cosine similarity as $\cos(\theta_i) = \frac{\mathbf{x}^\top \mathbf{W}_i}{\|\mathbf{x}\|\|\mathbf{W}_i\|}$. For the ground-truth class y, an angular margin m is added in the cosine space, modifying the similarity as:

$$\cos(\theta_y + m) = \cos(\theta_y)\cos(m) - \sqrt{1 - \cos^2(\theta_y)}\sin(m) \quad (4)$$

Finally the student logits are scaled by a factor s, based on:

$$z_{s,i} = \begin{cases} s \cdot \cos(\theta_y + m), & \text{if } i = y \\ s \cdot \cos(\theta_i), & \text{otherwise} \end{cases} \quad (5)$$

This formulation effectively increases the angular margin between classes in the normalized hypersphere, leading to improved inter-class separability and intra-class compactness.

C. Ground Truth Labels

To ensure task-specific accuracy, the student is trained on ground truth labels via cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(P_{s,i}) \quad (6)$$

where y_i is the one-hot ground truth label, and $P_{s,i}$ is the student's probability (with $T = 1$). This loss anchors the student's predictions to the dataset, balancing the teacher's influence with direct task learning [7].

D. Combined Objective

The student's training optimizes a composite loss:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{KD}} + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} \quad (7)$$

where $\alpha \in [0, 1]$ balances distillation and cross-entropy losses. Typical values include $\alpha = 0.5\check{}0.9$, $T = 2\check{}10$. This balance is often an empirical, enabling the student to synthesize teacher insights and ground truth effectively [7, 11, 12, 14]. These hyperparameters require careful tuning, often through validation, to reconcile objectives like teacher mimicry and task fidelity.

E. Experimental Setup

Our facial recognition system is based on the MobileFaceNet (student model), trained through knowledge distillation from a high-capacity InceptionResNetV1 (teacher model). Training was conducted using the Adam optimizer with a learning rate of 0.01 and weight decay of 10^{-4}.

While teacher training is resource-intensive, KD limits this cost to the student pre-training phase. The student requires minimal resources for fine-tuning or deployment, reducing production expenses.

In this work, we used the Labeled Faces in the Wild (LFW) dataset [15], to pre-train and evaluate the MobileFaceNet model. The LFW is a well established and commonly used dataset to evaluate face recognition models. It contains over 13 000 images of faces, with 5 749 identities, collected from the web, with three different splits: train, evaluation and test sets.

III. Results and Discussion

This section presents the results. First, a description of the evaluation metrics is given. Then, results are presented with an in-depth discussion. Finally, a comparison of the obtained results with MobileFaceNet against state-of-the-art models is presented.

A. Evaluation Metrics

We evaluated performance using the following standard metrics:

- **Accuracy** (%): Percentage of correctly classified pairs.
- **FAR False Acceptance Rate (%):** refers to the impostor accepted as genuine.
- **FRR False Rejection Rate (%):** is the genuine rejected as impostor.
- **EER Equal Error Rate (%):** is the point where FAR = FRR.
- **AUC (%):** Area Under the Receiver Operating Characteristics (ROC) Curve.

B. Obtained Results

Table. I
EXPERIMENTAL RESULTS OF THE PROPOSED METHOD

Accuracy (%)	FAR (%)	FRR (%)	EER (%)	AUC (%)
90.40	**0.87**	**18.33**	**18.66**	**96.25**

The experimental results illustrated in Table I, highlight the effectiveness of the proposed KD approach in training a lightweight face recognition model with high performance and low computational cost. The MobileFaceNet student achieved a top-1 accuracy of 90.4%, an AUC of 96.25% (Figure 2), and a FAR of 0.87%, demonstrating its capability to inherit rich representations from the InceptionResNetV1 teacher, which enhances generalization and avoids overfitting typical of smaller models trained from scratch.

This approach is highly practical for resource-constrained devices. The low FAR underscores the robustness of the learned embeddings, critical for security applications like biometric authentication, surpassing lightweight alternatives with less discriminative features. However, the high FRR and EER (around 18%) indicate a limitation, likely due to dataset biases or insufficient feature diversity in the student model. To reduce FRR and EER it is possible to explore the hyperparameter space by tuning T, α, and the learning rate with advanced regularization like dropout, batch normalization and early stopping. Overall, this method balances performance and efficiency, offering a scalable solution for real-world face recognition.

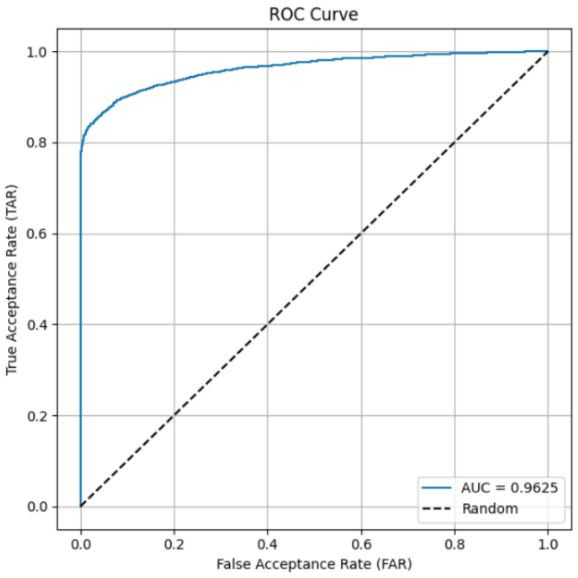

Fig. 2: ROC curve of MobileFaceNet trained via knowledge distillation.

C. Comparison with State-of-the-Art Models

Table. II
COMPARISON WITH LIGHTWEIGHT AND HEAVYWEIGHT SOTA MODELS

Model	Accuracy (%)	# of parameters (M)
Pyramid CNN [16]	85.5	-
DCMN [17]	90.00	0.5
ArcFace (ResNet50) [1]	98.2	23.5
CosFace (ResNet101) [2]	98.1	44.5
Ours (MobileFaceNet)	**90.4**	**0.99**

The distilled MobileFaceNet model is compared to state-of-the-art models in terms of accuracy, and number of trainable parameters in millions (M). Results are depicted in Table II. Compared to Pyramid CNN [16] and DMCN [17] models, our approach shows clear improvements in accuracy, making it suitable for embedded systems. While heavyweight models such as ArcFace [1] (23.5M) and CosFace [2] (44.5M) outperform ours in raw accuracy, they require significantly more computational resources, making them less practical for edge devices.

IV. CONCLUSION

In this work we demonstrated the efficacy of knowledge distillation (KD) in crafting efficient, high-performing models for face recognition, with broad implications for resource-constrained applications such as smartphone devices. Our hybrid KD approach, employing MobileFaceNet as the student model under an InceptionResNetV1 teacher, achieves 90.40% accuracy and 96.25% AUC, outperforming lightweight models while remaining viable for edge deployment. The methodology integrates soft labels, feature alignments, and ground truth learning, ensuring the student captures the teacher's nuanced knowledge, as evidenced by a robust False Acceptance Rate of 0.87%. Our appraoch has the advantages of computational efficiency, cross-architecture versatility, and it is tailored for real-time applications like biometric authentication and IoT systems.

The elevated False Rejection Rate suggests refinement opportunities, with future research exploring multi-teacher or self-distillation by leveraging an ensemble of teacher models that combine their knowledge into a single student, merging diverse strengths.

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Long Beach, CA, USA), pp. 4685–4694, 2019.

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, UT, USA), pp. 5265–5274, 2018.

[3] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Virtual Event), pp. 14220–14229, 2021.

[4] H. Cheng, M. Zhang, and J. Q. Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10558–10578, 2024.

[5] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," 2021.

[6] Y. Martinez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza, "Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, (Seoul, South Korea), pp. 2721–2728, 2019.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Neural Information Processing Systems (NeurIPS) Deep Learning Workshop*, (Montreal, QC, Canada), 2015.

[8] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3048–3068, 2022.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

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

[11] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations (ICLR)*, (San Diego, CA, USA), 2015.

[12] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2017.

[13] X. Peng, J. Zhou, and X. Wu, "Distillation-based cross-model transferable adversarial attack for remote sensing image classification," *Remote Sensing*, vol. 17, no. 10, 2025.

[14] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, South Korea), 2019.

[15] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *ECCV Workshop on Faces in Real-Life*, (Marseille, France), 2008.

[16] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou, "Learning deep face representation," 2014.

[17] W. Deng, J. Hu, N. Zhang, B. Chen, and J. Guo, "Fine-grained face verification: Fglfw database, baselines, and human-dcmn partnership," *Pattern Recognition*, vol. 66, pp. 63–73, 2017.

Hana Remma
She is currently pursuing a degree in Electronics as a State Engineer at the École Nationale Polytechnique in Algiers, Algeria. Her academic interests include machine learning, image processing, and embedded systems. This paper represents her first contribution to scientific research.

Chaimaa Ouarezki
She is currently a fourth-year student in electronics at the École Nationale Polytechnique in Algeria. Her academic journey is centered on developing a solid foundation in electronics. She is particularly interested in biomedical engineering, robotics, and image processing, where she seeks to apply machine learning techniques to solve real-world problems.

Youcef Ouadjer
Received the Master of science degree in Biomedical and Electrical Engineering, from University of Mouloud Mammeri, Tizi Ouzou, Algeria, in 2017. He is currently pusrsuing the Ph.D. degree with Department of Electronics, Ecole Nationale Polytechnique. His research interests include soft biometrics, data engineering, applied machine learning.

Mourad Adnane
Mourad Adnane received his Engineering degree in Electronics in 2003 from USTHB, Algiers, Algeria, and earned a Ph.D. in System Design Engineering in 2009 from Yamaguchi University, Japan. He is currently a professor at the National Higher School of Autonomous Systems Technology and leads the Embedded Systems team within the LDCCP Laboratory at École Nationale Polytechnique. His research interests include instrumentation, signal processing, pattern recognition, and machine learning, with a particular focus on applications in biomedical engineering.

Sid-Ahmed Berrani
Is a Professor in Computer Science at the National High School of Artificial Intelligence in Sidi-Abdallah, Algeria. He obtained an engineering degree in computer science from the University of Sidi Bel-Abbès (Algeria) in 2000 and a Ph.D. degree in computer science from the University of Rennes 1 in 2004. He has been a researcher at Orange Labs in France, the head the Multimedia Content Analysis and Indexing R&D unit of Orange Labs, and an associate professor at École Nationale Polytechnique (Algiers). His research activities focus on image and video indexing, machine learning, multidimensional data analysis and Artificial Intelligence. He has authored or co-authored over fifty scientific publications and has filed 13 patents