Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076623

ARTICLE

MobiIris: Attention-Enhanced Lightweight Iris Recognition with Knowledge Distillation and Quantization

Trong-Thua Huynh[1,*], De-Thu Huynh[2], Du-Thang Phu[1], Hong-Son Nguyen[1] and Quoc H. Nguyen[3]

[1]Faculty of Information Technology II, Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam
[2]School of Computer Science & Engineering, The Saigon International University, Ho Chi Minh City, Vietnam
[3]Institute of Digital Technology, Thu Dau Mot University, Ho Chi Minh City, Vietnam

*Corresponding Author: Trong-Thua Huynh. Email: thuaht@ptit.edu.vn

Received: 23 November 2025; Accepted: 06 January 2026; Published: 09 April 2026

ABSTRACT: This paper introduces MobiIris, a lightweight deep network for mobile iris recognition that enhances attention and specifically addresses the balance between accuracy and efficiency on devices with limited resources. The proposed model is based on the large version of MobileNetV3 and adds more spatial attention blocks and an embedding-based head that was trained using margin-based triplet learning, enabling fine-grained modeling of iris textures in a compact representation. To further improve discriminability, we design a training pipeline that combines dynamic-margin triplet loss, a staged hard/semi-hard negative mining strategy, and feature-level knowledge distillation from a ResNet-50 teacher. Finally, we investigate the use of post-training float16 quantization to reduce memory footprint and latency for deployment on mobile hardware. Experiments on the challenging CASIA-IrisV4-Thousand dataset show that the full-precision MobiIris model requires only 12 MB of storage and 27 ms inference latency, while achieving an EER of 1.409%, VR@FAR = 1% of 98.184%, and CMC@1 of 94.785%, closely matching a ResNet-50 baseline that is more than 7× larger and slower. Under post-training quantization, the model shrinks to 5.94 MB with 13 ms latency and maintains a competitive balance between accuracy and efficiency compared to other optimized variants. These results demonstrate that a coherent combination of lightweight architecture design, attention mechanisms, metric-learning objectives, hard negative mining, and knowledge distillation yields a practical iris recognition solution suitable for secure, real-time authentication on mobile and embedded platforms.

KEYWORDS: Iris recognition; lightweight architecture; model optimization; attention mechanism; knowledge distillation; model quantization

1 Introduction

Compared with other biometric features, the unique and stable nature of the iris pattern makes it a reliable and non-invasive means of identification [1]. This capability has proven effective across a wide range of security applications, such as mobile and physical access control, thereby contributing to the expansion of the global iris recognition market. Through miniaturized near-infrared sensors, iris recognition has been deployed on smartphones, which enables contactless authentication in response to the post-pandemic demand for touchless technologies [2]. Reflecting this momentum, Mordor Intelligence projects that this market will grow from USD 5.14 billion in 2025 to USD 12.92 billion by 2030, with a compound annual growth rate of 20.23% [3]. Given this trend, accurate and efficient iris recognition solutions are increasingly important.

However, achieving a balance between accuracy and efficiency remains challenging. This is mainly because widely used state-of-the-art deep learning models are built on large convolutional architectures that demand substantial computational and memory resources. For instance, VGG-16 contains over 138 million parameters and requires several gigabytes of memory during inference [4]. Moreover, DenseNet-121, compared with most ResNet architectures, is more parameter-efficient, with about 8 million parameters, yet still has a memory footprint of hundreds of megabytes [5,6]. This high computational requirement makes mobile deployment difficult for iris recognition, as small and fast models are critical for user experience and scalability. Moreover, iris acquisition in uncontrolled conditions poses challenges, such as illumination variation, occlusion, and reflection, which can significantly degrade recognition performance.

Over the years, iris recognition research has steadily progressed, moving from classical image processing and handcrafted machine learning features to deep learning methods [7]. Despite these improvements, the high computational demand of traditional models hinders their deployment on resource-constrained devices, thereby motivating growing interest in lightweight architectures. On the ImageNet dataset, GhostNet, a lightweight architecture with an improved feature generation module, achieved the highest classification accuracy among lightweight models, with competitive computational cost [8]. Similarly, [9] evaluated a set of deep learning models for iris recognition under constrained conditions, showing that MobileNetV3 models reduced parameters by 87%/83% and shortened inference time by 80%/56% compared to ResNet-50/DenseNet-201, respectively, on OpenEDS dataset. Beyond the native lightweight architectures, deep convolutional networks have also been adapted into lite versions to achieve comparable performance with lower resource usage. For example, ResNet-Lite applies model optimization techniques to compress ResNet-50 while maintaining near-original performance, achieving improvements of 5.40% and 7.13% on CIFAR-10 and Fashion-MNIST datasets, respectively [10]. In general, these findings show that lightweight architectures can deliver high accuracy with reduced computational cost.

While lightweight architectures establish a solid foundation for mobile iris recognition, further enhancements in performance can be obtained through optimization techniques. Introduced in [11], knowledge distillation involves transferring knowledge from a larger and more complex teacher model to a smaller and simpler student model. Despite being widely applied in various computer vision tasks, it is less explored in iris recognition. In [12], a comprehensive survey covering a wide range of models and knowledge distillation techniques reported accuracy improvements of over 5% on CIFAR-10 and 10% on CIFAR-100, and [13] underscored the effectiveness of this technique in resource-constrained scenarios. Based on these findings, [14] proposed a multimodal compression approach that incorporates knowledge distillation with additional optimization techniques. This study utilized a teacher model to guide the training of a VGG16 student model, thereby increasing accuracy by more than 1% (from 90% to 91%) while reducing model size from 528.95 to 90.22 MB and inference time from 0.92 to 0.58 s on CASIA-IrisV1 dataset. Generally, these results show that knowledge distillation enables smaller models to achieve performance comparable to larger models, which opens opportunities for further advances.

Despite recent notable progress, several important gaps persist in iris recognition research. First, although lightweight architectures and compression techniques have been widely adopted in other biometric modalities, such as face and fingerprint recognition, their use in iris recognition remains limited. Few studies have investigated their potential, thereby hindering the deployment of compact models in resource-constrained conditions. Second, while knowledge distillation has shown effectiveness in various computer vision tasks, its application combined with compression techniques in iris recognition remains relatively unexplored, limiting the development of compact and accurate models. Third, embedding-based feature

learning, which is a common approach in most biometric modalities, has not been fully utilized in iris recognition, as most studies rely on classification-based outputs. Finally, balancing accuracy and efficiency remains challenging, as most prior works focus on one aspect, and few address both within a unified framework.

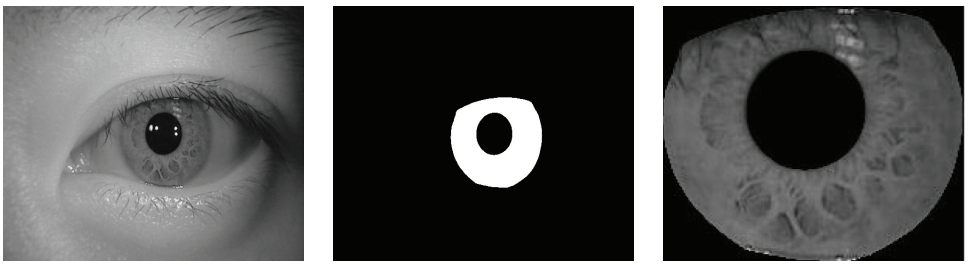
Motivated by these challenges, this study aims to balance accuracy and efficiency for practical deployment. Specifically, the main objectives are: (a) to propose a lightweight deep learning model optimized for mobile iris recognition; (b) to improve model performance through optimization methods; (c) to provide a comprehensive evaluation of model performance on a standard dataset. Accordingly, the primary contributions are: (a) a MobileNetV3-based model enhanced by spatial attention mechanism and task-specific embedding output module, featuring a model size under 10 MB and inference latency below 50 ms; (b) task-specific adaptation of hard negative mining strategy and knowledge distillation technique to improve model performance; (c) competitive performance against the ResNet-50 model on the CASIA-IrisV4-Thousand dataset, with an Equal Error Rate of no more than 5%, a Verification Rate of at least 95% at 1% False Acceptance Rate, and a rank-1 Cumulative Matching Characteristic of at least 90%.

2 Methodology

2.1 Dataset Preparation

For iris recognition tasks, datasets should be carefully prepared to ensure both diversity and consistency. This study employs the CASIA-IrisV4 dataset [15], which contains near-infrared images collected under various acquisition conditions. Specifically, the original data are reorganized into a subject-based directory structure in which each image folder belongs to a single subject. Since the focus of this work is on iris recognition, including feature extraction, encoding, and matching stages, iris segmentation is omitted. Instead, we propose a preprocessing procedure (Algorithm 1) that utilizes Open-Iris library to automatically detect and extract the iris region from each input image for recognition purposes, thereby ensuring that all modules operate solely on the normalized iris area [16].

For our experiments, we conducted tests on all CASIA-IrisV4 subsets. CASIA-IrisV4-Thousand was selected because it yielded the most representative recognition results. This subset also presents challenging acquisition conditions and is widely used in previous iris recognition benchmarks. In detail, it consists of 1000 subjects with a total of 20,000 images, about 20 images per subject. After processing with our proposed procedure (Algorithm 1), the final dataset contains 1997 identities, corresponding to the split of left and right eyes into separate identities, and 18,800 images in total, reflecting the removal of improperly processed images. Approximately 1200 images, accounting for roughly 6% of the dataset, were discarded due to segmentation failures, mainly caused by eyelid or eyelash occlusion, low iris–boundary contrast, or inaccurate pupil localization produced by Open-Iris. On average, each identity includes about 10 images ranging from 2 to 10. Example preprocessed iris region images are shown in Fig. 1.



(a) *Input Image* (b) *Input Mask* (c) *Output Image*

Figure 1: Iris images before and after preprocessing.

Algorithm 1: Dataset construction with Open-Iris Library

FOR EACH $file\_path$ **IN** all_files **DO**
Step 1. Load input image.
$eye\_side,\ eye\_image \leftarrow load\_image(file\_path)$
Step 2. Extract iris and pupil masks using Open-Iris library.
$iris\_mask,\ pupil\_mask \leftarrow openiris.segment(eye\_image)$
IF $exist(iris\_mask,\ pupil\_mask)$ **THEN**
Step 3. Create ROI mask and apply to the input image.
$roi\_mask \leftarrow iris\_mask - pupil\_mask$
$roi\_image \leftarrow eye\_image \times roi\_mask$
Step 4. Normalize ROI mask values to the 8-bit intensity range.
$min\_val,\ max\_val \leftarrow min(roi\_mask),\ max(roi\_mask)$
$norm\_mask \leftarrow (roi\_mask - min\_val)/(max\_val - min\_val) \times 255$
Step 5. Compute bounding box of iris region.
$xs,\ ys \leftarrow indices$ **WHERE** $norm\_mask > 0$
$min\_xy,\ max\_xy \leftarrow min(xs,\ ys),\ max(xs,\ ys)$
Step 6. Crop iris region using bounding box.
$iris\_image \leftarrow roi\_image[min\_xy,\ max\_xy]$
`// Discard improperly processed images.`
IF $euqal(size(iris\_image),\ size(eye\_image))$ **THEN SKIP**
Step 7. Export output image.
`// Separate left and right eyes into distinct identities.`
$export\_path \leftarrow create\_path(file\_path,\ eye\_side)$
$export\_image(export\_path,\ iris\_image)$
END IF
END FOR

Algorithm 1 presents the iris-image preprocessing pipeline, providing key operations, starting from loading the input image, extracting the relevant regions to create the region of interest mask, computing the iris bounding box to cropping the iris region, to generating the final output. To ensure uniformity across all samples, additional steps are included to discard improperly processed images and separate left and right eyes of each subject into independent identities, which play a crucial role in model training.

2.2 Model Architecture

To balance accuracy and efficiency, the large variant of MobileNetV3 is adopted as the backbone for its sufficient representational capacity. Architecturally, an embedding-based module generates compact feature embeddings, while Spatial Attention (SA) complements Squeeze-and-Excitation (SE) to enhance spatial features. As shown in Table 1, SA provides a more efficient alternative to Convolutional Block Attention Module (CBAM) and Non-Local Attention (NLA).

Table 1: Comparison of attention variants by functionality and complexity on MobileNetV3-L.

	Functionality	***Params*** $(\times 10^3)$	***FLOPS*** $(\times 10^6)$	***Description***
SA	*Local Spatial*	0.049	2.4–38.4	*Lightweight and Sufficient*
CBAM	*Channel & Spatial*	0.072–3.3	5.6–38.5	*Partially Redundant*
NLA	*Global Spatial*	1.73–76.8	384–14,752	*High Computational Cost*

2.2.1 Backbone Module

The Spatial Attention, illustrated in Fig. 2, functions similarly to the SE Attention; however, it is designed to focus on spatial features. Given an intermediate feature map F, spatial information is aggregated by applying average pooling and max pooling across the channel dimension, producing two spatial context descriptors, $F_{avg} \in R^{1\times h\times w}$ and $F_{max} \in R^{1\times h\times w}$.

$$M(F) = \sigma\Big(A^{n\times n}\big([F_{avg}; F_{max}]\big)\Big) \tag{1}$$

These context descriptors are then combined to generate the spatial attention map $M(F)(1)$ using the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. Here, $A^{n\times n}$ represents a $n \times n$ convolution kernel used to aggregate local spatial information, with smaller n capturing local details and larger n incorporating broader spatial context.

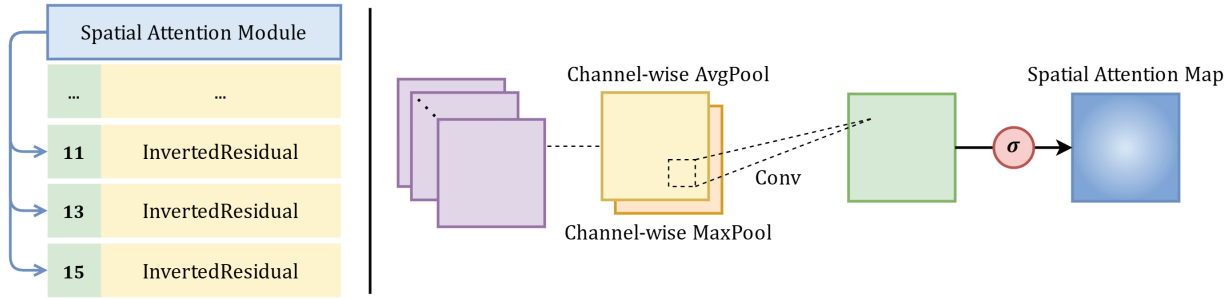


Figure 2: Spatial attention module for discriminative region highlighting.

2.2.2 Head Module

Instead of relying on closed-set classification, an embedding-based approach suited to the open-set nature of biometric recognition is adopted. As shown in Fig. 3, its design consists of a linear layer followed by batch normalization, producing compact feature embeddings. In related biometric tasks, embedding-based methods outperform classification-based ones: ArcFace achieved state-of-the-art performance, reporting 99.83% accuracy on LFW and 96.98% VR@FAR of 10^{-6} on MegaFace [17]; SphereFace improved verification from 98.71% using softmax to 99.42% on LFW [18]; and CosFace obtained an accuracy of 99.73% on LFW and 97.6% on YTF [19]. These methods train deep networks with margin-based losses, such as angular margin in SphereFace, cosine margin in CosFace, and additive angular margin in ArcFace, to project features into a hyperspherical space that enhances inter-class separability and intra-class compactness.

$$\ell_2(v) = \frac{v}{\|v\|_2} = \frac{[v_1, v_2, \ldots, v_n]}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}} \tag{2}$$

Let v be an embedding vector of dimension n. ℓ_2 normalization (2) is applied by dividing each element of v by its norm. After normalization, the resulting vectors satisfy unit length, $\|v\|_2 = 1$, giving all embeddings

the same magnitude. This property ensures that the model focuses on the direction of the embeddings rather than their scale, since the direction encodes the identity-related information while the magnitude mostly reflects feature scale.

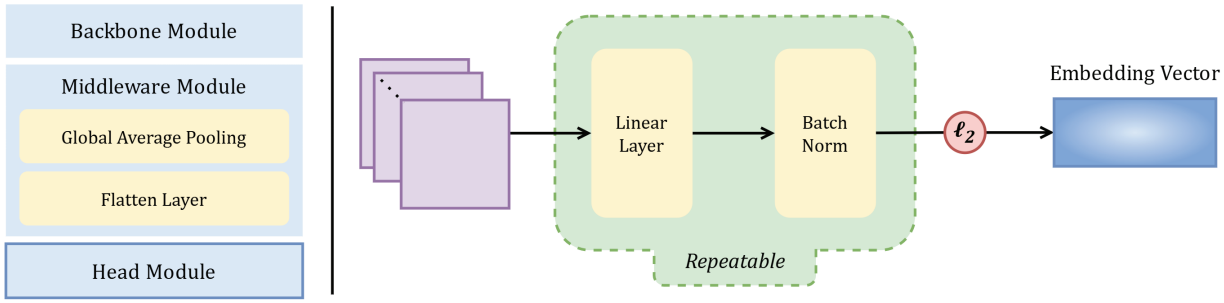


Figure 3: Embedding-based module for feature representation learning.

2.3 Model Optimization

Together, the lightweight backbone, spatial attention module, and embedding-based head form a cohesive feature extraction and representation framework, which is further reinforced during training through Hard Negative Mining and Knowledge Distillation, enabling the model to focus on challenging samples and to transfer discriminative knowledge from a stronger teacher model.

Algorithm 2 outlines the overall training optimization procedure under the triplet-learning paradigm, including triplet batch sampling and processing. The sampling stage constructs a triplet consisting of an anchor (the reference sample), a positive sample with the same label, and a negative sample with a different label that has not been paired with the selected anchor and positive samples in the current iteration. The processing stage extracts student and teacher embeddings and applies hard negative mining with knowledge distillation, depending on the current phase.

Algorithm 2: Training optimization with hard negative mining and knowledge distillation

WITH $dataset(\mathcal{D})$, $sample(a)$ **DO** `//Triplet Batch Sampling.`
  `//Create an anchor-positive-negative triplet batch.`
  $p \leftarrow random.choice(\{x \in \mathcal{D} \mid label(x) = label(a) \wedge x \neq a\})$
  $n \leftarrow random.choice(\{x \in \mathcal{D} \mid label(x) \neq label(a) \wedge x \text{ not paired with } (a, p)\})$
  RETURN $batch(a, p, n)$
WITH *batch(a, p, n)* **DO** `//Triplet Batch Processing.`
  `//Extract embedding vectors from student and teacher models.`
  $s\_embs \leftarrow MobileNetV3L(a), MobileNetV3L(p), MobileNetV3L(n)$
  $t\_embs \leftarrow ResNet50(a), ResNet50(p), ResNet50(n)$
  `//Apply Hard Negative Mining and Knowledge Distillation.`
  IF *warmup phase active* **THEN**
    $s\_embs[n] \leftarrow WarmHardNegativeMining(s\_embs)$
    $loss \leftarrow TripletLoss(s\_embs)$
  ELSE
    $s\_embs[n], t\_embs[n] \leftarrow SemiHardNegativeMining(s\_embs, t\_embs)$
    $loss \leftarrow TripletLoss(s\_embs) + DistilledLoss(s\_embs, t\_embs)$
  END IF

2.3.1 Triplet-Based Learning

Triplet-based learning is employed to learn discriminative embeddings by enforcing a relative distance constraint among anchor, positive and negative samples. Specifically, the model is trained to minimize the Euclidean distance d between the anchor a and positive p while maximizing the distance between the anchor a and negative n at least a margin m, as formalized in the triplet margin loss $L_{triplet}$ (3). This encourages the embedding space to preserve identity-level similarity, ensuring that samples of the same class are closer together than samples of different classes.

$$L_{triplet} = \frac{1}{N}\sum_{i=1}^{N}\max\left(d(a_i, p_i) - d(a_i, n_i) + m,\ 0\right) \tag{3}$$

To optimize the triplet margin loss, a margin scheduling strategy (5) is adopted, where the margin dynamically changes during training. The margin is initially fixed during a warm-hard phase and then gradually transitions to semi-hard phases using a cosine annealing schedule (4), facilitating effective integration with the hard negative mining strategy. Let $m(t)$ denote the $margin$ at epoch t, where M_{warm} is the fixed warm-hard margin, and M_{medi}, $M_{post} = [M_0,\ M_1]$ are the start and end margins for the first and second halves of the semi-hard phase, respectively.

$$CosAnneal = M_0 - \frac{1}{2}(M_0 - M_1)\left(1 - \cos\left(\pi\frac{t - T_{start}}{T_{end} - T_{start}}\right)\right) \tag{4}$$

$$m(t) = \begin{cases} M_{warm}, & 1 \le t \le T_{warm} \\ CosAnneal(M_{medi}), & T_{warm} < t \le T_{medi} \\ CosAnneal(M_{post}), & T_{medi} < t \le T_{post} \end{cases} \tag{5}$$

2.3.2 Hard Negative Mining

To enhance convergence and embedding discriminability, hard negative mining strategy (6) is applied during training. In the initial warm-up stage, $WarmHardNegativeMining$ is used, where semi-hard negatives that beyond the positive but within slightly loosened boundaries δ_{left} to avoid overly difficult cases are selected. In the remaining stages of the training, $SemiHardNegativeMining$ is used, where both δ_{left} and δ_{right} are gradually tightened to focus on negatives closer to the anchor, thereby progressively increasing the difficulty of the sampled pairs. When no samples satisfy the warm-hard or semi-hard condition, the top-k hardest negatives are used as substitutes. This progressive strategy enables the model to gradually transition from easier to more challenging samples.

$$\mathcal{N}_{hard} = \begin{cases} Warm\text{-}hard: & d(a,p) + loosen(\delta_{left}) < d(a,n_i) < d(a,p) + \delta_{right} \\ Semi\text{-}hard: & d(a,p) + tighten(\delta_{left}) < d(a,n_i) < d(a,p) + tighten(\delta_{right}) \\ Topk\text{-}hardest: & d(a,n_i) \le d(a,n_{i+1}) \le \cdots \le d(a,n_k) \end{cases} \tag{6}$$

2.3.3 Knowledge Distillation

To strengthen the discriminative capability of the student model, MobileNetV3-L, knowledge distillation is employed using a pretrained teacher model, ResNet-50, trained on the same dataset, as shown in Fig. 4. The student model is optimized with a combined objective L_{total} (7), including a triplet loss $L_{triplet}$ (3) and a cosine similarity-based loss $L_{distill}$ (8) scaled by a factor δ for each embedding type x, and an overall factor α (9).

$$L_{total} = L_{triplet} + \frac{\alpha}{3} \sum_{x \in \{a,\ p,\ n\}} \delta_x L^x_{distill} \tag{7}$$

$$L^x_{distill} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - cosine\_similarity(s^x_i,\ t^x_i)\right) \tag{8}$$

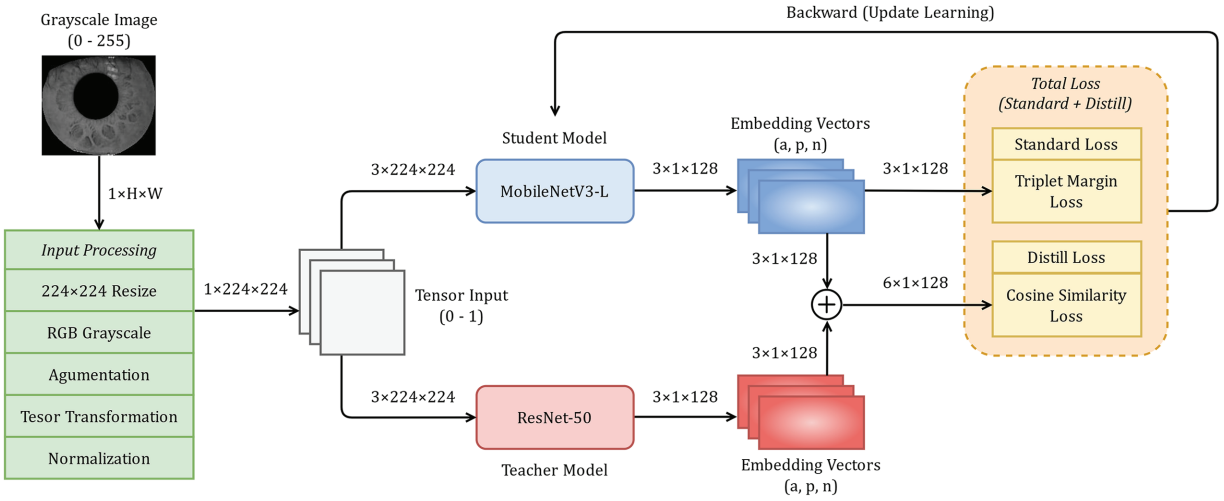


Figure 4: Knowledge Distillation between ResNet-50 teacher and MobileNetV3-L student.

Let $\alpha(t)$ denote the dynamic weighting factor at epoch t, with $\alpha_{semi} = [\alpha_0,\ \alpha_1]$ for controlling the contribution of teacher model during the semi-hard phase. During the initial warm-up phase, only $L_{triplet}$ is applied to allow the student model to first learn basic features without being influenced by the teacher model. Then, each $L^x_{distill}$ is added to progressively align the student representations with the teacher.

$$\alpha(t) = \begin{cases} 0, & 0 < t \leq T_{warm} \\ CosAnneal(\alpha_{semi}), & T_{warm} < t \leq T_{semi} \end{cases} \tag{9}$$

2.4 Evaluation Metrics

To evaluate recognition performance, two sets of metrics that capture different aspects of the model capabilities are used. These metrics offer a comprehensive assessment of how accurately the model can verify or identify subjects under various conditions. By integrating both threshold-based and ranking-based measures, the evaluation quantifies not only the accuracy but also the robustness of the model in challenging scenarios, such as hard negatives and unseen identities.

2.4.1 Similarity Metrics

A key step in recognition tasks is measuring the similarity between two normalized embedding vectors, $f_a,\ f_b \in \mathbb{R}^{dim}$, using either Cosine similarity (10) or Euclidean distance (11). The similarity scores, s_{ab} and d_{ab}, are used to compute threshold-based metrics such as EER and VR@FAR, for verification, and ranking-based metrics such as CMC@K, for identification.

$$s_{ab} = \frac{f_a \cdot f_b}{\|f_a\|_2 \, \|f_b\|_2} = \frac{\sum_{k=1}^{dim} f_{a,k} \, f_{b,k}}{\|f_a\|_2 \, \|f_b\|_2} \tag{10}$$

$$d_{ab} = \|f_a - f_b\|_2 = \sqrt{\sum_{k=1}^{dim} (f_{a,k} - f_{b,k})^2} \tag{11}$$

2.4.2 Verification Metrics

Verification metrics are used to evaluate the model's ability to determine whether pairs of samples correspond to the same or different identities. Given a similarity score s_{ab} or distance d_{ab}, a threshold τ is applied to classify a pair as a match or non-match.

- False Acceptance Rate, defined as $FAR(\tau) = \frac{N_{FA}(\tau)}{N_{imp}}$, measures the proportion of imposter pairs that are incorrectly accepted as matches, where $N_{FA}(\tau)$ denotes the number of imposter pairs whose similarity exceeds the threshold τ (or distance is below τ), and N_{imp} is the total number of imposter pairs.
- False Rejection Rate, defined as $FRR(\tau) = \frac{N_{FR}(\tau)}{N_{gen}}$, measures the proportion of genuine pairs that are incorrectly rejected as non-matches, where $N_{FR}(\tau)$ denotes the number of genuine pairs whose similarity is below the threshold τ (or distance is above τ), and N_{gen} is the total number of genuine pairs.
- Equal Error Rate, displayed as $\tau_{EER} \rightarrow FAR(\tau_{EER}) = FRR(\tau_{EER})$, represents the threshold τ where false acceptance rate FAR and false rejection rate FRR are equal, which reflects the difference between incorrectly accepting imposter pairs and rejecting genuine pairs.
- Verification Rate at a False Acceptance Rate, defined as $VR@FAR = \frac{N_{gen}(\tau_{VR})}{N_{gen}}$, measures the proportion of genuine pairs that model correctly verified at a defined FAR, where $N_{gen}(\tau_{VR})$ is the number of genuine pairs accepted at τ_{VR} corresponding to FAR and N_{gen} is the total number of genuine pairs.

2.4.3 Identification Metrics

Identification metrics are used to evaluate the model's ability to correctly identify a probe iris image from a gallery of known identities. In contrast to verification, which makes a binary decision, identification ranks all gallery images according to their similarity to the probe.

- Cumulative Matching Characteristic, defined as $CMC@k = \frac{N_{probe}(k)}{N_{probe}}$, measures the proportion that the correct identity of a probe appears is retrieved within the top-k ranked gallery results, where k denotes the rank threshold, N_{probe} is the total number of probe images, and $N_{probe}(k)$ is the number of probes whose correct match appears in the top-k gallery images.

3 Results and Discussion

3.1 Experimental Setup

Experiments are conducted on the CASIA-IrisV4-Thousand dataset, split into 3 subsets with a ratio of 70:15:15 at subject level to ensure that all images of the same subject are grouped together. As shown in Table 2, the splits contain 1,397, 300, and 300 subjects, respectively, corresponding to 13,150, 2838, and 2812 images. Afterwards, all grayscale images were resized to 224×224 and replicated across three channels for backbone compatibility, then scaled to the $[0,1]$ range, finally normalized with the standard ImageNet mean and standard deviation. For data generalization, minor augmentation randomly applied rotations of within ±5° and affine transformations with translation and scaling within ±5% to avoid excessive distortion.

Table 2: Overview of the Preprocessed Dataset.

	Subject	Image	Min Imgs/Sub	Max Imgs/Sub
Training	1397	13,150	2	10
Validation	300	2838	3	10
Testing	300	2812	3	10
Total	1997	18,800	–	–

The training process is conducted over 40 epochs with a batch size of 32, divided into 3 stages: a warm-hard negative stage for the first 8 epochs, and a semi-hard negative stage divided evenly into two parts covering the remaining 32 epochs. At the core of the training pipeline, the model is trained to minimize the $TripletMarginLoss$ (3) with a dynamic margin varying between 0.4 and 0.5. Additionally, the optimization is performed using $AdamW$ with different learning rates η and weight decays λ: a lower learning rate is applied to the backbone ($\eta : 5e^{-4} \rightarrow 1e^{-5},\ \lambda : 1e^{-4}$) to preserve pretrained features, while higher learning rates are used to the attention ($\eta : 1e^{-3} \rightarrow 1e^{-4},\ \lambda : 5e^{-5}$) and head ($\eta : 5e^{-3} \rightarrow 5e^{-4},\ \lambda : 5e^{-5}$) to learn task-specific representations. To leverage the pretrained backbone, its layers are gradually unfrozen for smooth fine-tuning, and two learning rate schedulers (12) are applied: a short linear decay during warm-up phase to stabilize early training, followed by a cosine decay to help the model converge smoothly in the remaining stages.

$$\eta(t) = \begin{cases} \eta_{\max} \dfrac{t}{T_{warm}}, & 0 < t < T_{warm} \\ \eta_{\min} + \dfrac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\dfrac{\pi(t - T_{warm})}{T_{semi}}\right)\right), & T_{warm} \leq t \leq T_{semi} \end{cases} \tag{12}$$

Regarding the hard negative mining strategy, the configuration defines the lower and upper bounds for negative sample selection as scales of the triplet margin. Specifically, these bounds are computed as the product of the margin and generated ratios, as formalized in Eq. (6), thereby selecting negatives according to the current discriminative ability of the model.

$$lower\_bound = \delta_{left} = ratio_{left} \times margin$$
$$upper\_bound = \delta_{right} = ratio_{right} \times margin$$

In the warm-hard stage, the ratios for the lower and upper bounds, $ratio_{\text{left}}$ and $ratio_{\text{right}}$, are fixed at 0.2 and 1.0, respectively. During the semi-hard stage, these ratios are adjusted using a cosine decay scheduler from 0.2 to 0.1 and 1.0 to 0.9. Meanwhile, the topk-hardest strategy is applied to select one of the 5 hardest samples if no negatives satisfy the bounds.

For the knowledge distillation technique, the configuration specifies a maximum distillation weight α of 0.5, with all component weights δ set to 1.0, as defined in Eq. (8). This setup enables the student model to effectively leverage the teacher's guidance across the anchor, positive, and negative components, facilitating the learning of more discriminative embeddings and improving overall representation quality during training.

3.2 Experimental Result

3.2.1 Original Performance

To evaluate the impact of the proposed optimizations, the baseline MobileNetV3-L (BASE) is compared with its optimized variants, and the state-of-the-art ResNet-50 (SOTA). Specifically, attention mechanism and knowledge distillation technique are incorporated into the baseline to create more robust variants. As observed, the individual variants show moderate improvements, whereas their combination (our proposal) yields a more substantial improvement and narrows the performance gap to the state-of-the-art. Table 3 presents a detailed comparison among the baseline, the optimized variants, and the state-of-the-art model.

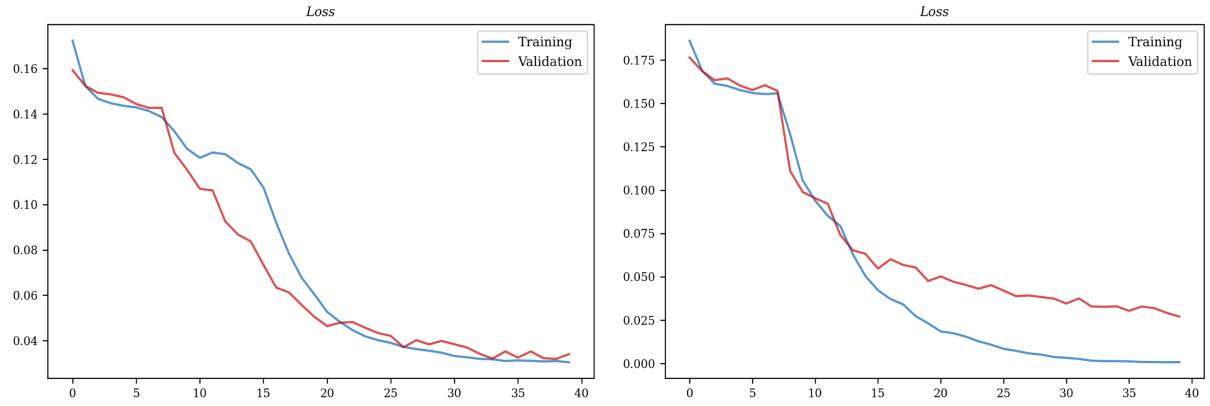
Table 3: Performance comparison across different optimization techniques at FAR = 1% and K = 1.

	***EER* (%)**	***VR@FAR* (%)**	***CMC@K* (%)**	***Size* (*MB*)**	***Latency* (*ms*)**
MobileNetV3-L (BASE)	1.699	97.678	93.073	11.9	25.2
Only Attention	1.484	97.993	93.949	12.0	27.1
Only Distilled	1.535	97.951	94.466	11.9	25.2
Attention & Distilled	1.409	98.184	94.785	12.0	27.1
ResNet-50 (SOTA)	1.567	98.092	95.103	90.9	188.1

The baseline achieves an EER of 1.699%, a VR@FAR of 97.678%, and a CMC@K of 93.073%, while remaining highly efficient at only 11.9 MB and 25.2 ms latency. In contrast, the state-of-the-art outperforms the baseline in accuracy, achieving an EER of 1.567%, a VR@FAR of 98.092%, and a CMC@K of 95.103%, at the expense of substantially larger size of 90.9 MB and higher latency of 188.1 ms. This comparison reflects the inherent trade-off between recognition accuracy and computational efficiency. To overcome this limitation, attention mechanism and knowledge distillation technique are applied, with the goal of pushing its accuracy closer to the state-of-the-art performance while retaining its compact-size and low-latency characteristics.

Compared to the baseline, both attention and distillation enhancements achieve consistent accuracy improvements with minimal computational cost. The attention-only variant reduces the EER to 1.484% and increases both the VR@FAR and CMC@K to 97.993% and 93.949%, respectively, as the attention mechanism helps the model capture more discriminative features. Similarly, the distilled-only variant achieves an EER of 1.535%, a VR@FAR of 97.951%, and a CMC@K of 94.466%, showing the benefit of knowledge transfer in enhancing feature discrimination. Taken together, these results confirm that each enhancement, regardless of the approach, independently strengthens the model's discriminative ability.

Based on these individual enhancements, combining attention and distillation further improves overall performance. Compared to the baseline, the combined variant slightly increases model size by 0.1 MB and latency by 2 ms, while achieving superior performance across all metrics, with an EER of 1.409%, a VR@FAR of 98.184%, and a CMC@K of 94.785%. Moreover, it remains over 7× smaller and faster than the state-of-the-art model, achieving comparable recognition accuracy. Additionally, the combined model exhibits clear convergence during training, as shown in Fig. 5. These results highlight the complementary advantages of these two techniques, since the attention strengthens feature discrimination and the distillation fosters generalizable representations. Collectively, this combination provides an excellent trade-off between accuracy and efficiency, making the model highly practical for mobile iris recognition.



(a) *Our proposal: Attention & Distilled* (b) *State-of-the-art Model*

Figure 5: Comparison of training and validation convergence.

3.2.2 Quantized Performance

To further examine the practicality of the proposed models for mobile deployment, we assess their performance under post-training quantization. As summarized in Table 4, the results highlight that reduced numerical precision substantially impacts recognition accuracy and computational efficiency. As expected, quantization causes a noticeable degradation in recognition performance across all models.

Table 4: Performance comparison across different optimization techniques after quantization.

	***EER* (%)**	***VR@FAR* (%)**	***CMC@K* (%)**	***Size* (*MB*)**	***Latency* (*ms*)**
MobileNetV3-L (BASE)	7.587	77.512	67.515	5.93	12.7
Only Attention	7.089	78.590	67.794	5.94	13.2
Only Distilled	6.417	81.086	72.611	5.93	12.7
Attention & Distilled	7.173	79.726	70.143	5.94	13.2
ResNet-50 (SOTA)	6.335	82.271	75.039	45.3	163.5

In this work, post-training quantization is applied to the final models due to its simplicity, compatibility with mobile deployment, and no retraining requirement. Specifically, the weights are converted from float32 to float16, while other tensors remain in full precision due to the model's sensitivity to numerical precision [20]. As float16 naturally preserves the dynamic range of the weights, converting them to half-precision reasonably reduces memory footprint and computational cost while minimizing precision loss.

The baseline model experiences a substantial accuracy drop after quantization, with an EER increasing from 1.699% to 7.587%, VR@FAR of 77.512% (a drop of about 20%), and CMC@K of 67.515% (a drop of about 27%), yet it retains a more compact size of 5.93 MB and lower latency of 12.7 ms. By comparison, the quantized state-of-the-art ResNet-50 remains more accurate but is significantly larger and slower, which is expected given the typical trade-off between accuracy and computational cost.

Meanwhile, the attention-only and distilled-only variants show slight overall improvements compared to the baseline, showing some resilience to quantization. Importantly, the combined variant achieves a balanced performance among all optimized models, with an EER of 7.173%, a VR@FAR of 79.726%, and a CMC@K of 70.143%, while maintaining a moderate size of 5.94 MB and reasonable latency of

13.2 ms. Overall, this indicates that integrating both enhancements effectively preserves discriminative features and provides a well-balanced trade-off between accuracy and efficiency, demonstrating its practicality for mobile deployment.

In general, the proposed optimizations enhance accuracy and maintain efficiency relative to the baseline; however, all models suffer a noticeable drop in recognition accuracy after quantization, with EER increasing, and VR@FAR and CMC@K decreasing across all variants. To address this, alternative strategies such as quantization-aware training or mixed precision could be considered to reduce the performance drop, though they involve additional training complexity and are left for future work. Consequently, this highlights the need for further development to better preserve accuracy and improve practicality for mobile deployment.

4 Conclusion

This study presents a lightweight MobileNetV3-L model for mobile iris recognition, optimized through attention mechanism, knowledge distillation, and quantization to achieve a practical balance between accuracy and efficiency. Comprehensive experiments demonstrate that the model achieves a compact size of 12 MB and a low latency of 27 ms, facilitating its adaptation for mobile applications through a further reduction to 5.94 MB and 13 ms after post-training quantization. Regarding recognition performance, the proposed combined variant, the main model of this study, delivers an EER of 1.409%, a VR@FAR=1% of 98.184%, and a CMC@K=1 of 94.785%, all meeting the respective targets of 5%, 95%, and 90% on the standard CASIA-IrisV4-Thousand dataset. Overall, the results highlight that the proposed strategies enable models to achieve performance comparable to state-of-the-art approaches.

Building on these findings, the practical value of this work lies in providing a mobile-ready iris recognition solution that balances recognition accuracy and computational efficiency, improving the overall system architecture to achieve robust and reliable performance, and demonstrating the framework's practicality through extensive experimental evaluation. By utilizing lightweight architectures and optimization techniques, the proposed framework facilitates deployment across a wide range of practical domains, including smart cities and mobile security. Despite these strengths, current evaluations are limited to standard datasets, and maintaining accuracy under aggressive quantization remains challenging. Consequently, future research built on this work can further expand this work through validation on larger and more diverse datasets, exploration of advanced optimization strategies, extension to multimodal biometric systems, and evaluation across different hardware platforms to enhance robustness, scalability and generalizability.

Acknowledgement: The authors extend their appreciation to the Posts and Telecommunications Institute of Technology (PTIT, Vietnam) for supporting this research.

Funding Statement: Not applicable.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Trong-Thua Huynh and De-Thu Huynh; methodology, Trong-Thua Huynh and Du-Thang Phu; software: Du-Thang Phu and De-Thu Huynh; validation, Quoc H. Nguyen and Hong-Son Nguyen; formal analysis, Trong-Thua Huynh and Quoc H. Nguyen, resources, Du-Thang Phu; data curation, Trong-Thua Huynh; writing—original draft preparation, Du-Thang Phu and De-Thu Huynh; writing—review and editing, Trong-Thua Huynh and Du-Thang Phu; visualization, De-Thu Huynh; supervision, Hong-Son Nguyen; project administration, Trong-Thua Huynh. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Grother P, Matey J, Tabassi E, Quinn G, Chumakov M. IREX VI: temporal stability of iris recognition accuracy. In: NIST Interagency/Internal Report (NISTIR) 7948. Gaithersburg, MD, USA: National Institute of Standards and Technology; 2013 Jul 11. doi:10.6028/NIST.IR.7948.
2. Emergen Research. Iris recognition market, by component, product, application, end-use, and region, forecast to 2034; 2025 Jul [cited 2025 Oct 15]. Available from: https://www.emergenresearch.com/industry-report/iris-recognition-market.
3. Mordor Intelligence Research & Advisory. Iris recognition market size & share analysis: growth trends & forecasts (2025–2030) [Internet]. 2025 Jul [cited 2025 Oct 15]. Available from: https://www.mordorintelligence.com/industry-reports/iris-recognition-market.
4. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014. doi:10.48550/arxiv.1409.1556.
5. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Piscataway, NJ, USA: IEEE; 2016. p. 770–8. doi:10.1109/CVPR.2016.90.
6. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Piscataway, NJ, USA: IEEE; 2017. p. 4700–8. doi:10.1109/CVPR.2017.243.
7. Nguyen K, Proença H, Alonso-Fernandez F. Deep learning for iris recognition: a survey. ACM Comput Surv. 2024;56(9):223:1–35. doi:10.1145/3651306.
8. Han K, Wang Y, Tian Q, Guo J, Xu C. GhostNet: more features from cheap operations. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 1580–9. doi:10.1109/CVPR42600.2020.00165.
9. Boutros F, Damer N, Raja KB, Ramachandra R, Kirchbuchner F, Kuijper A. On benchmarking iris recognition within a head-mounted display for AR/VR applications. In: Proceedings of the IEEE Int Joint Conference Biometrics (IJCB); 2020 Sep 28–Oct 1; Houston, TX, USA. p. 1–10. doi:10.1109/IJCB48548.2020.9304919.
10. Sumit S, Anavatti S, Tahtali M, Mirjalili S, Turhan U. ResNet-Lite: on improving image classification with a lightweight network. Comput Sci. 2024;246(9):1–10. doi:10.1016/j.procs.2024.09.597.
11. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531. 2015.
12. Gou J, Yu B, Maybank S, Tao D. Knowledge distillation: a survey. Int J Comput Vis. 2021;129(6):1789–819. doi:10.1007/s11263-021-01453-z.
13. Yin Y, He S, Zhang R, Chang H, Han X, Zhang J. Deep learning for iris recognition: a review. arXiv:2303.08514. 2023.
14. Latif SA, Sidek KA, Bakar EA, Hashim AHA. Online multimodal compression using pruning and knowledge distillation for iris recognition. J Adv Res Appl Sci Eng Technol. 2024;37(2):68–81. doi:10.37934/araset.37.2.6881.
15. Institute of Automation, Chinese Academy of Sciences (CASIA). CASIA Iris Image Database, Version 4.0. Beijing, China: Chinese Academy of Sciences; 2010.
16. Worldcoin AI. IRIS: iris recognition inference system of the worldcoin project [Computer software]. GitHub. 2023 [cited 2025 Oct 15]. Available from: https://github.com/worldcoin/open-iris.
17. Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. arXiv:1801.07698. 2019.
18. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L. SphereFace: deep hypersphere embedding for face recognition. arXiv:1704.08063. 2017.
19. Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, et al. Large margin cosine loss for deep face recognition. arXiv:1801.09414. 2018.
20. Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Piscataway, NJ, USA: IEEE; 2018. p. 2704–13. doi:10.1109/CVPR.2018.00286.