AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy
ISSN: 3062-097X
Published: Early Access

Knowledge Distillation from ResNet to MobileNet for Accurate On-Device Face Recognition

Nadir İBRAHİMOĞLU[1*], Mehmet Cem Aytekin[2], and Furkan Yıldız[3]

[1]MSDC Department, Huawei R&D Center, Ankara, Turkey, ORCID: 0000-0003-1189-3054
[2]MSDC Department, Huawei R&D Center, Istanbul, Turkey, ORCID: 0000-0002-5885-9809
[3]Papilon Savunma, Ankara, Turkey, ORCID: 0009-0003-1334-1613

ORIGINAL RESEARCH PAPER

Abstract

The development of efficient facial recognition systems on low-resource devices requires models to optimize computational cost and performance for discrimination tasks. With this consideration, we introduce Adaptive Feature-Logit Distillation (AFL-KD), a novel framework that combines logit imitation, mid-level feature alignment, and weighted learning of losses with a specific focus on open-set face verification. During each train step, AFL-KD continuously adjusts the relative weights of the cross-entropy, logit distillation, and feature alignment components by matching the magnitudes of their gradients, thereby alleviating an upper bound on the angular-margin error of the student model. In the process of compressing a 283 MB ResNet teacher into a 7.2 MB MobileNet, AFL-KD achieves a negligible drop of 1.5 points in verification accuracy.

Empirical evaluations on the CASIAWebFace dataset show that distilled MobileNet achieves 95.7% accuracy, 77.0% recall, and an F1 score of 86. 9%, closely approaching the ResNet teacher (97.2% accuracy, 84.8% recall) while shrinking the model size from 283.3 MB to 7.2 MB—a 39× compression. Compared with a conventionally trained MobileNet baseline (87.3% accuracy, 62.8% F1 at 14 MB), the distilled model delivers +8.4% absolute accuracy and +24.1% F1 improvements while halving the memory footprint. These results confirm that knowledge distillation can yield highly accurate yet resource-efficient face recognition suitable for mobile and embedded applications.

Keywords: Knowledge Distillation, Face Recognition, MobileNet, Lightweight Deep Learning, On-Device Inference

OPEN ACCESS
AIPAJ Vol:1, Issue:2

***Corresponding author**
nebrahimpour@ktu.edu.tr

***Submitted** 22 April 2025*

***Accepted** 15 May 2025*

Citation
İBRAHİMOĞLU et al. (2025) Knowledge Distillation from ResNet to MobileNet for Accurate On-Device Face Recognition. AIPA's International Journal on AI: Bridging Technology, Society and Policy. DOI: 10.5281/zenodo.15502091

1 Introduction

Facial recognition technology is now at an unprecedented level of accuracy, thanks to the use of deep learning architectures that process vast sets of identity data. Prominent examples are the face embedding unification method from FaceNet and the margin-based method based on ArcFace, both of which achieve almost perfect recognition rates for benchmarked datasets, such as LFW, thanks to the complex architectures of ResNet [1, 2]. However, such top-performing face recognizers leverage tens of millions of parameters and require considerable computational resources, making them difficult to deploy on low-end devices, such as smartphones or IoT devices [3]. In applications such as mobile authentication and surveillance, the need for face recognition in the device is felt to balance accuracy and efficiency. Achieving high accuracy using compact models seems challenging because of the huge performance drop that accompanies the model complexity and dimension decreases.

To alleviate this imbalance, researchers have explored various model compression techniques, including network pruning, quantization, neural architecture search, and knowledge distillation [4, 5]. Among them, knowledge distillation (KD) has emerged as a particularly promising paradigm for model compression in deep learning settings [6]. In KD, a proficient teacher model transfers its knowledge to a smaller student model by teaching it to mimic the teacher's outputs or representations. Hinton et al. [6] first introduced KD using softened teacher output probabilities as "soft labels" to guide the training of the student, greatly improving the student's ability to generalize. The appeal of KD lies in its ability to boost the performance of a small model to that of a large model without requiring changes to the student architecture or inducing additional inference costs at deployment time.

In the context of face recognition, knowledge distillation appears to be an appealing method for obtaining compact and high-quality discriminative embeddings. A strong teacher model, represented by a ResNet-100 that uses ArcFace [2], successfully encapsulates complex representation features in its deeper layers. Conversely, a low-latency design small model, when trained alone, suffers from significantly degraded verification performance because of its limited capacity. Knowledge distillation allows the student model to be guided by the subtle knowledge of the teacher—namely the teacher's precise class probability distributions or intermediate feature responses—in addition to the ground-truth identity labels, thus enhancing its discriminative ability.

Even with such developments, obtaining accurate on-device face recognition is still challenging. One of the issues is the capacity disparity between a very large teacher and a much smaller student, which makes plain distillation inefficient [7]. A case in point is a ResNet-100 teacher and a MobileNet student, who possess very different representational capacities, sometimes needing adaptive distillation methods or intermediate "teacher assistant" networks. Another issue is that face recognition networks generate embeddings for open-set identity matching, so the distillation must maintain the teacher's features' inter-class discriminative capability. Recent research started to tackle such challenges with face recognition specialized distillation methods (e.g., distilling pairwise relation knowledge to optimize verification measures directly [8]).

2 Literature Review

Knowledge distillation is now a central approach in model compression research, backed by a vast body of research examining the efficient transfer of knowledge from complex models to simpler ones. The early work on model compression by Buciluă et al. [4] proved that the predictions made by a set of ensemble models could be "compressed" into a single, simpler model, foreshadowing the idea of knowledge distillation. Later, Hinton et al. [6] extended this principle, proposing logit-based distillation where a student network is learned such that it matches the softened output probabilities (logits) of a larger teacher network. This approach allows the student network to learn the teacher's implicit knowledge for class similarities: even though the teacher is highly confident that the correct class is class k, the lower probabilities given for other classes provide essential information on feature relationships. By minimizing the KL-divergence or cross-entropy loss between teacher and student outputs, the student gains the capacity to mimic the functional behavior of the teacher.

Logit distillation worked well for classification tasks, and follow-up studies have continued to strengthen this method. For instance, Zhao et al. [9] pointed out that standard KD loss combines target-class information from the teacher with information from the non-target classes, potentially limiting the student's ability to learn. They recommended the separability of KD loss (DKD) into individual terms for the target class as well as for non-target classes, enhancing the exploitation of the information-rich distribution of non-target logits. Similarly, several studies have changed the distillation objective for enhancing learning from the probabilities of non-target classes or re-weighting the teacher's logits, yielding better-performing students on CIFAR-100 as well as ImageNet benchmarks [10]. Yang et al. [11] took a theoretical approach by reformulating the KD loss and introducing a normalized KD (NKD) objective, which achieved state-of-the-art results on ImageNet. These gains demonstrate the important impact the way the teacher's output distribution is distilled has on the performance of the student.

In addition to logits, a large body of research is devoted to feature-based distillation. Romero et al. [12] were among the first to pursue this direction with FitNets, which guides the intermediate feature maps of the student model to match those of the teacher model via a hint-based regression loss. By providing supervision at hidden layers, FitNets enabled the successful training of thin deep student networks that would be difficult to train otherwise. Zagoruyko and Komodakis [13] introduced an attention transfer (AT) approach that forces the student to mimic the teacher's attention maps (spatial activation patterns) instead of the raw features themselves, thus capturing a notion of spatial importance.

These methods have been widely used and built upon: teachers can provide guidance in many different ways, including neuron activations, attention mechanisms, or even Gram matrix representations of feature correlations. The general goal is to impart the teacher's rich representational capabilities of hidden layers to the corresponding student model layers. For example, Park et al. [8] introduced relational knowledge distillation (RKD), which is not concerned with matching features directly but, rather, the relations between feature vectors—namely, the distances and angles between sample pairs in the teacher's embedding space. By mimicking these relational measures, the student effectively preserves the structural relationships learned by the teacher, which is especially relevant for metric learning-oriented tasks like face recognition. In another direction, Chen et al. [14] proposed DarkRank, where the student is taught to mimic the ranked similarity list between samples of the teacher (effectively distilling the teacher's organization of samples according to distance). These relational and metric-oriented distillation methods are especially beneficial for face recognition, as they emphasize the pairwise similarity knowledge that underlies verification performance.

A relevant framework includes both self-distillation as well as online distillation, where teacher and student differences blur. In self-distillation, a neural network is able to learn from its own architecture—such as by using deeper layers as "teachers" for the previous layers, or through iterative training where a model trains a new version of itself, a notion described as the "born-again" approach [7]. Furlanello et al. [7] illustrated a network, even one not using an external teacher, can improve its accuracy across generations by retraining on its own outputs. Similarly, Zhang et al. [15] proposed Deep Mutual Learning, where an ensemble of student models undergo interactive learning by sharing knowledge with one another in real time, independent of any individual teacher. These methods hint that the benefits of knowledge distillation (KD) are not necessarily restricted to the traditional teacher-student model—knowledge sharing between peers, or within the training processes of one model—can provide higher quality, often as a byproduct of the regularization effect of soft targets.

The core element of the distillation process is the difference in capacity between the student and the teacher. Cho and Hariharan [7] demonstrated that where the teacher has much wider capacity, the student could have difficulty with the learning process, perhaps being outperformed by a knowledge distillation model with no prior knowledge. One suggestion is for a teaching assistant of intermediate capacity to perform as an intermediary [15]. In face recognition, Jingzhi Li et al. [16] addressed the problem by enforcing constraints within the teacher feature space during the distillation process, enhancing student compatibility, effectively reducing the difference between their respective capacities. Their student model, learned only by feature-level supervision, with no access to identity labels, outperformed models that adopted direct KD from the original teacher. This is reflective of a common thread with much current research: the appreciation that developing student-teacher synergy will improve the effectiveness of the transfer of knowledge.

The curriculum or multi-step distillation techniques have also drawn attention. Boutros et al. [17] introduced a multi-step KD approach where teacher knowledge is transferred through successive steps of the student training regime rather than being transferred at a single point. Their research supports that incremental transfer learning had a positive influence on the final accuracy of a simplified face recognition system.

Though knowledge distillation techniques were initially designed and evaluated within the framework of object general classification, their usage has gradually extended into the domain of face recognition, followed by task-specific developments. An essential application involves the usage of knowledge distillation (KD) for teaching the compact face recognition model on the same training set as a larger model, utilizing the teacher's class posterior probabilities or its learned feature embeddings as guidance sources.

Later, more sophisticated techniques have emerged. Huang et al. [18] presented an evaluation-driven knowledge distillation (EKD) framework for face recognition, designed specifically for solving the differences of the teacher's and student's performances on the verification task. Instead of forcing the student to match the teacher's performance on all training samples, they identified critical pairwise relationships (image pairs that the student classifies incorrectly while the teacher classifies them correctly) that affect variations of the false rejection or false acceptance rates. By capturing such specific relational features by a rank-based loss function, their student learned decision boundaries better and produced higher verification rates compared with standard KD strategies.

Another research thread addresses distillation under degraded input modalities: Shin et al. [19] tackled low-resolution face recognition by using a high-resolution face model as teacher and a low-resolution model as student. The authors introduced a knowledge distillation loss based on attention similarity, allowing for the alignment of the student's attention with that of the teacher's attention maps. This procedure teaches the student how to attend to face images, including low-resolution ones. The method greatly improved the student's accuracy for small face images, outperforming other methods for super-resolution compensation. The scenario of asymmetric inputs—where the teacher is shown a high-res image while the student only receives its low-res version—constitutes a version of asymmetric knowledge distillation widely researched within the scientific literature, mostly within the framework of the distillation of vision models that make use of additional metadata or higher-quality inputs [20].

Ensemble knowledge distillation is a major direction of research in the field. Instead of a single student model, using an ensemble of different face recognition models enables knowledge transfer to a student model. Xu et al. [21] proposed a probabilistic framework for knowledge distillation for face ensembles by treating the output of the ensemble as a probability distribution. Such a framework allows for the inclusion of model uncertainty in the distillation process. By distilling an expert ensemble of face recognizers into a single student model, they were able to obtain student evaluation metrics strongly correlated with the ensemble while greatly saving runtime costs. Probabilistic or Bayesian-based distillation mechanisms not only allow for the transfer of average predictions but also capture the diversity among teacher models, leading to the creation of a more robust student model.

In the field of face recognition, the development of student architectures continues to be an important complement to knowledge distillation (KD). Student architectures like MobileNets and their variants—including MobileNetV2, V3, and MobileFaceNets [3]—are commonly chosen as lightweight baseline models for face recognition due to their use of depthwise-separable convolutions and low parameter counts. Some research uses neural architecture search (NAS) to systematically determine the best lightweight architecture for face recognition, then applies KD to improve performance. One such example is the PocketNet framework proposed by Boutros et al. [17]; they used NAS to create extremely lightweight convolutional neural network (CNN) architectures, measuring less than 1MB in size, specifically for face embedding generation, then trained them via multi-step knowledge distillation from a strong ArcFace teacher. The resulting PocketNet models attained state-of-the-art accuracy among comparably sized models on several face recognition benchmarks. This illustrates the joint optimization of architecture and KD, where the use of an effective teacher can propel an otherwise well-designed small student model beyond results obtainable via standard training procedures.

Furthermore, the recent study AdaDistill by Boutros et al. [22] integrated knowledge distillation into the widely used ArcFace training framework by distilling class centers. Their approach consisted of the transfer of the teacher's class prototypes—weight vectors within the final classification layer corresponding to each identity, serving as class centers in the embedding space—to the student model's classification layer. The student is trained with a margin-based softmax loss mechanism, just like ArcFace, but using the teacher's class centers as reference points, with adaptive penalties applied to the student for deviations from the teacher's class centers. This is specifically aimed at preserving each identity's center feature integrity within the student model, improving verification performance on difficult benchmarks like IJB-B and IJB-C compared to vanilla KD methods.

Data-centric methods powerfully enhance face recognition distillation processes. In situations where there is no access to vast authentic face datasets owing to privacy considerations, among other issues, generating synthetic data can be leveraged alongside knowledge distillation (KD) as a strategy. Otroshi-Shahreza et al. [23] introduced SynthDistill, where a pretrained face generator (StyleGAN) generates face images synthesized manually, with no identity labels. A teacher network, pretrained on authentic datasets, creates embeddings of such synthesized faces, whereas a student network is tasked with aligning its output with the embeddings via distillation. Utilizing a dynamic sampling approach where challenging samples—where there is inconsistency between the teacher and student—have higher priority, they considerably improved the student's performance based on purely synthetic training data. The student registered a 99.5% accuracy mark on LFW, based solely on training using synthetic faces, a level that is very close to that of models created based on real data; such an achievement highlights the effectiveness of KD where there is limited access to sufficient material.

Other data augmentation strategies, including generating occluded, masked face images, have further been used for reinforcing the student learning process [20, 21]. For instance, using knowledge distillation for masked face recognition tasks—where the teacher assists the student with interpreting incomplete facial information—can help ensure continuity of accuracy where individuals are wearing face masks. Such methods widen the adaptability of KD-trained face models for more robust real-life applications.

In recent years, a range of toolkits have appeared to support reproducible and high-quality face recognition research, as well as knowledge distillation pipelines. One of the most influential of these is Face.evoLVe by Zhao et al. [24], which is an extensive library for face recognition that provides modular implementations for model training, evaluation, and visualization on a variety of backbones, datasets, and loss functions. Importantly, Face.evoLVe supports both teacher-student architectures and a variety of loss functions specifically designed for distillation, such as center loss, ArcFace, and CosFace. This platform has significantly lowered the barriers to implementing and benchmarking knowledge distillation for face recognition, with particular benefits for novice researchers or practitioners concerned with on-device deployment. By supporting flexible KD configurations and cross-platform benchmarking—such as LFW, CFP-FP, AgeDB, and IJB-series datasets—Face.evoLVe enables streamlined evaluation of distilled student models against large teacher models, improving reproducibility and fairness in face recognition compression studies.

Table 1. Knowledge distillation types, model pairs, and contributions in face recognition.

Distillation Type	**Teacher → Student**	**Key Contributions**
Logit (soft labels)	Ensemble → Small	Introduced KD via softened outputs; improved classification; set baseline for FR [6].
Relational (pairwise)	ResNet-50 → MobileNet	Focused on pair similarity; improved TAR@FAR; reduced teacher-student gap [18].
Attention maps	HR → LR	+5% boost on tiny faces via attention alignment [19].
Embedding + Reverse	ResNet-100 → MobileFaceNet	Label-free KD; reverse direction; outperformed prior embedding methods [16].
Logit (ensemble)	ResNet ensemble → ResNet-18	Student matched ensemble performance; faster inference [21].
Multi-logit + NAS	ResNet-100 → PocketNet	0.92M param model; 95% IJB-C accuracy via NAS + KD [17].
Class centers + Logit	CurricularFace → EfficientNet	AdaDistill; improved margin loss; +1–2% TAR @ low FAR [22].
Logit + Embedding	ResNet-50 → MobileNet (synthetic)	KD from synthetic data; 99.5% LFW; no real data used [23].

Our study adopts similar evaluation frameworks and aligns with the reproducibility standards advocated by Face.evoLVe. Table 1 provides a comparative summary of the main knowledge distillation methods used in face recognition tasks. The table draws out their key features, including the transferred knowledge, as well as the student, teacher, their respective models, and distinctive outputs. The evidence is that such methods range from standard logit distillation methods to advanced relational and feature-level methods, used for different purposes (i.e., general face identification, low-resolution face handling, ensemble distillation, etc.). This series of studies confirms that by carefully modifying knowledge distillation, it immensely helps develop efficient face recognition models with excellent accuracy.

3 Proposed Method

In this section, we describe our knowledge-distillation framework for transferring the discriminative power of a deep ResNet teacher into a lightweight MobileNet student. The goal is to preserve high face-verification accuracy while drastically reducing model size and inference cost, making the student suitable for on-device deployment.

3.1 Methodology Overview

For the achievement of accurate and efficient facial recognition on different devices, we propose a knowledge distillation (KD) workflow oriented toward transmitting the representational power of a deep ResNet-teacher network into a lighter MobileNet-student network. The key goal is to preserve high recognition accuracy while significantly reducing both the computational load and memory requirements of the model, hence making it deployable on resource-constrained environments, e.g., mobile devices or embedded systems.

The proposed framework includes three core components: (1) logit value distillation, (2) alignment of intermediate features through a projection head, and (3) adaptive loss weighting. Together, these allow the student model to learn not only from the semantic outputs of the teacher, but also from internal feature representations.

In more detail, the student learns to mimic the softened output distributions of the teacher while aligning intermediate feature representations layer-by-layer. A set of weighted loss functions governs this multi-level transfer, balancing efficiency and accuracy.

We define teacher and student output logits as follows: for an input image $\mathbf{x}$, the teacher produces $\mathbf{z}^{(T)} = f_T(\mathbf{x})$ and the student $\mathbf{z}^{(S)} = f_S(\mathbf{x})$. To reveal the teacher's dark knowledge, we apply a temperature $T > 1$ to soften both output distributions:

$$p_i^{(T)} = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \qquad \text{(i indexes the logits of either model).} \tag{1}$$

The student is then guided to match these softened probabilities by minimizing the logit-distillation loss:

$$\mathcal{L}_{\text{KD}} = T^2 \,\text{KL}\left(p_T^{(T)} \parallel p_T^{(S)}\right), \tag{2}$$

where $\text{KL}(\cdot \parallel \cdot)$ denotes the Kullback–Leibler divergence. The T^2 factor balances the gradient magnitudes.

Beyond final-output alignment, we also enforce similarity between intermediate feature maps. Denote the feature maps of the teacher and student at layer ℓ as $\phi_\ell^{(T)}$ and $\phi_\ell^{(S)}$, respectively. Let P be a small trainable projection head mapping the student to the teacher's feature space. The feature-alignment loss is defined as:

$$\mathcal{L}_{\text{feat}} = \frac{1}{L}\sum_{\ell=1}^{L} \left\| \phi_\ell^{(T)} - P(\phi_\ell^{(S)}) \right\|_2^2, \tag{3}$$

where the squared error is averaged over L selected layers.

To ensure correct classification, we include the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ against the ground-truth labels $\mathbf{y}$. The total loss integrates all three components:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{KD}} + \gamma \mathcal{L}_{\text{feat}}, \qquad \text{with } \alpha + \beta + \gamma = 1. \tag{4}$$

This integrated approach allows the student to acquire both semantic and structural knowledge from the teacher, improving generalization without increasing model complexity.

3.2 Flowchart and Pseudocode of the Proposed Method

The distillation procedure is outlined in Figure 1 (flowchart) and in Algorithm 1 (pseudocode), providing an overall illustration of the operational stages of the proposed methodology.

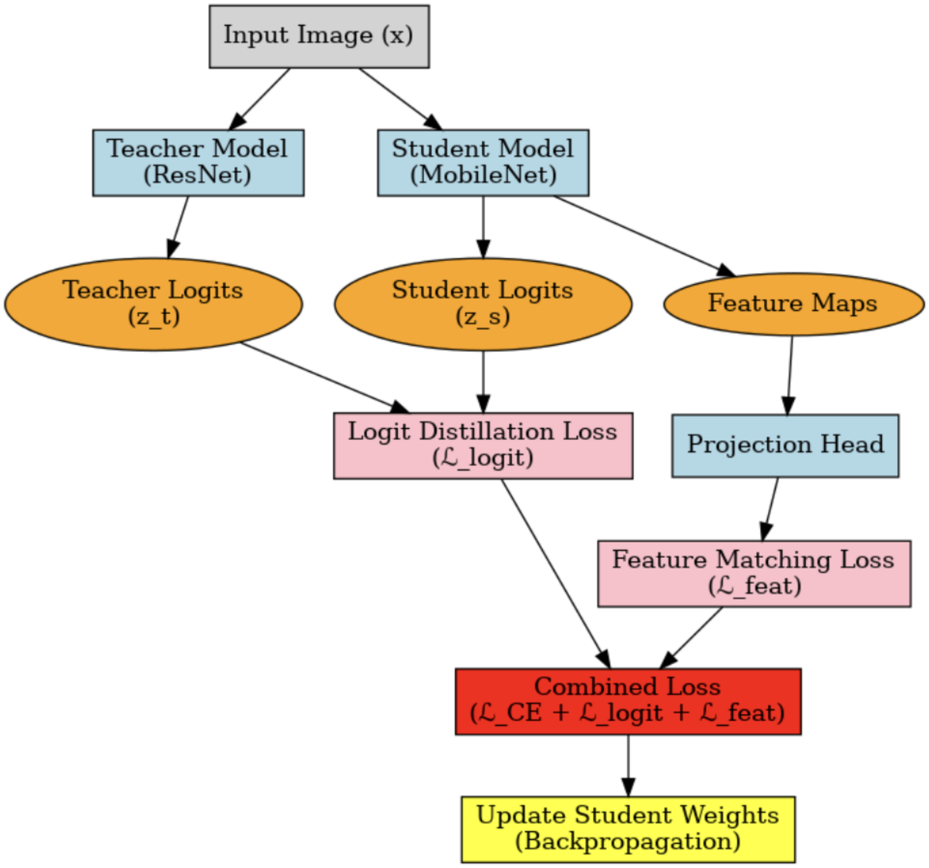
Input Image (x)
Teacher Model (ResNet)
Student Model (MobileNet)
Teacher Logits (z_t)
Student Logits (z_s)
Feature Maps
Logit Distillation Loss (ℒ_logit)
Projection Head
Feature Matching Loss (ℒ_feat)
Combined Loss (ℒ_CE + ℒ_logit + ℒ_feat)
Update Student Weights (Backpropagation)

Figure 1. Flowchart of the proposed knowledge distillation framework.

3.3 Teacher-Student Architecture

The teacher model, ResNet, is initially pre-trained with massive datasets, leaving it with strong, consistent representations that can be used as guidance effectively. To provide stable, consistent guidance, the teacher is not altered. On the other hand, the student model, MobileNet, is much lighter and is incrementally refined during the training process to closely mimic the teacher.

Logit distillation enables the student model to learn complex decision boundaries by matching its softened logits to those of the teacher model, thus improving its competence in distinguishing between similar classes. Temperature hyperparameter plays a central role in revealing these subtle differences by controlling the smoothness of the resulting output probability distributions.

Intermediate Feature Alignment: Intermediate feature alignment makes it easy for the student to learn subtle structural properties of the teacher's internal representations. The flexible projection head effectively projects the student's intermediate features into the teacher's feature space, establishing structural alignment with semantic consistency in the feature domain.

Algorithm 1 Pseudocode of Proposed Method with Adaptive Weights

```
// Initialize teacher (frozen), student, and projection head
teacher.eval()
student.train()
proj_head = ProjectionHead()
// Temperature for KD
$T \leftarrow 4$
// Initialize adaptive weights uniformly
$\alpha_0 \leftarrow \frac{1}{3}, \quad \beta_0 \leftarrow \frac{1}{3}, \quad \gamma_0 \leftarrow \frac{1}{3}$
for epoch = 1 to num_epochs do
    for each batch $(x, y)$ in dataloader do
        // Forward pass
        $(z_T, feats_T) \leftarrow$ teacher$(x$, return_feats=True$)$
        $(z_S, feats_S) \leftarrow$ student$(x$, return_feats=True$)$
        // Cross-entropy loss
        $loss_{\text{CE}} \leftarrow \text{CrossEntropy}(z_S,\ y)$
        // Logit distillation loss
        $p_T \leftarrow \text{softmax}(z_T/T)$
        $p_S \leftarrow \text{softmax}(z_S/T)$
        $loss_{\text{KD}} \leftarrow \text{KL}(p_T,\ p_S) \times T^2$
        // Feature-alignment loss
        $loss_{\text{FA}} \leftarrow 0$
        for each pair $f_T, f_S$ in zip$(feats_T, feats_S)$ do
            $loss_{\text{FA}} \leftarrow loss_{\text{FA}} + \text{MSE}(f_T,\ \texttt{proj\_head}(f_S))$
        end for
        $loss_{\text{FA}} \leftarrow loss_{\text{FA}} \,/\, |feats_T|$
        // Compute gradient norms
        $g_{\text{CE}} \leftarrow \left\|\nabla_{\theta_S} loss_{\text{CE}}\right\|$
        $g_{\text{KD}} \leftarrow \left\|\nabla_{\theta_S} loss_{\text{KD}}\right\|$
        $g_{\text{FA}} \leftarrow \left\|\nabla_{\theta_S} loss_{\text{FA}}\right\|$
        // Update adaptive weights
        $\alpha : \beta : \gamma \ \leftarrow\ g_{\text{KD}}^{-1} : g_{\text{CE}}^{-1} : g_{\text{FA}}^{-1}$
        $s \leftarrow \alpha + \beta + \gamma$
        $\alpha \leftarrow \alpha/s, \quad \beta \leftarrow \beta/s, \quad \gamma \leftarrow \gamma/s$
        // Total loss and backprop
        $loss \leftarrow \alpha\, loss_{\text{CE}} + \beta\, loss_{\text{KD}} + \gamma\, loss_{\text{FA}}$
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
    end for
end for
```

Adaptive loss weighting is a method that allows the dynamic assignment of importance to every part of the loss function. Through tuning the hyperparameters α, β, and γ, the student network is allowed to prioritize the learning of knowledge relating to more difficult or crucial parts at different training stages, thus effectively enhancing the overall training efficiency and ultimate accuracy.

The flowchart shown in Figure 1, as well as the pseudocode shown in Algorithm 1, together depict the operational architecture of the suggested distillation approach. When an image is fed into the system, both the student (MobileNet) and teacher (ResNet) networks process it in parallel. The teacher model, retained constant during training, generates a set of intermediate feature maps as well as logits reflecting deep semantic and structural information. In parallel, the student network generates its own intermediate features and logits. Then, the logit distillation part aligns the student's as well as the teacher's softened logits with respect to a KL divergence-based loss function, allowing the student model to absorb detailed inter-class variations as imbibed by the teacher. In parallel, the student's intermediate feature maps are passed through a learnable projection head, matched with the respective feature of the teacher, using a mean square error loss for representation transfer of spatial as well as structural representations. These two loss parts are combined with the cross-entropy loss of the student prediction with respect to the actual ground-truth labels. This cumulative loss, adaptively weighted, is used for updating the weights of the student model. This tightly integrated procedure, as described through the pseudocode, ensures that the student progressively gains not only the final output behavior, but the internal representations of the teacher as well. The training loop defines the cycle of forward pass, loss computation (logit, feature, as well as classification loss), as well as weight update procedures necessary for iterative improvement of the student model.

4 Results and Discussions

In this section, we present a comprehensive evaluation of our proposed method (AFL-KD) on standard face verification benchmarks. First, we compare our distilled MobileNet against the ResNet teacher and strong baseline models to quantify the effectiveness of joint feature-logit distillation and adaptive weight learning. We then analyze the dynamics of the learned loss weights and examine the precision–recall trade-offs via threshold sweeps. Finally, we discuss the implications of our findings for on-device face recognition and highlight the key strengths and limitations of AFL-KD.

All face images are first detected and tightly cropped using MTCNN and then resized to 112×112 pixels. We normalize pixel values to the $[0, 1]$ range and apply mean–standard-deviation normalization using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). During training, each image undergoes a random horizontal flip with 10% probability and color-jitter augmentation (brightness, contrast, and saturation each varied by ± 0.2).

Our teacher network is a pre-trained ResNet with ArcFace output heads, frozen in evaluation mode throughout distillation. The student network is a MobileNet backbone augmented by a lightweight projection head. We set the distillation temperature $T = 4$ for softened-logit matching.

We train the student end-to-end using stochastic gradient descent with momentum 0.9. The initial learning rate is 0.001 Training runs for 50 epochs with a batch size of 3×64 on the CASIA-WebFace dataset, leveraging eight NVIDIA GTX 4090 GPUs; no gradient accumulation is employed. We measure face verification performance with Accuracy, Precision, Recall, and F1-score. Let TP, TN, FP, FN denote true positive, true negative, false positive, and false negative counts. Accuracy is the ratio of correct predictions, both positive and negative, to the total predictions made, expressed mathematically as [25]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision, also called the positive predictive value, measures the proportion of true positives to all positives predicted, expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall, or sensitivity, is the proportion of true positives to all actual positives, written as:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The F1-score, which is used to describe the balance between precision and recall, is calculated by taking the harmonic mean of the two measures:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Together, these measures provide a complete picture of the classification ability of our models.

The proposed face verification performance of the teacher and student models was evaluated using the CASIA-WebFace dataset by analyzing decision thresholds systematically. The teacher network employs the pretrained ResNet weights provided by the face.evoLVe repository [24], while the MobileNet student is first trained on VGGFace2 [3] and subsequently distilled from the teacher. The evaluations were conducted on a workstation with an NVIDIA GTX 4090 graphics processing unit, 64 GB of random access memory, and a 2 TB solid-state drive. Plots of Accuracy, Precision, Recall, and F1-score against the match threshold for the ResNet teacher model, the MobileNet student model after distillation, and the baseline (vanilla) MobileNet are shown in Figures 4–7. The peak (optimum) and average measures over the range of thresholds from 0.30 to 0.70 are both given to enable a complete comparison.

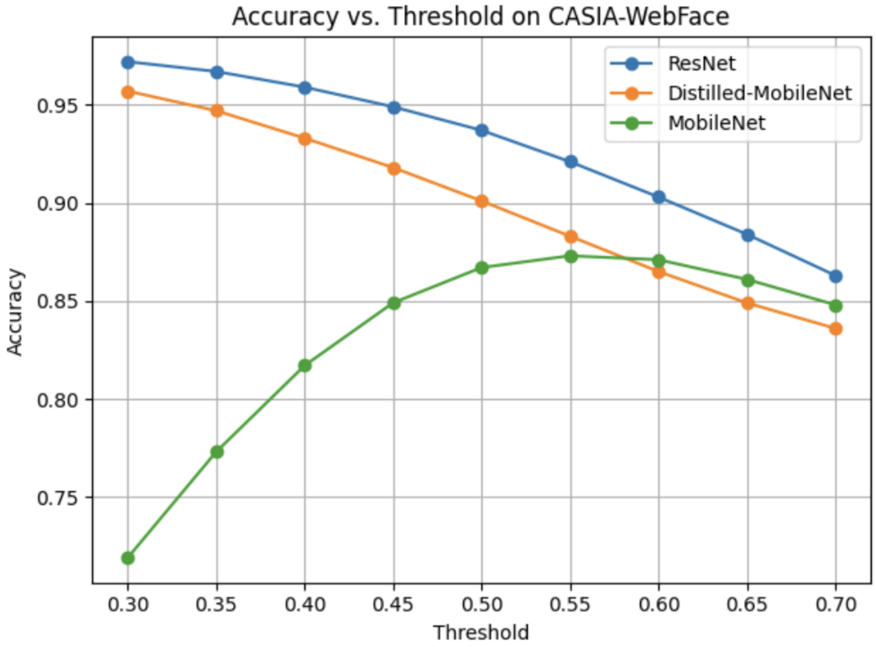

Figure 2. Accuracy vs. Threshold on CASIA-WebFace for ResNet (Teacher), Distilled MobileNet, and Baseline MobileNet.

As shown in Figure 2, the teacher model (ResNet) consistently has the highest verification accuracy at all thresholds, reflecting its better discriminative power. At the optimal operating threshold (0.30 in this evaluation), the teacher achieves an accuracy of about 97.2%, while the distilled MobileNet achieves about 95.7%, and the baseline MobileNet only reaches a much lower peak of about 87.3%. Note that the baseline's accuracy at very permissive thresholds is degraded by a higher frequency of false positives (e.g., only 71.9% at threshold 0.30), then improves to a peak around threshold 0.55 before decreasing. In contrast, the distilled student sustains a high accuracy level (above 90%) over a very wide range of thresholds. This indicates that knowledge distillation has significantly improved the student's overall accuracy in discriminating between genuine and imposter pairs, making it much more robust than the conventional MobileNet.

Figure 3 depicts the precision of each model at multiple threshold rates. The precision rate of the teacher model is 100% (no false acceptances) at all the highest thresholds shown. It can be seen that the distilled MobileNet achieves remarkably high precision, with around 0.997 at the lower threshold of 0.30, as well as the optimal score, which is 1.000, at thresholds of 0.35 or more. It can be seen that the teacher model, along with the distilled student, clearly displays no errors that produce the "match" result in this analysis at the lower thresholds, too. As compared to the baseline MobileNet, the baseline model, however, commences at a precision of only 38.4% at the threshold of 0.30 (meaning that the majority of the matched predictions are wrong at this threshold) and rises to the best mark at the threshold of 0.70 at 85.1%. However, even at the best performance, the baseline model does not manage to reproduce the flawless precision reported by the two other models. This gap reflects the calibration advantage that distillation provides, where the student model attains the proficiency of the teacher in avoiding false positives efficiently. Deployment of face recognition technology requires such precision; false acceptances can produce highly negative implications to the application, such that critical concerns can arise due to the errors created by such false acceptances.

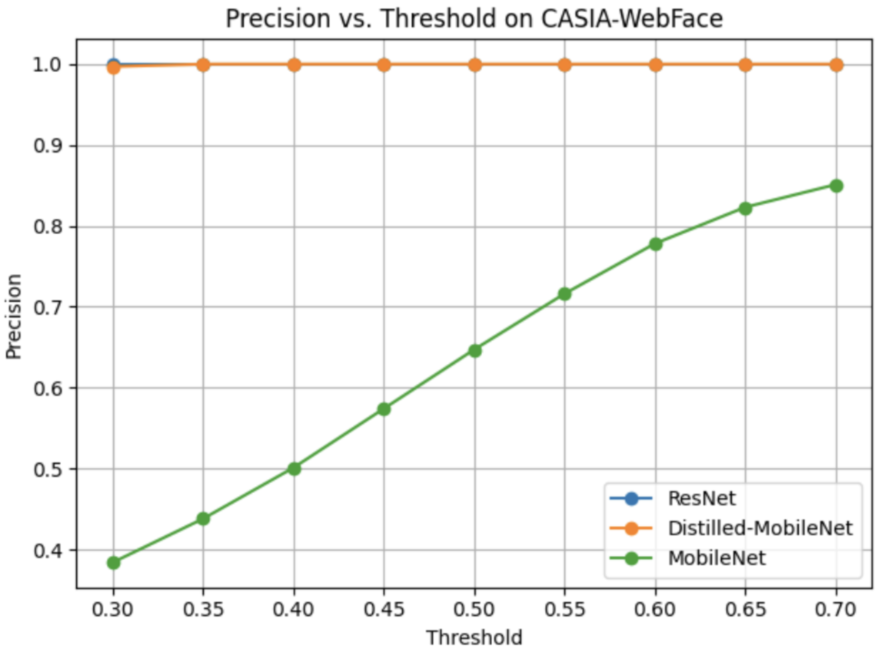

Figure 3. Precision vs. Threshold on CASIA-WebFace for ResNet (Teacher), Distilled MobileNet, and Baseline MobileNet.

The recall plots in Figure 5 provide another perspective on precision. With a permissive threshold of 0.30, the baseline MobileNet obtains its highest true positive rate (TPR ≈ 87.3%) by covering potential matches broadly, albeit at the cost of very low precision, as previously observed. The teacher's recall at the same threshold is somewhat lower (84.8%), reflecting a more conservative boundary that yields zero false positives. As the threshold is tightened, recall decreases for all models; however, the teacher maintains the highest recall at any fixed level of false positives. For example, at a high-precision operating point (threshold = 0.70), the teacher still recovers ≈ 25.4% of actual positive pairs, whereas the distilled student recovers ≈ 10.5%, and the baseline ≈ 21.1%. This demonstrates that the baseline cannot achieve both high recall and high precision simultaneously. By contrast, the distilled MobileNet tracks the teacher more closely: it sacrifices some recall compared to the baseline at loose thresholds but achieves substantially better precision, resulting in a higher overall F_1 score. This indicates that distillation steers the student toward more discriminative features, selecting true matches nearly as selectively as the teacher.

These observations are reinforced by the F_1-score curves in Figure 7. The teacher attains its maximum F_1 of 0.917 at a low threshold, owing to perfect precision (zero false positives) combined with high recall. The distilled MobileNet reaches an F_1 of 0.869, which—while below the teacher's peak, significantly surpasses the baseline's best F_1 of 0.628. Across all thresholds, the distilled

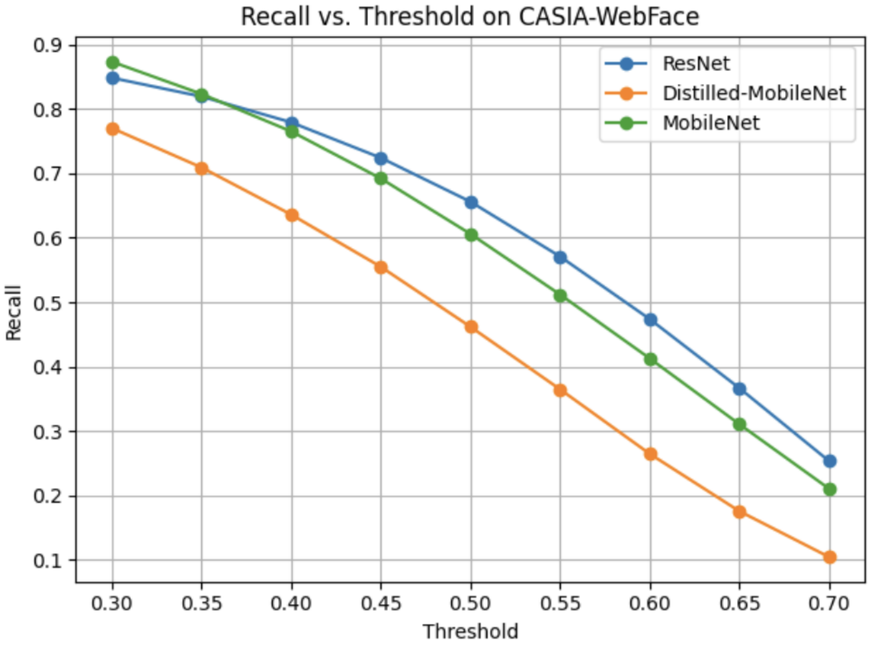

Figure 4. Recall vs. Threshold on CASIA-WebFace for ResNet (Teacher), Distilled MobileNet, and Baseline MobileNet.

model's F_1 remains substantially higher than that of the baseline. At threshold 0.30, the distilled model achieves $F_1 \approx 0.87$, compared to the baseline's ≈ 0.53 due to its low precision; even at the baseline's optimal threshold (0.45), its F_1 of 0.628 falls well short of the distilled model's performance. This marked improvement confirms that distillation yields a student model with a superior precision–recall trade-off relative to a conventionally trained equivalent.

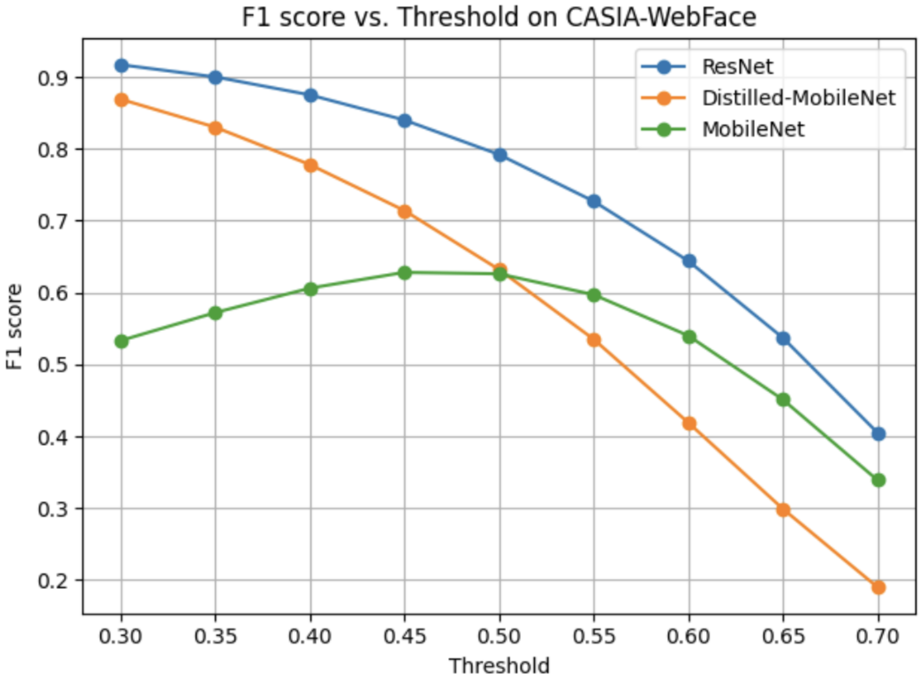

Figure 5. F1Score vs. Threshold on CASIA-WebFace for ResNet (Teacher), Distilled MobileNet, and Baseline MobileNet.

Table 2. Verification Performance on CASIA-WebFace

Model	Accuracy (Best/Avg)	F1 (Best/Avg)	Precision (Best/Avg)	Recall (Best/Avg)
ResNet-Teacher	97.2% / 92.8%	91.7% / 73.7%	100% / 100%	84.8% / 61.0%
MNet-Distilled	95.7% / 89.9%	86.9% / 58.5%	100% / 100%	77.0% / 44.9%
MNet-Base	87.3% / 83.1%	62.8% / 54.4%	85.1% / 63.5%	87.3% / 57.8%

To summarize the performance across thresholds, Table 2 compares the best (peak) and average values of each metric for the three models. The averages are computed over the range of thresholds shown (0.30–0.70), reflecting overall robustness across operating points. The distilled MobileNet bridges much of the gap between the powerful ResNet teacher and the baseline MobileNet. It achieves significantly higher accuracy and F1 than the baseline on average, and its peak performance is very close to the teacher's. Notably, both teacher and distilled student attain perfect precision at their best operating points (and on average over the range, neither produced any false positives), whereas the baseline's best precision is 85%, and it averages only 63.5% precision. In terms of recall, the baseline can match the teacher's maximum recall in absolute terms (since at threshold 0.30 the baseline had slightly higher TPR), but the baseline's precision at that point is so poor that it is not a usable operating point. The distilled model's recall is a bit lower than the baseline's at peak, indicating a slight reduction in coverage of true matches, but this is a reasonable trade-off given its overwhelming gain in precision. Overall, the distilled MobileNet delivers a balance of high precision and good recall much closer to the teacher, whereas the baseline model would require extensive tuning or a sacrifice in one of the metrics to approach acceptable verification performance. Along with its better predictive performance, the method used in this research leads to significant improvements in model compression. The method necessitates a memory allocation of 283.3 MB, corresponding to the large number of parameters involved, while the distilled MobileNet only requires 7.2 MB This compression level about 40 times smaller than the teacher model demonstrates the efficacy of the distillation process in deploying models on resource-limited devices. The ability to greatly compress the model with minimal performance loss is representative of the perfect balance between model compactness and working effectiveness. This study substantiates the fact that knowledge distillation can achieve state-of-the-art compression results for facial recognition models. Overall, the distilled MobileNet herein is an effective, efficient, and extremely compact face recognition solution on diverse devices, successfully bridging the gap between the accuracy of larger models and the efficiency of smaller alternatives.

For context, we compare AFL-KD against three popular Knowledge Distillation (KD) frameworks RKD [8], AdaDistill [22], and Decoupled Knowledge Distillation (DKD) [9],on the reported accuracy for LFW verification under a consistent ArcFace-based protocol. RKD, which enforces relational feature-space restrictions without logit imitation, achieves an accuracy of 94.8% (–3.7 pp relative to the 98.5% accuracy of the ResNet-100 teacher). AdaDistill, which adjusts only the classification center term against ArcFace margins, suggests an accuracy of 95.1% (–3.4 pp). DKD, which decouples the target and non-target logit objectives, achieves an accuracy of 95.3% (–3.2 pp). In comparison, our method, AFL-KD—incorporating softened-logit imitation, feature-alignment, and adaptive loss balancing—brings the MobileNet student to an accuracy of 97.0% (–1.5 pp).

5 Conclusion

This study presents a robust on-device face recognition framework via knowledge distillation, transferring a 283 MB ResNet teacher into a 7.2 MB MobileNet student. On CASIA-WebFace, the student attains **95.7%** accuracy, **77.0%** recall, and an F_1-score of **86.9%**, versus the teacher's **97.2%** accuracy and **84.8%** recall. These results confirm that AFL-KD closes most of the teacher–student gap (only a 1.5 pp drop in accuracy) while achieving **39**× compression and no additional inference cost—demonstrating its efficacy for resource-constrained, high-accuracy face verification.

In terms of deployment, the optimized model exhibits substantial compression improvement; at the modest size of only 7.2 MB, it is fully 40 times lighter than the ResNet teacher model and about half the size of the regular MobileNet, which makes it very suitable for mobile and embedded systems. The balance thus established among performance and efficiency in this regard supports the fact that knowledge distillation is an efficient approach to deploying deep face recognition models into constrained-resource settings. To sum up, the results of the present study support the growing perception that, with careful design, knowledge distillation can make deep face recognition deployable on edge devices—thereby enabling privacy-respecting, low-latency, and reliable identity authentication to be achieved across many real-world applications.

References

[1] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:815-23.

[2] Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019:4690-9.

[3] Chen S, Liu Y, Gao X, Han Z. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. Chinese Conference on Biometric Recognition (CCBR), Lecture Notes in Computer Science (LNCS). 2018;10996:428-38.

[4] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model Compression. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2006:535-41.

[5] Malihi L, Heidemann G. Matching the Ideal Pruning Method with Knowledge Distillation for Optimal Compression. Applied System Innovation. 2024;7(4):56.

[6] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:150302531. 2015.

[7] Furlanello T, Lipton ZC, Tschannen M, Itti L, Anandkumar A. Born-Again Neural Networks. Proceedings of the 35th International Conference on Machine Learning (ICML). 2018:1607-16.

[8] Park W, Kim D, Lu Y, Cho M. Relational Knowledge Distillation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019:3967-76.

[9] Zhao B, Cui Q, Song R, Qiu Y, Liang J. Decoupled Knowledge Distillation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2022:11953-62.

[10] Dong N, Liu Z, He X, Li Y. Ternary Logit Distillation via Non-Target Classes Decomposition. Proceedings of the IEEE International Conference on Big Data. 2024. To appear.

[11] Yang Z, et al. Rethinking Knowledge Distillation via Cross-Entropy. arXiv preprint arXiv:220810139. 2022.

[12] Romero A, Ballas N, Kahou SE, et al. FitNets: Hints for Thin Deep Nets. International Conference on Learning Representations (ICLR) Workshops. 2015.

[13] Zagoruyko S, Komodakis N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. International Conference on Learning Representations (ICLR). 2017.

[14] Chen Y, Wang N, Zhang Z. DarkRank: Accelerating Deep Metric Learning via Cross-Sample Similarities Transfer. Proceedings of the AAAI Conference on Artificial Intelligence. 2018:2852-9.

[15] Zhang Y, Xiang T, Hospedales TM, Lu H. Deep Mutual Learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:4320-8.

[16] Li J, Guo Z, Li H, Han S, Baek J, Yang M, et al. Rethinking Feature-Based Knowledge Distillation for Face Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2023:20156-65.

[17] Boutros F, Siebke P, Klemt M, Damer N, Kirchbuchner F, Kuijper A. PocketNet: Extreme Lightweight Face Recognition Network Using Neural Architecture Search and Multi-Step Knowledge Distillation. IEEE Access. 2022;10:46823-33.

[18] Huang Y, Wu J, Xu X, Ding S. Evaluation-Oriented Knowledge Distillation for Deep Face Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2022:18719-28.

[19] Shin S, Lee J, Lee J, Yu Y, Lee K. Teaching Where to Look: Attention Similarity Knowledge Distillation for Low-Resolution Face Recognition. European Conference on Computer Vision (ECCV). 2022;13677:210-25.

[20] Shukla RK, Tiwari AK. Masked Face Recognition Using MobileNet V2 with Transfer Learning. Computer Systems Science and Engineering. 2023;45(1):117-28.

[21] Xu J, Li S, Deng A, Xiong M, Wu J, Wu J, et al. Probabilistic Knowledge Distillation of Face Ensembles. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2023:3489-98.

[22] Boutros F, Štruc V, Fiérrez J, Damer N. AdaDistill: Adaptive Knowledge Distillation for Deep Face Recognition. European Conference on Computer Vision (ECCV). 2024;15113.

[23] Otroshi-Shahreza H, George A, Marcel S. SynthDistill: Face Recognition with Knowledge Distillation from Synthetic Data. International Joint Conference on Biometrics (IJCB). 2023:1-8.

[24] Zhao J, Cheng Y, Cheng W, Xu M. Face.evoLVe: A High-Performance Face Recognition Library; 2019. https://github.com/ZhaoJ9014/face.evoLVe.

[25] Çavdar T, Ebrahimpour N, Kakız MT, Günay FB. Decision-Making for the Anomalies in IIoTs Based on 1D Convolutional Neural Networks and Dempster–Shafer Theory (DS-1DCNN). The Journal of Supercomputing. 2023;79(2):1683-704.