中图法分类号:TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2026)08-2973-14

论文引用格式:Tang X, Zhang F F, Li G H, Ma Y X and Dong Z Y. 2026. Two-stage knowledge distillation method for lightweight face recognition. Journal of Image and Graphics, 31(8):2973-2986(唐鑫, 张飞飞, 李光辉, 马宇轩, 董正阳. 2026. 面向轻量级人脸识别的二阶段知识蒸馏方法. 中国图象图形学报, 31(8):2973-2986)[DOI:10.11834/jig.250454]

面向轻量级人脸识别的二阶段知识蒸馏方法

唐鑫[1],张飞飞[2],李光辉[1*],马宇轩[1],董正阳[1]

1. 江南大学人工智能与计算机学院,无锡　214122; 2. 江苏邦融微电子有限公司,苏州　215300

摘　要:**目的**　针对移动端与边缘设备计算资源受限的应用场景,现有高性能人脸识别模型虽然精度较高,但计算开销大,难以直接部署。知识蒸馏技术可将高性能教师模型的知识迁移至轻量化学生模型,从而在保持较高精度的同时降低计算复杂度。为进一步提升轻量化模型的人脸识别性能,提出一种新型两阶段对比式知识蒸馏框架TC-Face(two-stage contrastive knowledge distillation for face recognition)。**方法**　第1阶段,学生网络通过引入动量更新机制的对比学习,从教师网络的特征表示中学习,利用特征队列存储历史小批量特征,实现跨批次的知识传递,稳定蒸馏过程并增强特征区分性。为兼顾知识迁移与表征灵活性,第2阶段采用教师网络的分类器权重初始化学生分类器,从而在维持特征空间对齐的同时,赋予学生网络自主优化和拓展其判别能力的空间。**结果**　在IJB-C(IARPA janus benchmark)、IJB-B、MegaFace以及多个小规模验证基准上的综合实验表明,所提方法的识别精度都得到了提升,在IR100-MBF(IResNet100-MobileFaceNet)设置下,相较于最新的SOTA(state-of-the-art)方法UnifiedKD,在IJB-B和IJB-C数据集上,当假阳性率(false acceptance rate, FAR)为1E-5时,真接受率(true acceptance rate, TAR)都提升了1.89%。所有数据集上平均超过SOTA方法1.55%,并且在大规模测试集上的性能提升明显,在MegaFace数据集上平均准确率提升了2.72%,验证了所提方法在高难度人脸识别任务中的有效性与鲁棒性。**结论**　TC-Face框架通过两阶段训练策略,有效缓解了轻量化模型特征空间受限与精度下降的矛盾,兼顾了模型的高精度与轻量化特性,具有较高的实用价值和推广潜力。

关键词:人脸识别(FR);知识蒸馏(KD);对比学习;轻量化模型;动量更新

Two-stage knowledge distillation method for lightweight face recognition

Tang Xin[1], Zhang Feifei[2], Li Guanghui[1*], Ma Yuxuan[1], Dong Zhengyang[1]

1. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi　214122, China;

2. Jiangsu Bangrong Microelectronics Co., Ltd., Suzhou　215300, China

Abstract: **Objective** Face recognition (FR) has achieved remarkable progress in recent years mainly because of deep learning techniques and the availability of massive, annotated face datasets. Large-scale models, such as ResNet, demonstrate excellent discriminative capability, but their deployment on resource-limited devices, such as mobile and embedded systems, remains highly challenging because of prohibitive computational and storage demands. Lightweight networks, including MobileFaceNet, can reduce complexity and enable real-time applications but often suffer from a noticeable accuracy drop, especially under unconstrained and cross-domain conditions. To address this dilemma, researchers have introduced knowledge distillation (KD) as an effective strategy for compressing large teacher models into small student models

收稿日期:2025-09-28;**修回日期**:2025-11-22;**预印本日期**:2025-11-29

*通信作者:李光辉　ghli@jiangnan.edu.cn

基金项目:国家自然科学基金项目(62372214);苏州市科技计划项目(SGC2021070)

Supported by: National Natural Science Foundation of China (62372214); Suzhou Science and Technology Project (SGC2021070)

while retaining accuracy. However, conventional logit-based distillation is not very effective in FR because this task emphasizes discriminative feature embeddings rather than direct classification logits and faces additional challenges, including extremely large class space and flat soft targets. Recent studies have explored feature-based or relational distillation, but these approaches often couple student and teacher objectives too tightly, leading to gradient interference and unstable convergence. Furthermore, many methods rely solely on hard labels and overlook the rich relational information embedded in the training data. Motivated by these limitations, this study proposes a novel two-stage contrastive KD framework called TC-Face, which is designed to simultaneously enhance stability, efficiency, and recognition accuracy in lightweight FR models. **Method** The proposed TC-Face framework consists of two decoupled yet complementary stages. At stage one, we introduce a self-supervised contrastive distillation strategy. Instead of directly relying on logit outputs, the student network learns from the relational structure of teacher embeddings through a momentum-updated dynamic feature memory bank. This design allows the student to mimic fine-grained intersample relations derived from the teacher without being overwhelmed by noisy gradients. A difficulty-aware weighting mechanism is employed to further improve robustness. Each training sample is adaptively assigned a weight in accordance with its alignment difficulty: simple and overly hard samples are down-weighted, and moderately challenging samples contribute the most to optimization. This approach balances the knowledge transfer process and prevents overfitting to ambiguous or mislabeled identities. Moreover, momentum updating ensures that the teacher feature bank evolves smoothly, stabilizing supervision signals across iterations. The overall loss integrates embedding alignment, contrastive distribution matching, and adaptive weighting. At stage two, the student transitions from imitation to independent optimization. Specifically, the classifier parameters pretrained by the teacher are reused to initialize the student classifier, thereby reducing redundant training cost and accelerating convergence. With ArcFace-based angular margin loss, the student network now learns to refine its feature space independently; it explores discriminative embeddings while avoiding collapse into mere replicas of the teacher. The classifier is trained with frozen student parameters during the first epochs, followed by joint fine-tuning with shared learning rate schedules, to accelerate early-stage convergence. This two-stage decoupling ensures that the student absorbs structured relational knowledge from the teacher then refines its representation capacity through direct discriminative optimization. Compared with single-stage or tightly coupled KD, TC-Face balances imitation and independence, thus achieving training stability and strong recognition accuracy. **Result** We conduct extensive experiments on multiple public datasets to verify the effectiveness of TC-Face. Training is performed on MS1MV2 and MS1MV3 datasets (5.8 M and 5.1 M images, respectively), and evaluation covers widely used benchmarks, including labeled faces in the wild (LFW), AgeDB, cross-age labeled faces in the wild (CALFW), cross-pose labeled faces in the wild(CPLFW), celebrities in frontal-profile(CFP-FP), MegaFace, IJB-B, and IARPA janus benchmark(IJB-C). On the lightweight backbone MobileFaceNet (2.06 M parameters, 0.45 giga floating point operations per second), our method consistently outperforms prior distillation strategies. For example, the silhouette coefficient of embeddings trained with vanilla ArcFace is 0.200, whereas our method improves it to 0.236, surpassing even large models, such as ResNet18 (0.224). This result demonstrates superior intraclass compactness and interclass separability in the feature space. Ablation studies on the IJB-C dataset reveal the contribution of each component. Using vanilla KD with mean squared error loss yields true acceptance rate at false acceptance rate (TAR@FAR) = 1E−4 of 91.29% and TAR@FAR = 1E−5 of 79.79%. Introducing contrastive KD (CKD) improves performance to 92.81% and 84.24%. Our full method with adaptive weighting (CKD+) further increases accuracy to 93.07% and 88.33%. Varying the momentum update rate for the teacher feature bank shows that $\lambda = 0.999$ achieves optimal balance, reaching TAR@FAR = 93.45% at 1E−4 and 88.51% at 1E−5. These results indicate that dynamic memory updating and difficulty-aware reweighting are essential for optimal knowledge transfer. In large-scale evaluations, TC-Face demonstrates clear advantages over state-of-the-art methods. On MegaFace, our student model achieves Rank-1 accuracy of 94.2% and TAR@FAR = 1E−6 of 87.6%, outperforming AdaDistill-trained MobileFaceNet by more than 2% absolute margin. On IJB-B and IJB-C, where cross-pose and low-quality images present substantial challenges, TC-Face yields TAR improvements of 3%—5% over baseline KD methods. Moreover, despite the small capacity of the student network, our approach narrows the accuracy gap with teacher networks, such as ResNet50 and ResNet100, while retaining over 10× efficiency advantages in parameter size and floating point operations per second. Training efficiency is also substantially enhanced. Given that

stage one avoids training a large classifier and leverages precomputed teacher features, the number of trainable parameters is reduced by approximately 47.9 M, and the computation cost is reduced by around 91% compared with standard KD pipelines. Moreover, initializing the classifier at stage two with teacher parameters accelerates convergence, as shown by loss curves where TC-Face converges to a smaller training loss with fewer epochs compared with conventional approaches. These efficiency gains make the method highly suitable for edge deployment. **Conclusion** This work proposes TC-Face, a novel two-stage contrastive KD framework for lightweight face recognition. By decoupling imitation and discriminative optimization, TC-Face successfully mitigates the limitations of conventional KD methods. The first stage leverages a momentum-based feature bank and adaptive weighting to stabilize relational knowledge transfer, and the second stage enables independent embedding optimization with classifier parameter sharing. Extensive experiments across various benchmarks demonstrate that TC-Face not only improves accuracy by up to 5% TAR on challenging datasets, such as IJB-C, but also accelerates training and maintains efficiency suitable for mobile deployment. Compared with prior state-of-the-art KD methods, our approach achieves superior stability, faster convergence, and more discriminative embeddings, thus setting a new standard for lightweight FR training. In summary, TC-Face bridges the gap between large, high-accuracy teacher models and lightweight, deployable student models, providing a practical solution for real-world face recognition systems constrained by computational resources. Future work may extend this framework to other metric learning tasks, such as person re-identification or fine-grained visual categorization, where structured relational knowledge and decoupled training could similarly enhance lightweight model performance.

Key words: face recognition (FR); knowledge distillation (KD); contrastive learning; lightweight model; momentum updating

0 引　言

近年来，人脸识别(face recognition，FR)技术应用越来越广泛，例如，高光谱人脸识别通过引入丰富的光谱鉴别信息显著提升了跨谱条件下的鲁棒性(谢志华 等，2021)，而低质量三维人脸识别则利用软阈值去噪与视频帧融合有效缓解了真实场景中噪声和数据劣化带来的挑战(桑高丽 等，2023)。得益于高效的深度学习训练方法(Schroff等，2016；Wang等，2018；Deng 等，2019；Huang 等，2020；Kim 等，2022)以及大规模人脸数据集(Guo等，2016；Zhu等，2021)，人脸识别模型的准确率越来越高。现有的大规模数据集往往包含数百万身份，涵盖年龄、姿态和表情等多种变化，极大增强了模型的泛化能力。在这些数据的驱动下，具有数千万参数的大规模深度模型(如 ResNet(residual network)(He 等，2016))能够学习到高度判别性的特征，并在多个测试集上不断刷新性能纪录。

然而，将这些大型人脸识别模型直接应用于实际场景仍然面临挑战。现实中，很多应用需要将模型部署在移动端或边缘设备上，这些设备受限于存储与计算能力，难以支撑超大型深度模型的运行。为此，研究者提出了多种轻量化的人脸识别网络，例如 MobileFaceNet(MBF)(Chen 等，2018)，参数量不足100万，能够在移动端和嵌入式环境中实现实时验证。除此之外，GhostFaceNet(Alansari 等，2023)和 Edgeface(George 等，2024)等工作也在网络结构优化与操作简化方面做出了贡献。这类轻量化方法虽然能够有效降低模型复杂度和计算成本，但往往在识别精度上不如大型深度模型，尤其是在复杂场景或跨域测试下，性能劣化更为明显，并且出现认假率的提升。如何在有限的模型容量下仍保持高精度的人脸识别性能，成为亟需解决的问题。本文实验都是基于 MBF 实现的。

知识蒸馏(knowledge distillation，KD)作为一种经典的模型压缩与加速技术，为解决上述矛盾提供了可行方案。Hinton 等人(2015)最早提出通过蒸馏将教师模型中蕴含的知识迁移到学生模型中，其基本思想是利用教师网络的预测分布作为软标签，引导学生模型进行学习。后续研究(Kim 等，2021；Zhao 等，2022)在此基础上提出了多种改进方案，探索从分类层输出、隐藏层特征表示到样本关系结构的多层次蒸馏方式。事实证明，知识蒸馏能够在显著压缩模型规模的同时，保持甚至提升轻量化模型的性能，因此在人脸识别任务中得到了广泛应用。

然而，人脸识别的任务特性使得传统基于logit的蒸馏方法存在一定局限。与图像分类任务不同，人脸识别任务通常在训练集和测试集之间不存在身份交集，模型在推理阶段需要生成高质量的特征向量而非直接输出类别预测。与此同时，人脸数据集类别数量极大，教师网络输出的软标签往往过于平坦，导致蒸馏信号区分性不足，学生网络难以捕捉类别之间的细粒度差异。因此，近年来的研究逐渐转向特征层或特征之间的关系进行知识蒸馏。例如，Peng等人(2019)通过构建特征相关矩阵实现批级关系的迁移；Feng等人(2020)提出直接在特征空间进行对齐；Huang等人(2022)和Li等人(2023)进一步探索了基于样本关系的蒸馏方式。这些方法表明，相较于logit蒸馏，利用特征或关系进行知识迁移能够更好地适应人脸识别的需求。

除了从特征空间将教师网络和学生网络的特征进行对齐以外。人脸识别任务上的知识蒸馏工作将蒸馏损失与边际分类损失函数(如ArcFace)联合使用，使学生网络在训练中同时执行知识模仿与分类边界优化两项任务。虽然这种“双任务”机制在一定程度上能够提升模型判别性，但实验表明其往往加剧了训练的不稳定性，并未显著改善最终性能。Boutros等人(2024)提出在训练初期使用蒸馏，后期切换至边际损失，以缓解冲突；Mishra和Uikey(2025)则提出了统一的蒸馏框架，通过实例级嵌入对齐与几何关系保持，引入特征库与难例挖掘，进一步强化了蒸馏信号。然而，这些方法依然存在一定局限性：一方面，教师与学生在训练过程中的目标耦合过于紧密，容易造成梯度干扰与收敛不稳定；另一方面，现有蒸馏方法普遍依赖人工标签，未能充分利用样本间潜在的语义关系。

综上所述，尽管知识蒸馏在人脸识别中展现了良好潜力，但现有方法在训练稳定性与特征空间优化方面仍有改进空间。为此，本文提出了一种面向人脸识别的两阶段知识蒸馏框架TC-Face(two-stage contrastive knowledge distillation for face recognition)。第1阶段，引入基于对比蒸馏策略，结合动量更新机制构建动态特征库，使学生网络能够在教师网络的引导下稳定学习高质量特征表示；第2阶段，采用教师分类器参数初始化学生分类器，让学生网络继承教师网络的分类器权重，从而在加速收敛的同时，鼓励学生模型自主优化特征空间，避免单纯模仿教师网络特征库导致的学生网络特征库坍缩。通过这种解耦的两阶段训练策略，在多个基准数据集上实现了性能突破。

由图1可以直观地看出，传统的边际softmax损失与本文提出的对比损失方法在处理不同情况时的差异。

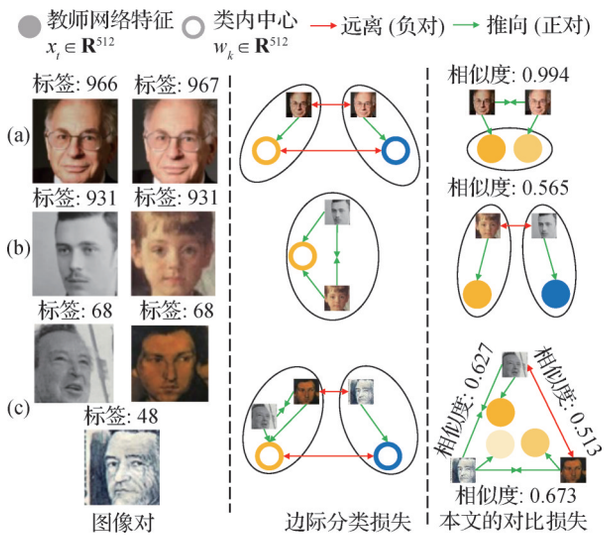


图1 传统人脸识别的训练方法与本文方法对比

Fig. 1 Comparison between traditional face recognition training methods and paper methods

在图1(a)中，尽管样本的标签不同，但它们实际上属于同一类别。传统方法往往会将这些同类样本与其对应的类别中心强行拉开，而本文方法则在教师网络特征表示的引导下对其进行合理对齐。在图1(b)中，样本的标签相同，但实际上它们属于不同类别。传统方法会错误地将这两个样本的特征尽量拉近，并一同推向类别中心；相比之下，本文方法不会被错误标签所误导，从而避免了这种不正确的特征对齐。在图1(c)中，当样本之间的类别关系难以判定时，传统方法依旧会依据给定标签进行硬性分类，而本文方法则利用教师网络的特征提示进行柔性分类，从而获得更加稳健的判别效果。

本文的主要贡献如下：1)提出了一种两阶段知识蒸馏框架TC-Face，在保证轻量化模型效率的同时显著提升识别性能。2)第1阶段利用对比蒸馏和动量特征库，确保学生网络高效稳定地模仿教师网络的特征表示；第2阶段通过教师分类器参数初始化学生分类器，加速模型收敛并促进学生网络探索更具判别性的特征空间。3)在多个公开基准数据集上

开展实验，结果表明本文方法在精度与稳定性方面均优于现有蒸馏方法。

1 本文方法

首先回顾传统的知识蒸馏方法及其在人脸识别中的改进，指出现有方法往往忽略训练样本之间的关系，从而限制了其在人脸识别任务中的效果。为解决这一问题，提出一种专门面向人脸识别的损失函数，有效捕捉样本间的关系信息。此外，强调采用两阶段训练策略对优化人脸识别模型的重要性。方法总体的框架如图 2 所示。

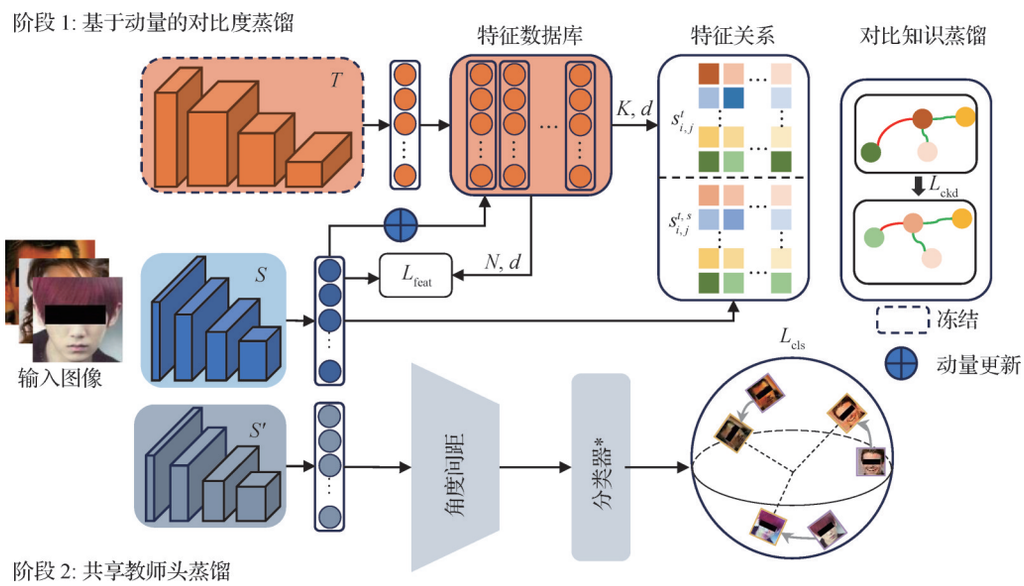


图 2 本文方法总体框架

Fig. 2 The overall framework of ours

1.1 知识蒸馏与能力差距

设教师网络为 T，学生网络为 S，大多数基于知识蒸馏的人脸识别方法的总损失函数可表示为

$$L = \alpha L_{\text{feat}} + \beta L_{\text{logit}} + \lambda L_{\text{cls}} \tag{1}$$

式中，α，β，λ 是权重系数，控制各项损失在总损失中的贡献比例。L_{cls} 通常为使用真实数据集标签计算的分类损失，常使用 ArcFace 作为损失函数；L_{feat} 是在特征层将教师网络和学生网络的特征进行对齐，常使用均方误差（mean squared error，MSE）、曼哈顿距离或余弦相似度（cosine similarity）等；L_{logit} 是在分类层将教师网络和学生网络的 logit（即概率预测）进行对齐，常使用 KL 散度（Kullback-Leibler divergence）。在人脸识别任务中，由于类别数量极大（常达数万甚至百万），直接在分类层进行蒸馏会导致梯度稀疏——因为教师模型对绝大多数非目标类别的输出 logit 接近很小，softmax 后概率趋近于 0，从而使得梯度几乎只集中在少数几个类别上。这种“极小梯度”现象会严重削弱蒸馏效果，甚至导致训练不稳定。所以不使用 L_{logit}，故将 β 设置为 0。第 1 阶段主要基于特征进行蒸馏学习，故只使用 L_{feat}；第 2 阶段基于数据集的标签进行学习，故使用 L_{cls}。本文尝试了将两个损失加权进行训练，发现损失发生了振荡并且对识别模型的精度提升不大（在后续的章节中会对实验进行详细描述）。在人脸识别中，常采用网络嵌入（embedding）的距离度量进行特征蒸馏。与 Li 等人（2023）和 Yu 等人（2023）的做法一样，本文在归一化后的嵌入特征上采用均方误差损失，其形式为

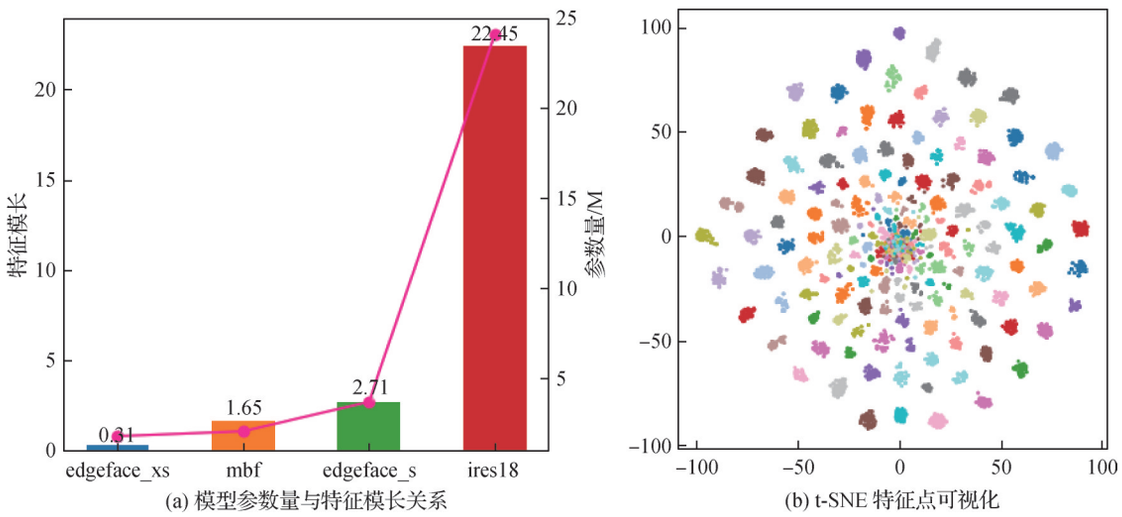
$$L_{\text{feat}}(\boldsymbol{f}_{\text{t}},\boldsymbol{f}_{\text{s}}) = \frac{1}{N}\sum_{i=1}^{N}\left\| \boldsymbol{f}_{\text{t}}^{i} - \boldsymbol{f}_{\text{s}}^{i} \right\|_2^2 \tag{2}$$

式中，$\boldsymbol{f}_{\text{t}}$ 和 $\boldsymbol{f}_{\text{s}}$ 分别表示标准化后由教师网络和学生网络提取的特征嵌入，N 为批大小。该损失等价于在超球面空间上最小化教师与学生特征向量之间的角距离。

人脸识别作为一种度量学习（Xing 等，2002），将样本映射到超球面空间，旨在学习适当的距离，但随着特征维度减小，人脸识别模型的准确性不可避免地下降。因此，需要在模型特征表示和准确性之间找

到平衡，本文方法在两阶段训练过程中解决此问题。

学生网络的容量限制是一个具有挑战性的问题。通过对使用传统方法（Deng 等，2019）训练的 MBF 模型推断的特征表示应用 t-SNE（t-distributed stochastic neighbor embedding）（Van Der Maaten 和 Hinton，2008）（见图 3（b）），观察到虽然大多数类别分离良好，但仍有许多样本集中在原点附近难以区分，这些模糊样本通常属于难以识别身份的个体或错误标记的类别。较大模型通常表现出更优越的分类性能，能更精确地区分样本。



(a) 模型参数量与特征模长关系　(b) t-SNE 特征点可视化

图 3　模型参数量与特征模长关系图以及特征点 t-SNE 可视化

Fig. 3　Relationship diagram between model parameters and feature modulus, and t-SNE visualization of feature points ((a) relationship diagram between model parameters and feature modulus; (b) t-SNE visualization of feature points)

为定量评估识别模型对样本的分类能力，本文引入轮廓系数（silhouette coefficient）（Rousseeuw，1987）作为衡量指标，如表 1 所示，其值介于−1 和 1 之间，越高越好。通过计算每个样本与其所属簇内其他样本的平均距离（类内紧密度）以及与最近邻其他簇样本的平均距离（类间分离度）来综合评估识别模型对样本特征的分类质量，轮廓系数定义为

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{3}$$

式中，$a(i)$ 表示样本 i 到同簇其他样本的平均距离，$b(i)$ 表示样本 i 到最近邻异簇样本的平均距离。更高的轮廓系数意味着样本在特征空间中具有更清晰的类内聚集性和类间可分性，从而反映更强的分类能力。实验结果表明，参数规模较小的模型往往具有较低的轮廓系数，反映出其较差的分类能力。通过轮廓系数进一步强化了模型容量与识别性能之间的权衡。

值得注意的是，如果学生网络仅在特征级别模仿教师网络，可能导致特征空间整体收缩，使得不同类别的特征彼此靠近，从而降低类间可分性，表现为轮廓系数下降。此外，仅依赖数据集标签的传统基于分类器的知识蒸馏方法，可能因数据中存在显著噪声（如错误标注或模糊身份）而误导容量有限的学生网络，使其难以学习到鲁棒的判别性特征。为解决上述问题，本文第 1 阶段采用基于教师网络特征关系的对比知识蒸馏（contrastive knowledge distillation，CKD）方法，通过保留教师网络中样本间的相对关系（如相似性结构），引导学生网络在受限容量下仍能学习到具有良好类间分离性和类内紧凑性的特征表示，从而提升轮廓系数和整体识别性能。

1.2　教师指导

鉴于人脸识别数据集本质上包含样本间的强相

表 1　使用 ArcFace 训练的常见人脸识别模型的轮廓系数

Table 1　Silhouette coefficient of common face recognition models trained using ArcFace

模型	分数
mbf	0.200
ires18	0.224
ires34	0.308
ires50	0.328
ires100	0.342

关性，对于小批量B中的每个样本$\boldsymbol{x}_i(i \in \{1, \cdots, N\})$，学生网络提取标准化后的特征嵌入$f_i^s$，从教师网络特征中维护动态特征库（$\{\boldsymbol{f}_j^t\}_{j=1}^K, j \in \{1, \cdots, K\}$），其中$K$表示特征数据库大小）。知识蒸馏的软目标仅从教师网络的内部特征相似性中得出。具体为

$$P_{ij} = softmax\left(\langle \boldsymbol{f}_i^t, \boldsymbol{f}_j^t \rangle\right) = \frac{\exp\left(\langle \boldsymbol{f}_i^t, \boldsymbol{f}_j^t \rangle\right)}{\sum_{k=1}^{K} \exp\left(\langle \boldsymbol{f}_i^t, \boldsymbol{f}_k^t \rangle\right)} \tag{4}$$

式中，P_{ij}表示样本i的教师特征与教师库中所有特征（索引$j \in \{1, \cdots, K\}$）之间的归一化相似性。$\langle \boldsymbol{a}, \boldsymbol{b} \rangle$代表向量$\boldsymbol{a}$，$\boldsymbol{b}$之间的余弦相似度。然后，学生网络通过计算其对同一教师特征库的预测相似性分布来模仿此教师派生分布，具体为

$$Q_{ij}^{(T)} = softmax\left(\langle \boldsymbol{f}_i^s, \boldsymbol{f}_j^t \rangle / T\right) = \frac{\exp\left(\langle \boldsymbol{f}_i^s, \boldsymbol{f}_j^t \rangle / \tau\right)}{\sum_{k=1}^{K} \exp\left(\langle \boldsymbol{f}_i^s, \boldsymbol{f}_k^t \rangle / \tau\right)} \tag{5}$$

式中，τ是温度缩放超参数。整体训练目标是最小化学生预测分布$Q_{ij}^{(T)}$与教师提供的软目标P_{ij}之间的交叉熵损失，具体为

$$L_{\mathrm{CKD}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} P_{ij} \log\left(Q_{ij}^{(T)}\right) \tag{6}$$

式中，温度参数τ仅应用于学生网络的特征相似度计算，而教师网络则保持原始相似度（即$\tau = 1$），这一设计选择源于对知识迁移过程的深入理论分析。作为已充分训练的模型，教师网络的特征空间已形成稳定且具有高度判别性的表示结构，直接使用原始相似度能够精确保留教师网络对样本间关系的内在刻画。相比之下，学生网络处于学习阶段，其特征表示能力有限且存在不稳定性，通过在学生端引入温度参数，可有效调节学生网络对教师知识的学习强度与范围，实现知识迁移过程的自适应控制。

实验结果表明，温度参数τ的合理设置对模型性能具有决定性影响。当$\tau > 1$时，softmax 函数输出的分布趋于平滑，促使学生网络以更均衡的方式关注各类样本关系，避免过度集中于少数困难样本；而当$\tau < 1$时，softmax 分布呈现更高的锐度，使学生网络过度聚焦于最具挑战性的样本关系。由于学生网络自身表征能力有限，若过度关注教师网络判定为困难的样本，将导致模型在这些样本上过拟合，反而损害泛化性能。针对人脸识别这一度量学习任务，所提出的方法在设置τ为 1.1 时，能够获得最优性能。这种设置既保持了样本关系的足够判别性，又通过适度平滑的分布避免了学生网络对难样本的过度关注，从而在特征表示的广度与识别精度之间实现优化平衡，有效满足人脸识别任务对精确区分不同身份间细微差异的要求。

教师网络生成的软标签蕴含丰富的细粒度信息，使学生网络能够学习更为全面和精确的特征表示。通过利用软标签，学生网络不仅摆脱了对刚性目标类别的依赖，还能通过教师网络的相似性信息逐步优化其特征表示，从而捕获更为精细的特征关系。该方法有效缓解了硬标签固有的信息压缩问题，尤其对于具有模糊性或噪声的复杂样本，软标签提供的平滑目标分布显著提升了学生网络从教师模型中吸收知识的效率。此外，软标签机制使教师网络能够将其对类间细微区别的理解有效传递给学生网络，促进学习过程中对特征空间广度和深度的深入探索，最终显著增强了模型的泛化能力和鲁棒性。第 1 阶段的损失函数可以归纳为

$$L = \alpha L_{\mathrm{feat}} + \beta L_{\mathrm{CKD}} \tag{7}$$

1.3 每个样本的动态权重

虽然先前所提出的方法最大化了从教师网络的知识转移并充分利用了数据集的内在相关性，但对于容量受限的学生网络 MBF 仍有一个很大的限制：没有自适应加权的均匀知识传播导致知识蒸馏的过程中优先考虑过于具有挑战性的样本（更多的是难以区分的样本）。为解决此问题，引入了一系列措施，动态调整学生在线学习中的困难样本，防止过拟合。为自适应地强调训练中有信息量的样本，引入了难度感知加权系数$D_i = \sum_{j=1}^{K} P_{ij} \log Q_{ij}^{(T)}$，对于每个样本$\boldsymbol{x}_i$，标准对比损失量化了学生和教师特征分布之间的差异（当$D_i = 1$时，学生网络的特征和教师网络的特征完美对齐），其中整体损失计算式为

$$L_{\mathrm{CKD}^*} = \frac{1}{N} \sum_{i=1}^{N} w_i D_i = \frac{1}{N} \sum_{i=1}^{N} \exp(-\mu D_i) D_i \tag{8}$$

式中，D_i为难度感知加权系数，$\mu > 0$是控制加权强度的温度超参数，样本权重定义为$w_i = \exp(-\mu D_i)$。该加权策略自动降低简单样本（$D_i \approx 0$，其中$w_i D_i \approx 0$）和困难样本（$D_i \gg 1/\mu$，其中$w_i \to 0$）的权重，优先考虑具有中等难度的样本（$D_i \approx 1/\mu$）。因此，学生网络将学习集中在教师提供的更有意义监督信号的实例上，增强对噪声或模糊特征的鲁棒性。

为进一步减轻嵌入空间中的过拟合，为对比蒸馏实现了基于动量的特征库更新机制。受He等人(2020)移动平均策略的启发，逐渐优化教师的特征库，而不是突然替换，增强对噪声或挑战性样本的稳定性。具体而言，对于教师库中的每个特征f_t，使用学生的对应特征f_s更新，具体为

$$f_t \leftarrow \lambda f_t + (1-\lambda) f_s \tag{9}$$

式中，$\lambda \in [0.9, 1.0]$是动量系数。较低的值更新教师特征库会导致教师特征库逐渐被学生网络的特征所替换，使得学生网络无法学习到教师网络的特征；过高的值则导致学生网络过于集中于教师网络中的结果，失去自主探索的能力。

1.4 两阶段训练

先前的方法通常将蒸馏损失与边际分类损失函数(如ArcFace)联合使用，在整个训练过程中都要求学生严格模仿教师网络的特征分布，教师与学生在训练过程中的目标耦合过于紧密，容易造成梯度干扰与收敛不稳定。然而单一使用蒸馏损失会限制学生模型对自身特征空间的探索。针对这一问题，提出了一种两阶段训练策略，旨在平衡继承教师知识和独立特征优化之间的关系。

第1阶段，训练的核心目标是最大化学生网络对教师网络知识的吸收。具体而言，学生网络在教师网络的监督下学习其特征表示及特征分布，主要通过嵌入损失和对比损失来完成。嵌入损失确保学生特征与教师特征在实例级保持一致，而对比损失则进一步建模样本之间的几何关系，使学生能够捕捉更精细的类间以及类内关系。值得注意的是，由于学生和教师在网络容量上存在显著差距，学生模型无法完全重现教师的高维分布，因此该阶段主要起到知识迁移和表征初始化的作用，而不是最终的特征空间优化。为了避免训练陷入过度模仿的局限性，在该阶段限制了学生探索自身特征空间的能力，使其尽可能集中于吸收教师提供的判别性知识。

第2阶段，训练范式回归到传统的人脸识别流程，学生网络不再依赖教师的嵌入指导，而是通过基于类别的损失函数(如ArcFace)独立优化特征空间。这一阶段的目标在于让学生充分释放其建模能力，在已有蒸馏初始化的基础上，自主构建最优的类间边界与类内紧致度。与部分先前方法不同，该方法并未将蒸馏过程与标准人脸识别训练混合，而是明确地将两者解耦。这样的设计能够有效避免蒸馏信号与分类监督之间的梯度干扰，使训练过程更为稳定和高效。

此外，为进一步加速第2阶段的收敛，并且提升最终训练出的识别模型的准确率，利用教师网络训练得到的分类器参数初始化学生网络的分类器。人脸识别模型中，分类器的参数代表着一个类别的中心，使用教师网络的参数初始化第2阶段学生网络训练的分类器，这样可以达到学生网络进一步模仿教师网络特征并且加速训练的收敛的目的。实验表明，这种策略使得收敛速度明显提升，同时在多个基准测试中均带来性能增益。如图2所示，冻结的教师网络T提取特征，用于计算学生网络S的嵌入损失。教师特征通过动量在特征库中更新，对比损失则鼓励学生与教师的特征空间对齐。S'为一阶段训练完成的学生网络。*表示分类器已使用教师分类器的参数初始化。实验结果表明，这种二阶段的训练策略不仅增强了学生网络对教师知识的利用效率，还提高了其在低维特征空间中寻找最优分布的能力，从而显著提升了最终的人脸识别准确率。其优点和机制将在第2.4节进行更为详细的分析。

2 实 验

本节讨论从全面比较和消融研究中获得的实验结果，以证明所提出TC-Face方法的有效性。

2.1 数据集与评估指标

使用MS1MV2(MS-Celeb-1M Version 2)和MS1MV3训练所提出模型，并与先进知识蒸馏(KD)方法进行公平比较。上述两个数据集都是MS-Celeb-1M的子集，分别包含580万和510万幅图像。

为全面评估模型，在7个基准数据集上进行了测试。其中，LFW(labeled faces in the wild)、CFP-FP(celebrities in frontal-profile)、CPLFW(cross-pose labeled faces in the wild)、AgeDB(age database)和CALFW(cross-age labeled faces in the wild)是人脸识别中最常用的基准数据集。尽管这些数据集在光照条件、姿态变化和年龄分布上有所不同，但它们通常保持高质量图像，适合评估标准设置下的人脸识别模型。此外，使用MegaFace和IJB-C(IARPA janus benchmark)、IJB-B数据集，这些数据集呈现更具挑战性的场景。这些数据集包含大量身份和更复杂的

现实世界变化。在人脸识别研究中，它们通常用于评估大规模场景下的模型性能，例如涉及数千甚至数万个身份的场景。在 LFW、AgeDB、CA-LFW、CPLFW 和 CFP-FP 数据集上，按照各自评估协议以验证准确率评估模型。对于 IJB-B 和 IJB-C 数据集，采用标准 1∶1 混合验证协议，并报告假阳性率（false acceptance rate，FAR）为 1E−4 和 1E−5 时的真接受率（ture acceptance rate，TAR）。对于 Megaface 基准，使用 Rank-1 识别准确率和 TAR = 1E−6 时的验证准确率报告识别性能。

本文比较了两个教师—学生对（IR100-MBF 和 IR50-MBF），并在各种基准上将提出的方法与几个 SOTA（state-of-the-art）竞争对手进行比较。为了公平比较，IR100- MBF 对在 MS1MV3 上进行训练，IR50-MBF 对在 MS1MV2 上训练。

2.2 实现细节

在人脸识别任务中，面部图像首先通过人脸检测模型处理以提取关键点，然后基于预定义的双眼坐标对齐到固定的 112 × 112 像素分辨率。所有训练和测试数据集均遵循此标准预处理流程。对于教师网络，采用 ResNet100（6 515 万参数和 12. 12 GFLOPs）和 ResNet50（4 359 万参数和 6. 34 GFLOPs），均使用默认 ArcFace 配置（尺度参数 $s = 64$，边距 $m = 0.5$）训练。特征维度设置为 512；对于学生网络，与先前的 SOTA 方法一致，采用 MFN（206 万参数和 0. 45 GFLOPs）。为确保公平比较，所有实验均在相同的条件下进行。实验使用的主要设备为：14th Gen Intel（R）Core（TM）i9-14900 中央处理器，Nvidia GeForce RTX 4090 显卡。由于本文方法侧重于从样本及其关系中学习，而非数据集标签，因此在第 1 阶段无需训练大型全连接分类器。这将可训练参数数量减少了约 4 793 万（以 MS1MV3 为例），显著加速了训练过程。此外，将教师网络推断的特征存储在磁盘上，避免在每次小批量中计算它们，上述举措可以将第 1 阶段训练的计算成本降低约 91%。对于学习率调度，使用带预热阶段的余弦学习率在前 5 个 epoch 从 1E−8 线性增加到 1E−1，然后使用余弦退火逐渐衰减到 1E−8。采用随机梯度下降（stochastic gradient descent，SGD）作为优化器，动量为 0. 9，权重衰减为 0. 000 5，这与 ArcFace 一致。鉴于学生网络的有限能力，仅应用水平翻转作为数据增强，概率为 0. 5。超参数配置如下：对比损失中的温度参数 T 设置为 1. 1。从教师特征库中每批检索的特征数量 K 等于 1 000，减少计算开销的同时保持较高的准确性，更新特征库的动量系数 λ 设置为 0. 999。考量到对于不同样本的均衡，CKD$^+$中的 μ 取值为 0. 9。对于平衡不同损失项，权重 α 和 β 分别分配为 0. 3 和 0. 7。第 2 阶段训练中，采用 ArcFace 配置，尺度参数为 64，边距为 0. 5。使用教师网络的分类器权重初始化全连接分类器。前 2 个 epoch，冻结学生网络参数，仅训练全连接层。之后，分类器和学生网络共享相同的学习率调度。此阶段的最大学习率设置为 1E−2，并随迭代轮次以固定衰减系数更新。

2.3 TC−Face 结果

TC-Face 使用对比损失取得了令人印象深刻的成果。如表 2 所示，其中，mbf_vanilla 表示使用 ArcFace（Deng 等，2019）训练的 MBF 模型，mbf_ada 表示使用 adadistill（Boutros 等，2024）训练的 MBF。从表 2 可以看出，本文方法训练的模型表现出更高的轮廓系数，证明了该方法在训练集上的最佳分类性能，它甚至超过了参数规模为其 10 倍的 ResNet18。

表 2 使用不同方法训练的 MBF 的轮廓系数

Table 2 Using different methods to train the simulation coefficients of MBF

模型	分数
mbf_vanilla	0.200
mbf_ada	0.210
mbf_ours	0.236
ires18	0.224

在所提出的方法中，不仅要求学生网络学习从教师网络派生的样本间关系，还要求模仿教师网络的特征表示。如图 4 所示，图 4（a）的柱状图展示了未归一化的学生和教师网络之间的特征比较，而图 4（b）的柱状图展示了归一化后的比较。虽然图 4（a）的柱状图显示学生和教师特征之间存在相当大的差异，但图 4（b）的柱状图表明两个网络的归一化特征几乎重叠。该方法在特征幅度和从教师模仿特征的保真度之间实现了平衡。

在训练的第 2 阶段，通过使用教师分类器参数初始化学生分类器。在前两个 epoch 中，冻结学生网络的梯度以加速收敛。如图 5 所示，在此梯度冻结期间，训练损失迅速下降。之后，当使用共享学习

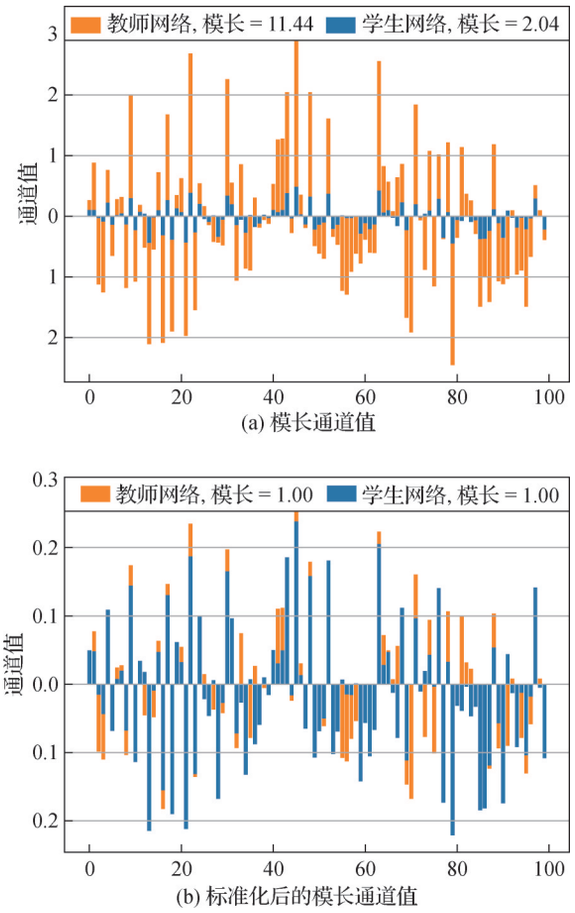


图4　学生与教师网络的特征对比

Fig. 4　Comparison of feature between student and teacher networks((a) modulus length channel value; (b) standardized modulus length channel value)

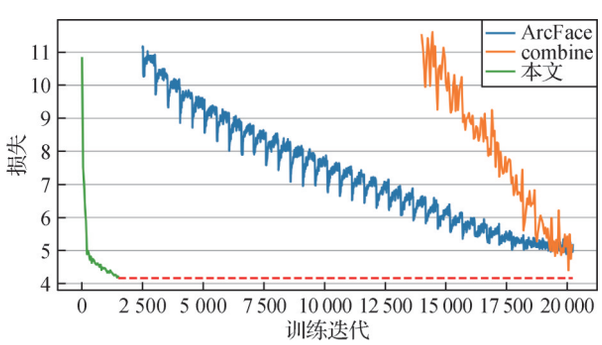


图5　不同方法训练损失值对比

Fig. 5　Comparison of training loss values of different methods

率启用学生网络和分类器的梯度时，可以观察到训练损失最初波动，然后稳定下降，导致快速收敛。与传统训练算法相比，该两阶段方法收敛到更低的损失，表明TC-Face允许从训练集中更彻底地学习。这不仅导致结果优于使用传统方法从头开始训练的性能，还降低了整体训练成本。

2.4　消融研究

2.4.1　第1阶段消融

在第1阶段训练上进行了一系列消融研究，并在IJB-C数据集（在TAR@FAR = 1E−4上报告，下同）上评估了该方法的准确性，如表3所示。首先，研究了不同知识蒸馏策略对模型准确性的影响。实验使用了仅MSE损失、CKD和CKD+。结果表明，与仅使用MSE损失相比，CKD将准确率提高了1.52%，而CKD+进而将准确率提高了1.78%。接下来，探索对比损失中队列长度的影响。较长的队列允许学生网络从批量中学习更多的目标，但也增加了计算成本。当队列长度设置为10 000时，准确率甚至下降了21.42%。实验结果表明，当队列长度设置为1 000时效果最好，拥有最高的准确率。最后，微调教师特征库的动量更新参数λ。较小的λ导致教师特征库更新过快，使得学生网络在第1阶段过多地自我探索特征空间，而不是从教师网络中学习知识。实验表明，将λ设置为0.999可获得最高准确性。

表3　消融实验结果（准确率）

Table 3　Result of ablation experiment (accurary)

类别	方法	IJB-C/%	
		1E−4	1E−5
	ResNet50(Teacher)	96.05	93.96
蒸馏方法	MFN + Vanilla KD(式(2))	91.29	79.79
	MFN + CKD(式(6))	92.81	84.24
	MFN + CKD$^+$(式(8))	93.07	88.33
特征库大小	MFN + CKD$^+$, k = 10 000	71.74	49.51
	MFN + CKD$^+$, k = 512(N)	92.81	85.93
	MFN + CKD$^+$, k = 1 000	93.07	88.33
更新比率	MFN + CKD$^+$, k = 1 000, λ = 0.9	73.64	57.00
	MFN + CKD$^+$, k = 1 000, λ = 0.99	92.48	85.16
	MFN + CKD$^+$, k = 1 000, λ = 0.999	93.45	88.51

2.4.2　第2阶段消融

对第2阶段训练进行了消融研究，如表4所示。结果表明，包含第2阶段训练对模型准确性有显著影响。在两种不同训练条件下评估了第2阶段的影响：使用MS1MV2和MS1MV3作为训练数据集，报告了IJB-C上FAR为1E−4时的人脸验证TAR。结果表明，在MS1MV2训练设置下，第2阶段训练将准确

性提高了约4.6%，而在MS1MV3下，它带来了1.1%的改进。这种增强使该方法能够超越其他基于KD的人脸识别训练方法。此外，观察到第1阶段训练后获得的特征幅度往往较低。这是因为第1阶段主要关注确保学生网络紧密模仿教师网络的特征。相比之下，第2阶段训练鼓励学生网络探索其自身特征空间，而不仅仅是模仿教师的特征。因此，特征幅度增加，进一步提高了准确性。同时，还探索了将阶段1和阶段2的损失结合到单一联合优化中进行训练。如图5所示，与单独使用ArcFace相比，联合损失收敛较慢，并达到比两阶段训练策略更高的最终损失值。

表4　不同实验设置和阶段的特征模长及IJB-C准确率

Table 4　Feature norm and IJB-C accuracy under different experimental settings and stages

实验	阶段	特征模长	IJB-C/%
IR50-MBF（MS1MV2）	第1阶段	1.37	85.79
	第2阶段	4.60	89.70
IR100-MBF（MS1MV3）	第1阶段	0.75	91.24
	第2阶段	4.39	92.17

2.5　与SOTA方法的比较

将第1阶段使用到式(6)方法称为TC-Face，使用到式(8)方法称为TC-Face+。本文报告了TC-Face和TC-Face+的离线性能，具体实现细节在2.2节进行了介绍。在表5和表6中，与通用KD方法(Peng等，2019；Park等，2019)和FR特定KD(Huang等，2022；Li等，2023；Boutros等，2024；Mishra和Uikey，2025)方法进行了比较。另外针对非知识蒸馏方法训练的模型进行了比较，分别是ArcFace(表中第2行)(Deng等，2019)和AdaFace(Kim等，2022)。IR50-MBF表中的实验数据引用Li等人(2023)和Boutros等人(2024)，而IR100-MBF表中的实验数据全部团队重新复现。

在两种实验设置中，TC-Face在大多数数据集始终取得最高性能，通常获得第1名或第2名。在LFW、CFP-FP、AgeDB-30、CA-LFW和CPLFW等小规模验证基准上，本文展示了在这些数据集上准确率的平均值(Avg)，TC-Face在这些数据集上领先。在IR100-MBF和IR50-MBF的设置下，TC-Face+在比

表5　IR100-MBF在测试集上的分数

Table 5　IR100-MBF score on the test set

/%

方法	平均值	IJB-C		IJB-B		MegaFace			
		1E−4	1E−5	1E−4	1E−5	Id	Ver	Id(R)	Ver(R)
IR100(teacher)	97.48	98.73	98.98	93.45	88.65	81.03	96.98	98.14	98.34
MBF(student)	94.19	89.36	81.92	87.28	74.88	75.74	91.02	91.15	92.95
FitNet(Romero等,2014)	95.81	93.58	89.85	<u>92.69</u>	86.58	79.16	92.24	95.26	95.81
KD(Hinton等,2015)	95.38	93.66	90.24	91.60	84.50	77.85	92.57	93.19	94.23
CCKD(Peng等,2019)	95.04	93.42	89.98	91.32	85.02	77.53	92.26	92.79	93.83
RKD(Park等,2019)	95.35	93.26	89.40	90.98	84.62	77.97	93.40	93.66	95.02
ShrinkTeaNet(Duong等,2019)	95.18	93.28	89.59	91.10	84.64	77.84	92.33	93.12	93.95
EKD(Huang等,2022)	95.05	93.23	89.50	91.05	83.50	77.82	92.40	93.42	93.01
TH-KD(Ben-Baruch等,2022)	95.33	93.38	88.81	92.46	86.04	78.94	94.14	95.07	95.27
AdaFace(Kim等,2022)	95.50	93.82	89.68	91.94	86.49	78.39	92.62	94.65	95.12
AdaDistill(Boutros等,2024)	95.33	93.69	90.43	91.56	85.91	77.03	91.79	92.50	93.50
UnifiedKD(Mishra和Uikey,2025)	95.48	93.87	90.28	91.64	86.07	77.34	91.92	92.98	94.12
TC-Face(本文)	<u>95.88</u>	<u>94.63</u>	<u>91.57</u>	92.61	<u>87.48</u>	<u>79.37</u>	<u>94.65</u>	<u>95.80</u>	<u>96.14</u>
TC-Face+(本文)	**96.17**	**94.75**	**92.17**	**92.93**	**87.96**	**79.53**	**94.99**	**96.22**	**96.50**

注：加粗、下划线字体表示各方法中的最优、次优结果。

表6　IR50-MBF在测试集上的分数

Table 6　IR50-MBF score on the test set

/%

方法	平均值	IJB-C		IJB-B		MegaFace			
		1E−4	1E−5	1E−4	1E−5	Id	Ver	Id(R)	Ver(R)
IR50(teacher)	96.78	96.05	93.96	93.45	88.65	80.62	96.83	98.14	98.34
MBF(student)	94.01	89.13	81.65	87.07	74.63	75.52	90.80	90.91	92.71
FitNet(Romero等,2014)	94.07	87.76	73.71	86.35	70.19	75.81	90.07	91.16	92.34
KD(Hinton等,2015)	94.00	88.37	80.39	86.08	74.30	75.81	90.80	91.02	92.41
CCKD(Peng等,2019)	94.18	87.99	78.75	86.53	72.38	75.73	90.63	91.17	92.05
RKD(Park等,2019)	94.44	89.65	83.21	87.25	73.21	75.75	91.21	91.44	93.08
ShrinkTeaNet(Duong等,2019)	94.14	87.80	79.78	88.53	75.23	75.55	90.56	90.73	92.03
EKD(Huang等,2022)	94.39	90.48	84.00	88.35	76.60	75.54	91.02	91.26	93.08
SH-KD(Ben-Baruch等,2022)	–	91.75	85.76	–	–	–	–	92.51	93.93
ReFO(Li等,2023)	–	91.26	85.56	–	–	–	–	92.38	93.02
ReFO+(Li等,2023)	–	92.41	87.02	–	–	–	–	92.41	93.75
AdaFace(Kim等,2022)	95.24	93.53	89.33	91.85	83.89	78.17	92.28	94.53	95.07
AdaDistill(Boutros等,2024)	95.43	93.27	89.32	91.21	84.13	76.39	91.54	92.12	93.32
UnifiedKD(Mishra和Uikey,2025)	95.52	92.96	89.21	91.44	83.16	75.56	92.34	92.72	94.58
TC-Face(本文)	**96.78**	93.78	89.70	91.72	83.43	78.61	94.73	95.69	96.31
TC-Face+(本文)	94.01	**93.93**	**89.74**	**92.01**	**83.95**	**79.28**	**94.80**	**95.87**	**96.52**

注:加粗、下划线字体表示各方法中的最优、次优结果,“–”表示引文中无数据。

UnifiedKD最高的基准上平均分别提高了0.24%和0.68。所提出的方法在IJB-C和IJB-B等大规模数据集上也表现出色。具体而言,在IR100-MBF设置下,TC-Face+在IJB-B上TAR@FAR=1E−5时获得1.89%的增益。在更具挑战性的MegaFace基准上,所提出的方法在所有评估指标上始终取得显著改进。TC-Face在平均上超越先前SOTA方法0.88%和1.23%,而TC-Face+进一步将性能增益推向1.18%和1.55%。

3　结　论

本文提出了一种面向轻量级人脸识别的两阶段对比知识蒸馏框架TC-Face。该方法首先通过动量更新的对比蒸馏,引导学生网络稳定学习教师模型的特征关系结构;随后,利用教师分类器权重初始化学生分类器,并结合ArcFace进行优化。实验结果表明,TC-Face显著提升了轻量化模型(如MBF)的性能。在IJB-B和IJB-C数据集上,当FAR = 1E−5时,TAR较当前最优方法UnifiedKD提升了1.89%,所有数据集上平均超过最优方法1.55%。消融实验验证了本文有效性。由于计算资源的限制,本文方法尚未在更多的轻量级网络上进行验证。后续研究方向将会在更多的轻量化网络上验证该方法的有效性。

参考文献(References)

Alansari M, Hay O A, Javed S, Shoufan A, Zweiri Y and Werghi N. 2023. GhostFaceNets: lightweight face recognition model from cheap operations. IEEE Access, 11: 35429-35446 [DOI: 10.1109/ACCESS.2023.3266068]

Ben-Baruch E, Karklinsky M, Biton Y, Ben-Cohen A, Lawen H and Zamir N. 2022. It's all in the head: representation knowledge distillation through classifier sharing [EB/OL]. [2022-04-05]. https://arxiv.org/pdf/2201.06945.pdf

Boutros F, Štruc V and Damer N. 2024. AdaDistill: adaptive knowledge distillation for deep face recognition//Proceedings of the 18th Euro-

pean Conference on Computer Vision. Milan, Italy: Springer: 163-182 [DOI: 10.1007/978-3-031-73001-6_10]

Chen S, Liu Y, Gao X and Han Z. 2018. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices// Proceedings of the 13th Chinese Conference on Biometric Recognition. Urumqi, China: Springer: 428-438 [DOI: 10.1007/978-3-319-97909-0_46]

Deng J K, Guo J, Xue N N and Zafeiriou S. 2019. ArcFace: additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4685-4694 [DOI: 10.1109/CVPR.2019.00482]

Duong C N, Luu K, Quach K G and Le N. 2019. ShrinkTeaNet: million-scale lightweight face recognition via shrinking teacher-student networks [EB/OL]. [2019-05-25]. https://arxiv.org/pdf/1905.10620.pdf

Feng Y S, Wang H, Hu H R, Yu L, Wang W and Wang S Y. 2020. Triplet distillation for deep face recognition//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, UAE: IEEE: 808-812 [DOI: 10.1109/ICIP40778.2020.9190651]

George A, Ecabert C, Shahreza H O, Kotwal K and Marcel S. 2024. EdgeFace: efficient face recognition model for edge devices. IEEE Transactions on Biometrics, Behavior, and Identity Science, 6(2): 158-168 [DOI: 10.1109/TBIOM.2024.3352164]

Guo Y D, Zhang L, Hu Y X, He X D and Gao J F. 2016. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 87-102 [DOI: 10.1109/FG.2019.8756593]

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2015-03-09]. https://arxiv.org/pdf/1503.02531.pdf

Huang Y G, Wang Y H, Tai Y, Liu X M, Shen P C, Li S X, et al. 2020. CurricularFace: adaptive curriculum learning loss for deep face recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5900-5909 [DOI: 10.1109/CVPR42600.2020.00594]

Huang Y G, Wu J X, Xu X K and Ding S H. 2022. Evaluation-oriented knowledge distillation for deep face recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18719-18728 [DOI: 10.1109/CVPR52688.2022.01818]

Kim M, Jain A K and Liu X M. 2022. AdaFace: quality adaptive margin for face recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18729-18738 [DOI: 10.1109/CVPR52688.2022.01819]

Kim Y, Park J, Jang Y, Ali M, Oh T H and Bae S H. 2021. Distilling global and local logits with densely connected relations//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6270-6280 [DOI: 10.1109/ICCV48922.2021.00623]

Li J Z, Guo Z D, Li H, Han S, Baek J W, Yang M, et al. 2023. Rethinking feature-based knowledge distillation for face recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20156-20165 [DOI: 10.1109/CVPR52729.2023.01930]

Mishra D and Uikey R. 2025. Unified knowledge distillation framework: fine-grained alignment and geometric relationship preservation for deep face recognition [EB/OL]. [2025-08-15]. https://arxiv.org/pdf/2508.11376.pdf

Park W, Kim D, Lu Y and Cho M. 2019. Relational knowledge distillation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3962-3971 [DOI: 10.1109/CVPR.2019.00409]

Peng B Y, Jin X, Li D D, Zhou S F, Wu Y C, Liu Y H, et al. 2019. Correlation congruence for knowledge distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5006-5015 [DOI: 10.1109/ICCV.2019.00511]

Romero A, Ballas N, Kahou S E, Chassang A, Gatta C and Bengio Y. FitNets: hints for thin deep nets [EB/OL]. [2014-12-19]. https://arxiv.org/pdf/1412.6550.pdf

Rousseeuw P J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20: 53-65 [DOI: 10.1016/0377-0427(87)90125-7]

Sang G L, Xiao S D and Zhao Q J. 2023. Soft threshold denoising and video data fusion-relevant low-quality 3D face recognition. Journal of Image and Graphics, 28(5): 1434-1444 (桑高丽, 肖述笛, 赵启军. 2023. 联合软阈值去噪和视频数据融合的低质量3维人脸识别. 中国图象图形学报, 28(5): 1434-1444) [DOI: 10.11834/jig.220695]

Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 815-823 [DOI: 10.1109/CVPR.2015.7298682]

Van Der Maaten L and Hinton G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(86): 2579-2605

Wang H, Wang Y T, Zhou Z, Ji X, Gong D H, Zhou J C, et al. 2018. CosFace: large margin cosine loss for deep face recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5265-5274 [DOI: 10.1109/CVPR.2018.00552]

Xie Z H, Li Y and Niu J Y. 2021. Hyperspectral face recognition based on partition bands optimal selection and deep features. Journal of Image and Graphics, 26(12): 2870-2878 (谢志华, 李毅, 牛杰一. 2021. 联合分块谱带优选和深度特征的高光谱人脸识别. 中国图象图形学报, 26(12): 2870-2878) [DOI: 10.11834/jig.200158]

Xing E P, Ng A Y, Jordan M I and Russell S. 2002. Distance metric learning, with application to clustering with side-information//Proceedings of the 16th International Conference on Neural Information Processing Systems. [s.l.]: MIT Press: #2968683 [DOI: 10.5555/2968618.2968683]

Yu Z P, Liu J H, Qin H Y, Wu Y C, Hu K, Tian J Y, et al. 2023. ICD-face: intra-class compactness distillation for face recognition//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20985-20995 [DOI: 10.1109/ICCV51070.2023.01924]

Zhao B R, Cui Q, Song R J, Qiu Y Y and Liang J J. 2022. Decoupled knowledge distillation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11943-11952 [DOI: 10.1109/CVPR52688.2022.01165]

Zhu Z, Huang G, Deng J K, Ye Y, Huang J J, Chen X Z, et al. 2021. WebFace260M: a benchmark unveiling the power of million-scale deep face recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10487-10497 [DOI: 10.1109/CVPR46437.2021.01035]

作者简介

唐鑫,男,硕士研究生,主要研究方向为生物特征识别和深度学习。E-mail:6233111055@stu.jiangnan.edu.cn

李光辉,通信作者,男,教授,主要研究方向为物联网、边缘计算、无损检测和集成电路设计验证。

E-mail: ghli@jiangnan.edu.cn

张飞飞,男,高级工程师,主要研究方向为图像处理算法的硬件加速和SoC芯片设计。E-mail:zhangfeifei@brmicro.com.cn

马宇轩,男,硕士研究生,主要研究方向为生物特征识别和深度学习。E-mail: 6233112034@stu.jiangnan.edu.cn

董正阳,男,硕士研究生,主要研究方向为生物特征识别、人脸表情识别和深度学习。

E-mail: 6233112011@stu.jiangnan.edu.cn