计算机科学
COMPUTER SCIENCE

基于无标签知识蒸馏的人脸识别模型的压缩算法

程祥鸣，邓春华

引用本文

程祥鸣，邓春华. 基于无标签知识蒸馏的人脸识别模型的压缩算法[J]. 计算机科学, 2022, 49(6): 245-253.

CHENG Xiang-ming, DENG Chun-hua. Compression Algorithm of Face Recognition Model Based on Unlabeled Knowledge Distillation[J]. Computer Science, 2022, 49(6): 245-253.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400023

基于无标签知识蒸馏的人脸识别模型的压缩算法

程祥鸣　邓春华
武汉科技大学计算机科学与技术学院　武汉 430065
武汉科技大学大数据科学与工程研究院　武汉 430065
武汉科技大学智能信息处理与实时工业系统湖北省重点实验室　武汉 430065
(781326019@qq.com)

摘　要　将人脸识别技术移植到移动设备上时,往往需要经过模型压缩等加速算法的处理。知识蒸馏是一种实际应用较广且易于训练的模型压缩方法,现有的知识蒸馏算法需要大量带标签的人脸数据,可能会涉及身份隐私泄露等安全问题。同时,大规模采集有标签人脸数据的成本较大,而海量可采集或生成的无标签人脸数据却无法利用。为解决上述问题,通过分析知识蒸馏在人脸识别任务中的特性,提出了一种无标签知识蒸馏的间接监督训练方法。该方法可以利用海量无标签的人脸数据,避免了隐私泄露等安全隐患问题。然而,无标签人脸数据集的数据分布无法预知,存在数据分布不均衡的问题,阻碍了间接监督算法的性能提升。文中进一步提出了一种人脸内容置换的数据增强方法,通过置换人脸部分内容来平衡人脸数据分布,同时增强了人脸数据的多样性。实验结果表明,人脸识别模型被大幅度压缩时,所提算法的性能达到了先进水平,并在 LFW 数据集上超越了大型网络。

关键词:人脸识别;知识蒸馏;模型压缩;间接监督;内容置换
中图法分类号　TP391

Compression Algorithm of Face Recognition Model Based on Unlabeled Knowledge Distillation

CHENG Xiang-ming and DENG Chun-hua
College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
Institute of Big Data Science and Engineering, Wuhan University of Science and Technology, Wuhan 430065, China
Hubei key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China

Abstract　When transplanting face recognition technology to mobile devices, it often needs to be processed by accelerated algorithms such as model compression. Knowledge distillation is a model compression method that has a wide range of practical applications and is easy to train. Existing knowledge distillation algorithms require a large amount of tagged face data, which may involve security issues such as identity privacy leakage. At the same time, the cost of large-scale collection of tagged face data is relatively high, while the massive amount of unlabeled face data that can be collected or generated cannot be used. In order to solve the above problems, this paper analyzes the characteristics of knowledge distillation in face recognition tasks, and proposes an indirect supervised training method of unlabeled knowledge distillation. This method can utilize massive amounts of unlabeled face data, thereby avoiding security risks such as privacy leakage. However, the data distribution of the unlabeled face data set is unpredictable, and there is the problem of uneven data distribution, which limits the performance of the indirect supervision algorithm. This research further proposes a data enhancement method for face content replacement, which balances the distribution of face data by replacing part of the content of the face, and at the same time enhances the diversity of face data. Sufficient experimental results show that when the face recognition model is greatly compressed, the performance of the algorithm in this research reaches an advanced level, and surpasses the large-scale network on the LFW data set.

Keywords　Face recognition, Knowledge distillation, Model compression, Indirect supervision, Content replacement

到稿日期:2021-03-31　返修日期:2021-09-04
基金项目:国家自然科学基金(61806150)
This work was supported by the National Natural Science Foundation of China(61806150).
通信作者:邓春华(dchzx@wust.edu.cn)

1 引言

近年来，计算机视觉领域的研究取得了较大的成果，其中深度学习方面的研究做了重要的贡献，这主要归功于3个方面的因素[1]：1)模型的设计越来越复杂；2)计算硬件资源越来越强；3)出现了大量标注的公共数据集。前两个因素，即模型复杂度和参数量增大的有效性已被证实，但在实际应用中大量计算资源有限的设备无法享受深度学习的发展带来的福利。因此，基于深度学习的轻量级网络模型成为了一个重要的研究方向。

为了实现网络模型的轻量化，一些研究者通过重新设计网络结构、参数和损失函数等，提出了较多轻量化网络模型，如SqueezeNet[2]，MobileNet[3]和ShuffleNet[4]等。同时，较多学者采用大型网络模型辅助训练小型网络模型的思想，来达到模型压缩、计算资源节能和速度提升等目的。现有的模型压缩主要从模型剪枝、权重共享、量化、核稀疏、二值化、Low-rank分解和知识蒸馏[5]等方面入手。其中知识蒸馏是一种比较容易训练的模型压缩方法，在工业界备受青睐。网络模型的知识蒸馏源于Hinton等提出的知识蒸馏方法[5]，该方法利用大型教师网络辅助小型学生网络进行联合训练，而学生网络结合教师网络输出的概率分布来进行学习[6]。相比人工标注的真实标签数据，这种概率分布可以提供更加丰富的监督信息。

对于第三个因素，文献[1]中的大量实验表明，任务复杂度与训练数据的数量级呈线性增长关系。深度学习使用反向传播算法来学习并更新模型内部表达上一层信息的参数，以从大量数据中发掘复杂的结构，其网络模型的参数优化来源于训练数据集的概率化信息反馈[7]。模型需要学习的参数越多，训练时需要的数据就越多。然而，在人脸识别领域中，大规模数据集的隐私安全已经不是一个新鲜的问题。例如，IBM于2019年1月发布的百万级别「人脸多样性」数据集引发了广泛的争议；微软官方也于2019年9月宣布删除3年前发布的世界上最大的公开人脸识别数据集——MS Celeb名人人脸数据集[8]。数据集的获取和使用不仅局限于技术层面，其包含的敏感数据信息也成为了一个社会关注的焦点。与此同时，网络上存在大量的无标签、不携带敏感信息的人脸图片，采用这些无标签图片来训练深度学习网络模型，既可以满足庞大的数量需求，又可以避免隐私泄露等安全性问题。

标签数据的数量很大程度上影响着模型的最终性能，如何训练无标签的数据集成为了一个亟待解决的科学问题。对于上述问题，一些学者从两个方面进行了研究：无监督人脸识别[9]和半监督人脸识别[10]。在无监督学习中，模型通过浅层特征对无标签数据进行聚类来完成整个学习过程[11]；在半监督学习中，模型通过建立无标签数据与学习任务之间的关联来进行训练。然而，受到具有强监督信号的数据标签的影响，无监督学习以及半监督学习与监督学习的性能差距较大[12]。相比无监督学习的浅层特征，深层特征更具有鲁棒性[13]。对于半监督学习来说，如何改进无标签数据和学习任务之间的有效关联成为一个具有挑战的问题。

基于上述问题和需求，本文研究了一种无标签知识蒸馏的间接监督训练方法。在训练过程中，该方法将学生网络拟合教师网络的输出作为间接监督信号的深层特征，这种间接监督方法比一般的监督学习和半监督学习更具优势。间接监督不仅可以避开监督学习的数据集标签需求，对无标签数据进行充分的利用，也可以在训练阶段避开半监督学习中需要建立的关联模型，使得监督信号更加鲁棒。然而，由于无标签数据分布未知，可能存在不均衡的问题，这限制了间接监督的性能发挥，因此，本文提出了一种人脸内容置换的数据增强方法，用于联合训练教师和学生网络模型。该方法能均衡训练数据的分布，进一步压缩学生网络模型的体积并提升其性能，达到使用更丰富的数据和压缩更小的模型的目的。

2 相关工作

2.1 人脸识别系统

经过数十年的发展，人脸识别系统的主流技术主要可分为两大类[13]：基于浅层表征的人脸识别方法[14]和基于深度学习的人脸识别方法[15-16]。

相比深层特征，浅层表征很难对人脸识别的许多规律或规则进行显性的描述[13]。基于深度学习的人脸识别方法通过深层网络提取和表示人脸的特征[17]，并以此作为人脸识别的度量。它的识别过程可以描述为：

$$D[F(\boldsymbol{I}_i),F(\boldsymbol{I}_j)] \tag{1}$$

其中，$\boldsymbol{I}_i$和$\boldsymbol{I}_j$表示两张人脸图像，F表示人脸特征提取和编码表示过程，D表示提取完的特征之间的相似度量。

随着深度学习的发展，较多优秀的基于深度学习的大型人脸识别模型逐步被提出。但这些大型的模型拥有动辄千万甚至上亿个参数，其需要的计算资源非常庞大，在越来越普及的嵌入式和移动设备中部署这些网络是一个很大的挑战。因此，轻量化的人脸识别模型逐渐开始被重视。为了达成模型轻量化的目标，研究者们探索了两个方向。一个是重新对模型框架进行设计，其中较为成功的网络有Mobilefacenet，该方法仅有120万个参数，但在LFW上同样达到了99.51%的识别精度；另一个方向则是对现有网络进行模型压缩的处理，其中的知识蒸馏是一种易于训练并且简洁的压缩方法，也是本文研究的重点。

2.2 知识蒸馏方法

近年来，模型压缩加速技术逐渐被研究者重视。其中，知识蒸馏可以有效地从大型“教师”模型中提取知识来训练小型的“学生”模型，让学生网络获得更好的性能。

在典型的图像分类任务中，知识蒸馏方法将网络最后一层的输出当作教师模型提供的知识，而非知识蒸馏方法的训练样本并不提供这种知识，这种知识被称为软目标(soft targets)，与之相对，非知识蒸馏训练方法中的one-hot标签被称为硬目标(hard targets)。以手写数字识别为例，与数字3相比，数字2的形状更类似于数字7，因此其作为数字7的软目标的概率比作为数字3的概率更高，如图1所示。

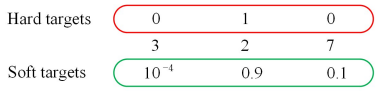


图 1 手写数字识别[5]

Fig. 1 Handwritten digit recognition[5]

在 Hinton[5]等对知识蒸馏进行系统诠释的基础上，较多优秀的知识蒸馏方法被提出。同时，基于数据隐私性、对抗学习和跨模态学习等方面的考虑，无标签的知识蒸馏开始引起学者们的关注[18-19]。而当无标签知识蒸馏方面的研究具体到人脸识别任务上时，人脸数据的隐私性成为更需要被关注的方面。在一般的无标签知识蒸馏工作中，研究者需要对蒸馏过程中的伪标签构成进行探索[19]，而在人脸识别任务中，存在一个天然可以作为伪标签的监督信号，即人脸表征[13]。这种监督信号不同于依赖部分手工先验知识产生的伪标签，其可以简洁地运用在无标签人脸知识蒸馏的过程中。基于该考虑，本文在无标签人脸知识蒸馏方向上进行探索。

3 主要工作

本文针对无标签非敏感数据的模型训练问题，提出了一种无标签人脸知识蒸馏的间接监督训练方法。在此训练过程中，首先由教师网络输出人脸表征信息，接着由学生网络将该表征信息当成间接监督信号进行模型训练，以此完成训练过程。该训练过程中的学生网络无需额外标签信息的辅助，该方法既可以解决一般半监督学习中的伪标签生成不准确的问题，又避免了人脸数据集的安全隐私问题。对于无标签数据使用中出现的数据量的规模问题和数据分布不均衡问题，本文通过人脸内容置换的数据增强方法 FaceMix 来增强数据，将无标签人脸图像进行置换以产生新的样本数据，用新的数据和新的分布来解决该问题。对于产生的新样本的监督问题，在无标签的间接监督中，我们直接采用教师网络提供的监督信号进行间接监督，以此完成对无标签的人脸数据集的扩充过程和无标签知识蒸馏的间接监督训练全过程。

3.1 无标签人脸知识蒸馏的间接监督训练方法

本文提出的无标签人脸知识蒸馏的间接监督方法脱离了原始的知识蒸馏方法，是基于响应的知识蒸馏方法，并采用离线蒸馏的训练策略来完成整个训练过程。相比一般的目标识别任务，人脸识别任务中的类别显然更多，原始的知识蒸馏方法并不能很好地适用于人脸识别任务，因为这会使得 soft targets 的维度过大，从而导致该方法在人脸识别上的收敛性能差[20]。由于人脸识别任务的特性，知识蒸馏在人脸识别任务上有了另一种解法。在人脸识别任务中，我们将人脸图像输入神经网络中，经过神经网络的处理后，可以得到一个特定维度的特征向量，该向量可以很好地表征人脸数据，使不同人脸的两个特征向量距离尽可能大，同一张人脸的两个特征向量距离尽可能小，这样就可以通过特征向量来进行人脸识别，该特征向量即人脸表征。在人脸识别的知识蒸馏任务中，将人脸表征当成 soft targets 不仅可以使学生网络学习到教师网络传递过来的知识，而且解决了人脸识别带来的 soft targets 维度过大问题。

图 2 给出了本文用作知识蒸馏中骨架网络的 SE-ResNet 模块[21]，本文使用的教师网络层数有 50 层，学生网络的网络结构与教师网络相同，层数减少为 18 层。

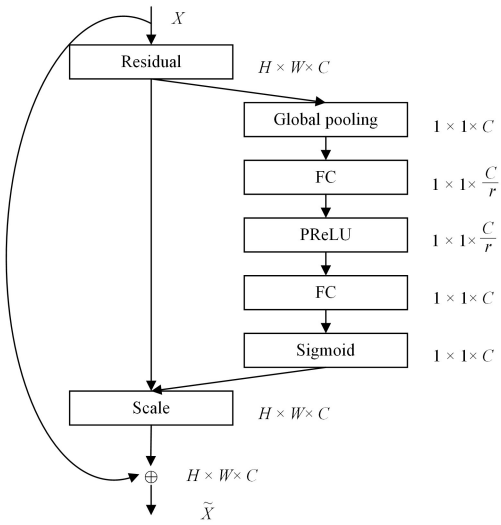


图 2 SE-ResNet 模块[21]

Fig. 2 SE-ResNet module[21]

在知识蒸馏过程中，将输入的数据同时提供给教师网络和学生网络，教师网络和学生网络分别输出维度为 512 维的人脸表征 $\boldsymbol{F}_t$ 和 $\boldsymbol{F}_s$。接着分为两个 loss 进行计算损失，第一个 loss 使用 KL 散度 loss 作为距离度量标准，输入为两个人脸表征 $\boldsymbol{F}_t$ 和 $\boldsymbol{F}_s$，如式(2)所示：

$$KLDloss=\sum_{i\in I}\boldsymbol{F}_{s_i}\log\left(\frac{\boldsymbol{F}_{s_i}}{\boldsymbol{F}_{t_i}}\right) \tag{2}$$

第二个 loss 使用 Arcface[22]中使用的 loss，即 additive angular margin loss。损失函数的输入为学生网络输出的人脸表征 $\boldsymbol{F}_s$ 和数据集提供的真实标签信息，这里的真实标签属于 Hard targets，该损失函数如式(3)[22]所示：

$$Arcloss=-\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j-1,j\neq y_i}^{n}e^{s\cos\theta_j}}$$

$$\boldsymbol{W}_{y_i}=\frac{\boldsymbol{W}_{y_i}}{\|\boldsymbol{W}_{y_i}\|},\boldsymbol{F}_{s_i}=\frac{\boldsymbol{F}_{s_i}}{\|\boldsymbol{F}_{s_i}\|},\cos\theta_{y_i}=\boldsymbol{W}_{y_i}^{\mathrm{T}}*\boldsymbol{F}_{s_i} \tag{3}$$

其中，$\boldsymbol{W}$ 为网络输出层的参数。最后通过设置一个超参数 λ 来控制这两个 loss 在总 loss 中的影响占比，λ 为一个取值范围为(0,1)的实数。由此，知识蒸馏最终的损失函数如式(4)所示：

$$Total_{\mathrm{loss}}=(1-\lambda)*Arcloss+\lambda*KLDloss,\lambda\sim(0,1) \tag{4}$$

在训练过程中，λ 的值越接近 1，学生网络得到的 soft targets 的监督越多，学生网络能学习到的教师网络的知识也就越多，训练模型的精度也能更快到达最优精度附近。一般情况下，在训练的前期，设置较大的 λ 值以快速学习教师网络的知识；在训练的后期，减小 λ 的值，使用 hard targets 帮助网络鉴别困难样本，以此获得较好的结果。

上述有数据标签监督的基于知识蒸馏的人脸识别训练方法中，学生网络可以通过知识蒸馏学习教师网络传递过来的知识。但由于人脸数据集的隐私性和安全性问题已经跳出学术界，成为了社会共同关注的问题，人脸数据集的采集所引发的社会关于人权隐私的讨论愈发激烈，这或增大新数据集的获取难度，产出的新公共数据集的数据量或在一定程度上

无法满足训练所需要的大量数据的要求。与此同时，大量的研究者采用众多不同的网络模块、网络层数、损失函数和训练策略，在使用现存的大型人脸识别数据集的情况下，训练出了许多高精度的预训练模型。在知识蒸馏的概念被提出后，这些预训练模型的强大潜力被挖掘出来，它们可以作为指导更小型网络进行训练的一种监督，并且这些模型可以轻松地被获取，不同于人脸数据集的获取会牵扯到安全隐私问题。

如前文所述，在人脸识别领域有一个人脸表征的概念，它可以用于表示一张人脸的图像，同一个人的不同人脸图像提取出来的表征极为相似，与之相反，不同人脸图像的表征有明显的差异。如图 3 所示，数字 0 和 1 代表不同的人脸表征，可以看出，不同的人脸表征可以被明显区分开。这代表在训练过程中，这种教师网络生成的具有差异的人脸表征可以被当成一种监督信号来使用。

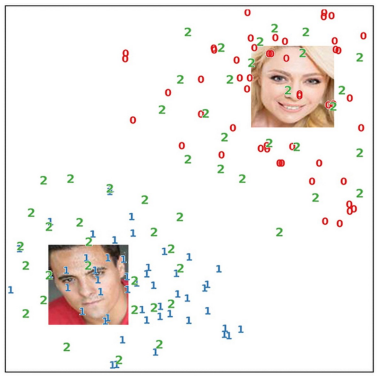


图 3 不同人脸的人脸表征分布 t-SNE 图

Fig. 3 t-SNE diagram of the distribution of face representations of different faces

在此基础上，使用教师网络产生的知识来进行训练是可行的。在仅使用教师网络输出的人脸表征时，模型的训练不仅可以避开监督学习的标签需求，对无标签数据进行充分的利用，还能充分发挥高精度预训练模型的潜能。在此基础上，本文对上述有数据标签监督的基于知识蒸馏的人脸识别训练流程进行了修改，在去掉了数据集提供的真实标签后进行训练，此做法即可解决在训练过程中需要数据集提供的真实标签的问题。此外，文献[23]提出，余弦公式与常用于人脸识别的相似度度量是匹配的。本文将网络的损失改为余弦损失，将教师网络和学生网络输出的人脸表征 $\boldsymbol{F}_t$ 和 $\boldsymbol{F}_s$ 输入余弦相似度损失函数进行计算，如式(5)所示：

$$Cosloss=1-\frac{\boldsymbol{F}_s \cdot \boldsymbol{F}_t}{\|\boldsymbol{F}_s\| \times \|\boldsymbol{F}_t\|} \tag{5}$$

其他部分则与之前的有数据标签监督的基于知识蒸馏的人脸识别训练的教师网络结构和学生网络结构相同。

相比现存的无标签知识蒸馏方法[18-19]，本文方法率先将无标签知识蒸馏用于人脸图像识别领域。同时，不同于现有方法通过引入新的网络结构、使用较难训练的 GAN 网络等来实现无标签知识蒸馏，本文方法直接使用人脸识别任务中存在的间接监督信号——人脸表征来完成整个过程，使得本文方法在结构简洁性和易于训练上优于现有的无标签知识蒸馏网络。

3.2 人脸内容置换的数据增强方法 FaceMix

本节中，本文将进一步在数据方面对无标签知识蒸馏方法进行探索。如文献[1]所述，任务表现与训练数据的数量级呈线性增长关系。在完成无标签网络训练的基础框架后，本文通过数据增强方法进一步改进训练效果。

在一般的训练任务中，对数据进行增强是极为常见的做法，而在采集的无标签数据集中无标签数据分布未知，这些训练数据的分布可能存在不均衡的问题，因此会对学习过程造成一定的影响，也限制了间接监督算法的性能发挥。如图 3 所示，代表两类人脸的数字 0 和 1 在特征空间中的分布并不均衡，常用的数据增强方法通过旋转、翻折等操作进行数据的增广。但这些做法并不对数据中语义结构的分布进行扩充，而更偏向于一般特征[24]，对于数据分布方面的改进效果并不明显。由于数据增强中的混合样本数据增强(Mixed Sample Data Augmentation，MSDA)[24]具有可以生成新的数据分布的特性，这一特性符合拟合增量的经典定义，使得该数据增强方法在一般训练过程中十分有效。增广后的新训练数据不仅提升了数据规模，而且使特征空间中的分布更加平滑，能更加充分地分布在数据的特征空间中，从而更好地发挥间接监督的性能。因此，本文对无标签数据和混合样本数据增强的结合进行了更进一步的探索。

在 MSDA 的一系列方法中，研究者们倾向于混合不同的图像样本产生更加自然的图像样本，并通过插值、二进制遮掩和二进制替换等方式进行改进[24-25]。但相比在其他图像领域进行混合样本数据增强，人脸识别领域具有一定的优势。完整的人脸识别任务包含 3 个部分：人脸检测、人脸对齐和人脸识别。其中，人脸对齐可以将检测到的一些角度不正的人脸图片进行矫正，使其脸部关键点对齐。对于其图片来说，在图片中的关键点的位置是相符的，这使得图片进行混合样本数据增强时，可以做到替换的图片区域都是相互对应的，使得我们使用二进制替换的图片会更加自然，如图 4 所示。

图 4 不同增强场景下的混合样本生成图

Fig. 4 Generation diagram of mixed samples in different enhancement scenarios

因此，本文使用 MTCNN[26]对齐和二进制人脸内容置换结合的 FaceMix 数据增强方法，首先将采集到的人脸图片进行对齐操作，通过 MTCNN 检测得到人脸的 5 个关键点，将图像进行仿射变换，变换矩阵如式(6)所示：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{6}$$

对齐后对图片的随机裁剪区域进行二进制替换，生成样本如式(7)所示：

$$\boldsymbol{x}_N = \boldsymbol{M} * \boldsymbol{x}_A + (1 - \boldsymbol{M}) * x_B \tag{7}$$

其中，$\boldsymbol{x}_N$ 表示生成的新样本，$\boldsymbol{x}_A$ 和 $\boldsymbol{x}_B$ 为两个不同的训练样本，$M \in \{0,1\}^{W \times H}$ 表示将图像部分区域进行裁剪和填充的二进制掩码，* 表示将图像的部分区域进行逐像素相乘，1 是所有元素都为 1 的二进制掩码。首先对需要裁剪的区域进行随机采样，以得到裁剪框 $\boldsymbol{B} = (x, y, w, h)$，其中，4 个参数如式(8)所示：

$$\begin{aligned} &x \sim Unif(0, W), w = W\sqrt{1-\lambda} \\ &y \sim Unif(0, H), h = H\sqrt{1-\lambda} \\ &\lambda \sim Beta(\alpha, \alpha) \end{aligned} \tag{8}$$

其中，x, y, w, h 分别为裁剪框的中心点的横纵坐标以及框的宽度和高度，W 和 H 分别为图片的宽度和长度，通过随机 λ 值来对裁剪框参数进行生成。

确定好裁剪框 $\boldsymbol{B}$ 后，将掩码 $\boldsymbol{M}$ 中的裁剪区域置 0，其他区域置 1。在完成掩码的采样后，将样本 A 中的剪裁区域移除，将样本 B 中的剪裁区域进行裁剪然后填充到样本 A 中，这样就完成了新图片的生成过程。

本文方法中，我们将一个 *batch* 中已经对齐的图片随机配对成 *batch*/2 对，然后对每对人脸的内容进行替换，具体流程如图 5 所示。而对于新样本的监督信号的生成，本文则从假设教师网络输出的人脸表征可以明显区分不同人脸的这一基础出发。将新生成的人脸图片 Mixface 作为区别于 Face1 与 Face2 的一张新的人脸送入教师网络进行训练，把教师网络生成的知识作为标签，并以此标签为训练学生网络的间接监督信号，利用卷积神经网络的特征提取能力和知识蒸馏的知识监督能力来更好地完成训练，以获得令人满意的结果。如图 3 所示，数字 2 表示新生成的人脸图片的表征分布，新分布区别于原有数据的表征分布，填充了原有数据分布中的空白区域，扩充了训练过程中无标签训练数据的数据规模并均衡了数据的分布情况。最后，将人脸内容置换的数据增强方法 FaceMix 与无标签的基于知识蒸馏的人脸间接监督训练方法相结合，就得到了本文设计的整个人脸识别的训练流程。本文方法首先使用原始数据集进行 FaceMix 置换增强，然后将增强的数据送入网络中进行训练并完成整个过程，其他配置均与上述无标签的基于知识蒸馏的人脸识别训练方法相同，如图 6 所示。

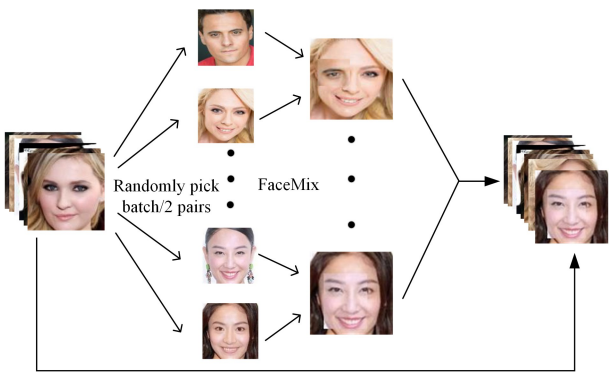


图 5 人脸内容置换增强流程图

Fig. 5 Flowchart of face content replacement and enhancement

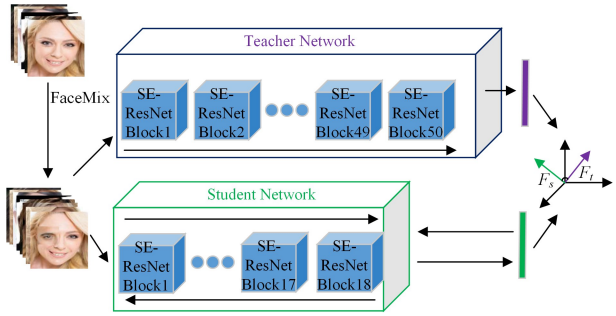


图 6 无标签数据增强的基于知识蒸馏的人脸识别训练结构图

Fig. 6 Face recognition training structure diagram based on knowledge distillation by unlabled data enhancement

4 实验结果与分析

4.1 数据集介绍

MS1M 数据集[8]由微软亚洲研究院推出。该数据集包含 10 万位名人，每位名人大概 100 张图片，合计 1 000 万张图片。MS1M_arcface 是 Insightface 团队针对 MS1M 数据集进行清理后制作的数据集，清理后的数据集包含 8.5 万位名人，共计 580 万张图片[22]。本文截取数据集 MS1M_arcface 中的前 1 万位名人，共 77 万张图片，组成数据集 MS1M_arcface10K，以此数据集来模拟中小型无标签数据集进行训练的情况。

AgeDB(Age Database 数据集包含 6 000 对，共 440 个 ID，12 240 张不同姿态、表情、年龄和性别的图片。此数据集使用年龄差为 30 的数据，命名为 Agedb_30。

LFW(Labled Faces in the Wild)人脸数据库是目前人脸识别的常用数据集，其中包含的人脸图片均来源于生活中的自然场景。该数据集共包含 13 000 多张人脸图像，主要用于研究非受限场景下的人脸识别问题。

CALFW (Cross-Age LFW)数据集是基于 LFW 数据集标注的跨越年龄的数据集，它加强了数据集的类内年龄差异性。该数据集包含5 749 个人，共计 13 233 张图片。

CPLFW (Cross-Pose LFW)数据集是基于 LFW 数据集标注的跨姿态数据集，它加强了数据集的类内姿势变化性。该数据集包含 5 749 人，共计 13 233 张图片。

CFP (Celebrities in Frontal-Profile in the Wild)数据集同时收集了人的正脸图片和侧脸图片，包含 500 人，共计 7 000 张图片。在该数据集中提供了两种评估方案，即 Frontal-Frontal (FF)人脸验证和 Frontal-Profile (FP) 人脸验证，根据两种方案的不同，该数据集分别被命名为 CFP_FF 和 CFP_FP。

4.2 实验细节

在数据集的选择上，本文使用 MS1M_arcfac，MS1M_arcface10K，AgeDB_30，LFW，CALFW，CPLFW，CFP_FF 和 CFP_FP 8 种数据集进行训练和测试。其中，图片大小均为 112×112 像素。

在软硬件系统的选择上，实验使用的训练平台为 Pytorch，所有的模型都在 Ubuntu 系统上进行训练，该系统环境拥有的硬件配置为单块 GTX-2080TI 显卡和单颗 Intel Xeon E5-2683

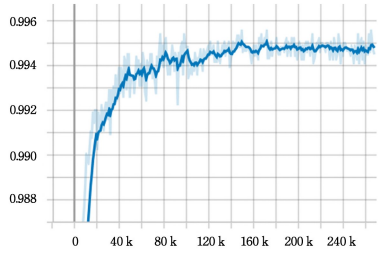
v3@2.00GHz CPU。

在参数选择上，本文将 batch size 设置为 80，初始学习率设置为 0.001，并在第 12，15 和 18 个 epoch 将学习率下降为 1/10，将动量设置为 0.9，并在网络中使用 Dropout，将 drop_ratio 设置为 0.6，一共训练 50 个 epoch，并采用 SGD 优化器进行训练。

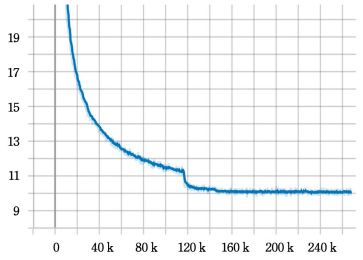
在模型训练上，本文使用 SE-ResNet100 结构和 SE-ResNet50 结构作为教师网络，使用 SE-ResNet18 结构作为学生网络。在教师网络的训练中，教师网络在 MS1M_arcface 数据集上使用 arcface 的损失函数[22]进行训练。在学生网络的一般训练中，学生网络在 MS1M_arcface10K，AgeDB_30 和 LFW 数据集上使用 arcface 的损失函数进行训练。在学生网络的知识蒸馏训练中，学生网络同样在 MS1M_arcface10K，AgeDB_30 和 LFW 数据集上使用余弦相似度 loss 进行训练，分别采用原始的知识蒸馏方法 KD、无标签的知识蒸馏方法 NoLabel_KD、结合 CutMix[25]的无标签知识蒸馏方法和结合人脸内容置换的数据增强方法 FaceMix 的无标签知识蒸馏方法进行训练。

4.3 结果分析

图 7 为使用本文方法进行训练的准确率变化曲线和训练集 loss 变化曲线。如图 7 所示，使用本文方法进行训练时，模型在 LFW 数据集上的准确率呈上升趋势，而训练集 loss 呈下降趋势。



(a)训练时模型的 accuracy 变化曲线(LFW 数据集)



(b)训练时模型的 loss 变化曲线

图 7 训练过程中的可视化曲线

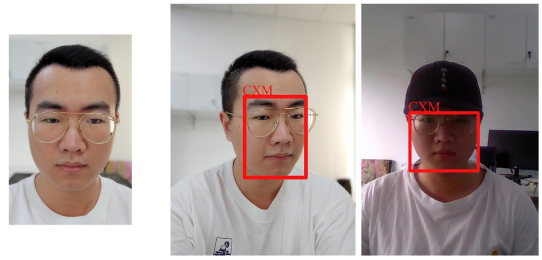
Fig. 7 Visualization curve of training process

为了验证本文方法在一般的人脸识别训练任务上的优劣性，本文选取 SE-ResNet18 作为骨架网络，并将使用 softmax 变体损失函数进行训练的经典人脸识别训练方法、本文方法以及作为教师模型的 SE-ResNet50 网络进行对比。结果如表 1 所列，数据与环境设置情况相同时，本文方法在 LFW 测试集上优于非知识蒸馏的人脸识别训练方法。使用本文方法在中小型无标签数据集 MS1M_arcface10K 下训练的 SE-ResNet18 模型的准确度甚至超过了在大型有标签数据集 MS1M_arcface 训练下作为教师网络的 SE-ResNet50 模型。将本文方法的人脸识别模型和 MTCNN[26]人脸检测模型相结合得出的人脸识别整体框架的可视化结果如图 8 所示，图 8(a)给出了人脸识别模型的注册人脸，图 8(b)给出了识别结果。

表 1 本文方法与不同人脸识别方法在 LFW 数据集上的结果对比

Table 1 Comparison between the method in this paper and different face recognition methods on the LFW data set

Method	*Accuracy*/%
W&F Norm Softmax[27] (SE-ResNet18)	97.48
Cosface[23] (SE-ResNet18)	99.16
Arcface[22] (SE-ResNet18)	99.17
Arcface[22] (SE-ResNet50)	99.52
Ours(SE-ResNet18)	**99.55**



(a)注册人脸 (b)识别结果

图 8 人脸识别可视化图

Fig. 8 Visualization of face recognition

为了验证本文方法在多个训练集上的有效性，本文分别使用 AgeDB_30 和 LFW 数据集进行训练，并使用 CALFW，CPLFW，CFP_FF 和 CFP_FP 等数据集进行结果测试。结果如表 2、表 3 所列，在使用 AgeDB_30 和 LFW 等训练集进行训练时，本文方法训练出的模型精度也优于非知识蒸馏方法训练出的模型精度。

表 2 非知识蒸馏训练方法和本文方法在各个验证集上取得的准确度

Table 2 Accuracy of the non-knowledge distillation training method and the method in this paper on each validation set

(单位：%)

Backbone	Method	CALFW	CPLFW	CFP_FF	CFP_FP
SE-ResNet18	NOKD	90.98	82.22	97.97	88.41
	Nolabel_KD				
	FaceMix(Ours)	**93.05**	**83.98**	**98.79**	**90.07**

注：训练采用 AgeDB_30 数据集

表 3 非知识蒸馏训练方法和本文方法在各个验证集上取得的准确度

Table 3 Accuracy of the non-knowledge distillation training method and the method in this paper on each validation set

(单位：%)

Backbone	Method	CALFW	CPLFW	CFP_FF	CFP_FP
SE-ResNet18	NOKD	90.02	80.52	97.13	85.16
	Nolabel_KD				
	FaceMix(Ours)	**93.12**	**84.25**	**98.76**	**88.91**

注：训练采用 LFW 数据集

同时，为了验证本文方法的各部分模块对整体的贡献度，对不同模块构成的训练方法进行了对比实验。

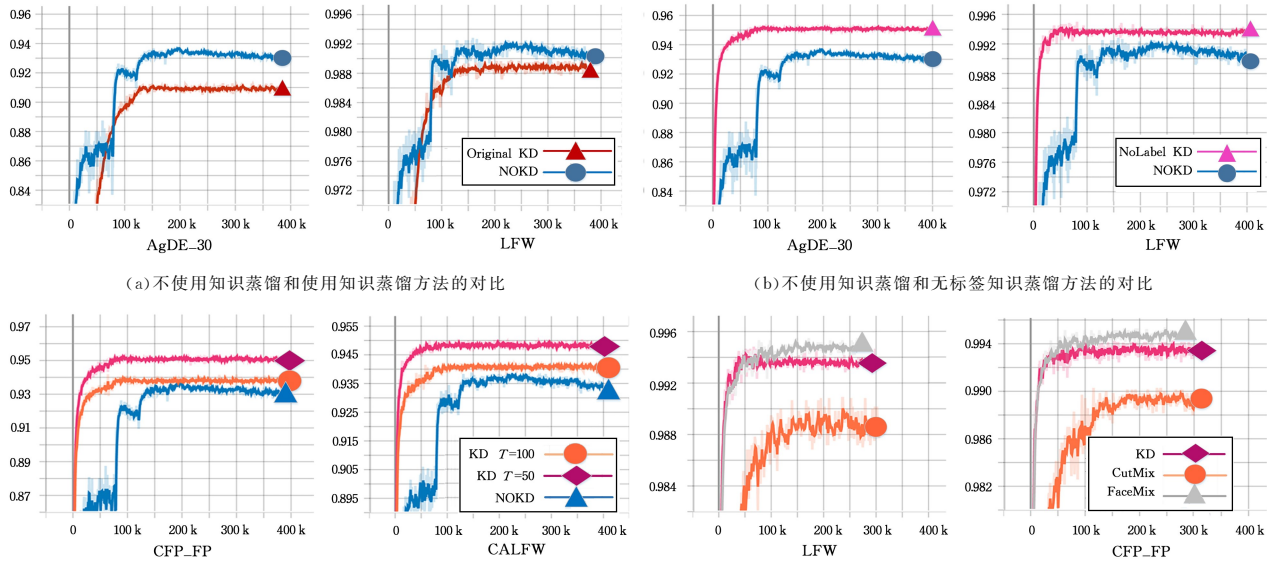
首先，将不使用知识蒸馏的方法和使用无标签知识蒸馏的

方法进行对比。如图 9(a)所示，网络的 BackBone 均为 SE-ResNet18，在知识蒸馏过程中以 SE-ResNet50 为教师网络。以 AgeDB_30 和 LFW 数据集的测试数据为例，图表的 x 轴表示训练迭代的步数，图表的 y 轴表示训练出来的模型在当前数据集下测试出的准确度，图表中以三角形标记的曲线代表使用原始知识蒸馏方法 Original KD 的训练过程，以圆形标记的曲线代表不使用知识蒸馏方法的训练过程。不同于在图像分类中使用知识蒸馏的情况，相比不使用知识蒸馏的方法，将原始的知识蒸馏方法使用到人脸识别训练中的做法并没有使性能得到提升，甚至造成了模型性能的严重下降。当数据集提供真实标签时，真实标签计算的 loss 和教师网络计算的 loss 会相互影响，这种影响是非正面的，通过知识蒸馏方法训练的模型反而较难去拟合教师网络的原本输出，以至于无法得到一个满意的结果。

其次，将不使用知识蒸馏的方法和无标签知识蒸馏方法 NoLabel_KD 进行对比。如图 9(b)所示，网络的 BackBone 均为 SE-ResNet18，在知识蒸馏过程中以 SE-ResNet50 为教师网络。以 AgeDB_30 和 LFW 数据集的测试数据为例，图表中以三角形标记的曲线代表使用无标签知识蒸馏方法 NoLabel_KD 的训练过程，以圆形标记的曲线代表不使用知识蒸馏的训练过程。在训练过程中，相比不使用知识蒸馏的方法，无标签的知识蒸馏方法有较大的提升。在监督信号仅由教师网络产生知识的前提下，使用无标签知识蒸馏的方法进行人脸训练可以使训练出的模型的准确率更高，并且能以一个较快的训练速度达到模型的最佳结果。

然后，将通过无标签知识蒸馏并使用不同网络深度的教师网络训练出的学生网络和不使用知识蒸馏方法训练出来的网络进行对比。如图 9(c)所示，不使用知识蒸馏的网络和知识蒸馏的学生网络的 BackBone 均为 SE-ResNet18，在知识蒸馏过程中以 SE-ResNet50 或 SE-ResNet100 为教师网络。以 CFP-FP 和 CALFW 数据集的测试数据为例，图表中以圆形标记的曲线代表教师网络为 SE-ResNet100 的知识蒸馏训练过程，以菱形标记的曲线代表教师网络为 SE-ResNet50 的知识蒸馏的训练过程，以三角形标记的曲线代表不使用知识蒸馏的训练过程。在训练过程中可以明显看出，虽然两种教师网络训练下的 SE-ResNet18 网络模型均超过了不使用知识蒸馏训练的网络模型，但规模更大的教师网络 SE-ResNet100 训练下的准确度却没有 SE-ResNet50 训练下的模型的准确度高。这个结果和文献[28]在图像分类领域做的探索是一致的，在人脸识别任务上，规模越大的教师网络会提供更为 Harder 的人脸表征，但这种人脸表征对于一个 Soft targets 驱动的知识蒸馏方法来说，不是一个符合蒸馏原理的选择。

最后是将无标签知识蒸馏分别与同为混合样本数据增强的 CutMix 方法和 FaceMix 方法进行对比的结果。如图 9(d)所示，训练中的学生网络的 BackBone 均为 SE-ResNet18，在知识蒸馏过程中以 SE-ResNet50 作为教师网络。以 LFW 和 CFP_FF 数据集的测试数据为例，图表中以菱形标记的曲线代表使用无标签知识蒸馏训练过程，以圆形标记的曲线代表使用 CutMix 方法的无标签知识蒸馏训练过程，以三角形标记的曲线代表使用 FaceMix 方法的无标签知识蒸馏的训练过程。可以看出，使用 CutMix 方法的无标签知识蒸馏得到的结果，相较于不使用混合样本数据增强的无标签知识蒸馏得到的结果，其模型的准确率有一定的损失。而使用 FaceMix 方法的无标签知识蒸馏方法得到的结果相较于不使用混合样本数据增强的无标签知识蒸馏得到的结果，其模型的准确率有一定的提升。这表明，在人脸识别任务上，FaceMix 方法优于同为混合样本数据增强的 CutMix 方法，该方法有利于人脸识别的知识蒸馏训练。



(a)不使用知识蒸馏和使用知识蒸馏方法的对比

(b)不使用知识蒸馏和无标签知识蒸馏方法的对比

(c)不同的教师网络对于学生网络训练的提升程度

(d)结合不同的增强方法对学生网络训练的提升程度

图 9 不同方法在训练过程中 accuracy 变化曲线的对比

Fig. 9 Comparison of accuracy change curve of different methods during training

表4列出了本文的基于无标签的人脸知识蒸馏和FaceMix数据增强方法相结合的训练方法与其他训练方法在各个数据集上的准确度。如表4所列，本文方法具有一定的优势，学生网络压缩为教师规模的55%时，相比不使用知识蒸馏方法得到的训练模型，本文方法在CALFW，CPLFW，CFP_FF和CFP_FP数据集上的准确率分别提升了1.32%，2.15%，0.25%和0.41%。而相比原始的知识蒸馏方法，本文方法的准确率分别提升了2.25%，5.36%，0.62%和6.83%。

表4 不同方法在各个验证集上取得的准确度

Table 4 Accuracy of different methods on each validation set

Backbone	# of Params/$\times 10^6$	Method	CALFW/%	CPLFW/%	CFP_FF/%	CFP_FP/%
SE-ResNet50	43.79	Teacher	95.57	91.07	99.62	95.04
SE-ResNet18	24.1$\times 10^6$	NOKD	93.75	86.88	99.28	91.23
		Original_KD	92.82	83.67	98.91	84.81
		Nolabel_KD	94.93	88.39	99.38	91.14
		Nolabel_KD CutMix	93.23	82.64	98.93	84.26
		Nolabel_KD FaceMix(Ours)	**95.07**	**89.03**	**99.53**	**91.64**

注：训练采用MS1M_arcface和MS1M_arcface10K数据集

结束语 本文在综合考虑人脸识别在边缘设备上的部署问题、人脸数据集的标签问题带来的隐私安全性问题、无标签人脸数据的使用问题和人脸识别任务特性后，以传统的知识蒸馏框架为基础，提出了一种基于无标签知识蒸馏的间接监督和人脸内容置换的数据增强方法FaceMix相结合的模型压缩训练方法。该方法使小型模型能够更好地模拟大型模型，同时有效地利用了无标签数据以及避免了人脸数据集的隐私安全问题，通过增强数据来扩充数据并均衡数据的分布情况，以达到更好的效果。本文不仅从实验的角度证明了一般混合样本数据增强方法在人脸知识蒸馏训练中的不足，也在各个数据集上验证了本文方法的有效性。同时，不仅在人脸数据集上存在隐私安全问题，其他的数据也存在相应的敏感数据问题，如医疗数据中的患者数据信息。在未来对无标签蒸馏训练的研究中，将会在无标签人脸知识蒸馏方面和数据增强方面进行更深入的探索，并将基于无标签知识蒸馏的间接监督训练的研究扩展到其他领域。

参 考 文 献

[1] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 843-852.

[2] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[J]. arXiv: 1602.07360, 2016.

[3] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[4] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[5] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503.02531, 2015.

[6] WANG R Z, GAO J, HUANG S H, et al. Malicious Code Family Detection Method Based on Knowledge Distillation[J]. Computer Science, 2021, 48(1): 280-286.

[7] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[8] GUO Y, ZHANG L, HU Y, et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition[C]// European Conference on Computer Vision. 2016: 87-102.

[9] DATTAS, SHARMA G, JAWAHAR C V. Unsupervised learning of face representations[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 135-142.

[10] YU H, FAN Y, CHEN K, et al. Unknownidentity rejection loss: Utilizing unlabeled data for face recognition[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE Computer Society, 2019: 2662-2669.

[11] HUANG C, LOY C C, TANG X. Unsupervised learning of discriminative attributes and visual representations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5175-5184.

[12] SCHMARJE L, SANTAROSSA M, SCHRÖDER S M, et al. A survey on semi-, self-and unsupervised techniques in image classification[J]. arXiv: 2002.08721, 2020.

[13] MASI I, WU Y, HASSNER T, et al. Deep face recognition: A survey[C]// 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2018: 471-478.

[14] JI C M, SONG T C. Sparse Representation-Based Classification Under Optimization Forms for Face Recognition[J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(2): 120-126.

[15] KE P F, CAI M G, WU T. Face Recognition Algorithm Based on Improved Convolutional Neural Network and Ensemble Learning[J]. Computer Engineering, 2020, 46(2): 262-267, 273.

[16] TAO S F, LI Y F, HUANG Y F, et al. Face Detection Algorithm Based on Deep Residual Network and Attention Mechanism[J]. Computer Engineering, 2021, 47(11): 276-282.

[17] PARKHI O M, VEDALDI A, ZISSERMAN A. Deepface recog-

nition[C]// Proceedings of the British Machine Vision Conference. 2015:41.1-41.12.

[18] NOROOZI M,VINJIMOOR A,FAVARO P,et al. Boosting self-supervised learning via knowledge transfer[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9359-9367.

[19] VONGKULBHISAL J,VINAYAVEKHIN P,VISENTINI-SCARZANELLA M. Unifying heterogeneous classifiers with distillation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3175-3184.

[20] LUO P,ZHU Z,LIU Z,et al. Face model compression by distilling knowledge from neurons[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016:3560-3566.

[21] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.

[22] DENG J,GUO J,XUE N,et al. Arcface:Additive angular margin loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:4690-4699.

[23] WANG H,WANG Y,ZHOU Z,et al. Cosface:Large margin cosine loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5265-5274.

[24] DEVRIES T,TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv:1708.04552,2017.

[25] YUN S,HAN D,OH S J,et al. Cutmix:Regularization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019:6023-6032.

[26] ZHANG K,ZHANG Z,LI Z,et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters,2016,23(10):1499-1503.

[27] WANG F,XIANG X,CHENG J,et al. Normface:L2 hypersphere embedding for face verification[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017:1041-1049.

[28] MIRZADEH S I,FARAJTABAR M,LI A,et al. Improved knowledge distillation via teacher assistant[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:5191-5198.

CHENG Xiang-ming, born in 1996, postgraduate. His main research interests include computer vision and model compression.

DENG Chun-hua, born in 1984, Ph.D, associate professor. His main research interests include computer vision and machine learning.

(责任编辑:柯颖)