基于 InsightFace 的改进教室人脸识别算法及其应用

田曦初，苏寒松，刘高华*，刘腾腾
天津大学电气自动化与信息工程学院，天津 300072

摘要 针对教室场景小人脸识别准确率低的问题，基于 InsightFace 算法，结合 MobileFaceNet 结构和 DenseNet 结构，提出一种将通道相加和通道级联结合起来的 Dual-MobileFaceNet 轻量级网络结构，提高了识别速度和小人脸识别准确率。同时，提出一种双重分类算法，进一步改善了 InsightFace 算法的识别分类能力，在 LFW 数据集上准确率达 99.46%。最后将所提算法移植在 Jetson TX2 嵌入式开发板上，在 8 人、18 人教室场景下识别准确率分别达 96.24%、94.68%，每帧识别速度分别约为 0.14 s、0.29 s。相比其他大型网络，所提网络更具实时性和有效性。所提算法为教室人脸识别、无感知考勤系统提供了有效思路。

关键词 机器视觉；人脸识别；卷积神经网络；教室场景；深度学习
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.221501

Improved Classroom Face Recognition Algorithm Based on InsightFace and Its Application

Tian Xichu, Su Hansong, Liu Gaohua*, Liu Tengteng
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Aiming at the problem of low recognition accuracy of small faces in classroom scene, this paper proposes a lightweight network structure (Dual-MobileFaceNet) combining channel addition and channel concatenation based on the InsightFace algorithm by integrating the MobileFaceNet and DenseNet structures, so as to improve the recognition speed and the recognition accuracy of small faces. Meanwhile, a double classification algorithm is proposed to improve the identification and classification ability of the InsightFace algorithm. The proposed algorithm achieves an accuracy of 99.46% on LFW dataset. Finally, the proposed algorithm is transplanted to Jetson TX2 embedded development board. In 8- and 18-people classrooms, the recognition accuracy of the proposed algorithm is 96.24% and 94.68%, and the recognition speed of each frame is 0.14 s and 0.29 s, respectively. Compared with other large networks, the proposed network is more realistic and efficient. The proposed algorithm provides an effective concept for the classroom face recognition and non perception attendance system.

Key words machine vision; face recognition; convolutional neural network; classroom scene; deep learning
OCIS codes 150.1135; 100.4996; 330.5000; 100.2960

1 引 言

深度学习在计算机视觉领域取得了惊人的成就，已经产生了很多性能优越的深度学习神经网络算法，这些算法应用于交通监控[1]、手写识别[2]等方面。在人脸识别领域，近几年涌现出 DeepFace、DeepID、FaceNet、SphereFace、CosFace、InsightFace[3-8]等优秀算法。DeepFace 算法[3]使用深度卷积神经网络对对齐后的人脸进行多类的分类学习。随后 DeepID 系列[4]可随训练人脸种类增加而提升识别能力。但上述两种算法均使用 Softmax 损失函数，学习到的人脸特征区分度不高、人脸识别

收稿日期：2020-02-17；**修回日期**：2020-03-11；**录用日期**：2020-03-25
基金项目：广州市科技计划(201802020008)
* **E-mail**：suppig@126.com

效果差。为解决此问题，一些新型损失函数被提出。FaceNet 算法[5]采用 Triplet 损失函数监督人脸特征学习，获得较好效果，但提取人脸三元组时既耗时又不稳定。SphereFace 算法[6]将损失转向球面角度，提出 A-Softmax 损失函数。CosFace 算法[7]改进了基于 A-Softmax 的 Large Margin Cosine 损失（LMCL）函数，将余弦距离作为分类依据。InsightFace 算法[8]进行了进一步的改进，给出了一种用角度距离作分类依据的 ArcFace 损失函数，相比余弦距离，其对角度的影响更加直接，减小了计算开销、加大了类间距、提高了准确率，是目前最先进、最优秀的人脸识别算法之一，在 LFW[9]、AgeDB[10]等开源数据集上能达到 90% 以上的识别准确率。然而以单一损失函数作分类依据时，其识别性能有局限，难以获得进一步提升。故本文提出一种双重分类器的算法，改善了原 InsightFace 算法[8]的识别分类策略，进一步提升了人脸识别的准确率，并在公开数据集上进行了实验分析。

上述各种算法大多是在大型公开数据集上进行实验的，关于教室等人员众多、前后排人脸差距较大、小人脸复杂的特殊场景的研究较少；且主要使用参数量很多的深度神经网络在价格高昂、性能优越的服务器上进行实验研究，在计算能力有限的嵌入式设备中难以应用。针对以上问题，本文模拟真实教室场景，制作了人脸数据集和测试视频，并结合 MobileFaceNet 结构[11]和 DenseNet 结构[12]的优势，提出一种名为 Dual-MobileFaceNet 的轻量级网络结构，用来替换 InsightFace 算法[8]中的 ResNet 结构[13]，大量减少参数量的同时提升了小人脸的识别准确率，使算法应用于嵌入式设备成为可能。最终，参考文献[14-15]，将改进算法移植到英伟达公司（NVIDIA）发布的 Jetson TX2 人工智能嵌入式开发平台上，并在自制实时视频上进行实验分析，验证了所提算法在教室人脸识别方面的优越性，也为教室人脸识别应用提供了思路。

2 基本原理

2.1 人脸识别过程

人脸识别算法主要包括人脸检测、人脸图像预处理、人脸图像特征提取、人脸识别 4 个步骤。其中预处理步骤主要是对人脸图像进行对齐裁剪等，使结果便于后续的特征提取和识别工作。参考文献[16]，使用多任务卷积神经网络（MTCNN）[17]人脸检测算法进行人脸检测、尺度放缩及人脸对齐的预处理工作，这可为后面人脸识别准备数据。

人脸图像特征提取效果的好坏主要由卷积神经网络决定，而人脸识别的效果则与分类策略密切相关。

2.2 网络结构

课堂实际应用对网络参数量和实时性均有要求，而文献[8]中所采用的基准结构 ResNet 是大型网络，难以保证实时性，因此参考了 MobileFaceNet，其是专门为嵌入式设备所设计的一种轻量级网络结构。在骨架网络中使用深度可分离卷积，并使用全局深度卷积（GDConv）替代全局平均池化方式，从而大大减少参数量。

全局深度卷积层的输出可表示为

$$G_r=\sum_{i,j}K_{i,j,r}\times F_{i,j,r},\tag{1}$$

式中：F 为输入特征图的尺寸，K 为深度可分离卷积核，二者尺寸均为 $W\times H\times R$，表示特征图宽、高、通道数相乘；输出特征图 G 的尺寸为 $1\times1\times R$；r 为当前通道；(i,j)为空间宽高尺寸。最终全局深度卷积层计算开销与全局平均池化层的计算开销之比为

$$\frac{W\cdot H\cdot R}{W\cdot H\cdot R\cdot Q}=\frac{1}{Q},\tag{2}$$

式中：Q 为滤波器数量。由此可知，极大地减少了参数量。将文献[11]中的瓶颈结构称为移动块结构。

MobileFaceNet 模型轻便，且准确率稍逊大型网络。但教室人脸识别的重要问题在于教室后排的小人脸识别。在对经过前向传播的特征进行逐层提取过后，小人脸信息所剩无几，且其信息大多分布在底层，因此若直接使用该网络仍会丢失很多重要信息。为了强化对底层特征的提取，参考了 DenseNet[12]中的密集结构，其拼接操作在通道维度上直接拼接，可以完整保留传入特征的信息和低层信息，因此拼接操作可保存更多低层原始特征信息。为了减少参数量，每个密集块结构中采用了深度可分离卷积结构。假设特征图用 x_l 表示，密集块共 l 层，则密集连接块层间变换可表示为

$$x_l=H_l([x_0,x_1,\cdots,x_{l-1}]),\tag{3}$$

式中：$[x_0,x_1,\cdots,x_{l-1}]$为从第 0 到第 $l-1$ 层特征图的拼接融合，为密集连接，共有 $l(l+1)/2$ 个连接；$H_l(\cdot)$包含 8 个非线性变换，依次为 1×1 卷积、批量归一化（BN）、带参线性整流函数（PReLU）、3×3 深度可分离卷积、BN、PReLU、1×1 深度可分离卷积、BN。将变换 $H_l(\cdot)$称为密集

块结构。

结合上述，所设计的具体网络结构如下。1）针对步长为 1 的 Mobile-Block，每一个底层块结构（Block）的输入都与更高层块结构的输入连接并进行元素加操作，遇到步长为 2 的块结构时停止，如图 1 所示。此结构包含普通卷积（Conv）、逐点卷积（pw_Conv）、深度可分离卷积（dw_Conv）操作。2）针对步长为 2 的 Mobile-Block，输入一条经 1×1 卷积层和 2×2 平均池化层（Avg_pooling）的分支，将其与输出相加融合，如图 2 所示。3）因拼接和元素相加的侧重点不一样，拼接通过更多通道数保存更多低层数据，元素相加则使每个通道信息更丰富。因此便融合这两种操作，提出由一个 Mobile-Block 和一个 Dense-Block 并行传输组成的双重块（Dual-Block）结构。此 Dual-Block 结构仅在步长为 1 时进行，步长为 2 时仅保留 Mobile-Block 结构。至此，便得到本文的基准网络结构，命名为 Dual-MobileFaceNet，其结构如图 3 所示。

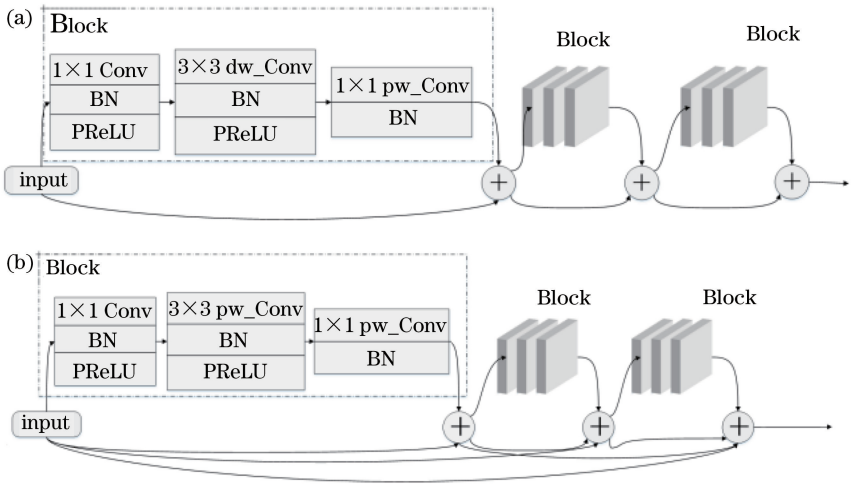

图 1 步长为 1 时的 Mobile-Block 结构。（a）MobileFaceNet；（b）Dual-MobileFaceNet

Fig. 1 Mobile-Block structure with stride of 1. (a) MobileFaceNet; (b) Dual-MobileFaceNet

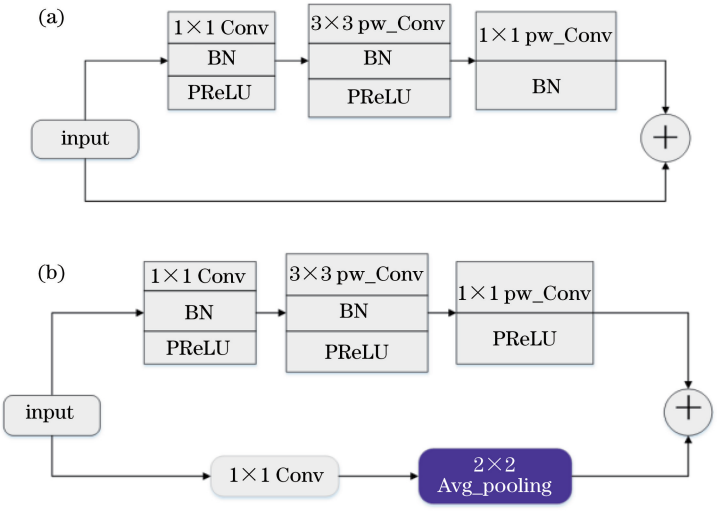

图 2 步长为 2 时的 Mobile-Block 结构。（a）MobileFaceNet；（b）Dual-MobileFaceNet

Fig. 2 Mobile-Block structure with stride of 2. (a) MobileFaceNet; (b) Dual-MobileFaceNet

图 3 是由 3 个连续的 Dual-Block 和 1 个步长为 2 的 Mobile-Block 组成的网络结构示意图，完整的网络结构如表 1 所示。除了 Linear 结构（Linear_Conv、Linear_pw_Conv）是由卷积层和 BN 组成的，其余所有操作（operation）结构都是由卷积层、BN 和 PReLU 共同构成的。s 代表步长，Block 结构中的步长为 2，指的是第一个卷积层的步长，其余层的步长始终为 1；n 代表重复次数；pad 代表填充 0 的数目；k 代表密集块的增长率。针对几组连续 Dual-Block 结构，为了有更好的底层提取效果且避免参数量过多，设置第一组 $k=48$，第二组 $k=32$，其余 $k=16$。

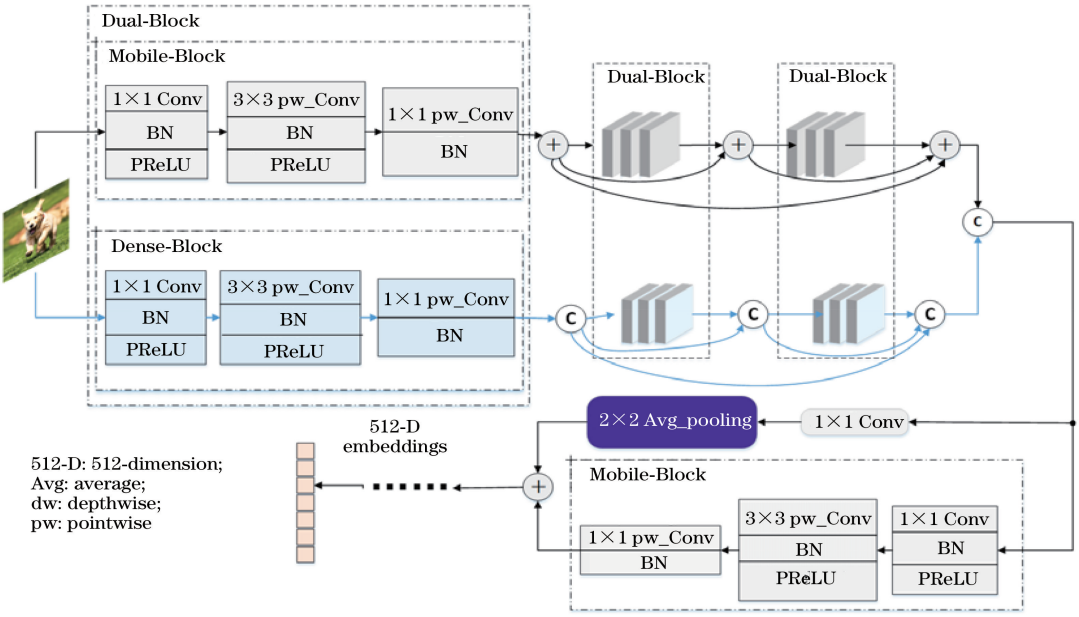

图 3 Dual-MobileFaceNet 结构示意图

Fig. 3 Schematic of Dual-MobileFaceNet structure

表 1 Dual-MobileFaceNet 网络结构

Table 1 Network structure of Dual-MobileFaceNet

Input size/Number of channels	Type	Output size/Number of channels	Operation	s	n	Pad
112×112/3	Convolution	56×56/64	3×3 Conv	2	1	1
56×56/64	Convolution	56×56/64	3×3 dw_Conv	1	1	1
56×56/64	Dual-Block	56×56/128+2k	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	1	2	
56×56/128+2k	Mobile-Block	28×28/64	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	2	1	1
28×28/64	Dual-Block	28×28/128+6k	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	1	6	
28×28/128+6k	Mobile-Block	14×14/128	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	2	1	1
14×14/128	Dual-Block	14×14/256+4k	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	1	4	
14×14/256+4k	Mobile-Block	7×7/128	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	2	1	1
7×7/128	Dual-Block	7×7/256+2k	[1×1 pw_Conv 3×3 dw_Conv 1×1 Linear_pw_Conv]	1	2	
7×7/256+2k	Convolution	7×7/512	1×1 pw_Conv	1	1	0
7×7/512	Convolution	1×1/512	7×7 Linear_Conv	1	1	0
1×1/512	Convolution	1×1/128	1×1 Linear_pw_Conv	1	1	0

2.3 双重分类策略

InsightFace 算法核心是贡献了一种全新的损失函数 ArcFace，效果优于 FaceNet、SphereFace 等诸多算法，因此，采用此算法并对其进行改进。

ArcFace 损失函数虽然可以很好地扩大类间距离、缩小类内距离，但若将其单独作为分类依据，会使网络无法对图形相邻像素点间联系进行学习，且阈值、超参数的选取对最终的结果有很大影响。其虽然对公共人脸数据库有很好的泛化能力，但对特定的或背景相似的情境，识别效果欠佳。

因此，在测试阶段提出了一种基于双重分类器的算法。此算法除了用 ArcFace 损失函数计算角度距离（作为分类依据），还训练了一个 2 层全连接层组成的分类器，称其为平行分类器，双重分类器示意图如图 4 所示。其中 $\cos\theta_{yi}$ 为归一化特征向量和归一化权重乘积的结果，其大小取决于特征向量和权重间的角度，$\cos\theta_{yi}$ 越接近 1 则两人脸相似度越高。再用反余弦函数直接求得角度距离 θ_{yi}，在角度空间进行 ArcFace 损失函数运算，在角度空间中最大化分类界限，获得更好的识别效果。和 ArcFace 损失函数的分类依据不同，全连接层分类器虽然不能扩大类间距离，但可以在特定数据集上提高模型非线性表达能力和学习能力。将两个分类器的结果组合起来共同作为分类的最终标准。为减少参数量也为防止过拟合现象，要加丢弃（Dropout）层，否则对小数据集会失去效果。

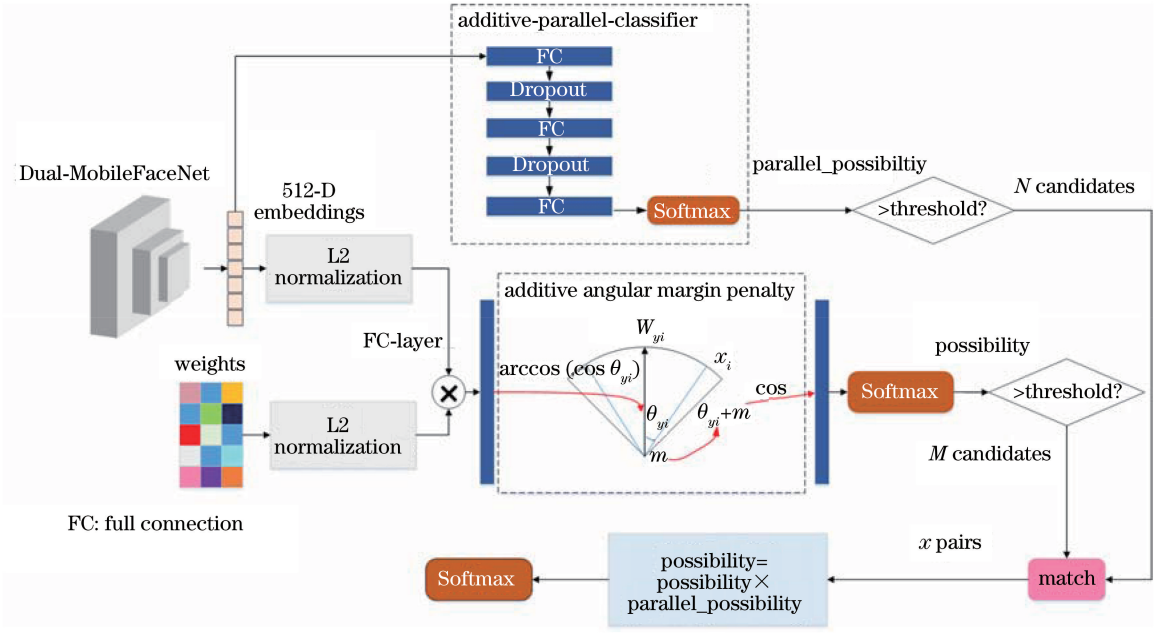

图 4 双重分类器结构示意图

Fig. 4 Schematic of double classifier structure

传入图片经 Dual-MobileFaceNet 提取到的特征得到 512 维（512-D）特征向量，然后这些向量传入两路分类器：一路是原 ArcFace 论文中[8]的结构，由 ArcFace 函数算出分类置信度；另一路传入平行分类器结构。之后 ArcFace 函数结合 Softmax 层算出各类概率，经过一个阈值判断输出 M 个候选识别结果；平行分类器同样输出各类预测概率，经过某一阈值判断后输出 N 个候选识别结果。接下来匹配对应的人名，能匹配成功 x 对（x pairs），其中 $x<\min(M,N)$，再将 x 对概率值组成数组，对每对进行内部相乘，得到最终的 x 个判决值，经过 Softmax 操作输出最大的概率，其所对应的类便是最终判决结果。

3 实验分析与讨论

3.1 实验设置

3.1.1 预处理

进行人脸识别网络的训练和测试前，首先要进行人脸检测和对齐工作。用 MTCNN 算法[17]识别和检测人脸，并用相似变换方式将人脸对齐居中，裁剪至 112×112 的大小。

3.1.2 数据集介绍

本次实验训练数据集分为大型公开数据集和自制数据集两部分。在大型数据集 MS1MV2[18]上进行预训练，该数据集总共有 85742 个身份，580 万张图片。

接下来在公开数据集上进行实验时不需再微调，在自制教室视频上进行实验时需微调。针对教室场景，在光照条件相近的情况下制作了多张人脸数据库来模拟教室学生，总共 57 个人共 6750 张人脸图片，平均每人约 110 张图片，其中男生 30 人，女生 27 人，避免性别不平衡影响实验效果的问题。每张照片包含不同角度、不同大小的单一人脸，便于后续训练时学到足够多的信息，可以识别具有不同姿态的人脸。实验数据的纯度对于实验结果至关重要，因此在数据集图片中，一定要保证左右眼、鼻尖、左右嘴角 5 个部位清晰，且每张照片仅有一张清晰无遮挡人脸。同时，需进行左右翻转、裁剪等数据增强手段来弥补数据集较小的缺点，按照 1∶10 扩充数据集图片。部分数据样本如图 5 所示。

图 5 自制训练数据集图例

Fig. 5 Examples of self-made training dataset

测试数据方面，在训练过程中采用的验证集是公开数据集，包括 AgeDB-30[10]、CFP[19]、LFW[9]。在 Jetson TX2 测试过程中，模拟了真实的教室环境和云台摄像头的位置，分别录制 8 人和 18 人的视频用于测试。

3.1.3 教室场景介绍

测试视频是在模拟教室场景下录制的。模拟教室宽 6 m、长 10 m。摄像云台型号为 HS-HDC500C，安装在教室前墙正中间高 2.8 m 处。第一排座位距前墙约 2 m，相邻排座位间距为 0.8 m，每排放置 4 个座位。教室实景图与示意图如图 6 所示。

3.1.4 Jetson TX2 配置

测试视频实验将在 NVIDIA 公司开发的 Jetson TX2 人工智能嵌入式开发板上进行。Jetson TX2 包含具有多个 ARM(advanced RISC machine)架构的中央处理器(CPU)。和服务器计算机架构不同，开发时必须用官方提供的 Jetpack 工具包来完成相应 Ubuntu 系统及 CUDA、cudnn、OpenCV 等深度学习依赖库的安装。需要特别注意的是，Jetpack 版本对应的各个安装包版本均不一样，如 Jetpack4.2 对应的是 Ubuntu18.04 版本，因此需使用契合自己需求的版本。

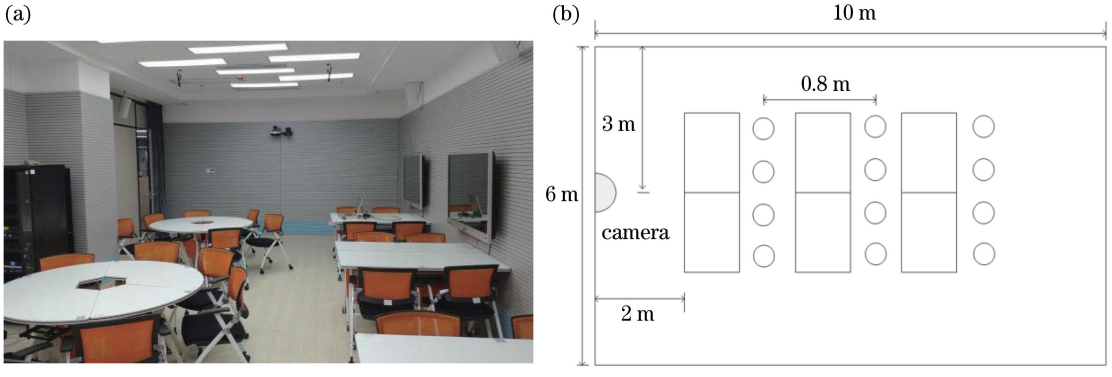

图 6 教室场景图。(a)实景图；(b)示意图

Fig. 6 Classroom scene. (a) Real scene; (b) sketch map

考虑到 Jetson TX2 平台计算能力有限，先在服务器上训练算法模型，之后把模型移植到开发板上测试，因此需将 TX2 开发板与服务器环境设置成一致，否则无法进行移植工作。本实验在 Jetpack3.2 安装的 Ubuntu16.04 系统环境及 CUDA 9、cudnn7.1.5 显卡依赖环境上进行，使用

Python 编程语言和 MXNet1.5.0 深度学习框架。

3.1.5 训练细节

在公开数据集上进行实验前，在 MS1MV2 数据集上直接训练基于双重分类器的整个神经网络，初始学习率为 0.01，在第 10 万和第 16 万次迭代后分别降至前次迭代的 1/10，训练在 18 万次迭代后停止。

在云台录制视频上进行实验前，预训练时以 ArcFace 为损失函数，不训练双重分类器中的平行分类器结构。在 MS1MV2 数据集上进行实验，初始学习率为 0.01，在第 10 万和第 16 万次迭代后分别降至前次迭代的 1/10，训练在 18 万次迭代后停止。之后固定整个神经网络参数不变，在自制数据集上训练平行分类器，平行分类器选择交叉熵为损失函数。因数据集较小，迭代 100 轮即可。随后将训练好的模型移植到 Jetson TX2 上进行测试。至此实验的前期准备工作已经完成，下面进行实验结果分析。

3.2 实验结果分析

3.2.1 不同网络结构实验结果与分析

表 2 是基于 InsightFace 算法的不同网络模型在公开数据集上的实验结果，测试集选取 AgeDB 数据集、CFP 数据集的正面照与形象照子集(CFP-FP 和 CFP-FF)、LFW 数据集、跨年龄 LFW 数据集(CALFW)。所有的网络模型均是在 MS1MV2 数据集上训练获得的。由表 2 可以得出：所提网络结构尺寸为 8.8 MB，小于 MobileNet-v1 结构[20]，与 MobileNet-v2 结构[21] 相当，稍大于 ShuffleNet 结构(1×表示通道数不变，$g=3$ 表示 3 组通道)[22] 和 MobileFaceNet 结构；识别准确率优于其他轻量级网络，说明融合不同方式操作可以加强对低层特征的提取能力；识别速度仅次于 ShuffleNet 和 MobieFaceNet，具有很好的实时性。和大型网络相比，Dual-MobileFaceNet 虽然识别准确率稍逊一些，但差距不大，识别速度约是 ResNet-101 的 7 倍，DenseNet-201 的 3 倍，具有有明显的优势，故所提结构较好地平衡了识别速度与识别准确率间的关系。

表 2 不同网络实验结果对比

Table 2 Comparison of experiment results of different networks

Network	Recognition accuracy /%					Speed /(frame·s^{-1})	Model Size /MB
	AgeDB	CFP_FP	CFP_FF	LFW	CALFW		
ResNet-101[13]	97.28	95.11	99.65	99.71	96.65	42.64	250
ResNet-50[13]	96.03	94.06	99.62	99.52	95.36	70.84	174.5
DenseNet-201($k=32$)[12]	96.68	94.83	99.62	99.68	96.04	100.17	161.8
DenseNet-169($k=32$)[12]	95.38	93.66	99.01	98.86	95.28	120.34	114.4
ShuffleNet(1×,$g=3$)[22]	89.27	89.09	97.75	98.70	93.06	410.78	7.4
MobileNet-v1[20]	88.65	88.54	97.06	98.43	93.01	206.64	13.7
MobileNet-v2[21]	88.81	88.53	97.36	98.38	92.88	230.71	8.6
MobileFaceNet[11]	92.95	89.46	98.03	98.96	93.89	432.41	4.1
Dual-MobileFaceNet	93.94	91.16	98.68	99.18	94.02	326.35	8.8

3.2.2 不同算法实验结果与分析

不同算法在验证集上的实验结果如表 3 所示。网络结构均采用 Dual-MobileFaceNet，统一在 MS1MV2 数据集上训练。Double classifier 即所提基于双重分类策略的算法，O-Double classifier 表示平行分类器不含丢弃层。对比有、无丢弃层结果，发现有丢弃层时效果优势明显，而无丢弃层时算法效果和 InsightFace 算法相当。说明无丢弃层时，算法会使平行分类器的泛化性变弱、双重分类策略失效，这证明了丢弃层结构的必要性。和其他算法相比，所提算法在 LFW、CFP-FP、AgeDB-30 数据集上均取得了最好的识别效果，证明了融合两个不同损失分类策略的算法可以改善单一损失函数分类的局限性，使不同分类策略相互补足，共同产生最终结果，从而提高准确率。

表 3 不同算法的识别准确率对比

Table 3 Recognition accuracy comparison of different algorithms %

Algorithm	LFW	CFP-FP	AgeDB-30
DeepFace[3]	95.53	87.46	89.61
Deep FR[23]	96.04	88.26	90.13
DeepID2[4]	96.14	87.85	90.26
FaceNet[5]	96.95	88.20	90.69
SphereFace[6]	97.58	90.03	91.84
CosFace[7]	98.43	90.75	92.33
InsightFace[8]	99.18	91.16	93.94
O-Double classifier	99.12	91.21	93.22
Double classifier	99.46	93.33	95.88

3.2.3 云台视频实验结果与分析

从实际教室场景应用出发，云台视频测试在 Jetson TX2 开发平台上进行，在服务器上训练好神经网络后将其移植到 Jetson TX2 上进行测试。为控制计算开销和保证视频检测流畅，测试视频每 3 帧检测摄像头并传入帧一次。Jetson TX2 接线方式如图 7 所示。

表 4 是基于双重分类策略的不同网络模型在 Jetson TX2 上对 8 人、18 人测试视频的识别准确率和模型复杂度进行对比的结果。其中 FLOPS (floating-point operations per second)为浮点运算数，用来衡量模型的复杂度。

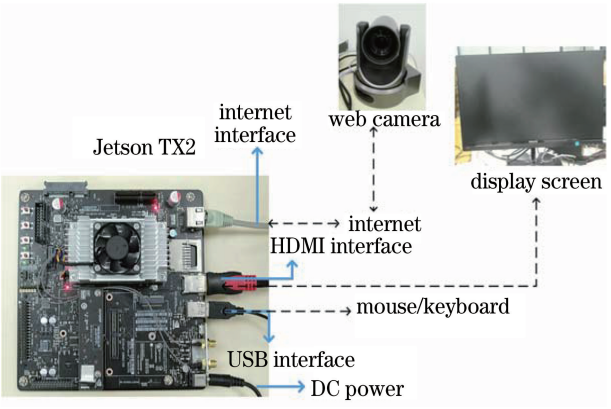

图 7 Jetson TX2 接口连接

Fig. 7 Interface connection of Jetson TX2

表 4 不同网络结构在云台视频上的实验结果

Table 4 Experimental results of different networks on pan tilt video

Network	Recognition accuracy /%		Speed /(frame·s^{-1})		FLOPS/10^6
	8-people	18-people	8-people	18-people	
ResNet-101[13]	97.08	94.14	2.16	4.37	22.69×10^3
ResNet-50[13]	95.96	91.51	1.28	2.61	12.34×10^3
DenseNet-201[12]	96.78	94.98	1.16	2.36	8.5×10^3
DenseNet-169[12]	95.27	91.69	0.89	1.81	6.6×10^3
ShuffleNet[22]	92.05	87.53	0.12	0.26	591
MobileNet-v1[20]	91.12	85.60	0.16	0.35	1.1×10^3
MobileNet-v2[21]	91.96	86.33	0.13	0.28	1.0×10^3
MobileFaceNet[11]	92.83	88.77	0.10	0.21	439.8
Dual-MobileFaceNet	96.24	94.68	0.14	0.29	1.0×10^3

准确率方面，和其他类型网络相比，Dual-MobileFaceNet 在 8 人测试视频上的准确率仅稍次于 ResNet-101 结构和 DesnseNet-201 结构，优于其他轻量级网络；在 18 人测试视频上仅稍次于 DenseNet-201，但模型尺寸却小很多。这是因为 Dual-MobielFaceNet 用 Mobile-Block 和 Dense-Block 并行的方式将低层特征和复用融合前层特征有效结合，提升了网络的学习能力，在小人脸信息较少的情况下也可以提取出丰富的强区分度特征；又延用了深度可分离卷积的思想，在学习到丰富的信息同时大大减少了网络参数量。同时，所提网络的浮点运算数小，与 MobileNet-v2 相当，在实现良好特征提取能力同时保证了网络结构的轻便。

在识别速度方面，ResNet 等网络的弊端也由此可见，每帧已经需要 1 s 以上甚至 2 s 以上，且模型计算量巨大，当面对多人教室时，完全无法承担实时监督的任务。而 Dual-MobileFaceNet 识别速度在 8 人视频下每帧平均 0.14 s，18 人视频下每帧平均 0.29 s，仍可满足实时性要求，为课堂实时人脸识别、课堂无感知考勤等功能的实现提供了可能。

表 5 是不同算法的识别准确率结果，网络均采用 Dual-MobileFaceNet。双重分类算法在 8 人、18 人视频上均取得了最佳的识别效果，且在 18 人视频上优势更明显。证明在小人脸信息有限的情况下，用结合不同分类器的双重分类策略进行预测，可以有效改善单分类依据的局限性，获得更好效果。图 8 为所提双重分类算法在测试视频上的识别效果。图 9 是 InsightFace 算法和双重分类算法在 8 人视频上的准确率混淆矩阵，颜色越深表示准确率越高，可看出，双重分类算法对每个人的识别准确率高于 InsightFace 算法。

表 5 不同算法的识别准确率

Table 5 Recognition accuracy of different algorithms %

Algorithm	8-people	18-people
DeepFace[3]	87.53	83.67
Deep FR[23]	88.54	84.27
DeepID2[4]	88.94	84.25
FaceNet[5]	89.35	85.33
SphereFace[6]	90.58	87.68
CosFace[7]	91.83	90.75
InsightFace[8]	93.69	91.68
Double classifier	96.24	94.68

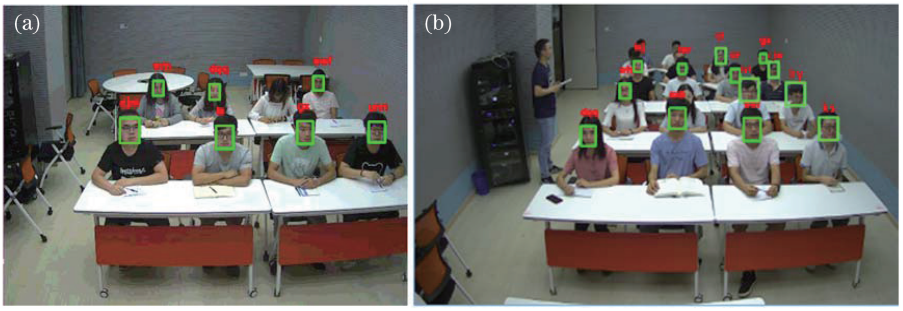

图 8 所提算法识别结果。(a) 8 人视频;(b) 16 人视频

Fig. 8 Recognition results of proposed algorithm. (a) 8-people video; (b)16-people video

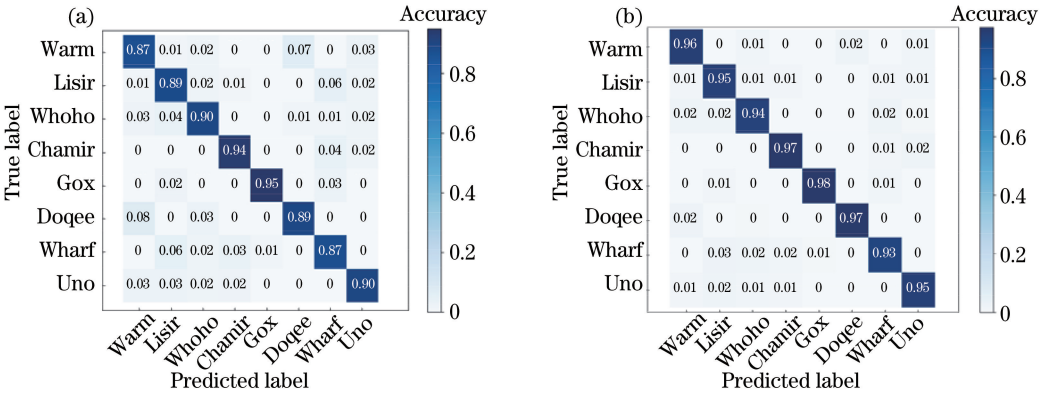

图 9 8 人视频的识别准确率混淆矩阵。(a) InsightFace;(b) Double classifier

Fig. 9 Recognition accuracy confusion matrix of 8-people video. (a) InsightFace; (b) Double classifier

分析算法对不同尺寸人脸识别情况。统计模拟教室场景下人脸图片的大小,第一排人脸平均尺寸最大,约 125 pixel×125 pixel;第二排人脸平均尺寸中等,约 110 pixel×110 pixel;后三排人脸尺寸相差不大,平均尺寸最小,约 100 pixel×100 pixel。考虑数据的均衡,定义尺寸大于 120 pixel×120 pixel 的为大人脸图片;尺寸介于 120 pixel×120 pixel 和 105 pixel×105 pixel 之间的为中等人脸图片;尺寸小于 105 pixel×105 pixel 的为小人脸图片。图片尺寸如图 10 所示,图中数值为虚线参照框的大小。

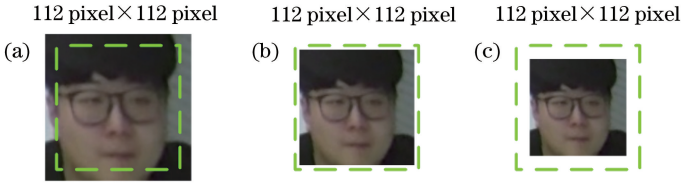

图 10 不同人脸尺寸示意图。(a)大人脸;(b)中人脸;(c)小人脸

Fig. 10 Diagram of different face sizes. (a) Big face; (b) medium face; (c) small face

图 11 是基于双重分类算法的不同网络结构实验结果。可以看出:Dual-MobileFaceNet 对大、中人脸识别效果与 ResNet-101 相当,优于其他两种轻量级网络;对小人脸识别效果最佳,能达 91% 的高准确率。分析折线走向,其他三种网络准确率随人脸尺寸的减小下降明显,而 Dual-MobileFaceNet 下降趋势平缓。原因在于不同尺寸人脸放缩至 112 pixel×112 pixel 统一大小时,中、小人脸图片质量无法保证与大人脸图片质量相同,会因放大而引入失真,使图片模糊、信息更抽象。Dual-MobileFaceNet 可以融合更丰富的语义信息,学习更抽象的人脸特征,从而面对小人脸时仍可提取出高区分度特征。

图 12 是不同算法对不同尺寸人脸的识别准确率。可看出,各算法对大人脸识别效果差距不大,随着人脸尺寸缩小,双重分类算法优势扩大。实验结果表明,随着有效信息不断减少,双重分类算法可以有效互补不同分类思想,避免了单一损失分类的局

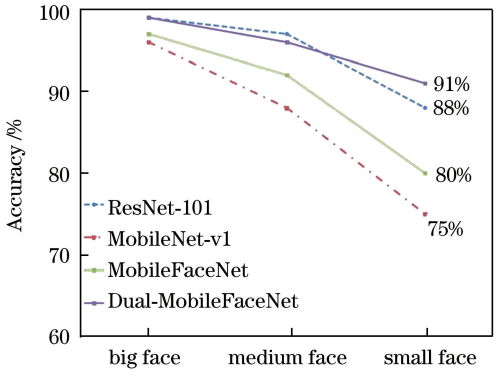

图 11 不同网络对不同尺寸人脸的识别准确率

Fig. 11 Recognition accuracy of different networks for different sizes of faces

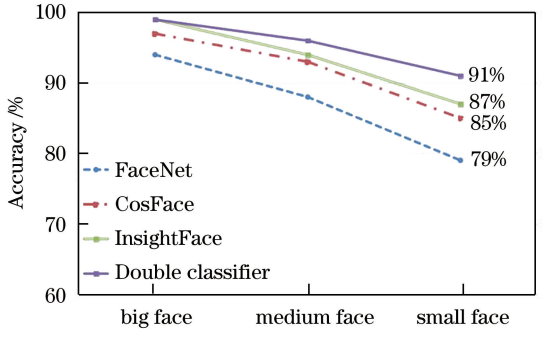

图 12 不同算法对不同尺寸人脸的识别准确率

Fig. 12 Recognition accuracy of different algorithms for different sizes of faces

限，从而得到更可靠的识别性能。

4 结 论

针对教室场景人脸识别，提出了一种新的轻量级网络模型并命名为 Dual-MobileFaceNet，该网络提升了对人脸图像特征的提取能力。同时，提出了双重分类算法，进一步提升了算法对小人脸的识别效果。最后，将算法移植在 Jetson TX2 嵌入式开发板上，并通过在公开数据集和自制云台实时视频上的一系列对比实验，验证了所提算法的可行性和优越性，为实际教室场景人脸识别的应用提供了参考。后续研究工作可以考虑进一步改进网络结构、研究更高效的双重分类策略、怎样更高效利用 Jetson TX2，以能同时进行若干个教室的人脸识别。

参 考 文 献

[1] Tong Y, Yang H C. Real-time traffic sign detection method based on improved convolution neural network [J]. Laser & Optoelectronics Progress, 2019, 56(7): 071003.
童英, 杨会成. 基于改进卷积神经网络的实时交通标志检测方法[J]. 激光与光电子学进展, 2019, 56(7): 071003.

[2] Fang D B, Feng G, Cao H Y, et al. Handwritten formula symbol recognition based on multi-feature convolutional neural network [J]. Laser & Optoelectronics Progress, 2019, 56(7): 072001.
方定邦, 冯桂, 曹海燕, 等. 基于多特征卷积神经网络的手写公式符号识别[J]. 激光与光电子学进展, 2019, 56(7): 072001.

[3] Taigman Y, Yang M, Ranzato M, et al. DeepFace: closing the gap to human-level performance in face verification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1701-1708.

[4] Sun Y, Wang X G, Tang X O. Deep learning face representation by joint identification-verification[C]// Conference and Workshop on Neural Information Processing System, December 8-11, 2014, Montreal, Canada. New York: Curran Associates, 2014, 27: 1097-1105.

[5] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 815-823.

[6] Liu W Y, Wen Y D, Yu Z D, et al. SphereFace: deep hypersphere embedding for face recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6738-6746.

[7] Wang H, Wang Y T, Zhou Z, et al. CosFace: large margin cosine loss for deep face recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 5265-5274.

[8] Deng J K, Guo J, Xue N N, et al. ArcFace: additive angular margin loss for deep face recognition [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 4685-4694.

[9] Huang G B, Learned-miller E. Labeled faces in the wild: updates and new reporting procedures [J].

University of Massachusetts Amherst Technical Report, 2014: 17716267.

[10] Moschoglou S, Papaioannou A, Sagonas C, et al. AgeDB: the first manually collected, in-the-wild age database[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1997-2005.

[11] Chen S, Liu Y, Gao X, et al. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices[M]//Zhou J, Wang Y H, Sun Z A, et al. Biometric recognition. Lecture notes in computer science. Cham: Springer, 2018, 10996: 428-438.

[12] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.

[13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[14] Goyal M, Reeves N D, Rajbhandari S, et al. Robust methods for real-time diabetic foot ulcer detection and localization on mobile devices[J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23(4): 1730-1741.

[15] Lim B, Yang B, Kim H. Real-time lightweight CNN for detecting road object of various size[C]//2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), April 10-12, 2018, Miami, FL, USA. New York: IEEE, 2018: 202-203.

[16] Yang H X, Chen F, Gan W F. Deep face recognition algorithm based on multitask learning[J]. Laser & Optoelectronics Progress, 2019, 56(18): 181005.

杨恢先, 陈凡, 甘伟发. 基于多任务学习的深层人脸识别算法[J]. 激光与光电子学进展, 2019, 56(18): 181005.

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

[18] Guo Y D, Zhang L, Hu Y X, et al. MS-Celeb-$_1$M: a dataset and benchmark for large-scale face recognition[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9907: 87-102.

[19] Sengupta S, Chen J C, Castillo C, et al. Frontal to profile face verification in the wild[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV), March 7-10, 2016, Lake Placid, NY, USA. New York: IEEE, 2016.

[20] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2020-02-16]. https://arxiv.org/abs/1704.04861.

[21] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-2, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 4510-4520.

[22] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 6848-6856.

[23] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C]//2015 British Machine Vision Conference (BMVC), September 7-10, 2015, Swansea, England. [S.l.: s.n.], 2015.