第 36 卷 第 3 期
2021 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 36 No. 3
Jun. 2021

DOI:10.13364/j.issn.1672-6510.20200109

基于 M-Xception 网络的戴口罩人脸表情识别

韦赛远，林丽媛，张怡然
（天津科技大学电子信息与自动化学院，天津 300222）

摘 要：目前，国内外对于有遮挡的人脸表情识别研究较少，其中戴口罩的人脸表情识别（face emotion recognition with mask，FERM）应用场景复杂、数据集缺乏、识别准确率低，因此提出一种改进的 Xception 网络模型 M-Xception Net（Modified Xception Net），并建立一个 FERM 数据集. M-Xception 网络模型具有轻量级特性，参数量较少，对细微表情信息敏感，适用于场景复杂、分辨率较低表情数据集，并利用 Dlib 的 Face-Mask 技术，对 FER2013 表情数据进行戴口罩的遮挡处理，生成 FERM 数据集. 将 M-Xception 网络模型在 FERM 数据集上进行测试，结果表明，戴口罩人脸表情识别准确率达到 88.95%，高于直接利用 Xception 网络进行戴口罩表情识别的准确率 84.94%，并且缩短了训练时间.

关键词：表情识别；M-Xception 网络；遮挡
中图分类号：TP391.4 文献标志码：A 文章编号：1672-6510（2021）03-0072-05

Facial Expression Recognition with Mask Based on M-Xception Net

WEI Saiyuan，LIN Liyuan，ZHANG Yiran
（College of Electronic Information and Automation，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：At present，there is little research on masked facial expression recognition with occlusion at home and abroad. For the masked facial expression recognition（face emotion recognition with mask，FERM）with the complex application scenarios，lack of data set，and low recognition accuracy rate，thus this article proposes an improved M-Xception Net（Modified Xception Net），and establishes a data set FERM. It has lightweight network characteristics with fewer parameters，more sensitive to subtle expression information. Therefore，it is more suitable for complex scenes and emoticon datasets of lower resolution. The Face-Mask technology of Dlib is applied to mask FER2013 expression data and generates FERM expression data set. The experiments on FERM data set show that the M-Xception network model has better recognition performance with the 88.95% accuracy，which is higher than the 84.94% accuracy and shorter training time with using Xception net directly.

Key words：facial expression recognition；M-Xception Net；occlusion

人脸识别技术作为身份认证的重要工具，具有非接触式、成本低、方便快捷的特点[1]. 同样，人脸表情识别技术由于其高信息量、情感交互的作用，吸引更多研究者的关注. 随着计算机技术的不断发展，基于深度学习的人脸表情识别技术正在得到充分挖掘和应用. 但是，由于口罩遮挡人脸表情的绝大部分信息，使得戴口罩的人脸表情识别具有高度复杂性，所以戴口罩的人脸表情识别技术的相关研究较少，识别的准确率一直偏低. 目前，普遍关于人脸表情识别的研究都集中在无遮挡情况，其中王韦祥等[2]提出改进的双通道卷积神经网络模型，在 JAFFE 数据集和 CK+数据集上均取得不错的识别效果，但是没有考虑到光照、遮挡对识别的影响. 李勇等[3]提出一种跨连接的 LeNet-5 结构，学习低层次特征以弥补表情数据集样本不足带来的问题，但是该结构关注低层次特征而忽略了各层特征的相互关联，导致模型泛化性不好. 王建霞等[4]提出一种改进的跨连接 VGG[5]网络，将网络中的低层次特征与高层次特征进行融合以提

收稿日期：2020–06–28；**修回日期**：2020–12–17
基金项目：天津市教委科研计划项目（2019KJ211）；天津科技大学大学生实验室创新基金（1902A202）
作者简介：韦赛远（1997—），男，安徽合肥人，本科；**通信作者**：林丽媛，讲师，linly@tust.edu.cn

高鲁棒性，并引入 Inception 网络结构加快收敛速度，在 FER2013 识别率(无口罩)达到 72.76%，但该方法仅考虑了表情的类内差距较大的情况，对于混合表情识别率较差.

本文提出了一种 M-Xception 网络(Modified Xception Net)模型，实现戴口罩有遮挡的人脸表情识别. 该网络简化了 Xception 冗杂的参数结构，保留残差机制和可分离卷积特征，注重各层特征的关联性，对细微表情信息的提取尤其敏感. 同时，为了防止过拟合现象，在全连接层引入了 Dropout 技术. 实验结果表明，改进后的网络能有效提高戴口罩人脸表情识别的准确率，达到更好的分类效果.

1　M-Xception 网络模型

1.1　M-Xception 网络模型的训练测试框架

Xception[6] 网络利用“极致的 Inception”模块(Block)以减少网络参数，并结合类似于 ResNet[7]的残差机制以保证网络的稳定性. 本文在 Xception 网络基础上，结合戴口罩的人脸表情识别(face emotion recognition with mask，FERM)受限于遮挡的问题，在简化模型的基础上注重细微特征(如眉毛、眼神)的提取，搭建了一个输入尺寸为 48 × 48 × 1 的 M-Xception 网络模型，其训练测试框架如图 1 所示.

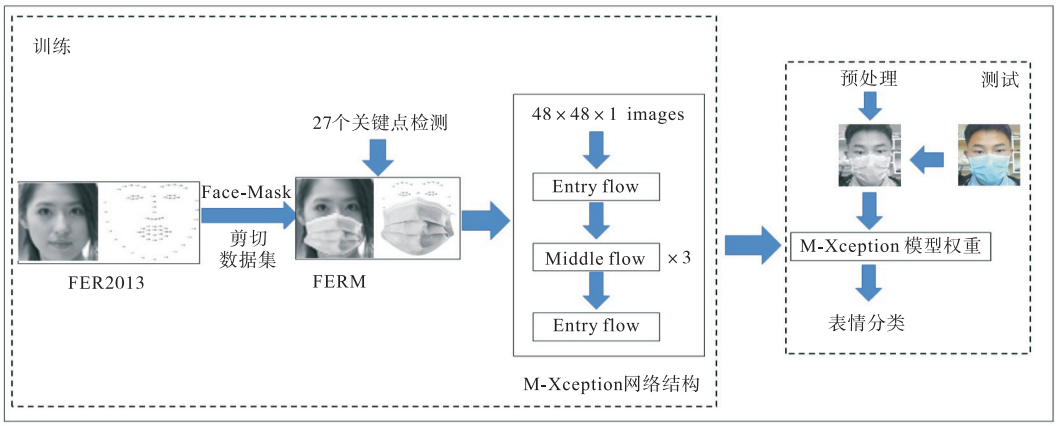


图 1　M-Xception 网络模型的训练测试框架

Fig. 1　Training and testing framework of M-Xception net model

在训练部分，未戴口罩的人脸关键特征点有 68 个，首先用 Face-Mask 技术确定鼻子、嘴巴等部位的特征位置，并加上口罩进行特征遮挡处理. 对添加口罩遮挡的数据集进行筛选，去除标签丢失、标签错误、非人脸等不符合要求的图片，生成新的数据集 FERM. 其中与表情相关的特征点为 27 个，然后将此数据集投入 M-Xception 网络进行训练.

测试网络结构时，对图片进行预处理后，采用训练好的模型进行测试，得到表情的分类结果.

1.2　M-Xception 网络结构

M-Xception 网络模型的网络结构(图 2)共分为 3 个部分：Entry flow、Middle flow 和 Exit flow.

首先，在 M-Xception 网络结构的 Entry flow 输入部分，与 Xception 通过 2 个 stride 分别为 2、1 的卷积层不同，M-Xception 网络将输入特征图通过 2 个 stride = 1 的标准卷积层，目的是在保留原始特征位置信息的同时加深网络深度.

其次，通过 2 个深度可分离卷积 Block 代替 Xception 网络中“极致的 Inception”，以减少参数. 假设输入特征图大小为 $M\times M$，通道数为 $C(C>1)$，标准卷积的规格是 $k\times k\times C\times C'$($k$ 取 3，C'表示特征图经过卷积后的输出通道数)，输出大小为 $M_1\times M_1$，标准卷积如图 3(a)所示，则其参数量为

$$n_1 = k\times k\times C\times C' \tag{1}$$

若以上情况使用“极致的 Inception”，其结构如图 3(b)所示，其参数量为

$$n_2 = C\times C' + k\times k\times C' \tag{2}$$

$$n_1 - n_2 = k\times k\times C\times C' - C\times C' - k\times k\times C' = C'(8C-9) \tag{3}$$

由于 $C>1$ 且为整数，所以 $n_1 - n_2>0$，因此在特征提取过程中可知参数量 $n_1>n_2$. 随着层数的增多，“极致的 Inception”一般可比标准卷积减少约 90% 的计算量. 本文所使用的深度可分离卷积，其结构如图 3(c)所示，与图 3(b)相似，区别在于将“极致的 Inception”结构中 1 × 1 和 3 × 3 卷积进行颠倒可得到深度可分离卷积. 在上述假设中，参数量为

$$n_3 = k\times k\times C + C\times C' \tag{4}$$

在本文的网络结构中，$C'>C$（除最后一层分类），故参数量 $n_2>n_3$. 因此，即 M-Xception 网络放弃使用 Xception 网络的“极致的 Inception”，改用深度可分离卷积模块.

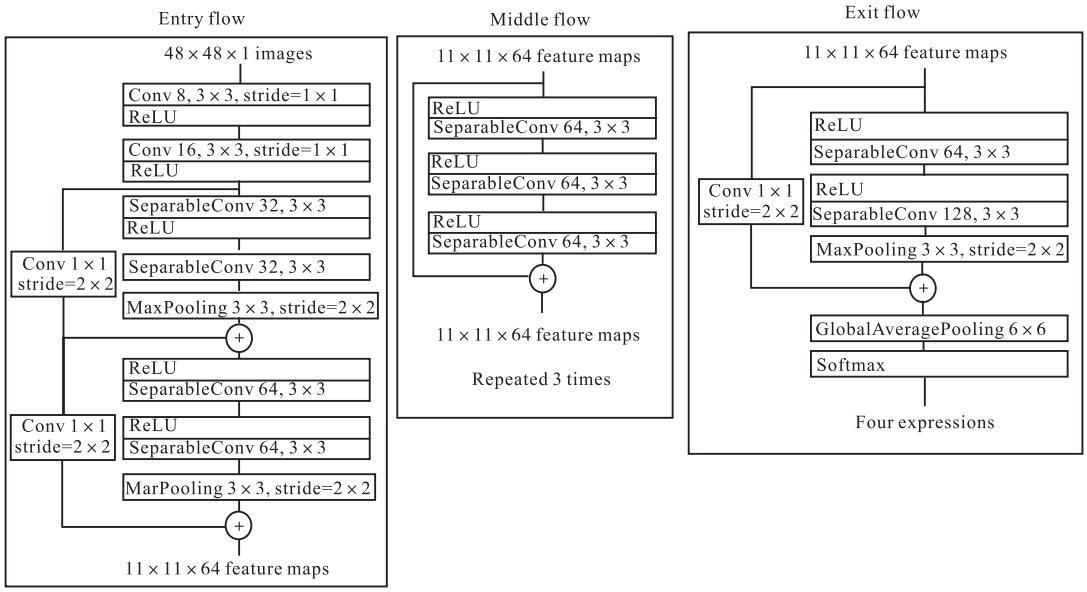
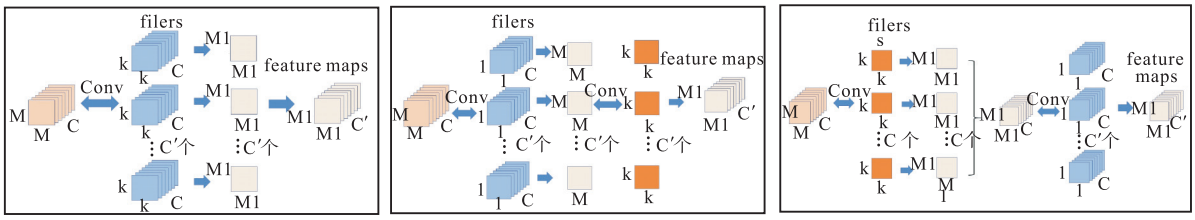


图 2　M-Xception 网络模型的网络结构

Fig. 2　Structure of M-Xception network



（a）标准卷积　　（b）极致的 Inception　　（c）深度可分离卷积

图 3　标准卷积、“极致的 Inception”和深度可分离卷积示意图

Fig. 3　Diagrammatic drawing of standard convolution，“extreme Inception” and deep separable convolution

最后，把 Xception 的 Middle flow 部分提前 1 个 Block 并将重复 8 次减少至 3 次，目的是减少参数、简化模型，实验证明对准确率没有影响. 经过 1 个深度可分离卷积 Block，进行全局平均池化后通过 Softmax[8]完成分类. 整个 M-Xception 网络结构，除了深度分离卷积层 dw 和 pw 之间使用线性激活函数（保留信息特征）之外，其他卷积层之间均添加 BN[9]层、ReLU[10]激活函数，防止数据发散并增强模型的非线性表达能力.

分别将 Xception、M-Xception 网络模型在没有口罩遮挡的数据集 FER2013 上进行训练、测试，结果见表 1. M-Xception 网络模型具有更加简洁、快速、高效等轻量级的特性. 改进前后两种网络模型在 FER2013 数据集上的准确率均为 68% 左右，准确率都不高，这是因为 FER2013 数据集存在标签缺失、错误以及非人脸表情图片等问题，而由人类进行主观识别的准确率也只有 $(65\pm5)\%$左右. 因此，后续戴口罩实验先去掉了 FER2013 数据集的标签缺失、错误以及非人脸表情等图片，再进行 Face-Mask 口罩遮挡处理，具体实验过程详见 2.2 节.

表 1　Xception 网络改进前后对比

Tab. 1　The comparison of before and after Xception network improvement

模型	参数	准确率/%	epoch 时间/s
Xception	9.2×10^5	68.36	32.825
M-Xception	0.8×10^5	68.41	21.210

2　实　验

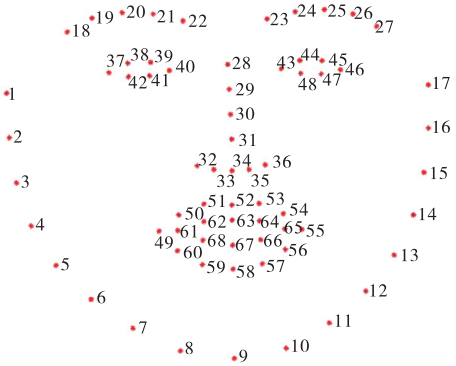
2.1　实验环境

在 PC 端上的实验以深度学习框架 Keras 为基础，使用 TensorFlow 作为其后端，编程语言采用 Python 3.6. 实验使用的 GPU 为 NVIDIA GeForce

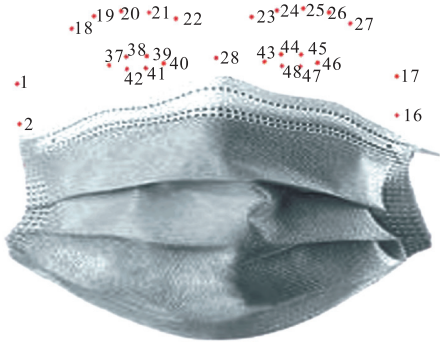
RTX 2080 Ti，其显存大小为 11 GB，CPU 为 Intel Xeon CPU E5-2678 v3 六核，其内存大小是 62 GB，在 Windows 10 64 位操作系统上进行.

2.2　数据集

有关表情识别的现有开源数据集有 CK+、JAFFE、HUMAINE Database、MMI、NVIE 和 FER2013 等. 由于 FER2013 数据集更加符合实际生活场景，由 7 种表情共 35 886 张人脸表情图片组成，数据规模较大，在全脸未遮挡的情况下，嘴巴动作、双眼、眉毛形态等使人脸的表情信息量较为丰富. 佩戴口罩前后的人脸关键点信息如图 4 所示. 利用 Dlib[11]的 Face-Mask 技术，对 FER2013 无遮挡的表情数据进行戴口罩的遮挡处理，处理前后图片对比如图 5 所示.



(a) 全脸条件下的人脸关键点分布



(b) 佩戴口罩后的人脸关键点

图 4　佩戴口罩前后的人脸关键点信息

Fig. 4　Facial key points before and after wearing a mask

由于口罩遮挡，鼻子、嘴巴以及脸上部分轮廓的特征几乎完全损失，导致表情中的微笑与中性、厌恶与生气、恐惧与惊讶的差异较小，使得这些表情在有遮挡的情况下难以区分，因此对数据集进行裁剪处理：第一，去除戴口罩条件下的相似表情类别中的一类，保留中性、悲伤、惊讶、生气表情数据；第二，去除非正脸、模糊、标签丢失、标签错误、非人脸等不合格图片. 处理后的数据集使用 FER+标签完成数据标注. 最后得到 3 840 张戴口罩的 4 类表情数据作为实验数据集 FERM，其表情类别数量分布见表 2.

(a) 图片处理前

(b) 图片处理后

图 5　Face-Mask 处理前后

Fig. 5　Before and after Face-Mask treatment

表 2　4 类表情数量分布

Tab. 2　Number distribution of the four types expressions

表情	中性	悲伤	惊讶	生气
数量/幅	1 120	980	1 050	690

2.3　图像预处理

为了提高模型的泛化性能，采用框架 Keras 的 ImageDataGenerator()工具进行数据增强，设置图片随机转动角度为 10°，图片随机水平、竖直偏移的幅度为 0.1，并进行随机缩放和水平翻转，以避免出现过拟合现象.

2.4　实验结果与分析

实验均采用 Adam[12]优化器优化损失，epoch 为 100，batch_size 为 64，将训练集和验证集的比例设为 9∶1. 按照表 1 所设计的网络模型进行训练，将预处理后的图片输入网络，M-Xception 网络模型的实验结果验证集准确率相比较于 CNN、mini_Xception 和 Xception(同等输入规模)网络的戴口罩有遮挡表情

识别的平均准确率见表 3. 由表 3 可知，M-Xception 网络加上 Dropout 技术的准确率最高，为 88.95%，说明本文改进的模型具有更加理想的识别效果. M-Xception + Dropout 网络模型准确率比 M-Xception 网络模型高，可见 Dropout 有效地防止了过拟合，提高了验证集准确率. M-Xception 网络模型的准确率比 Xception 网络模型的高，且 M-Xception 网络模型耗费时间比 Xception 网络模型的短，表明改进后的模型既节省了模型参数又缩短了训练时间，同时还提高了准确率，进一步证明 M-Xception 网络模型模型对于较小输入特征具有良好的应用价值. 使用 M-Xception + Dropout 模型权重在全数据集中进行准确率的测试，得到混淆矩阵如图 6 所示，结果表明模型的准确率均较高.

表 3　不同模型的平均准确率对比

Tab. 3　Comparison of average accuracy of different models

模型	平均准确率/%	epoch 时间/s
CNN	83.38	1.296
mini_Xception	84.68	2.052
Xception	84.94	4.644
M-Xception	88.83	2.538
M-Xception + Dropout	88.95	2.376

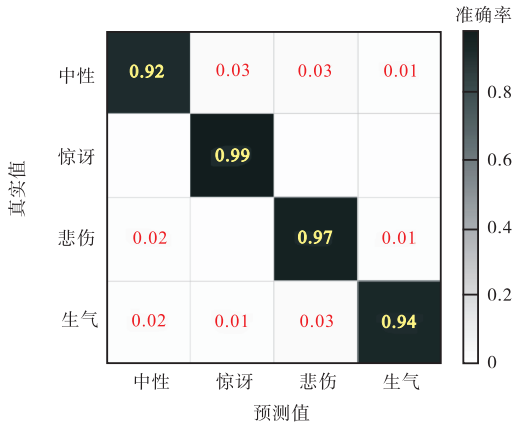


图 6　M-Xception + Dropout 模型权重下的混淆矩阵

Fig. 6　Confusion matrix under the weight of M-Xception + Dropout model

3　结　语

本文提出了一种 M-Xception 网络模型用于戴口罩人脸表情的识别. 在 M-Xception 网络模型中，通过减少原网络 Xception 模型网络层数和使用深度可分离卷积，保证了网络的轻量级特性，同时保留 Xception 模型将低层次特征与高层次特征进行融合的特性，提高网络的特征提取能力，并加入 Dropout 技术防止过拟合现象，最后采用 Softmax 分类器对提取的特征进行分类. 实验结果表明本文改进的模型可以在戴口罩有遮挡的 FERM 数据集上达到较高的识别准确率，超过了现有的有遮挡表情识别方法. 由于现实生活中表情识别的干扰因素众多(光照、侧脸等)，因此如何建立自然场景下人脸有遮挡表情数据集，并进一步分析被遮挡的有限关键点的人脸表情识别，将是后续工作的研究重点.

参考文献：

[1] 杨巨成，代翔子，韩书杰，等. 人脸识别活体检测综述[J]. 天津科技大学学报，2020，35(1)：1–9.

[2] 王韦祥，周欣，何小海，等. 基于改进 MobileNet 网络的人脸表情识别[J]. 计算机应用与软件，2020，37(4)：137–144.

[3] 李勇，林小竹，蒋梦莹. 基于跨连接 LeNet-5 网络的面部表情识别[J]. 自动化学报，2018，44(1)：176–182.

[4] 王建霞，陈慧萍，李佳泽，等. 基于多特征融合卷积神经网络的人脸表情识别[J]. 河北科技大学学报，2019，40(6)：540–547.

[5] Simonyan K，Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2020–06–28]. https://arxiv.org/abs/1409.1556.

[6] Chollet F. Xception：Deep learning with depthwise separable convolutions[EB/OL]. [2020–06–28]. https://arxiv.org/abs/1610.02357.

[7] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[J]. Conference Proceedings，2016，1：770–778.

[8] Grave E，Joulin A，Moustapha C，et al. Efficient softmax approximation for GPUs[EB/OL]. [2020–06–28]. https://arxiv.org/pdf/1609.04309.pdf.

[9] Ioffe S，Szegedy C. Batch Normalization：Accelerating deep network training by reducing internal covariate shift[EB/OL].[2020–06–28]. https://arxiv.org/abs/1502.03167.

[10] Glorot X，Bordes A，Bengio Y. Deep sparse rectifier neural networks[J]. Journal of Machine Learning Research，2011，15：315–323.

[11] Kazemi V，Sullivan J. One millisecond face alignment with an ensemble of regression trees[J]. Conference Proceedings，2016，1：1867–1874.

[12] Kingma D P，Ba J. Adam：A method for stochastic optimization[EB/OL]. [2020–06–28]. https://arxiv.org/abs/1412.6980.

责任编辑：郎婧