第48卷 第2期
Vol.48 No.2
计 算 机 工 程
Computer Engineering
2022年2月
February 2022
·图形图像处理·
文章编号：1000-3428(2022)02-0180-06 文献标志码：A 中图分类号：TP391

基于多通道注意力机制的人脸替换鉴别

武 茜，贾世杰
（大连交通大学 电气信息工程学院，辽宁 大连 116028）

摘 要：基于深度学习的人脸替换技术取得快速发展，但由DeepFake自动生成的人脸替换图片有可能危害人们的隐私安全。针对DeepFake图片鉴别问题，建立一种基于多通道注意力机制的深度学习鉴别网络模型。将Xception网络作为基础特征提取器，在多通道注意力模块中通过矩阵相乘的思想融合全局和局部的注意力表示，以减少重要信息损失。设计损失函数时添加中心损失，从而提高特征区分度。在训练过程中利用注意力图来引导训练图像的裁剪和去除，以达到数据增强的目的。实验结果表明，相比Xception、B4Att方法，在FaceForensics++数据集上该网络模型对DeepFake的检测精度分别提高0.77和0.45个百分点，在Celeb-DF数据集上分别提高5.30和4.68个百分点。

关键词：人脸替换；多通道注意力机制；图片鉴别；Xception网络；深度学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：武茜，贾世杰.基于多通道注意力机制的人脸替换鉴别[J].计算机工程，2022，48(2)：180-185，193.
英文引用格式：WU Q，JIA S J.Face swapping detection based on multi-channel attention mechanism[J].Computer Engineering，2022，48(2)：180-185，193.

Face Swapping Detection Based on Multi-Channel Attention Mechanism

WU Qian，JIA Shijie
（School of Electrical Information Engineering，Dalian Jiaotong University，Dalian，Liaoning 116028，China）

【Abstract】In recent years，the face swapping technology based on deep learning has developed rapidly，but the images of face swapping automatically generated by DeepFake may endanger people's privacy and security.To detection the DeepFake images，a deep learning-based network model using the multi-channel attention mechanism is designed.The model employs Xception network as the basic feature extractor，and the idea of matrix multiplication is used to combine global and local attention representations in the multi-channel attention module to reduce information loss.Then a center loss is introduced into the design of the loss function，so the feature discrimination can be improved.At the same time，in the training process，the attention map is used to guide the cropping and erasing of training images to achieve data enhancement.The experimental results show that the detection accuracy of the network model for DeepFake is 0.77 percentage points higher than that of Xception，and 0.45 percentage points higher than that of B4Att on the FaceForensics++ dataset.The accuracy of the proposed model is 5.30 percentage points higher than that of Xception，and 4.68 percentage points higher than that of B4Att on Celeb-DF dataset.

【Key words】face swapping；multi-channel attention mechanism；image detection；Xception network；deep learning
DOI：10. 19678/j. issn. 1000-3428. 0060739

0 概述

近年来，基于深度学习的人脸替换技术取得了较快的发展[1]。人脸伪造主要体现在修改身份、转移表情、生成全新人脸这3种情况，其中，修改身份即为人脸替换，DeepFake是人脸替换中的主要方法。DeepFake起源于2017年Reddit论坛中的一个匿名代码，是基于深度学习的人脸替换方法，其简单易用且没有违和感[2]。DeepFake以自编码器（AutoEncoder，AE）为核心结构，利用编码器提取面部图像的潜在特征，然后使用解码器重建面部图像，从而将目标人物的面部图像替换到原视频中的人物上。为了在原图像和目标图像之间交换面部，需要2个编码器-解码器对，每个编码器-解码器对都在人物图像集上进行训练，并且编码器的参数在2个网络之间共享。目前，人脸替换鉴别方法主要分为两类。

基金项目：辽宁省教育厅科学研究项目（JDL2019006）。
作者简介：武 茜（1994—），女，硕士，主研方向为深度学习、换脸鉴别；贾世杰，教授、博士。
收稿日期：2021-01-29 **修回日期：**2021-03-01 **E-mail：**panicwq@126.com

第一类方法利用DeepFake视频中前后帧的时间信息进行鉴别：GÜERA[3]研究发现DeepFake视频前后相邻帧之间包含不一致的时序性内容，其提出利用CNN和LSTM检测假视频的方法；LI等[4]研究发现DeepFake假视频的人物眨眼频率低于真实视频，因此，将裁剪后的眼部区域序列分配到长期循环卷积网络（LRCN）[5]中进行动态预测；张怡暄等[6]研究发现DeepFake视频中人脸区域的帧间差异明显大于真实视频，其利用视频相邻帧中人脸图像的差异特征进行预测；陈鹏等[7]利用全局时序特征和局部空间特征来发现伪造人脸视频；LI等[8]利用DeepFake视频相邻帧上的抖动来检测视频真伪，并解决了训练不能很好收敛的问题。上述方法容易被DeepFake技术所借鉴并进行改进，因此，方法的时效性通常较弱[9]。

第二类方法提取DeepFake的图像特征信息进行鉴别：YANG等[10]提出一种鉴别方法，该方法利用由头部方向和位置组成的三维头部位姿之间的差异，将提取的特征输入SVM分类器进行分类，但是实际情况中三维头部姿态获取效率低；AFCHAR[11]利用神经网络中层语义信息，使用具有少量层的神经网络来学习真假人脸图像内在特征的不一致性；LI等[12]针对DeepFake会留下特殊伪影的现象，利用深度学习网络来检测DeepFake伪影；NGUYEN等[13]利用胶囊网络（Capsule-Net）来检测DeepFake，并在FaceForensics++数据集[14]上进行评估。文献[11-13]方法虽然在各自的数据集上具有有效性，但泛化能力弱，对于高质量的DeepFake图像检测效果不佳。

BONETTINI等[15]将简单的注意力机制引入卷积神经网络中，在FaceForensics++和DFDC数据集上进行评估，结果表明，注意力机制对于鉴别DeepFake具有有效性。因此，本文提出一种基于多通道注意力机制的人脸替换鉴别方法。对现有的注意力模型进行扩展，设计一种多通道注意力模块，根据矩阵相乘的思想融合全局和局部的注意力表示，在注意力模块连接主网络的方式上借鉴残差神经网络（ResNet）[16]的跳跃连接方法，以减少重要信息损失。在训练过程中，通过由多通道模块生成的注意力图来引导图像裁剪和去除，从而实现数据增强。

1　本文方法

1.1　网络结构

本文方法的整体网络框架如图1所示。将图片$\boldsymbol{I}$输入特征提取器，得到特征$\boldsymbol{F}$，通过多通道注意力模块得到注意力图$\boldsymbol{A}$，特征图$\boldsymbol{F}$与每个通道的注意力图$\boldsymbol{A}$按元素相乘得到特征矩阵$\boldsymbol{T}$，然后通过全连接层得到概率P，从而区分输入图片是否为DeepFake所生成。

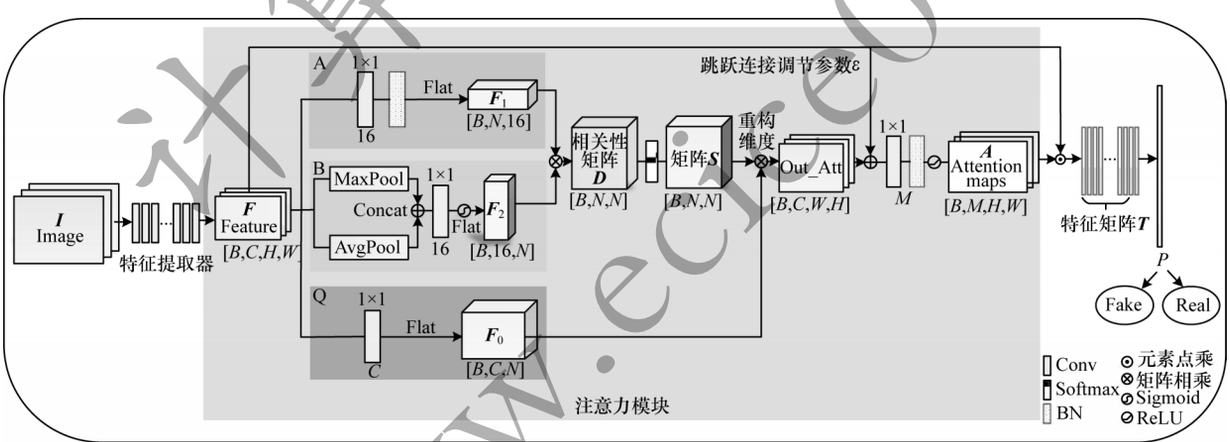


图1　整体网络框架

Fig.1　Overall network framework

1.1.1　特征提取器

本文利用Xception[17]网络作为特征提取器。与常规网络卷积操作相比，Xception的参数量和运算成本较低，且其可以更好地提升网络训练效率，在同等参数量以及大规模数据集上，效果优于Inception-v3。此外，在给定硬件资源的情况下，Xception可以有效提高网络效率和性能。

1.1.2　多通道注意力模块

本文多通道注意力模块以矩阵相乘的方式融合全局和局部注意力表示，再以跳跃连接的方法与主网络连接。具体的注意力模块结构如图1中的浅色阴影区域所示，整个注意力模块分为A、B、Q这3个分支：

1）A分支为全局注意力表示。注意力表示方法将特征图通过16个1×1卷积核的卷积层，获得全局注意力表示，此时更加突出重要的权重，最终得到注意力特征图$\boldsymbol{F}_1$。

2）B分支为局部注意力表示。该分支采用CBAM[18]空间注意力表示，空间特征图$\boldsymbol{F}_2(\boldsymbol{F})$的计算过程如下：

$$\boldsymbol{F}_2(\boldsymbol{F})=\boldsymbol{F}_2\cdot\sigma\left(f^{1\times1}\left[\mathrm{MaxPool}(\boldsymbol{F});\ \mathrm{AvgPool}(\boldsymbol{F})\right]\right) \tag{1}$$

其中：$\boldsymbol{F}$为输入的特征图；MaxPool和AvgPool分别为最大和平均池化层；$f^{1\times1}$为1×1大小的卷积核；σ为Sigmoid激活函数。

对特征图分别进行基于通道的最大池化和平均池化，在通道上做拼接操作，再经过Sigmoid激活函数得到特征图$\boldsymbol{F}_2$。因为最大和平均池化会造成一定的信息损失，所以这里将其称为局部注意力表示。

3）Q分支得到经过2 048个1×1卷积核卷积后的特征图$\boldsymbol{F}_0$，其对Feature多增加一层卷积映射，使网络学到更多的参数。将$\boldsymbol{S}_{i,j}$权重应用到$\boldsymbol{F}_0$上，即每一个元素点都与整个Feature相关，相关性来自于$\boldsymbol{D}$矩阵。Q分支的输出$\boldsymbol{O}$计算公式如下：

$$\boldsymbol{O}=\sum_{i=1}^{N}\boldsymbol{S}_{i,j}\cdot\boldsymbol{F}_0(x_i) \tag{2}$$

其中：$\boldsymbol{F}_0(x_i)$为Q分支的卷积操作表示。

如图2所示，将$\boldsymbol{F}_0[C\times N]$矩阵与$\boldsymbol{S}[N\times N]$矩阵的转置相乘，得到输出$\boldsymbol{O}[C\times N]$。输出$\boldsymbol{O}$中的第$i$行第$j$列的元素表示被矩阵$\boldsymbol{S}$对应第$j$列元素加权之后的Feature在第$i$个通道的值。然后对输出$\boldsymbol{O}$进行维度重构，使输出$\boldsymbol{O}$恢复为$C\times W\times H$尺寸。为了减少训练时间，在输出后再加一个简单的卷积层，最终得到$\boldsymbol{A}$。

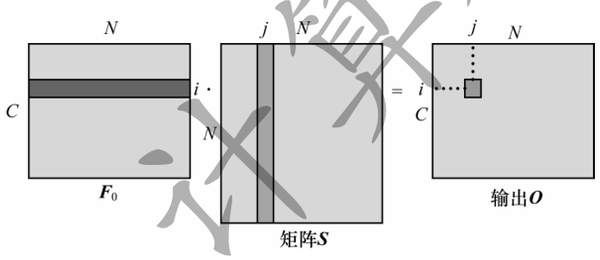


图2 输出$\boldsymbol{O}$的计算过程示意图

Fig.2 Schematic diagram of calculation process of output $\boldsymbol{O}$

本文方法使用矩阵相乘的方式来结合全局和局部注意力表示。为了满足矩阵相乘的条件，将$\boldsymbol{F}_1$和$\boldsymbol{F}_2$的尺寸从2个维度（$H\times W$维）压缩为1个维度（N维），使尺寸从$[B\times16\times H\times W]$分别变为$[B\times N\times16]$和$[B\times16\times N]$，其中，$N=H\times W$，$H$和$W$分别为Feature的长和宽。$\boldsymbol{F}_1$和$\boldsymbol{F}_2$矩阵相乘后得到$\boldsymbol{D}$矩阵，大小为$[B,N,N]$，其可以看作一个相关性矩阵，即$\boldsymbol{F}$上各个元素点之间的相关性表示。通过上述过程，既突出了重要权重又避免了重要信息的损失。矩阵$\boldsymbol{D}$的计算公式如下：

$$\boldsymbol{D}_{i,j}=f_1(x_i)^{\mathrm{T}}\cdot f_2(x_j) \tag{3}$$

其中：f_1和f_2分别为由分支A和B进行的卷积操作；$\boldsymbol{D}_{i,j}$可以理解为矩阵$\boldsymbol{D}$中第i行第j列的元素值，表示$\boldsymbol{F}_2$中第j个元素对$\boldsymbol{F}_1$中第i个元素的影响。

为了防止梯度爆炸问题，将$\boldsymbol{D}$矩阵逐行通过Softmax函数得到$\boldsymbol{S}$矩阵。矩阵$\boldsymbol{S}$的计算公式如下：

$$\boldsymbol{S}_{i,j}=\frac{\exp(\boldsymbol{D}_{i,j})}{\sum_{i=1}^{N}\exp(\boldsymbol{D}_{i,j})} \tag{4}$$

其中：$\boldsymbol{S}_{i,j}$各行元素之和为1；$\boldsymbol{S}$矩阵中第i行元素代表Feature中所有位置的元素对第i个元素的影响，这种影响即为权重。

在连接主网络的方式上，本文借鉴残差神经网络的跳跃连接，引入调节参数ε，使输出$\boldsymbol{O}$的权重需要通过反向传播来更新，具体计算公式如下：

$$y=\varepsilon\boldsymbol{S}+\boldsymbol{F} \tag{5}$$

在初始阶段，ε为0，输出y直接返回输入的$\boldsymbol{F}$，随着训练的进行，输出y逐渐学习到要将经过注意力机制的$\boldsymbol{F}$加在原始$\boldsymbol{F}$上，从而强调了需要施加注意力的部分$\boldsymbol{E}$。

1.1.3 输出层

将特征图$\boldsymbol{F}$与每个通道的注意力图按元素相乘，具体计算公式如下：

$$\boldsymbol{T}_i=\boldsymbol{A}_i\odot\boldsymbol{F},\ i=1,2,\cdots,M \tag{6}$$

其中：$\boldsymbol{A}_i$为注意力图；M为注意力图的个数。

相乘之后以拼接的方式得到特征矩阵$\boldsymbol{T}$，$\boldsymbol{T}$中的每一行代表一张图像的所有特征，然后将$\boldsymbol{T}$特征矩阵输入线性分类层进行二分类，最终得到概率P从而判断输入图像的真假。

1.2 损失函数设置

本文方法的损失函数表达式如下：

$$L=L_{\mathrm{c}}+L_{\mathrm{c}} \tag{7}$$

其中：L_{c}为交叉熵损失；L_{c}为中心损失。

L_{c}的计算方式如下：

$$L_{\mathrm{c}}=-\sum_{i=1}^{N}\left(y_i\log_a y_i'+(1-y_i)\log_a(1-y_i')\right) \tag{8}$$

其中：y_i为真实标签值；y_i'为预测标签值。

L_{c}借鉴了中心损失[19]的原理，将原来中心损失的类中心替换成不同特征的特征中心，使同一类别中同一部分的特征尽可能地接近。L_{c}的计算方式如下：

$$L_{\mathrm{c}}=\sum_{i=1}^{N}\left\|y_i'-c_{y_i}\right\|_2^2 \tag{9}$$

其中：y_i'为网络预测值；c为设置的标签空间特征中心。c的初始值设置为0，按照以下滑动平均公式来更新：

$$c_{y_i}\leftarrow c_{y_i}+\delta\cdot(y_i'-c_{y_i}) \tag{10}$$

其中：δ在实验中初始值取0.05。

1.3 训练过程

本文网络的训练过程使用迁移学习中的微调（Fine-tuning）技术。使用Xception网络在ImageNet数据集上的预训练模型，去掉原来的全连接层，添加新的模块和全连接层，在原有参数的基础上训练整个网络，从而提高实验效率。同时，本文利用细粒度分类WSDAN网络[20]中的训练方式，通过每一个轮

次训练好的注意力图来引导一个轮次图像的裁剪和去除，然后进入网络进行训练，从而实现数据增强。具体过程如下：

输入图像经过特征提取和注意力网络后输出$\boldsymbol{A}$，尺寸为$B\times M\times W\times H$，选取$M$张图像中权重较高的2张图像分别用作图像裁剪和图像去除，经过归一化处理得到$\boldsymbol{A}_1'$和$\boldsymbol{A}_2'$。归一化计算公式为：

$$\boldsymbol{A}_k'=\frac{\boldsymbol{A}_k-\min(\boldsymbol{A}_k)}{\max(\boldsymbol{A}_k)-\min(\boldsymbol{A}_k)} \tag{11}$$

当k=1时，$\boldsymbol{A}_1$用作裁剪得到mask，计算公式如下：

$$\boldsymbol{C}_m(i,j)=\begin{cases}1, \boldsymbol{A}_1(i,j)\geqslant\theta_c\\0, \boldsymbol{A}_1(i,j)<\theta_c\end{cases} \tag{12}$$

其中：$\boldsymbol{C}_m$为得到的mask；θ_c为设定的裁剪阈值。$\boldsymbol{C}_m$为不规则形状的mask，取能包含该mask的最小的矩形边界（bounding box），将该矩形边界坐标覆盖至原图，并将该边界区域放大至原图大小，即放大已经受到关注的区域，就可得到裁剪图片以继续参与训练。

当k=2时，$\boldsymbol{A}_2$用作去除得到mask，计算公式如下：

$$\boldsymbol{E}_m(i,j)=\begin{cases}1, \boldsymbol{A}_2(i,j)<\theta_e\\0, \boldsymbol{A}_2(i,j)\geqslant\theta_e\end{cases} \tag{13}$$

其中：$\boldsymbol{E}_m$为得到的mask；θ_e为设定的去除阈值。$\boldsymbol{E}_m$与原图进行对应元素相乘，得到去除后的图片，使得已经受到关注的区域被消除，保留没有受到关注的区域，且去除后的图片继续参与训练。

2　实验结果与分析

2.1　实验设置

2.1.1　环境设置

本文实验在单机PC端训练完成，实验环境设置如下：处理器为Intel® Core™ i7-6700HQ CPU @2.60 GHz，显卡为NVIDIA GeForce GTX 950 M 4 G，操作平台为Windows 10，软件平台为Python3.6，主要依赖库为CUDA 9.0、cuDNN 7.6。

2.1.2　数据集构成

本文实验只针对由自编码器生成的换脸视频（以下称为DeepFake）。目前关于鉴别DeepFake的数据集质量不统一，FaceForensics++（以下简写为FF++）的DeepFake数据集中有些视频生成效果不佳，人眼就能识别出DeepFake视频，因此，本文对FF++（c40）中DeepFake数据集进行重新人工筛选，将有明显生成痕迹的假视频剔除，并在此基础上扩增数据集，分别由Celeb-DF数据集[21]、DFD（DeepFake-Detection）数据集[14]、网络收集构成。为了降低原始数据的复杂度，提升模型训练稳定性，本文对原视频做预处理，利用MTCNN[22]进行人脸检测，把裁剪出的人脸作为输入图片，最终数据集中训练集总共有17 200张图片，真假图片各占一半，为8 600张，测试集总共有4 300张，真假图片各占一半，都为2 150张。具体的训练集、测试集构成如表1所示。

表1　数据集信息

Table 1　Datasets information

图片类型	训练集	测试集
真实图片	FF++数据集（5 600张）	FF++数据集（1 400张）
	Celeb-DF人脸数据集（2 240张）	Celeb-DF人脸数据集（560张）
	CelebA人脸数据集（296张）	CelebA人脸数据集（74张）
	网络整理（200张）	网络整理（50张）
	FFHQ数据集（160张）	FFHQ数据集（40张）
	DFD数据集（104张）	DFD数据集（26张）
Fake图片	FF++数据集（5 944张）	FF++数据集（1 486张）
	Celeb-DF数据集（2 056张）	Celeb-DF数据集(514张)
	DFD数据集（400张）	DFD数据集（100张）
	网络整理（200张）	网络整理（50张）

2.1.3　实验参数设置

模型基于Pytorch 1.1.0深度学习框架搭建网络架构，训练方法为随机梯度下降法（SGD），初始学习率设为0.001，动量设置为0.95，权重衰减为0.000 01，batch size为8，输入图像大小为300×300，总共进行30轮的训练，$\theta_c\in(0.4,0.6)$，$\theta_e\in(0.4,0.7)$。在训练过程中，采用微调的训练方式，提取Xception除全连接层外的最后一层，即得到的特征图数量为2 048。

2.1.4　评估标准

本文实验采用的评估标准为精度（Accuracy），其定义如下：

$$A_{\text{Accuracy}}=\frac{T_{\text{TP}}+T_{\text{TN}}}{T_{\text{TP}}+T_{\text{TN}}+F_{\text{FP}}+F_{\text{FN}}} \tag{14}$$

其中：T_{TP}为真阳性；T_{TN}为真阴性；F_{FP}为假阳性；F_{FN}为假阴性。本文实验中将人脸真图定义为正类，人脸假图定义为负类。

2.2 对比实验结果

本文方法和其他鉴别方法在测试集上的测试精度对比如表2所示。从表2可以看出，与其他基于深度学习的检测方法相比，本文方法测试精度最高，测试精度相比Xception[14]方法提高了2.63个百分点，相比B4Att[15]方法提高了1.35个百分点，充分验证了本文方法的有效性。

表2 6种方法的测试精度对比

Table 2 Comparison of test accuracy of six methods %

方法	测试精度
Meso-Net[11]	94.75
Capsule-Net[13]	97.33
Xception[14]	96.61
S-MIL-T[8]	97.56
B4Att[15]	97.89
本文方法	99.24

在FF++(c40)、Celeb-DF、DFD[23]数据集上分别进行测试对比，结果如表3所示。从表3可以看出：各方法在Celeb-DF和DFD数据集上的测试精度均低于FF++数据集；本文方法在Celeb-DF和DFD数据集上的测试精度能达到97.85%和92.17%，且在FF++数据集上，本文方法的测试精度相比B4Att[15]提高了0.45个百分点，在挑战性相对较高的Celeb-DF和DFD数据集上，测试精度分别提高4.68和3.59个百分点，本文方法整体性能优于其他对比方法，泛化能力更强；S-MIL-T是基于视频的检测方法，相比其余基于图片的检测方法，其只在Celeb-DF数据集上表现突出。

表3 在FF++、Celeb-DF、DFD数据集上的测试精度对比

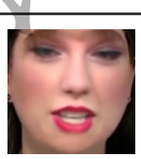
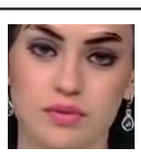
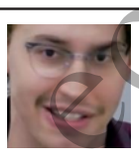
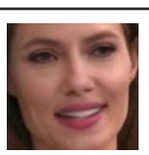
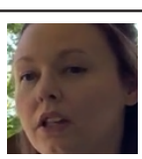
Table 3 Comparison of test accuracy on FF++, Celeb-DF and DFD datasets %

方法	FF++(c40)	Celeb-DF	DFD
Meso-Net[11]	99.12	86.54	83.22
Capsule-Net[13]	99.20	93.48	84.14
Xception[14]	99.21	92.55	87.94
S-MIL-T[8]	97.14	98.84	85.11
B4Att[15]	99.53	93.17	88.58
本文方法	99.98	97.85	92.17

本文还在具有代表性的测试图例上进行实验对比，结果如表4所示，其中，第一、第二幅图为FF++数据集，第三、第四、第五幅图为Celeb-DF数据集，最后一幅图为DFD数据集，表格内“√”代表该网络能正确判断该图为DeepFake图片，“×”代表网络将图片误判为真图。表4中给出的例子实际均为DeepFake图片，从第一幅图片的测试结果可以看出，生成效果不佳的DeepFake图片有明显的伪影边界，表中方法均能鉴别出该图为DeepFake图片，但随着DeepFake图片质量的提升，其他方法会出现误判的情况，而本文方法仍然能够正确地鉴别出该图为DeepFake图片。

表4 DeepFake图片的鉴别结果

Table 4 Identification results of DeepFake pictures

方法						
Meso-Net[11]	√	×	×	×	×	×
Capsule-Net[13]	√	√	√	×	×	×
Xception[14]	√	√	×	×	×	×
S-MIL-T[8]	√	×	×	√	√	×
B4Att[15]	√	√	√	×	×	×
本文方法	√	√	√	√	√	√

2.3 消融实验结果

对本文所设计的模型进行消融实验，测试精度对比情况如表5所示。其中：Base model为直接使用Xception网络进行分类鉴别的模型；+Attention为在Base model上添加本文注意力机制的模型；eraser mask、crop mask分别为注意力引导的图像去除、裁剪的模型；最后4行All代表本文模型在不同的预训练模型(ResNet101[16]、VGG19[24]、Inception-v3[25]、Xception)上进行测试。从表5可以得出：

1)在Xception网络的基础上加入本文设计的Attention模块，测试精度有2.27个百分点的提升；在基础模型上添加中心损失，测试精度也有0.44个百分点的提升；添加A、B不同分支的注意力表示，对基础模型的测试精度分别有2.13和2.18个百分点的提升，即局部和全局注意力表示均能发挥一定作用，将它们相结合后精度能够进一步提升。

2)在多通道注意力模块的基础上引入注意力引导的图像裁剪和去除，测试精度能够提升0.28个百分点；裁剪的作用(提升0.22个百分点)比去除的作用(提升0.08个百分点)更明显，即对于鉴别DeepFake，

裁剪的图像有利于网络提取到更细节的特征。

3）在不同的特征提取网络的基础上，本文设计的多通道注意力模块的测试精度都能达到97%以上，其中Xception网络效果最好。

表5　消融实验结果

Table 5　Results of ablation experiment　%

模型	测试精度
Base model（Xception）	96.61
+center Loss	97.05
+Attention A	98.74
+Attention B	98.79
+Attention	98.88
+Attention + eraser mask	98.96
+Attention + crop mask	99.10
+Attention + eraser&crop mask	99.16
All（ResNet101）	97.47
All（VGG19）	98.23
All（Inception-v3）	98.47
All（Xception）	99.24

3　结束语

本文针对DeepFake图片鉴别问题，建立一种基于多通道注意力模块的鉴别网络模型。将注意力模块添加到现有的预训练模型中，融合全局和局部注意力表示以避免重要信息损失。在训练过程中使用注意力引导的图像裁剪和去除的训练方式，从而起到数据增强的作用。在FF++、Celeb-DF和DFD数据集上的实验结果表明，该模型泛化能力较强，测试精度优于B4Att、S-MIL-T等方法。但是，本文模型难以直接对输入视频进行鉴别，也未利用视频中各帧之间的相关性信息，对以上问题进行研究以提升模型的检测性能将是下一步的研究方向。

参考文献

[1] 暴雨轩，芦天亮，杜彦辉. 深度伪造视频检测技术综述[J]. 计算机科学，2020，47(9)：289-298.
BAO Y X，LU T L，DU Y H. Overview of DeepFake video detection technology[J]. Computer Science，2020，47(9)：289-298. (in Chinese)

[2] MIRSKY Y，LEE W. The creation and detection of DeepFakes：a survey[J]. ACM Computing Surveys，2021，54(1)：1-41.

[3] GÜERA D，DELP E J. DeepFake video detection using recurrent neural networks[C]//Proceedings of 2018 IEEE International Conference on Advanced Video and Signal Based Surveillance. Washington D. C.，USA：IEEE Press，2018：1-6.

[4] LI Y，CHANG M C，LÜ S. In ictu oculi：exposing AI created fake videos by detecting eye blinking[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Washington D. C.，USA：IEEE Press，2018：1-7.

[5] DONAHUE J，ANNE HENDRICKS L，GUADARRAMA S，et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2015：2625-2634.

[6] 张怡暄，李根，曹纭，等. 基于帧间差异的人脸篡改视频检测方法[J]. 信息安全学报，2020，5(2)：49-72.
ZHANG Y X，LI G，CAO Y，et al. A method for detecting human-face-tampered videos based on interframe difference[J]. Journal of Cyber Security，2020，5(2)：49-72. (in Chinese)

[7] 陈鹏，梁涛，刘锦，等. 融合全局时序和局部空间特征的伪造人脸视频检测方法[J]. 信息安全学报，2020，5(2)：73-83.
CHEN P，LIANG T，LIU J，et al. Forged facial video detection based on global temporal and local spatial feature[J]. Journal of Cyber Security，2020，5(2)：73-83. (in Chinese)

[8] LI X，LANG Y，CHEN Y，et al. Sharp multiple instance learning for DeepFake video detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York，USA：ACM Press，2020：1864-1872.

[9] TOLOSANA R，VERA-RODRIGUEZ R，FIERREZ J，et al. DeepFakes and beyond：a survey of face manipulation and fake detection[J]. Information Fusion，2020，64：143-184.

[10] YANG X，LI Y，LÜ S. Exposing deep fakes using inconsistent head poses[C]//Proceedings of 2019 IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2019：8261-8265.

[11] AFCHAR D，NOZICK V，YAMAGISHI J，et al. Mesonet：a compact facial video forgery detection network[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Washington D. C.，USA：IEEE Press，2018：1-7.

[12] LI Y，LÜ S. Exposing DeepFake videos by detecting face warping artifacts[EB/OL]. [2020-12-26]. https://arxiv. org/pdf/1811. 00656. pdf.

[13] NGUYEN H H，YAMAGISHI J，ECHIZEN I. Capsule-forensics：using capsule networks to detect forged images and videos[C]//Proceedings of 2019 IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2019：2307-2311.

[14] ROSSLER A，COZZOLINO D，VERDOLIVA L，et al. FaceForensics++：learning to detect manipulated facial images[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2019：1-11.

[15] BONETTINI N，CANNAS E D，MANDELLI S，et al. Video face manipulation detection through ensemble of CNNs[EB/OL]. [2020-12-26]. https://arxiv. org/pdf/2004. 07676. pdf.

[16] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：770-778.

[17] CHOLLET F. Xception：deep learning with depthwise separable convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2017：1251-1258.

（下转第193页）

（上接第185页）

[18] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 3-19.

[19] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 499-515.

[20] HU T, QI H, HUANG Q, et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification [EB/OL]. [2020-12-26]. https://arxiv. org/pdf/1901. 09891v2. pdf.

[21] LI Y, YANG X, SUN P, et al. Celeb-DF: a large-scale challenging dataset for deepfake forensics [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 3207-3216.

[22] 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.

[23] DOLHANSKY B, BITTON J, PFLAUM B, et al. The deepfake detection challenge (DFDC) dataset [EB/OL]. [2020-12-26]. https://arxiv. org/pdf/2006. 07397. pdf.

[24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020-12-26]. https://arxiv. org/pdf/1409. 1556. pdf.

[25] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception architecture for computer vision [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2818-2826.

编辑　吴云芳