计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300258

基于 i_ResNet34 模型和数据增强的深度伪造视频检测方法

暴雨轩　芦天亮　杜彦辉　石　达
中国人民公安大学信息网络安全学院　北京 100038
(412851819@qq.com)

摘　要　针对深度伪造视频检测存在的面部特征提取不充分的问题,提出了改进的 ResNet(i_ResNet34)模型和 3 种基于信息删除的数据增强方式。首先,优化 ResNet 网络,使用分组卷积代替普通卷积,在不增加模型参数的前提下提取更丰富的人脸面部特征;接着改进模型虚线残差结构的 shortcut 分支,通过最大池化层完成下采样操作,减少视频帧中人脸面部特征信息的损失,然后在卷积层后引入通道注意力层,增加提取关键特征通道的权重,提升特征图的通道相关性。最后,利用 i_ResNet34 模型对原数据集及 3 种基于信息删除的数据增强方式扩充后的数据集进行训练,其在 FaceForensics++的两类数据集 FaceSwap 和 Deepfakes 上的检测准确率分别达到了 99.33%和 98.67%,优于现有的主流算法,从而验证了所提方法的有效性。

关键词:深度伪造;深度学习;残差网络;特征提取;数据增强;人工智能安全
中图法分类号　TP309;TP18

Deepfake Videos Detection Method Based on i_ResNet34 Model and Data Augmentation

BAO Yu-xuan,LU Tian-liang,DU Yan-hui and SHI Da
College of Information and Cyber Security,People's Public Security University of China,Beijing 100038,China

Abstract　Existing Deepfake videos detection methods are weak in extracting facial feature. Therefore,this paper proposes an improved ResNet(i_ResNet34) model and three data augmentation methods based on information dropping. Firstly,the ResNet is optimized by using the group convolution to replace the ordinary convolution to extract more sufficient facial features without increasing model parameters. Then,max pooling layer is used to the down sampling in the shortcut branch of the dashed residual structure of the model whichis improved,so that loss of facial feature information decreases in video frames. Then,the channel attention layer is introduced after the convolution layer to increase the weight of the channel which extracts the key features and improves the channel correlation of the feature map. Finally,the i_ResNet34 model is implemented to train the original dataset and the expanded dataset with three data augmentation methods based on information dropping,achieving 99.33% and 98.67% detection accuracy on FaceSwap and Deepfakes datasets of FaceForensicans++ respectively,superior to the existing mainstream algorithms,thus verifying the effectiveness of the proposed method.

Keywords　Deepfake,Deep learning,Residual network,Feature extraction,Data augmentation,Artificial intelligence security

1　引言

“深度伪造”(deepfake),又称“换脸”,指通过深度学习等技术手段,将源人脸替换到目标人脸来制造网络虚假音视频。最初该技术被用于制作色情视频[1],后被用来传播虚假新闻、制造政治矛盾等,给个人、社会、国家带来潜在威胁。同时,随着生成式对抗网络技术的发展,生成的虚假视频更加难以分辨[2]。因此,检测深度伪造视频已成为主要的研究方向。

早期的检测方法利用视频属性并结合统计学方法来实现,也有相关学者通过计算机视觉方法提取人脸面部特征向量并结合机器学习算法进行检测。近年来,越来越多的研究者通过深度学习对伪造视频进行检测。然而,现有深度学习检测方法很难在检测时间和检测准确率之间达到平衡,对伪造视频人脸空间特征的提取不够充分,且分类效果不够好。鉴于此,本文提出了一种新的深度伪造视频检测方案。本文的主要工作和贡献如下:

(1)在数据处理阶段,提出 3 种基于信息删除的数据增强方式来扩充数据集的多样性,使模型更深入地学习伪造人脸面部空间特征,增强模型的泛化性;

(2)在模型训练阶段,对 ResNet34 网络从 3 个方面进行改进,在尽可能少增加模型参数量的情况下大大提升了模型分类的准确率;

(3)将改进模型在公开的深度伪造数据集 FaceForensics++及扩充后的数据集中进行训练,并采用融合策略选取最佳结

到稿日期:2021-03-25　返修日期:2021-04-29　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2017YFB0802804);中国人民公安大学 2020 年基本科研业务费重大项目(2020JKF101)
This work was supported by the National Key R&D Program of China(2017YFB0802804) and 2020 Fundamental Research Funds for the Central Universities of PPSUC(2020JKF101).
通信作者:芦天亮(lutianliang@ppsuc.edu.cn)

果，最终所提方法在 FaceSwap 和 Deepfakes 两类数据集上分别取得 99.33%和 98.67%的准确率，优于大多数主流检测算法。

2 相关工作

2.1 基于视频帧内差异的检测方法

基于视频帧内差异的检测方法指利用传统算法、机器学习算法或深度学习模型对深度伪造视频单帧的人脸空间特征进行学习，再对视频各帧的预测结果做综合决策。该方法可以充分提取伪造人脸的面部空间特征，但由于忽略了视频本身属于时序数据，因此不能充分利用伪造视频的时序特征。

传统算法更多地考虑了视频本身的属性，如帧率、亮度等。Koopman 等[3]利用相机拍摄过程中的感光噪声差异，计算真假视频的归一化互相关分数并做出区分。Li 等[4]利用 Lambert 算法计算视频帧的二维光照方向，通过判断待测视频二维光照方向的角度变化是否平滑来确定视频的真伪。

机器学习算法常配合人工提取特征手段，通过提取能够表示人脸面部特征的一维向量，对真假人脸做出分类。Matern 等[5]利用伪造人脸存在的双眼颜色差异较大、鼻边出现阴影、牙齿没有几何规则等问题，通过颜色直方图、颜色聚合向量等计算机视觉方法提取人脸五官的特征向量，将其放入 KNN 算法做出分类。Yang 等[6]发现在篡改人脸过程中原人脸图像头部姿态会发生变化，将篡改部分和整个人脸面部的特征向量差作为分类标准，使用 SVM 算法做出分类。Durall 等[7]通过离散傅里叶变换提取视频帧的二维功率谱，并利用方位平均将二维功率谱压缩为一维特征向量，使用逻辑回归算法做出分类。

然而，传统算法和机器学习算法在区分深度伪造视频时，常需要人工提取面部特征，且由于分类器的局限性很难提取到图像深层次的像素级空间特征，因此随着研究的不断深入，卷积神经网络逐渐被用于检测深度伪造视频。Rahmouni 等[8]使用特殊池化层计算图像的统计特征，以区别深度伪造视频。Afchar 等[9]认为视频帧低层的图像噪声特征会随视频压缩而退化，而高层语义特征又难以分辨，因此提出结合 Inception 模块的 MesoNet 以提取图像中间层特征，从而对伪造视频做出分类。Zhou 等[10]提出一种双流 CNN，分别使用 RGB 卷积层和噪声卷积层提取视频帧的像素特征和噪声特征，再将两类特征融合，以提升检测准确率。Nguyen 等[11]在视频帧经 VGG16 提取特征后，使用“胶囊网络”(capsule network)对伪造人脸的面部姿态(位置、色相、纹理)等细节信息进行学习，以提取更丰富的面部特征。国内方面，Wu 等[12]在 Xception 网络引入多通道注意力模块以减少信息损失，同时添加中心损失函数以提高真假人脸的区分度；Hu 等[13]提出适用于深度伪造人脸的交并比算法，并结合全卷积网络(Fully Convolutional Network，FCN)进行不同数据集的跨库检测，提升了模型的泛化性。

2.2 基于视频帧间差异的检测方法

基于视频帧间差异的检测方法通常依据伪造视频出现的时序伪造特征对其做出分类。此种方法能够充分利用视频帧的时序特征，但对帧的长度敏感，针对播放时间较短的视频无法有效提取时序特征，且缺少对伪造人脸局部细节特征的学习。

Sabir 等[14]采取循环卷积策略，首先利用 CNN 提取视频每帧人脸面部特征，再放入循环神经网络(Recurrent Neural Network，RNN)对面部特征的时序变化进行学习，从而鉴别伪造视频。此外，一些方法发现真假人脸的眨眼频率会随视频播放出现差异，Li 等[15]将此作为鉴别依据，使用 VGG16 提取人脸眼部可区分特征，再利用 LSTM 学习真假人脸眼部眨眼频率特征差异。Amerini 等[16]使用 PWC-Net 和 LK 算法按视频播放顺序提取 RGB 帧的光流矢量，利用 VGG16 捕获真假人脸在光流矢量方向、大小的差异，从而做出区分。国内方面，Zheng 等[17]利用基于编码-解码结构的 LSTM 网络，并结合注意力机制共同处理伪造视频的帧间差异特征，更有效地实现了对视频帧的时序特征关联融合；Zhang 等[18]利用孪生网络提取真假视频前后连续帧的深度特征，并分别计算其相似度，从而发现两类视频在时序特征上的差异。

3 基于 i_ResNet34 和数据增强的深度伪造视频检测方法

为进一步提升伪造视频检测准确率，本文基于帧内差异的检测方法，提出了基于 i_ResNet34 模型和数据增强的深度伪造视频检测方法，其分为数据处理、模型训练、综合决策 3 个阶段，以更深入地学习伪造视频的空间分布特征，如图 1 所示。

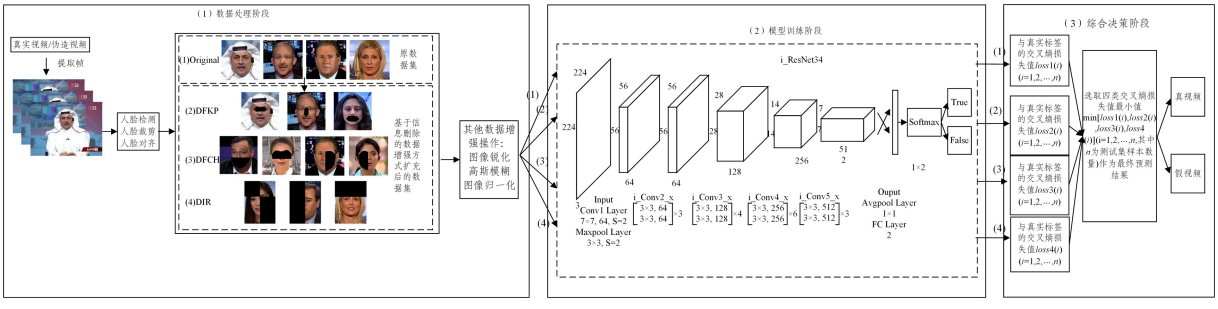


图 1 基于 i_ResNet34 和数据增强的深度伪造视频检测方法

Fig. 1 Deepfake videos detection method based on i_ResNet34 model and data augmentation

3.1 数据预处理

3.1.1 提取帧和人脸

本文首先提取真假视频相同帧位置的人脸并使用 SSIM 算法评价真人脸与伪造人脸图像的差异，发现伪造人脸区域常集中在人脸中心的五官区域，如图 2 所示。因此，可通过人脸检测算法定位并裁剪视频帧人脸头像部分，这样可以使模

型更关注人脸面部特征，以提升检测效率。

图 2 真假人脸面部差异

Fig. 2 Facial differences between true and false faces

本文首先使用 python 开源计算机视觉工具包 cv2 的 VideoCapture()函数等时间间隔抽取每个视频 30 帧，再使用 RetinaFace[19] 人脸检测模型依据检测到的人脸五官关键点定位人脸面部矩形，将人脸面部矩形的长宽各扩展 1.3 倍并进行人脸对齐、人脸裁剪等处理，然后将其输入到 i_ResNet34 模型。

3.1.2 基于信息删除的数据增强方式

正则化(regularization)是深度学习的一项重要技术，指在训练过程中对模型设置约束规则来减少过拟合，包括随机池化、数据增强(data augmentation)等。

数据增强由于只需要对数据进行处理，而不用改变网络结构，因此被用于许多任务中。常见方法有随机裁剪和信息删除。

信息删除方法以文献[20]、文献[21]为代表，指通过删除输入图像的信息来减少数据的过拟合，以增强模型对图像局部空间特征的学习。本文针对深度伪造人脸的特殊性提出 3 种基于信息删除的数据增强方式。

(1)删除面部关键部位信息(Dropout Facial Key Parts，DFKP)

首先利用 MTCNN 人脸关键点提取器提取人脸图像面部 5 个关键点(双眼含两个关键点、鼻、嘴含两个关键点)，其坐标依次记为 $(X_i,Y_i)(i=1,2,3,4,5)$，并对原数据集各取 1/3 做随机删除眼部、鼻部、嘴部信息的操作，作为第一种扩充方式，如图 3 所示。

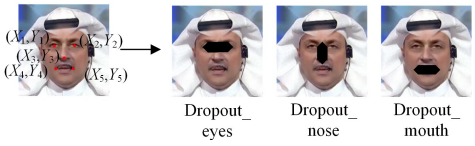


图 3 删除面部关键部位信息

Fig. 3 Dropping information about key facial parts

定义 (X_m,Y_m) 和 (X_n,Y_n) 为平面空间的两点，对于 3 个部位的信息删除方式，(X_m,Y_m) 依次取 (X_1,Y_1)，$\left(\frac{X_1+X_2}{2},\frac{Y_1+Y_2}{2}\right)$，$(X_4,Y_4)$，对应 (X_n,Y_n) 取 (X_2,Y_2)，(X_3,Y_3)，(X_5,Y_5)，以两点欧氏距离为长，距离的 1/4 为宽做计算机视觉的二值膨胀操作，将膨胀区域的像素值全部设为 0，以达到信息删除的目的，宽度计算如式(1)所示：

$$width=\frac{\sqrt{(X_m-X_n)^2+(Y_m-Y_n)^2}}{4} \tag{1}$$

(2)删除面部凸包信息(Dropout Facial Convex Hull，DFCH)

使用 shape_predictor_68_face_landmarks 人脸关键点定位器定位人脸 68 个关键点，并取最外围的 27 个关键点，找到其图像中心 (X_1,Y_1)，并以该坐标为分割线做图像分割，水平分割线与上下面部关键点分别组成面部上部凸包和面部下部凸包，垂直分割线与左右面部关键点分别组成面部左部凸包和面部右部凸包。如图 4 所示，将 4 种凸包区域像素值设为 0，以达到信息删除的目的。

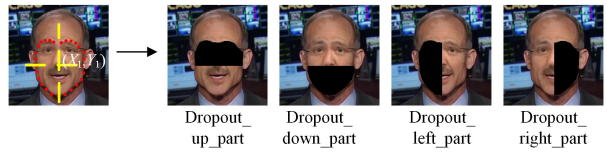


图 4 删除面部凸包信息

Fig. 4 Dropping information about facial convex hull

对原数据集各取 1/4 做随机删除 4 种面部凸包信息的操作，作为第二种扩充方式。

(3)删除图像矩形信息(Dropout Image Rectangle，DIR)

如图 5 所示，以图片宽度的一半对图像做矩形分割，分割为 4 个区域，并保持水平线距顶端的距离也为图片宽度的一半。

图 5 删除图像矩形信息

Fig. 5 Dropping information about image rectangle

第一种形式将(1)(3)区域像素值设为 0，使模型学习伪造人脸局部特征；由于伪造区域常为人脸面部中心，第二种形式将(1)(2)区域像素值设为 0，能够让模型更关注伪造部分与额头真实部分交界处的像素差异，也能让模型充分学习伪造部分与脸部两侧及脸部下方的空间分布关系；第 3 种形式将(1)(4)区域像素值设为 0，使模型学习左下伪造区域与右上未伪造区域的空间分布关系，增加数据集多样性。对原数据集各取 1/3 做随机删除 3 种图像矩形信息的操作，作为第三种扩充方式。

3.1.3 其他数据增强方式

在通过 3 种信息删除方式扩充数据集后，进一步对原数据集和扩充后的数据集采取其他数据增强操作，以减少模型的过拟合。

通过适当改变图像的色相饱和度值，以进一步丰富数据集的颜色通道；通过锐化使部分图像变得清晰；对图像添加高斯模糊，以降低图像噪声；对图像做数据标准化与归一化，实现数据中心化，增强模型的泛化能力。

3.2 改进的 ResNet34 模型

He 等[22] 提出的 ResNet 曾获 2015 年 ImageNet 视觉识别竞赛冠军。ResNet 在普通卷积结构中引入残差机制，通过堆叠残差结构使神经网络能够在输入特征的基础上更容易学习到新的特征，同时减少了神经网络在反向传播时梯度直接传递的次数，从而可以在一定程度上加深神经网络的层数。ResNet 家族主要有 ResNet18，ResNet34，ResNet50，ResNet152 等。

为提升分类效率，本文使用参数量较少的 ResNet34 进行实验。如图 6 所示，图 6(a)给出了 ResNet34 的 Conv3_x 的

虚线残差结构，通过设置 *Stride*=2 的下采样层使特征图尺寸减半；图 6(b)给出了 ResNet34 的 Conv3_x 的实线残差结构，在保持特征图尺寸不变的前提下进行特征提取。

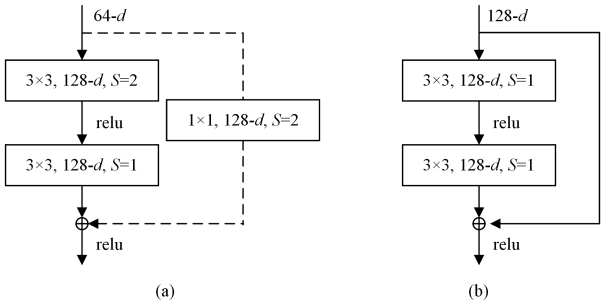


图 6 ResNet34 的 Conv3_x 结构

Fig. 6 Conv3_x structure of ResNet34

原 ResNet34 模型无法充分提取伪造人脸面部更深层次的特征，不能有效区分真假人脸关键特征，且模型泛化能力不足。因此，本文从采用分组卷积、使用最大池化层做下采样、引入通道注意力层 3 个方面对原模型进行改进。在模型主干网络首先使用分组卷积代替普通卷积，以提取更丰富的人脸面部特征图，待连接好各卷积分组后引入通道注意力机制，增加提取到伪造人脸面部细节特征通道的权重，同时在模型 shortcut 分支使用最大池化层代替普通卷积做下采样，减少图像像素信息的损失，最大化保留模型提取到的能够区分真假人脸的关键特征，最终将模型主干网络和 shortcut 分支获得的特征图做矩阵求和运算，作为模型的最终输出。改进后的i_ResNet34(Improved_ResNet34)的 i_Conv3_x(Improved_Conv3_x)的残差结构如图 7 所示。

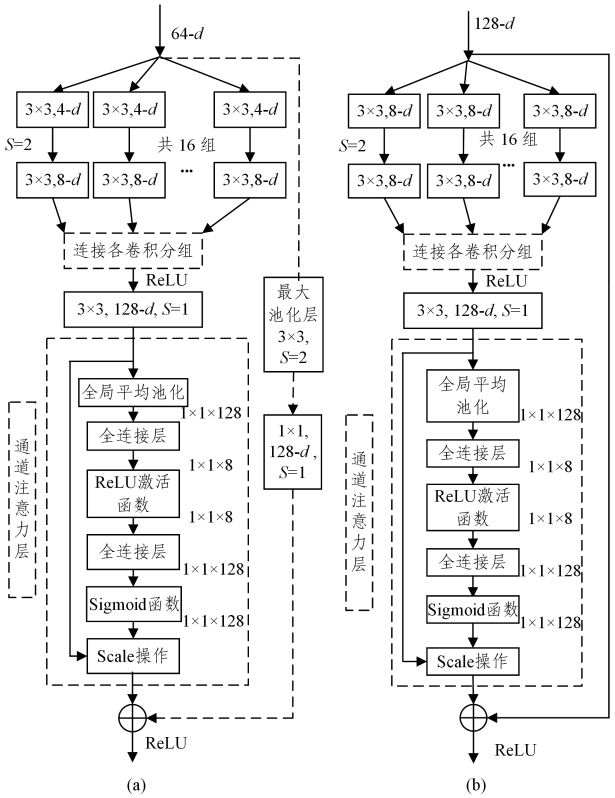


图 7 i_ResNet34 的 i_Conv3_x 结构

Fig. 7 i_Conv3_x structure of i_ResNet34

3.2.1 分组卷积

Xie 等[23]采用分组卷积(Group Convolution，GC)对 ResNet50 进行优化，以提升网络性能。

采用分组卷积前后特征提取的变化情况如图 8 所示。图 8(a)为普通卷积操作，每个卷积核做卷积操作时的运算量为 $k^2\times C\times W'\times H'$；图 8(b)为分组卷积操作，将特征图和卷积核尺寸均分为 g 组，各卷积核在对应组内做卷积操作，运算量为 $k^2\times\frac{C}{g}\times W'\times H'\times g$，与普通卷积相同。因此，采用分组卷积可以在不增加卷积运算量的前提下生成做普通卷积操作 g 倍的特征图，从而提取更丰富的图像特征。

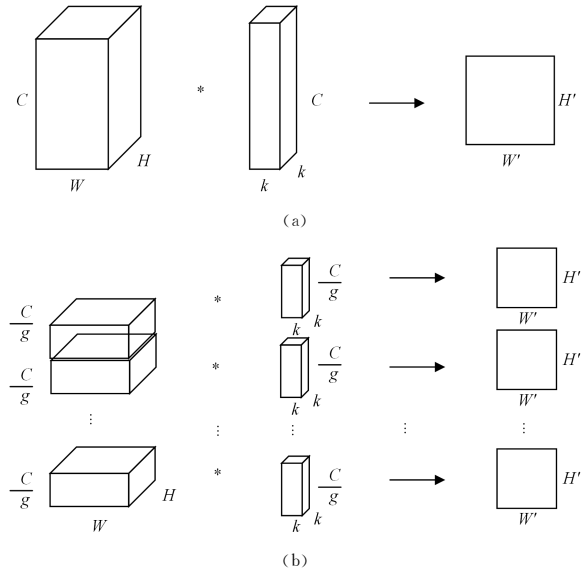


图 8 采用分组卷积前后特征提取的变化情况

Fig. 8 Feature extraction changes before and after using group convolution

深度伪造人脸与真实人脸图像十分相似，仅靠肉眼观察是无法准确区分的。原模型在提取人脸面部特征时获得的卷积特征图过于单一，无法为模型后续检测工作提供有效依据。为此，本文将分组卷积用于对 ResNet34 模型的改进，使模型在中间层生成更丰富的真假人脸卷积特征图，以更充分地描述真假人脸特征差异，从而为检测深度伪造人脸提供更充分的中间数据支撑。

为便于整合好的卷积分组与后续的通道注意力层更好地连接，本文仅在每次卷积操作时做一次卷积分组，待连接好各卷积分组后再做一次普通卷积操作。

经实验验证分析，最终选取的分组组数为 16。由于特征图经过 Conv2_x 通道数不发生变化，因此 ResNet34 的 Conv2_x 到 Conv5_x 的虚线残差结构每个分组卷积核的 in_channel，out_channel 分别为 4,4,8,16 和 4,8,16,32，实线残差结构每个分组卷积核的 in_channel，out_channel 均为 4,8,16,32。

3.2.2 下采样引入最大池化层

在 ResNet34 由卷积结构 Conv*i*_x 向 Conv(i+1)_x(i=2,3,4)转变时，先使用虚线残差结构做下采样，使特征图减半，再使用实线残差结构做特征提取。

如图 9(a)所示，原模型的虚线残差结构通过在 shortcut 分支设置步长为 2 的 1×1 卷积核完成下采样操作。虽然下

采样操作能够通过减小图片尺寸来降低后续计算量，但在使用 1×1 卷积核完成步长为 2 的卷积操作时仅能够考虑特征图部分像素信息，会忽略对其他关键特征的提取，从而造成 3/4 的图像像素信息损失，如图 10(a)所示。深度伪造人脸与真实人脸外观极为相似，差异仅存在于伪造区域与原人脸拼接的部位，因此仅考虑 1/4 的特征图像素信息使得模型无法有效提取到人脸更深层次的空间像素特征，导致检测准确率下降。

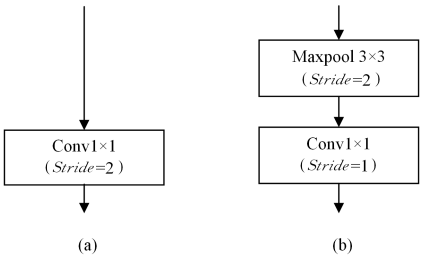


图 9 改进前后虚线残差结构 shortcut 支路的变化情况

Fig. 9 Changes of shortcut branch of dashed residual structure before and after improving model

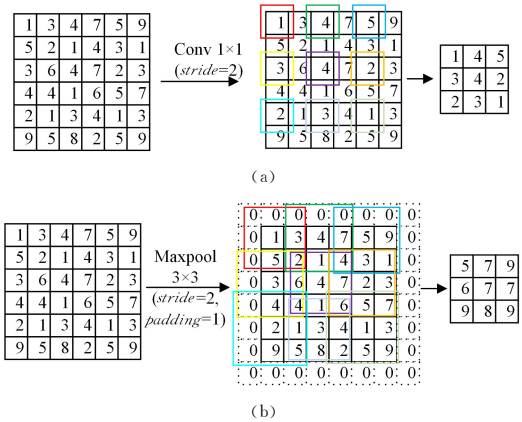


图 10 改进前后提取特征图像素信息变化情况

Fig. 10 Changes of extracting feature map pixel information before and after improving model

为此，本文在原模型虚线残差结构 shortcut 分支使用[1×1，S=1]的卷积核代替原[1×1，S=2]的卷积核，并在卷积核前增加一个[3×3，S=2]的最大池化层，代替卷积核做下采样。实线残差结构 shortcut 支路保持不变，如图 9(b)所示。

改进后的特征图像素信息提取情况如图 10(b)所示。最大池化层能够充分考虑特征图的每个像素点，并保留卷积核提取到的某个面部特征(眼、鼻子、嘴巴)的最大值，通过进一步强化该关键特征，使模型更充分地学习伪造人脸和真实人脸的特征差异，从而减少人脸图像像素信息损失，以增强平移不变性，提升网络识别性能。之后再通过[1×1，S=1]的卷积操作改变特征图的通道数，实现特征图空间尺寸变换操作和通道数变换操作的分离。

3.2.3 通道注意力机制

如式(2)所示，设映射 $F_{tr}:\boldsymbol{X}\rightarrow U$ 表示经传统卷积操作后特征图的变化。令 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_C]$ 表示 C 个卷积核集合，$\boldsymbol{v}_c=[\boldsymbol{v}_c^1,\boldsymbol{v}_c^2,\cdots,\boldsymbol{v}_c^{C'}]$ 为第 c 个卷积核，$\boldsymbol{X}=[\boldsymbol{x}^1,\boldsymbol{x}^2,\cdots,\boldsymbol{x}^{C'}]$，$\boldsymbol{X}\in R^{H'\times W'\times C'}$，$U\in R^{H\times W\times C}$，$\boldsymbol{u}_c\in R^{H\times W}$，$*$ 代表卷积操作。传统卷积操作的每个卷积核在提取特征时仅将原特征图的各个通道卷积结果做简单相加，导致各个通道之间的特征关系与卷积核学习到的空间关系混淆在一起。

$$\boldsymbol{u}_c=\boldsymbol{v}_c*\boldsymbol{X}=\sum_{s=1}^{C'}\boldsymbol{v}_c^s*\boldsymbol{x}^s \tag{2}$$

为解决此问题，SENet[24] 通过 Squeeze(sq)和 Excitation(ex)操作增加表征能力强的通道的权重，使模型提取特征的指向性增强。如式(3)所示，sq 操作利用全局平均池化将特征图通道的空间特征平均为一个值，使特征图通道的相关性增强，空间分布的相关性减弱。

$$z_c=F_{sq}(\boldsymbol{u}_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j),z\in R^C \tag{3}$$

ex 操作如式(4)所示，其中 W_1，W_2 为全连接层，分别起到降维和升维作用；δ 为 ReLU 激活函数，σ 为 sigmoid 函数，以得到各个通道的权重系数，从而在降低复杂度的前提下学习到各个通道之间的非线性关系。

$$s_c=F_{ex}(z_c,W)=\sigma(g(z_c,W))=\sigma(W_2\delta(W_1z_c)) \tag{4}$$

最终将特征图各个通道的原始特征乘以各通道学习到的权重，获得最终的特征图，如式(5)所示，其中 F_{scale} 为特征图与权重的点乘操作。

$$\hat{\boldsymbol{x}}_c=F_{\text{scale}}(\boldsymbol{u}_c,s_c)=s_c\boldsymbol{u}_c \tag{5}$$

原模型在提取人脸面部特征时，仅对各个通道的卷积特征图赋予相同的权重，导致模型无法重点关注能够区分真假人脸的关键区域和关键通道。

为此，本文在卷积操作层后加入通道注意力层，该层通过为特征图每个通道分配不同的权重，增加提取到伪造人脸面部细节特征的通道的权重并进一步强化，使网络关注更重要的通道，抑制表征面部特征能力差的通道，从而将真假人脸图像数据中的关键特征，如眼、鼻等五官特征差异和伪造人脸在面部拼接部分的特征差异标识出来，从而较原模型能够更有效地识别出伪造人脸。

图像矩阵经过通道注意力层后，再与改进的 shortcut 分支提取到的特征图做矩阵求和运算，作为最终的特征输出，如图 7(a)所示。

3.3 融合策略

在模型训练阶段，使用本文提出的 i_ResNet34 模型对原数据集及 3 种基于信息删除的数据增强方式扩充的数据集分别进行训练，得到 4 组权重结果。在测试阶段保持测试集不做数据增强操作，使用 i_ResNet34 模型分别加载 4 组权重进行预测，选取与测试集真实标签的交叉熵损失(CrossEntropy Loss)最小的结果作为最终结果。融合策略如图 11 所示。

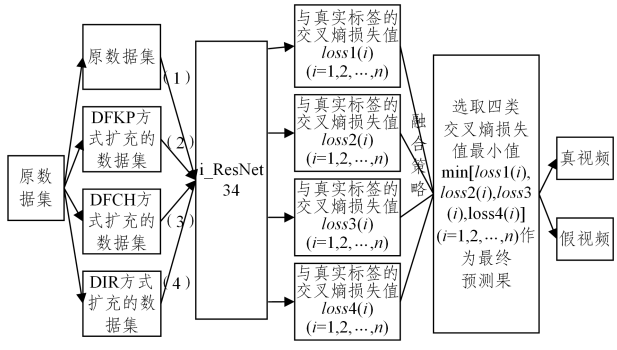


图 11 融合策略

Fig. 11 Fusion strategy

4 实验结果与分析

4.1 实验环境

本文实验环境如表 1 所列。

表 1 实验环境

Table 1 Environment of experiment

实验环境	版本
操作系统	Ubuntu 16.04 64bit
内存/GB	16
CPU	2.2GHz Intel(R)
GPU	Nvidia 1080Ti 11G
Anaconda3	4.7.10 版本 64bit
Pytorch	1.5.0

4.2 数据集介绍

FaceForensics++数据集[25]选取 YouTube 上的 1000 个公开的源人脸视频，每个视频 10 s 到 15 s 不等，采取 FaceSwap，Deepfakes，Fac2Face，NeuralTextures 4 种方式生成伪造视频。其中 FaceSwap 和 Deepfakes 为人脸身份替换，Fac2Face 和 NeuralTextures 为人脸表情替换。由于深度伪造常指人脸身份替换，因此本文仅对 FaceSwap 和 Deepfakes 两类数据集各 1000 个视频进行测试。

FaceForensics++数据集为模拟视频在网络中传输存在压缩的情形，使用 H.264 编码方式对原样本集做不同程度的压缩。Raw 表示原始视频，C23 表示压缩程度为高质量(high quality)，C40 表示压缩程度为低质量(low quality)。由于 Raw 版本数据集所占存储空间过大，且现有算法对于 Raw 版本数据集的检测准确率均比较高，而 C40 版本数据集视频分辨率过低，均不利于实验，因此本文实验选择 C23 版本数据集，并按 7:1:2 的比例将数据集划分为训练集、验证集、测试集。

4.3 实验设置

为实现实验的公平比较，针对两种不同的数据集分别训练一个二分类器，用于区分结果为真视频或深度伪造视频。

在将训练好的模型权重用于测试数据样本时，在原模型的全连接层后加入 Softmax 函数，将模型输出的 1 维二分类向量各元素的数值转换到(0,1)区间内，且所有元素和为 1，以便于更直观地观察分类结果，如式(6)所示。

$$\mathrm{Softmax}(z_i)=\frac{e^{z_i}}{\sum_{c=1}^{C}e^{z_c}} \tag{6}$$

其中，z_i 为全连接层第 i 个节点的输出，C 分类数。本文中 $C=2$，设真视频标签为 0，假视频标签为 1，以 0.5 为阈值评估分类效果。

本文实验使用准确率(Accuracy，Acc)来评价模型分类的精度，使用交叉熵损失函数(CrossEntropy Loss)来评价模型分类结果与实际标签之间的误差。准确率的计算公式如式(7)所示：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

其中，TP 代表实际是真人脸，且被预测为真人脸的数量；FP 代表实际是深度伪造人脸，但被预测为真人脸的数量；FN 代表实际是真人脸，但被预测为深度伪造人脸的数量；TN 代表实际是深度伪造人脸，且被预测为深度伪造人脸的数量。本文只取每个视频的部分帧进行测试，将视频帧的人脸图像预测结果取平均值作为该视频的预测结果，再运用式(7)求出对真假视频的分类准确率。

交叉熵损失的计算公式如式(8)所示：

$$loss(p,q)=-\sum_{i=1}^{n}p(x_i)\log(q(x_i)) \tag{8}$$

其中，$p(x)$ 和 $q(x)$ 分别为样本的真实概率分布与预测概率分布。通过设置 log 函数实现对分类错误样本的惩罚，使得样本预测概率与实际概率的差值越大，惩罚越大。

4.4 实验结果分析

本文实验所有代码全部在 Pytorch 深度学习框架下实现，使用 SGD 算法对模型进行优化，学习率初始化为 0.01，每经过 4 个 Epoch 调整为原来的 0.2，迭代次数 Epoch 为 20。

4.4.1 视频帧数选取

本文实验基于选取的视频帧数展开，选择正确的帧数有助于提升检测准确率。如图 12 所示，经实验验证，对于 300 帧的视频，选取帧数过少会导致样本取样不充分，使模型无法充分学习视频帧的空间特征，而选取帧数过多会导致样本冗余，使模型过拟合。实验在选取 30 帧时检测准确率最高，之后曲线收敛。因此，本实验对每个视频等间隔选取 30 帧作为实验样本。

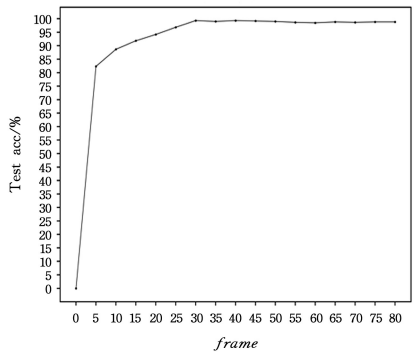


图 12 准确率随视频帧数的变化折线图

Fig. 12 Broken line graph of changes of accuracy with number of video frames

4.4.2 模型改进前后对比

为详细研究每种改进策略产生的性能增益，本文以 ResNet34 模型为基础网络，逐步添加分组卷积模块(GC)、最大池化模块(MP)和通道注意力模块(SENet)，分别计算其在 FaceSwap 和 Deepfakes 数据集上的检测准确率和交叉熵损失，实验结果如表 2 所列。

表 2 每种改进策略产生的性能增益对比

Table 2 Comparison of performance gains generated by each improved strategy

Model				Acc/%		CrossEntropy Loss	
ResNet34	GC	MP	SENet	FS	DF	FS	DF
√	—	—	—	93.67	94.17	0.1823	0.1923
√	√	—	—	94.50	94.83	0.1537	0.1614
√	√	√	—	96.17	95.83	0.1165	0.1376
√	√	√	√	**98.33**	**97.17**	**0.0822**	**0.1135**

由表 2 可知，由于 GC 和 MP 模块只对人脸特征图局部空间关键特征进行强化，因此仅能为模型带来准确率提升

0.6%～1%的性能增益；而通道注意力层对所有通道的特征图予以充分考虑并进行赋予权重的综合决策，因此能为模型带来准确率提升近 2%的性能增益。同时，随着各个模块的叠加，交叉熵损失逐渐减小，这说明模型预测的结果与真实样本的结果的差异越来越小，从另一方面说明了 i_ResNet34 相较于 ResNet34 在每一步改进策略所产生的性能增益。

为了更直观地探讨 3 种改进策略产生的性能增益，本文使用模型提取的中间层特征图和 CAM 热力图进行对比，结果如图 13、图 14 所示。

(a)Conv1×1(*Stride*=2)　　(b)Maxpool3×3(*Stride*=2)

图 13　改进下采样操作前后模型中间层特征图的变化情况

Fig. 13　Changes of feature map of middle layer of model before and after improving down sampling

图 14　每种改进策略产生的 CAM 热力图性能增益(电子版为彩色)

Fig. 14　CAM heatmaps performance gain generated by each improved strategy

图 13(a)和图 13(b)分别表示采用[1×1，S=2]的卷积核和[3×3，S=2]的最大池化层做下采样操作。可以直观地看出，图 13(b)较图 13(a)提取的人脸面部特征图像素更高，且能够明显突出人脸面部五官特征，从而验证了、了模型改进之后能够有效减少人脸图像像素信息丢失，并最大化保留能够区分真假人脸的关键面部特征。

图 14 使用类激活映射(Class Activation Mapping，CAM)可视化每种改进策略产生的性能增益，红色部分表示模型侧重关注的区域。原 ResNet34 模型由于更多地关注图像的背景信息，从而忽略了对人脸面部特征的学习，导致检测准确率偏低。在使用分组卷积(GC)代替普通卷积后，能够使模型在中间过程生成更丰富的人脸特征图，因此模型能够在学习过程中逐渐将关注区域转移到人脸面部，但仍存在少量背景噪声。深度伪造技术在拼接目标人脸和源人脸时常会在眉毛上部留下拼接痕迹，且伪造人脸的制作依赖生成网络的训练，因此得到的面部五官空间分布较真实人脸存在差异，同时生成的五官也会出现局部视觉伪影或扭曲等不规则形状。当继续增加通道注意力层后，模型对提取到人脸面部关键特征的通道赋予更高的权重，使得模型的感兴趣区域进一步转移到人脸的五官(眼、鼻、嘴)和能够区分出真假人脸的拼接部位上来，从而使得检测准确率大幅提升。

此外，为实现公平比较，本文以 ResNet34 模型为基础，与本文改进后的 i_ResNet34 模型、目前在 FaceForensics＋＋数据集测试准确率最高的 Xception 模型[26]，以及对 i_ResNet34 使用 3 种基于信息删除的数据增强方式扩充的样本学习后的 i_ResNet34-DA(i_ResNet34-Data Augmentation)方法在两个数据集上展开对比实验，准确率和交叉熵损失结果如图 15－图 18 所示。

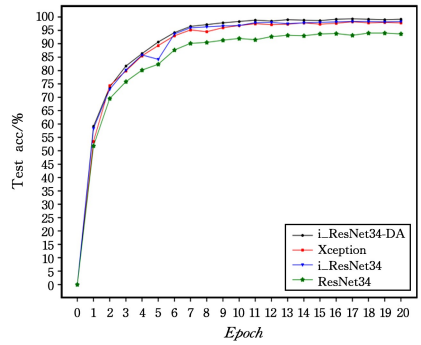


图 15　4 种方法的准确率在 FaceSwap 数据集上的对比实验结果

Fig. 15　Experimental results of accurancy of four methods on FaceSwap dataset

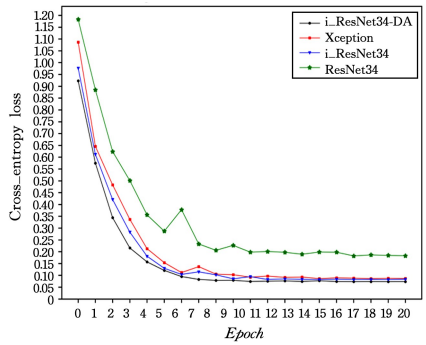


图 16　4 种方法的交叉熵损失在 FaceSwap 数据集上的对比实验结果

Fig. 16　Experimental results of CrossEntropy loss of four methods on FaceSwap dataset

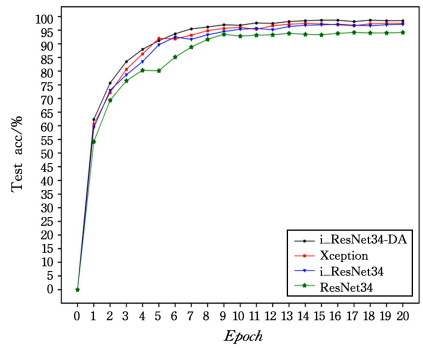


图 17　4 种方法的准确率在 Deepfakes 数据集上的对比实验结果

Fig. 17　Experimental results of accurancy of four methods on Deepfakes dataset

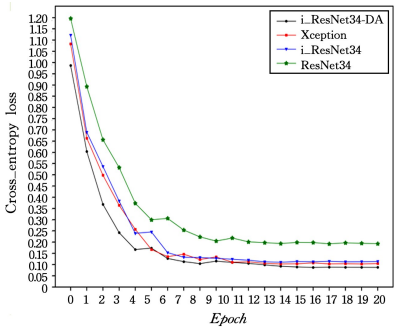


图 18 4 种方法的交叉熵损失在 Deepfakes 数据集上的对比实验结果

Fig. 18 Experimental results of CrossEntropy loss of four methods on Deepfakes dataset

从实验结果可以观察到：1）根据 ResNet34 和 i_ResNet34 的结果对比，证明了本文从 3 个角度对原模型进行改进后，检测准确率提升了近 4.5%。2）根据 i_ResNet34 和 Xception 的结果对比，对于 FaceSwap 数据集而言，i_ResNet34 在训练时间为 Xception 的一半且参数量仅为 Xception 的 0.4 的情况下，能够在准确率上超过 Xception，且预测的损失值低于 Xception，从而在节约时间和参数的前提下提升了检测效果；对于 Deepfakes 数据集，虽然改进的 i_ResNet34 检测准确率略低于 Xception，但与其基本持平，使短时间内检测出大多数深度伪造视频成为可能。3）根据 i_ResNet34，Xception 和 i_ResNet34-DA 的结果对比，证明了模型对扩充的样本集进行学习能够提取到人脸面部更丰富的空间特征，从而成为最优的检测方法。最终，本文提出的 i_ResNet34-DA 方法相比 Xception 模型准确率能够提升近 1.2%，且预测结果与真实标签损失值更小，更接近真实样本标签，从而验证了本文所提方法的有效性。

另外，本文针对 3 种扩充数据集的方式以及不同的融合策略进行了对比实验，结果如表 3 所列。其中，Original，DFKP，DFCH 和 DIR 分别表示 i_ResNet34 对原数据集和 3 种扩充后的数据集进行学习，√ 表示不同程度的融合策略，FS 和 DF 分别代表 FaceSwap 和 Deepfakes 数据集。由实验结果可以得出，i_ResNet34 在原数据集训练的基础上，再对任意一个扩充后的数据集进行训练都能够带来约 0.3%～0.7%的检测准确率的性能增益，且至少检测准确率不会降低。i_ResNet34 同时对原数据集及 3 种扩充方式得到的数据集进行训练时，能够得到最优的检测效果。

表 3 3 种扩充数据集的方式以及不同结合策略的对比实验结果

Table 3 Comparison of experimental results of three ways expanding dataset with different combination strategies

Xception Net	i_ResNet34				Acc/%		CrossEntropy Loss	
	Orignal	DFKP	DFCH	DIR	FS	DF	FS	DF
√	—	—	—	—	98.17	97.50	0.0861	0.1027
	√	—	—	—	98.33	97.17	0.0822	0.1135
	√	√	—	—	98.67	97.83	0.0769	0.0952
	√		√	—	98.50	98.17	0.0773	0.1044
	√	—	—	√	98.33	98.00	0.0781	0.1037
	√	√	√		98.67	98.33	0.0745	0.0898
	√	√	—	√	99.00	98.17	0.0736	0.0916
	√	—	√	√	99.17	98.50	0.0748	0.0903
	√	√	√	√	**99.33**	**98.67**	**0.0732**	**0.0874**

4.4.3 与其他主流算法的对比

本文实验选取其他前沿的深度伪造视频检测方法进行比较，包括 Durall R[7]，MesoNet[9]，Sabir E[14]，Xception[26]，实验对比结果如表 4 所列，其结果证明了利用 CNN 并采取对数据集多样性扩充的方法能够使算法更充分地学习伪造人脸面部特征，优于其他算法。

表 4 各方法在两类数据集上的准确率

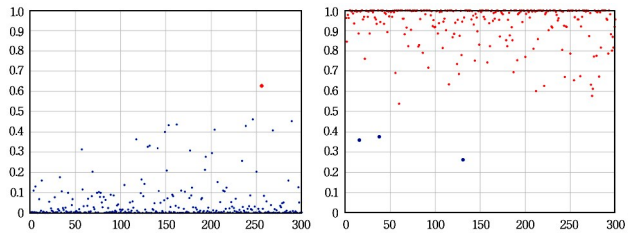
Table 4 Accuracy of each method on FaceSwap and Deepfakes

（单位：%）

	FaceSwap	Deepfakes
Durall 等[7]	90.00	90.40
MesoNet[9]	93.43	95.26
Sabir 等[14]	96.30	96.90
Xception[26]	98.17	97.50
i_ResNet34-DA	**99.33**	**98.67**

4.4.4 分类效果

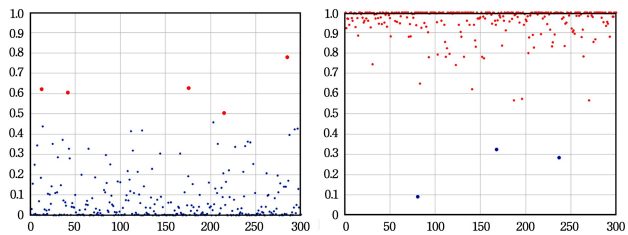
以 0.5 为阈值，将 Softmax 函数的分类结果转化为 0-1 散点分布，其中，真视频的区间为[0，0.5]，伪造视频的区间为[0.5，1]，将本文所提方法（i_ResNet34-DA）的预测结果以散点图的形式直观地展现出来，将分类错误的样本用不同颜色的点表示出来，如图 19、图 20 所示。可以看出，本文所提方法能够将大部分真视频预测在[0，0.3]区间内，将大部分假视频预测在[0.7，1]区间内，基本实现了对深度伪造视频的准确检测。



（a）对 FaceSwap 数据集 300 个真样本的测试结果　（b）对 FaceSwap 数据集 300 个假样本的测试结果

图 19 FaceSwap 数据集的样本测试结果（电子版为彩色）

Fig. 19 Sample test results of FaceSwap dataset



（a）对 Deepfakes 数据集 300 个真样本的测试结果　（b）对 Deepfakes 数据集 300 个假样本的测试结果

图 20 Deepfakes 数据集的样本测试结果（电子版为彩色）

Fig. 20 Sample test results of Deepfakes dataset

结束语 针对现有深度伪造检测方案存在的人脸空间特征提取不够充分、检测准确率不够高以及样本分类效果不够明显的问题，本文从采用分组卷积、使用最大池化层做下采样以及引入通道注意力层 3 个方面对 ResNet34 网络进行改进，同时提出 3 种基于信息删除的数据增强方式。经实验验证，本文所提方法能够对深度伪造人脸面部特征进行充分提取，并较主流检测算法提升了检测准确率和分类效果。

由于深度学习模型对特定数据分布具有依赖性，因此针

对跨数据库的检测会成为深度伪造视频检测未来的难题。

参 考 文 献

[1] BBC Bitesize. "Deepfakes: What are They and Why Would I Make One?" [OL]. http://www.bbc.co.uk/bitesize/articles/zfkwcqt.

[2] BAO Y X, LU T L, DU Y H. Overview of Deepfake Video Detection Technology[J]. Computer Science, 2020, 47(9): 283-292.

[3] KOOPMAN M, RODRIGUEZ A M, GERADTS Z. Detection of Deepfake Video Manipulation[C]// The 20th Irish Machine Vision and Image Processing Conference (IMVIP). 2018: 133-136.

[4] LI J C, LIU B B, HU Y J, et al. Deepfake Video Detection Based on Consistency of Illumination Direction[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 760-767.

[5] MATERN F, RIESS C, STAMMINGER M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]// Proceedings of 2019 IEEE Winter A pplications of Computer Vision Workshops (WACVW). IEEE, 2019: 83-92.

[6] YANG X, LI Y, LYU S. Exposing Deepfakes Using Inconsistent Head Poses[C]// Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8261-8265.

[7] DURALL R, KEUPER M, PFREUNDT F J, et al. Unmasking deepfakes with simple features[J]. arXiv: 1911.00686, 2019.

[8] RAHMOUNI N, NOZICK V, YAMAGISHI J, et al. Distinguishing computer graphics from natural images using convolution neural networks[C]// IEEE Workshop on Information Forensics and Security. 2017: 1-6.

[9] AFCHAR D, NOZICK V, YAMAGISHI J. et al. Mesonet: a compact facial video forgery detection network[C]// IEEE International Workshop on Information Forensics and Security (WIFS'18). 2018: 1-7.

[10] ZHOU P, HAN X, MORARIU V I, et al. Learning Rich Features for Image Manipulation Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1053-1061.

[11] 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 (ICASSP). IEEE, 2019: 2307-2311.

[12] WU X, JIA S J. Face swapping detection based on multi-channel attention mechanism[J/OL]. Computer Engineering: http://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSJC20210309002&v=g9BjGJf5ZLXy78n4jlJAkrMIf9WfK22uyNI%25mmd2FoMZhYd%25mmd2BZyJoAIHxgsHFuBZk4eeGN.

[13] HU Y J, GAO Y F, LIU B B, et al. Deepfake Videos Detection Based on Image Segmentation withDeep Neural Networks[J]. Journal of Electronics & Information Technology, 2021, 43(1): 162-170.

[14] SABIR E, CHENG J, JAISWAL A, et al. Recurrent Convolutional Strategies for Face Manipulation Detection in videos[J]. Interfaces (GUI), 2019, 3: 1.

[15] LI Y, CHANG M C, LYU S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking[C]// 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.

[16] AMERINI I, GALTERI L, CALDELLI R, et al. Deepfake Video Detection through Optical Flow based CNN[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 1205-1207.

[17] ZHENG B W, XIA H W, CHEN R D, et al. Exposing DeepFake Videos Based Convolutional LSTM Network[J/OL]. Laser & Optoelectronics Progress. http://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JGDJ2021031100H&v=2fqOuK4zqEKYz%25mmd2BwfP0UoP60YtASzh6HtS%25mmd2B3KmaItdtD1HZNzgPIh1HjtsAoOg9bl.

[18] 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.

[19] DENG J, GUO J, ZHOU Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv: 1905.00641, 2019.

[20] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 13001-13008.

[21] CHEN P, LIU S, ZHAO H, et al. Gridmask data augmentation [J]. arXiv: 2001.04086, 2020.

[22] HE K M, ZHANG X Y, RENS Q, et al. Deep residual learning for image recognition[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[23] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1492-1500.

[24] 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.

[25] ROSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: Learning to detect manipulated facial images [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1-11.

[26] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017.

BAO Yu-xuan, born in 1997, master. His main research interests include cyber security and artificial intelligence.

LU Tian-liang, born in 1985, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include cyber security and artificial intelligence.