

基于噪声注意力的伪造人脸检测方法

张博林^{1,2,3}, 朱春陶^{1,2,3}, 殷琪林^{1,2,3}, 付婧巧^{1,2,3}, 刘凌毅^{1,2,3},
刘佳睿^{1,2,3}, 刘红梅^{1,2,3}, 卢伟^{1,2,3}

(1. 中山大学计算机学院, 广东 广州 510006; 2. 广东省信息安全技术重点实验室, 广东 广州 510006;
3. 机器智能与先进计算教育部重点实验室, 广东 广州 510006)

摘要: 随着人工智能和深度神经网络的不断发展, 图像生成与编辑变得越来越容易, 恶意运用图像生成工具进行篡改伪造的现象层出不穷, 这对多媒体安全以及社会稳定造成了极大威胁, 因此研究伪造人脸的检测方法至关重要。人脸篡改伪造的方式和工具多种多样, 在篡改的过程中可能留下不同程度的篡改痕迹, 而这在图像噪声中都有一定程度上的反映。从图像噪声的角度出发, 通过噪声去除的方式挖掘反映伪造人脸篡改痕迹的噪声成分, 进一步生成噪声注意力, 指导主干网络进行伪造人脸检测。使用 SRM 滤波监督噪声去除模块的训练, 并将噪声去除模块所得到的噪声再次加入真实人脸图像中, 形成一对有监督的训练样本, 通过自监督的方式对噪声去除模块进行加强指导, 实验结果说明噪声去除模块得到的噪声特征具有较好的区分度。在多个公开数据集上进行了实验, 所提方法在 Celeb-DF 数据集上达到 98.32% 的准确率, 在 FaceForensics++ 数据集上达到 94% 以上的准确率, 在 DFDC 数据集上达到 92.61% 的准确率, 证明了所提方法的有效性。

关键词: Deepfake 检测; 图像噪声; 注意力机制; 篡改痕迹

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023062

Noise-attention-based forgery face detection method

ZHANG Bolin^{1,2,3}, ZHU Chuntao^{1,2,3}, YIN Qilin^{1,2,3}, FU Jingqiao^{1,2,3}, LIU Lingyi^{1,2,3},
LIU Jiarui^{1,2,3}, LIU Hongmei^{1,2,3}, LU Wei^{1,2,3}

1. School of Computer Science and Engineering, Sun Yat-sen University, GuangZhou 510006, China

2. Guangdong Province Key Laboratory of Information Security Technology, GuangZhou 510006, China

3. Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, GuangZhou 510006, China

Abstract: With the advancement of artificial intelligence and deep neural networks, the ease of image generation and editing has increased significantly. Consequently, the occurrence of malicious tampering and forgery using image generation tools is on the rise, posing a significant threat to multimedia security and social stability.

收稿日期: 2022-04-11; 修回日期: 2023-05-26

通信作者: 刘红梅, isslhm@mail.sysu.edu.cn

基金项目: 国家自然科学基金 (U2001202, 62072480)

Foundation Item: The National Natural Science Foundation of China (U2001202, 62072480)

引用格式: 张博林, 朱春陶, 殷琪林, 等. 基于噪声注意力的伪造人脸检测方法[J]. 网络与信息安全学报, 2023, 9(4): 155-165.

Citation Format: ZHANG B L, ZHU C T, YIN Q L, et al. Noise-attention-based forgery face detection method[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 155-165.

Therefore, it is crucial to research detection methods for forged faces. Face tampering and forgery can occur through various means and tools, leaving different levels of forgery traces during the tampering process. These traces can be partly reflected in the image noise. From the perspective of image noise, the noise components reflecting tampering traces of forged faces were extracted through a noise removal module. Furthermore, noise attention was generated to guide the backbone network in the detection of forged faces. The training of the noise removal module was supervised using SRM filters. In order to strengthen the guidance of the noise removal module, the noise obtained by the noise removal module was added back to the real face image, forming a pair of supervised training samples in a self-supervised manner. The experimental results illustrate that the noise features obtained by the noise removal module have a good degree of discrimination. Experiments were also conducted on several public datasets, and the proposed method achieves an accuracy of 98.32% on the Celeb-DF dataset, 92.61% on the DFDC dataset, and more than 94% on the FaceForensics++ dataset, thus proving the effectiveness of the proposed method.

Keywords: Deepfake detection, image noise, attention mechanism, tempering artifact

0 引言

随着移动网络时代的到来以及社交媒体的发展,图像和视频在人们日常生活中随处可见,成为人们获取信息的重要载体。然而,随着图像编辑技术和视频篡改技术的发展,人们常说的“眼见为实”已经成为历史,伪造视觉信息在社交网络和互联网上急剧增加,影响了我们对事物和现象的判断。视觉内容可以很容易地创建和编辑,同时不留下明显的篡改痕迹,人工智能和深度学习技术,尤其是变分自编码器(VAE, variational auto-encoder)^[1]和生成对抗网络(GAN, generative adversarial network)^[2]等生成模型的出现,让虚假图像和视频的生成更加容易并且更加真实,而人脸作为人们的“社交名片”,在人的交流和社会关系中起着重要的作用,所以人脸伪造也受到越来越多关注。然而,伪造技术很可能由于某些恶意的目的被滥用,造成严重的安全和道德问题。因此,为了减轻对公共安全和个人隐私的不利影响,开发有效的解决方案来对抗这些人脸伪造攻击是至关重要的。

现在主流的伪造人脸方法多是基于GAN(如STGAN^[3]、StarGAN^[4]、StyleGAN2^[5]等)的,GAN在合成高质量虚假图像和自然地编辑图像方面表现出强大的能力,人类很容易被这些合成的假图像欺骗,这种技术也被称为深度伪造(Deepfake)。人脸整体合成、人脸属性编辑、人脸表情操作和人脸深度伪造是4种不同的基于GAN的典型假

面孔合成方式。人脸整体合成意味着一个人脸图像可以完全用GAN合成,而被合成的人脸在世界上并不存在。人脸属性编辑主要篡改人脸上的单个多个属性,如头发、眼睛、性别等。人脸表情操作会改变一个人的面部表情或交换多人的面部表情。Deepfake也被称为身份交换,它通常在不同的人之间交换合成的人脸,并广泛应用于制作假视频,是目前公认危害最大、研究最广泛的一类篡改方式^[6],也是本文的研究对象。一些研究者借助深度神经网络学习一个通用的分类器(如DenseNet^[7]、Xception^[8]),通过大量的数据使模型对真假人脸进行建模,从而实现真假人脸分类,这类方法往往不是针对伪造人脸检测专门设计的,可解释性不强。

另一些研究者关注Deepfake本身生成的过程,注意到Deepfake生成结果中存在人工伪影等不完美的信息,这些缺陷凭人眼难以分辨,但是通过机器学习可以较好地进行判别。这些缺陷包括一些生理上的缺陷,如瞳孔颜色不一致^[9],生成视频中的人物没有眨眼、呼吸等特征^[10];在混合边缘处容易留下轮廓伪影^[11];更深层次的如在颜色空间上的差异^[12]、光照不一致带来的伪影^[13]、GAN的使用所带来的指纹印记^[14]等。还有学者从频域特征着手,发掘频域空间中真假人脸的不一致成分,如离散余弦变换(DCT, discrete cosine transform)空间^[15-16]。这些篡改痕迹存在于人脸的各个部位,可以看作伪造人脸的内生特征。然而,相较于人脸本身的语义信息,这些篡改痕迹

较为微弱,神经网络更容易捕捉到的是人脸本身的语义信息,而较难关注到这些对于真假判别来说至关重要的篡改痕迹。一些现有的工作使用诸如富隐写分析模型(SRM, steganalysis rich model)^[17]、DCT 等方法来使神经网络关注到这些篡改痕迹,取得了一定的效果。

在一般的图像去噪任务中,噪声被认为是区别于图像主成分的、需要去除的成分,而在伪造人脸检测任务中,这种区别于人脸本身的特征会在图像噪声中留下痕迹,而图像噪声一定程度上反映了伪造人脸的过程中所引入的篡改痕迹。本文从图像噪声的角度出发,提出了一种基于噪声注意力的伪造人脸检测方法,使神经网络关注到存在篡改痕迹的重点区域。该方法使用噪声去除模块对输入的人脸图像进行噪声去除,同时使用SRM滤波进行监督训练,将去除的噪声经过激活作为注意力来加强对输入图像的引导,让主干网络完成伪造人脸检测。相较过去的一些篡改痕迹挖掘方法,本文通过引入一种自监督的噪声去除训练方式,可以更好地挖掘伪造人脸的篡改痕迹。本文将提取出来的噪声特征再次融合到真实人脸中,形成一对有监督的训练样本,通过自监督的方式让噪声去除模块能够发现真假人脸之间的差异性,挖掘伪造人脸的篡改痕迹。所提的基于噪声注意力的伪造人脸检测方法在 Celeb-DF^[18]数据集上达到了98.32%的准确率,在 Faceforensic++^[19]数据集上也有94%以上的准确率。另外,为了证明得到的噪声特征反映了伪造人脸中的不完美信息,具有区分真假人脸的能力,本文将噪声去除模块得到的噪声特征直接输入主干网络中,在 Celeb-DF 数据集上达到了97.06%的准确率。

1 相关工作

1.1 伪造人脸篡改痕迹

深度伪造经常会产生人类难以识别的纹理,但能较容易地被机器和取证分析识别,如不一致性、背景不规则和 GAN 指纹等。检测一个人行为的波动、生理信号、相干性和视频帧同步都是时间伪影的例子。Agarwal 等^[20]提出了一种将面部识别的静态生物识别与面部表情和头部运动的时间行为生物识别相结合的方法; Nirkin 等^[21]利

用两个网络分别识别人的面孔和上下文(如这个人的头发、耳朵和脖子),通过两者之间的差异来识别假面孔。有研究从 GAN 的不完美设计出发,为区分真脸和假脸提供了信号。McCloskey 等^[22]探索了 GAN 生成器的结构缺陷,关注生成器的归一化过程,这将减少饱和及未曝光像素的频率; Marra 等^[23]提出了 GAN 指纹,之后 Yu 等^[14]研究了 GAN 指纹的图像属性,并将其进行分类。分析生物信号中的伪影也得到了研究人员的突出关注,如视觉音频不一致^[24-25]、视觉不一致(如眨眼频率^[26]、人脸轮廓与背景不一致^[27]、眼睛、牙齿和面部特征^[28])和视频中的生物信号(如心跳节律^[29])。伪造人脸大多存在不同种类的篡改痕迹,这些篡改痕迹存在于人脸的各个部位,表现为人脸纹理缺失、瞳孔颜色不一致、边缘模糊等,在图像噪声中会有一定程度的反映。

1.2 伪造人脸篡改检测技术

为了应对深度伪造技术对社会公共安全的严重威胁,研究者使用不同技术提出了多种伪造人脸篡改检测技术,主要可分为两类:基于传统方法的检测技术和基于深度神经网络的检测技术。

基于传统方法的检测技术往往关注并利用像素之间的相关性,在检测过程中找出真假像素之间的不一致,这类方法往往具有较强可解释性。Rodriguez 等^[30]验证了光响应不一致性(PRNU, photo response non-uniformity)分析方法在检测深度伪造视频方面的有效性,分析显示真实和深度伪造视频之间的平均归一化互相关得分有统计学上的显著差异。Li 等^[31]评估了深度网络生成图像的统计属性,如 HSV 和 YCbCr 颜色空间中相邻像素之间的相关性,用来区分深度伪造图像。基于传统方法的检测技术具有很强的可解释性,参数量也更低。

基于深度神经网络的检测技术是由数据驱动的,使用网络来分析空域特征从而提高检测效果,体现了卷积神经网络(CNN, convolutional neural networks)的有效性。Do 等^[32]使用经典的 CNN 结构 VGG-Net^[33]进行训练来检测伪造图像; Tariq 等^[34]使用了更多的基于 CNN 的模型来检测虚假人脸图像,包括 VGG16、ResNet^[35]、DenseNet^[7]、Xception^[8]等。此外,Nguyen 等^[36]提出了一种基

于胶囊网络的检测方法,将深度伪造结果的异常外观当作关键特征,由胶囊网络输出其判定结果。Sabir 等^[37]将循环卷积和人脸对齐方法相结合,通过对齐的人脸和双向循环神经网络来建模,利用微观、中观和宏观 3 种尺度特征进行篡改检测。Hsu 等^[38]使用双流网络的架构,通过结合改进版的 DenseNet 主干网络和暹罗网络(孪生神经网络 Siamese),得到一个通用伪特征网络(CFFN, common fake feature network)。Gandhi 等^[39]提出了一种通过在深度伪图像中添加对抗性扰动来提高现有深度伪模型性能的方法,同时引入 CNN 正则化和深度图像先验(DIP, deep image prior),提高了基于 CNN(ResNet 和 VGG)的深度伪检测器的鲁棒性。龚晓娟等^[40]融合了通道注意力和空间注意力来让网络捕获伪造人脸的异常特征。Wu 等^[41]使用 Xception 网络来获取图像的空间特征和分析特征,使用 RNN 来挖掘时序特征,最

后将所有提取的信息结合起来进行 Deepfake 检测。Liu 等^[42]利用全局纹理数据,提高了现有 CNN 在识别合成假脸方面的鲁棒性和泛化能力,实验显示所提出的 Gram-Net 对降采样、JPEG 压缩、模糊和噪声等扰动攻击有较显著的抵抗力。基于深度神经网络的检测技术相比基于传统方法的检测技术准确率更高,但是缺乏足够的可解释性。本文从图像噪声的角度出发,使用噪声注意力来加强神经网络对于篡改痕迹的关注。

2 本文方法

本文所提出的伪造人脸检测方法包括噪声去除模块与检测模块两部分,输入人脸通过噪声去除模块得到对应的噪声特征,经过注意力激活模块生成噪声注意力,从而指导主干网络对输入图像进行检测分类,整体框架如图 1 所示。

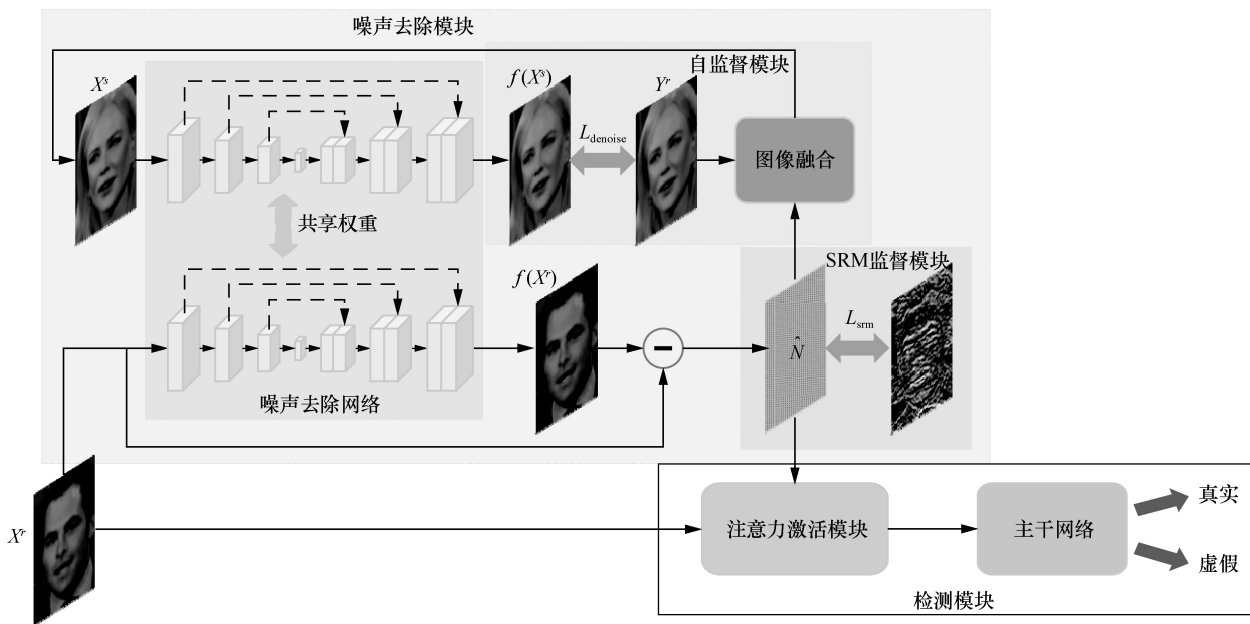


图1 所提基于噪声注意力的伪造人脸检测框架

Figure 1 The proposed noise-attention-based forgery face detection framework

2.1 噪声去除模块

图像噪声是一些捕捉亮度或颜色信息的随机变化的高频信号,它们的分布受到数码相机的图像传感器或电路的影响,因此由不同设备处理或来自不同来源的图像具有不同的噪声模式。噪声可以被视为一种图像的内在特性,存在于很多数

字图像领域。由于篡改操作破坏了原始图像特征的一致性,在噪声空间中往往会留下不同种类的痕迹。伪造人脸也包含和真实人脸不一致的成分,这种不一致包括人脸纹理缺失、瞳孔颜色不一致、边缘模糊等篡改痕迹,这些痕迹会在图像噪声中有所体现,噪声去除模块就是要挖掘出反映伪造

人脸中的篡改痕迹的噪声成分。

2.1.1 噪声去除网络

噪声去除模块的主体是噪声去除网络, 本文使用 U-Net^[43]作为噪声去除网络, 由于其对称的下采样-上采样结构以及跳跃连接的设计, 在医学影像分割和图像去噪方面取得了不错的效果。对于一张输入的人脸图像, 噪声去除网络会输出一张相同尺寸的图像, 反映篡改痕迹的噪声成分大幅减少, 将其视作一张干净的人脸图像, 对输入和输出计算残差就得到了对应的噪声特征, 即

$$\hat{N} = X^r - f(X^r) \quad (1)$$

其中, X^r 是含噪声的输入人脸图像, $f(\cdot)$ 代表噪声去除网络, $\hat{N}$ 就是得到的噪声成分。

2.1.2 SRM 监督模块

伪造人脸检测与一般的图像分类任务不同, 图像的语义内容往往不是本文所关注的, 相反需要抑制图像语义, 让网络关注到更细节的信息, 如图像的边缘、纹理特征等, 希望得到的噪声成分 $\hat{N}$ 也有这样的效果。为此使用 SRM^[17]对输入图像 X_r 进行滤波, 指导噪声 $\hat{N}$ 的训练。SRM 滤波能从图像中提取出高频噪声, 这种高频噪声会去除图像的颜色内容, 同时保留人脸的内在特征, 让噪声特征能够反映篡改痕迹。SRM 滤波的目的是减少 SRM 滤波相似性损失。

$$L_{\text{srn}} = \|f_{\text{srn}}(X^r) - \hat{N}\|_2 \quad (2)$$

L_{srn} 刻画了输入图像的 SRM 滤波和噪声特征的近似程度, 该损失越小代表两者越接近。

随着模型的不训练, 噪声去除模块能够越来越好地保留图像细节同时去除噪声, 得到的噪声中结构信息越来越少, 但 SRM 滤波相似性损失会对这个过程起到阻碍作用, SRM 滤波中的结构信息对得到的噪声会起到负面作用。为了减轻这种负面影响, 逐渐减少 SRM 滤波相似性损失 L_{srn} 的频率, 即计算 L_{srn} 的时间间隔 t 越来越大, 具体表示如下。

$$T = \left\lfloor \frac{t}{500} \right\rfloor + 1 \quad (3)$$

其中, T 表示训练的迭代次数, $\lfloor \cdot \rfloor$ 是向下取整函数。

2.1.3 自监督模块

为了找到真假人脸之间的差异性, 提取出人

脸图像中反映篡改痕迹的噪声成分, 让噪声去除模块有更好的效果, 本文将提取出来的噪声再次加入干净的图像, 即真实人脸 Y^r 中。

$$X^s = Y^r + \hat{N} \quad (4)$$

新生成的人脸 X^s 和原始输入的人脸 X^r 所含的噪声类似, 但是图像内容并不同。更重要的是, X^s 拥有明确的标签 Y^r , 两者形成了一对样本, 从而对噪声去除模块进行有监督的加强训练, 将 Y^r 再次输入噪声去除网络中得到去噪结果 $f(X^s)$ 。这样就能从多角度加强对噪声去除模块的监督, 对应的目标函数为

$$L_{\text{denoise}} = \|f(X^s) - Y^r\|_2 \quad (5)$$

L_{denoise} 反映了新生成的人脸与加入噪声前人脸的相似程度, 网络训练希望减小该损失来提高噪声去除模块的能力。

2.1.4 图像融合

在将提取噪声加入 Y^r 时, 考虑到噪声可能含一些原始输入图像的结构信息, 这样会导致合成的图像质量太低, 甚至可能影响噪声去除模块的性能, 本文使用一个二元掩膜 m 对噪声进行过滤, 之后与真实人脸图像进行融合, 式(4)变为

$$X^s = Y^r + \hat{N} \odot m \quad (6)$$

其中, $\odot$ 代表对应元素相乘, 掩膜 m 的尺寸和噪声 $\hat{N}$ 一致。二元掩膜的每一个元素都是以 0.5 的概率为 1, 0.5 的概率为 -1, 这样就破坏了噪声 $\hat{N}$ 中的残留结构信息从而生成质量更高的图像 X^s 。这样得到的 X^s 有利于噪声去除模块学习噪声去除和图像细节恢复, 另外, 由于二元掩膜是随机生成的, 所以噪声和掩膜的乘积能得到多种噪声, 一定程度上起到了数据增强的作用。

2.2 检测模块

检测模块将得到的噪声特征经过注意力激活生成噪声注意力, 从而指导主干网络对输入人脸图像进行检测分类。

2.2.1 注意力激活模块

噪声去除模块得到的噪声特征包含输入人脸图像中的噪声成分, 反映了其中的可疑篡改痕迹。为了让主干网络更关注到这些痕迹, 将噪声 $\hat{N}$ 经过核大小为 1×1 的卷积层进行信息处理, 使用 Sigmoid 激活函数进行激活形成噪声注意力。

$$M_{\text{att}} = \text{Sigmoid}(f_{\text{conv}}(\hat{N})) \quad (7)$$

得到的噪声注意力 M_{att} 与输入图像对应元素相乘, 通过噪声注意力来指导主干网络对输入人脸图像进行分类, 让主干网络关注到输入人脸图像中的噪声成分。

$$O = f_{\text{bn}}(M_{\text{att}} \odot I + I) \quad (8)$$

其中, I 是注意力激活模块的输入特征, O 是输出特征, f_{bn} 为归一化操作。

2.2.2 主干网络

检测模块的目的是将输入的样本通过一系列处理最终进行分类, 判别真假。这里选择 Xception 作为主干网络, 它的设计基于深度可分离卷积, 即将普通卷积分解为两个过程: 空间卷积和点态卷积。空间卷积在每个输入通道上独立执行, 点态卷积使用核大小为 1×1 的逐点卷积, 这样不仅减少了参数的数量, 而且减少了计算的数量。Xception 由 Entry、Middle、Exit 这 3 个流组成, 在伪造人脸检测方面取得了良好的性能, 常被研究人员用于主干网络。为了减少模型过拟合的风险, 同时减少模型的参数量, 将 Middle 流中原本的 8 个块进行简化, 仅保留 1 个块形成简化版的 Xception 来实现检测分类的任务。

作为二分类问题, 传统分类模型往往使用交叉熵函数作为损失函数, 而本文受到人脸识别领域工作的启发, 使用 AM-Softmax 损失函数^[44]作为分类损失目标函数 L_{cls} , 该函数提出了一种加性间隔, 不仅能够区分不同类的训练样本, 还能够在此基础上缩小类内距离, 增大类间距离, 从而提升模型的表现。最终整个框架的损失函数由三部分构成。

$$L = \lambda_1 L_{\text{srn}} + \lambda_2 L_{\text{denoise}} + \lambda_3 L_{\text{cls}} \quad (9)$$

其中, λ_1 、 λ_2 、 λ_3 为超参数, 训练过程中和网络参数一起更新, L_{srn} 、 L_{denoise} 和 L_{cls} 分别是 SRM 监督模块、自监督模块以及检测模块的损失, 三者共同指导网络的训练。

3 实验及分析

3.1 数据集及实验细节

为了验证所提检测方法的有效性, 本文在公开数据集 FaceForensic++^[19] (以下简称 FF++) 和

Celeb-DF^[18] 上进行实验。

FF++ 数据集从 YouTube 上筛选了 1 000 个真实视频作为样本, 在此基础上使用 FaceSwap、Deepfake、Face2Face 以及 NeuralTextures 这 4 种先进的自动深度伪造方法进行篡改, FF++ 数据集由于数据量大、使用的深度伪造方法多而受到很多研究工作的关注。在压缩级别方面, FF++ 有 3 个版本, 即 C0 无压缩、C23 低压缩(HQ)和 C40 高压缩(LQ), 随着压缩等级的提高, 检测出篡改痕迹的难度也随之加大。无压缩的视频质量有限, 而且在生活中并不常见, 因此本文关注低压缩和高压缩版本。

Celeb-DF 数据集包含了从 YouTube 上筛选的 590 个真实视频和在此基础上对应生成的 5 639 个伪造视频, 该数据集的伪造方法仅限于 DeepFake, 但是数据集的视觉效果相比以前的数据集有了较大提升, 包括将人脸分辨率提升到 256×256 像素; 减少了边缘处的伪影, 更好地融合了伪造人脸的轮廓部分; 建立了真实人脸和伪造人脸的颜色转换算法, 使得合成的效果更好; 减少了合成视频的抖动, 视频在时间轴上变得更加连续。该数据集由于上述改进有了更好的视觉效果, 也被很多研究者用作测试数据集。

本文框架是基于 CNN 搭建的, 在模型实现上使用 U-Net 当作噪声去除模块, 将 Xception 模型的 Middleflow 仅保留一个 block 形成简化版的 Xception, 作为主干网络用于对输入图像进行处理并最终分类。本文使用 PyTorch 库搭建网络模型, 并用 Adam 优化器和大小为 16 的批大小来训练网络, 初始学习率设置为 0.001, 并随着训练轮次增加而衰减。在数据集的处理上, 本文将视频分帧保存, 在此基础上使用 DLIB^[45] 工具来帮助提取面部区域并进行对齐, 形成的人脸区域图像尺寸为 256×256 。实验的评估指标为测试数据集上的准确率 (ACC, accuracy) 以及 (AUC, area under curve), 两者的值越大, 表示分类的效果越好。

3.2 对比实验及分析

本文将所提出的检测方法和一些对比方案进行了比较: Conv^[46] 是一种新型的卷积来约束图像内容, MesoNet^[47] 主要基于图像的中层语义进行检测, Xception^[8] 和 EfficientNet-B4^[48] 都是具有代表性的卷积神经网络框架, 在图像分类上取得了

很好的效果, F3-Net^[16]使用 DCT 估计频率统计量, SPSL^[49]将空间像素和相位谱相结合来捕捉人脸伪造的上采样痕迹。

本文首先在 FF++数据集 (Deepfake 伪造方法) 上的两种不同压缩率下进行实验, C23 代表低压缩, C40 代表高压缩, 检测准确率如表 1 所示。从结果可以看出, 所提方法检测准确率比 MesoNet 高出了 15.37% 和 24.54%, 比 Xception 高出了 2.74% 和 8.15%, 说明所提噪声特征能够反映伪造人脸与真实人脸的不一致性, 得益于自监督模块, 将噪声再次加入真实人脸中进行自监督的训练, 增强了对噪声去除模块的指导, 弥补了高压缩率带来的细节信息损失, 本文所提方法在两种情况下都达到了最好的检测效果。

表 1 不同方法在 FF++数据集上的检测准确率 (ACC)
Table 1 Accuracy of detection on FF++ dataset with different methods

方法	ACC	
	FF++C23	FF++C40
BayarConv	82.97 %	66.84 %
MesoNet	83.10 %	70.47 %
Xception	95.73 %	86.86 %
EfficientNet-B4	96.63 %	86.67 %
F3-Net	97.52 %	90.43 %
SPSL	—	93.48 %
本文	98.47%	95.01%

本文在 Celeb-DF 数据集上也进行了对比实验: C3D (convolutional 3D)^[50]、I3D (inflated 3D convnet)^[51]使用了 3D 卷积, GFFD (generalizing face forgery detection with high-frequency features)^[52]使用了多尺度的高频噪声特征。Celeb-DF 数据集使用了一些后处理手段使得伪造人脸的视觉效果更好, 本文方法能够有效地捕获到可疑篡改痕迹并取得最好的检测效果, 结果如表 2 所示。

同时, 在 DFDC 数据集^[53]上进一步验证所提方法的有效性, 检测结果如表 3 所示。DFDC 数据集包含 1 133 个真实视频和 4 080 个虚假视频, 使用的伪造技术更加丰富多样, 包括 FSGAN、Style GAN 等。DFDC 数据集数量更多且更加生活化, 视觉效果也更好, 人脸所占视频的部分比较小, 较难被检测, 因此更具挑战性。RFM

(representative forgery mining)^[54]使用了基于注意力的数据增强方法, Multi-attention^[55]使用了多注意力指导网络关注重点区域, GFFD 引入了多尺度的高频噪声, 与对比模型相比, 本文方法通过噪声注意力的引入, 取得了最好的检测效果。

表 2 不同方法在 Celeb-DF 数据集上的检测结果
Table 2 Result of detection on Celeb-DF dataset with different methods

方法	ACC	AUC
MesoNet	—	53.6 %
I3D	76.08 %	83.00 %
C3D	78.67 %	84.00 %
Xception	89.55 %	89.91 %
GFFD	94.14 %	98.30%
本文	98.32%	97.77 %

表 3 不同方法在 DFDC 数据集上的检测结果
Table 3 Result of detection on DFDC dataset with different methods

方法	ACC
Xception	88.91 %
Multi-attention	91.02 %
RFM	90.01 %
GFFD	92.33 %
本文	92.61%

3.3 消融实验及分析

本文在 Celeb-DF 数据集上进行了消融实验来探讨各个部分的有效性, 在保持其他结构不变的情况下, 首先去除 SRM 监督模块进行实验, 没有 SRM 的指导, 噪声去除模块不能很好地抑制图像的内容信息, 产生的噪声注意力会对输入图像有所误导, 导致方法性能下降; 接下来去除自监督模块, 也有同样的性能下降现象, 这说明将噪声特征再次加入噪声去除模块的自监督方式提高了噪声去除模块对于篡改痕迹的捕捉能力; 最后去除 SRM 监督模块和自监督模块, 仅保留噪声去除网络和主干网络, 结果证明盲目的噪声去除很难关注到伪造人脸的篡改痕迹, 对于主干网络并没有实质性的提升, 消融实验结果如表 4 所示。

为了进一步证明通过噪声去除的方式得到的噪声特征反映了真假人脸的不一致性, 本文把噪声去除模块得到的噪声特征不经过注意力激活,

直接当作主干网络的输入进行分类检测，主干网络还是用简化的 Xception 网络。在 Celeb-DF 数据集上进行了实验，作为对比，本文将原始人脸图像经过 SRM 滤波后的结果当作主干网络的输入，结果显示仅用 SRM 滤波得到的特征不具有区分度，而通过噪声去除得到的噪声特征体现了伪造人脸的篡改痕迹，具有较高的区分度，实验结果如表 5 所示。

表 4 消融实验结果
Table 4 Result of ablation experiment

方法	ACC
无 SRM 监督模块	96.96 %
无自监督模块	96.77 %
无 SRM 监督模块和自监督模块	96.03 %
本文	98.32 %

表 5 特征对比实验结果
Table 5 Result of feature comparison experiment

方法	ACC
SRM 滤波	96.43 %
噪声特征	97.40 %

与此同时，将这两种特征进行可视化，以进一步说明本文所提的噪声特征具有区分真假人脸

的特性。虚假与真实人脸的特征对比如图 2 所示，本文选取了两张虚假人脸与两张真实人脸放在第一列，第二列是人脸经过 SRM 滤波所得到的特征图，第三列是本文提取的噪声特征，由于特征图有 3 个通道，为了更直观地比较，将每个通道单独提出并进行展示。从图 2 中可以看出，SRM 滤波能够提取图像的高频信息，抑制语义内容，但对于真假人脸提取的特征相差不大。而本文经噪声去除网络提取的噪声特征对于真假人脸所得到的特征相差较大，更具有区分性。所以使用 SRM 滤波结果作为主干网络的输入并没有取得很好的效果，这也印证了特征对比实验的结果，说明了本文基于噪声特征生成注意力来指导主干网络进行检测的有效性。

3.4 模型表现及分析

为了进一步表明所提方法的有效性，本文在不同数据集下进行了实验，并测试虚警率（FPR, false positive rate）和漏检率（FNR, false negative rate），其中虚警率代表分类器正确预测正样本纯度的能力，即负样本被预测为正样本占总负样本的比例，漏检率代表模型正确预测负样本纯度的能力，即正样本被预测为负样本占总正样本的比例，两者值越小，表示模型性能越好。从表 6 可以

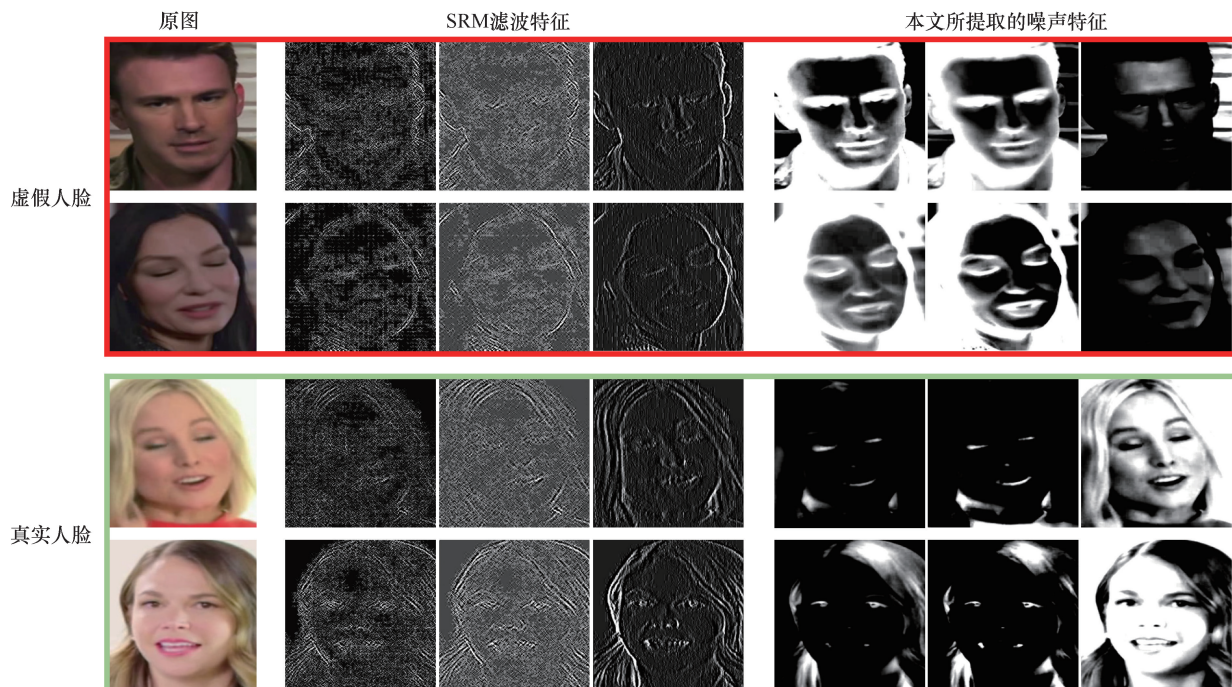


图 2 虚假与真实人脸的特征对比
Figure 2 Comparison of features between fake and real faces

看出, 本文所提方法在不同数据集以及不同篡改方式下都能取得较好的检测效果, 在 Deepfake 上效果最好, 达到了 98.47%, 在 Face2Face 和 FaceShifter 数据集上分别达到了 98.33% 和 98.25%, 在 Celeb-DF 数据集上达到了 98.32%, 在 NeuralTextures 数据集上, 由于篡改方式局限于人脸的嘴部区域, 伪造区域较小从而更具挑战性, 仅达到 92.93%。同时可以看到, 各种数据集上的虚警率和漏检率都保持在较低的水平, 在实际应用中也很重要。

同时, 本文对所提方法的效率进行了分析, 由于不同机器环境会影响到模型的训练时间以及检测时间, 且会受到机器当前任务量多少的影响, 所以直接统计模型的训练时间和检测时间不具有很高的参考价值, 为此, 本文统计了模型的参数量以及计算量, 其中参数量就是网络模型中需要训练的参数总数, 计算量即 FLOP 每秒浮点运算的次数, 两者能反映出网络模型的运行效率。如表 7 所示, 本文所提方法参数量仅有 9.72 MB, 比 Xception 少了 $\frac{1}{2}$, 不到 F3-Net 的 $\frac{1}{4}$, 计算量上也仅有 5.08 GB, 但是取得了最好的检测效果, 体现了所提方法的高运行效率。

表 6 所提方法在不同数据集上的检测结果
Table 6 Detection results of the proposed method on different datasets

数据集	ACC	FPR	FNR
Deepfake	98.47 %	1.73 %	1.33 %
Face2Face	98.33 %	2.81 %	0.52 %
FaceShifter	98.25 %	2.33 %	1.17 %
NeuralTextures	92.93 %	9.55 %	5.07 %
Celeb-DF	98.32 %	0.36 %	4.10 %
DFDC	92.61 %	7.95 %	7.29 %

表 7 不同方法的计算开销比较结果
Table 7 Comparison results of computational overhead of different methods

方法	参数量	FLOP	ACC
Xception	20.81 MB	8.44 GB	95.73%
F3-Net	41.63 MB	16.91 GB	97.52%
本文	9.72 MB	5.08 GB	98.47%

4 结束语

本文提出了一种基于噪声注意力的伪造人脸

检测方法, 通过噪声去除的方式来得到反映人脸图像篡改痕迹的噪声成分, 在此基础上生成噪声注意力来指导简化的主干网络关注到篡改痕迹突出的区域, 更好地检测伪造人脸; 还使用 SRM 滤波来监督噪声去除模块的训练, 并引入一种自监督的噪声去除训练方式加强噪声去除模块的训练。实验结果表明, 噪声去除模块能够得到具有区分度的噪声特征, 提高了网络的分类能力, 所提出的伪造人脸检测方法在多个公开数据集上取得了很好的检测性能。未来, 将进一步研究如何提高噪声特征的鲁棒性, 以适用于更多实际应用场景。

参考文献:

- [1] XU W, KESHMIRI S, WANG G. Adversarially approximated auto encoder for image generation and manipulation[J]. IEEE Transactions on Multimedia, 2019, 21(9): 2387-2396.
- [2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural Information Processing Systems. 2014: 27.
- [3] LIU M, DING Y, XIA M, et al. Stgan: a unified selective transfer network for arbitrary image attribute editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3673-3682.
- [4] CHOI Y, CHOI M, KIM M, et al. Stargan: unified generative adversarial networks for multi-domain image-to-image translation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8789-8797.
- [5] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8110-8119.
- [6] 王鹏程, 郑海斌, 邹健飞, 等. 面向商用活体检测平台的鲁棒性评估[J]. 网络与信息安全学报, 8(1): 180-189.
WANG P C, ZHENG H B, ZOU J F, et al. Robustness evaluation of commercial liveness detection platform[J]. Chinese Journal of Network and Information Security, 2022, 8(1): 180-189.
- [7] WAHEED A, GOYAL M, GUPTA D, et al. An optimized dense convolutional neural network model for disease recognition and classification in corn leaf[J]. Computers and Electronics in Agriculture, 2020, 175: 105456.
- [8] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.
- [9] MATERN F, RIESS C, WINTER M S 2019 I, et al. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). 2019: 83-92.
- [10] GÜERA D, DELP E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE International Conference on

- Advanced Video and Signal Based Surveillance (AVSS). 2018: 1-6.
- [11] NIRKIN Y, WOLF L, KELLER Y, et al. Deepfake detection based on the discrepancy between the face and its context[J]. arXiv Preprint arXiv:2008.12262, 2020.
 - [12] MCCLOSKEY S, ALBRIGHT M, ACST H. Detecting GAN-generated imagery using color cues[J]. arXiv Preprint arXiv:1812.08247, 2018.
 - [13] PENG B, WANG W, DONG J, et al. Optimized 3D lighting environment estimation for image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2016, 12(2): 479-494.
 - [14] YU N, DAVIS L S, FRITZ M. Attributing fake images to GANs: Learning and analyzing GAN fingerprints[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7556-7566.
 - [15] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//International Conference on Machine Learning. 2020: 3247-3258.
 - [16] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//European Conference on Computer Vision. 2020: 86-103.
 - [17] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
 - [18] LI Y, YANG X, SUN P, et al. Celeb-df: a large-scale challenging dataset for deepfake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3207-3216.
 - [19] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. Faceforensics++: learning to detect manipulated facial images[C]//International Conference on Computer Vision (ICCV). 2019: 1-11.
 - [20] AGARWAL S, FARID H, EL-GAALY T, et al. Detecting deepfake videos from appearance and behavior[C]//2020 IEEE International Workshop on Information Forensics and Security (WIFS). 2020: 1-6.
 - [21] NIRKIN Y, WOLF L, KELLER Y, et al. Deepfake detection based on discrepancies between faces and their context[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, (8).
 - [22] MCCLOSKEY S, ALBRIGHT M. Detecting GAN-generated imagery using saturation cues[C]//2019 IEEE International Conference on Image Processing (ICIP). 2019: 4584-4588.
 - [23] MARRA F, GRAGNANIello D, VERDOLIVA L, et al. Do GANs leave artificial fingerprints?[C]//2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). 2019: 506-511.
 - [24] AGARWAL S, FARID H, Fried O, et al. Detecting deepfake videos from phoneme-viseme mismatches[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 660-661.
 - [25] BHATTACHARYA U, MITTAL T, CHANDRA R, et al. Step: spatial temporal graph convolutional networks for emotion perception from gaits[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 1342-1350.
 - [26] 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). 2018: 1-7.
 - [27] LI Y, LYU S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv Preprint arXiv:1811.00656, 2018.
 - [28] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). 2019: 83-92.
 - [29] QI H, GUO Q, JUEFEI X F, et al. Deephythm: exposing deepfakes with attentional visual heartbeat rhythms[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 4318-4327.
 - [30] 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.
 - [31] LI H, LI B, TAN S, et al. Identification of deep network generated images using disparities in color components[J]. Signal Processing, 2020(174): 107616.
 - [32] DO N T, NA I S, KIM S H. Forensics face detection from GANs using convolutional neural network[C]//ISITC. 2018: 376-379.
 - [33] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint arXiv:1409.1556, 2014.
 - [34] TARIQ S, LEE S, KIM H, et al. Detecting both machine and human created fake face images in the wild[C]//Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. 2018: 81-87.
 - [35] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
 - [36] NGUYEN H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019: 2307-2311.
 - [37] SABIR E, CHENG J, JAISWAL A, et al. Recurrent convolutional strategies for face manipulation detection in videos[J]. Interfaces (GUI), 2019, 3(1): 80-87.
 - [38] HSU C C, ZHUANG Y X, LEE C Y. Deepfake image detection based on pairwise learning[J]. Applied Sciences, 2020, 10(1): 370.
 - [39] GANDHI A, JAIN S. Adversarial perturbations fool deepfake detectors[C]//2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1-8.
 - [40] 龚晓娟, 黄添强, 翁彬, 等. 基于双层注意力的 Deepfake 换脸检测[J]. 网络与信息安全学报, 2021, 7(2): 151-160.
 - GONG X J, HUANG T Q, WENG B, et al. Deepfake swapped face detection based on double attention[J]. Chinese Journal of Network and Information Security, 2021, 7(2): 151-160.
 - [41] WU X, XIE Z, GAO Y, et al. Sstnet: detecting manipulated faces through spatial, steganalysis and temporal features[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020: 2952-2956.
 - [42] LIU Z, QI X, TORR P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8060-8069.
 - [43] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015: 234-241.
 - [44] WANG F, CHENG J, LIU W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

- [45] KING D E. Dlibml: a machine learning toolkit[J]. The Journal of Machine Learning Research, 2009, 10: 1755-1758.
- [46] BAYAR B, MEMBER S, STAMM M C. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2691-2706.
- [47] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security(WIFS). 2018: 1-7.
- [48] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. 2019: 6105-6114.
- [49] LIU H, LI X, ZHOU W, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 772-781.
- [50] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.
- [51] CARREIRA J, ZISSERMAN A, COM Z, et al. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
- [52] LUO Y, ZHANG Y, YAN J, et al. Generalizing face forgery detection with high-frequency features[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16317-16326.
- [53] DOLHANSKY B, HOWES R, PFLAUM B, et al. The deepfake detection challenge (DFDC) preview dataset[J]. arXiv Preprint arXiv:1910.08854, 2019.
- [54] WANG C, DENG W. Representative forgery mining for fake face detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14923-14932.
- [55] ZHAO H, ZHOU W, CHEN D, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2185-2194.

[作者简介]



张博林（1999-），男，陕西宝鸡人，中山大学硕士生，主要研究方向为深度伪造检测。



朱春陶（1997-），女，河南周口人，中山大学硕士生，主要研究方向为深度伪造检测。



殷琪林（1995-），男，江苏泰州人，中山大学博士生，主要研究方向为数字多媒体取证。



付婧巧（1987-），女，四川南充人，中山大学硕士生，主要研究方向为自然语言处理。



刘凌毅（1998-），男，江西瑞金人，中山大学硕士生，主要研究方向为深度伪造检测。



刘佳睿（1997-），男，山东招远人，中山大学硕士生，主要研究方向为深度伪造检测。



刘红梅（1969-），女，内蒙古丰镇人，中山大学副教授，主要研究方向为图像/视频处理、多媒体内容安全。



卢伟（1979-），男，河南南阳人，中山大学教授、博士生导师，主要研究方向为多媒体信息安全、数字取证、信息隐藏、隐私保护、多媒体大数据智能。