计算机科学
COMPUTER SCIENCE

基于双流网络结构的深度伪造人脸的检测方法

李颖, 边山, 王春桃, 黄琼

引用本文

李颖, 边山, 王春桃, 黄琼.基于双流网络结构的深度伪造人脸的检测方法[J]. 计算机科学, 2022, 49(11A): 220100106-9.

LI Ying, BIAN Shan, WANG Chun-tao, HUANG Qiong. Detection of Deepfakes Based on Dual-stream Network [J]. Computer Science, 2022, 49(11A): 220100106-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100106

基于双流网络结构的深度伪造人脸的检测方法

李 颖[1] 边 山[1,2,3] 王春桃[1,2] 黄 琼[1,2]
1 华南农业大学数学与信息学院 广州 510642
2 广州市智慧农业重点实验室 广州 510642
3 广东省信息安全技术重点实验室 广州 510006
(spade@stu.scau.edu.cn)

摘 要 深度伪造技术(Deepfake)是一种基于生成对抗网络(Generative Adversarial Networks,GAN)的深度网络模型,可以利用源和目标人脸生成高度逼真且难以鉴别的人脸视频。如果不法分子借此技术制造虚假视频并在互联网上传播谣言,将会侵犯个人肖像权,造成不良的社会影响,甚至引发严重的司法纠纷。面对深度伪造技术带来的严重威胁,国内外众多研究机构高度关注深度伪造检测技术的研究并提出了若干检测方法。现有的检测方法在高质量视频上可以取得良好的检测效果,然而日常应用中的视频通常会通过社交软件从而被压缩为低质量视频,在此类低质量数据集中,现有的大多数伪造人脸检测方法的准确率有着明显的下降,并且现有方法在跨库情况下的检测性能也不够理想。文中针对现有工作的局限性,提出了一种注意力机制下基于Xception模型的双流网络结构。该网络结构中包含了使用多重注意力机制的RGB分支,以及用于捕捉低质量视频伪影效应的频率域分支。通过研究发现,真实图像与伪造图像之间的微小差别更多地集中在局部位置,因此多重注意力机制下的RGB分支将使得模型关注人脸的不同区域,并在注意力图的指导下得到由低层纹理特征及高层语义特征聚合的全局特征。频率域分支引入离散余弦变换作为频域变换手段,为图像提供与RGB分支互补的特征表示,此分支能够反映细微的伪造痕迹或者压缩误差。为了验证该网络结构的有效性,所提算法在FaceForensics++,Celeb-DF以及DFDC 3个公开数据集上进行了大量对比实验。实验结果表明,所提算法在低质量视频集上的性能优于现有的检测算法,并且所提模型在跨库场景下具有更好的检测性能,即验证了文中提出的注意力机制下的RGB和频率域双流特征的结合可以提高检测模型在低质量视频集及跨库情形下的鲁棒性。

关键词:深度伪造;视频取证;双流网络;注意力机制;RGB分支;频率域分支
中图法分类号 TP391

Detection of Deepfakes Based on Dual-stream Network

LI Ying[1],BIAN Shan[1,2,3],WANG Chun-tao[1,2] and HUANG Qiong[1,2]
1 College of Mathematics and Informatics,South China Agricultural University,Guangzhou 510642,China
2 Guangzhou Key Laboratory of Intelligent Agriculture,Guangzhou 510642,China
3 Guangdong Provincial Key Laboratory of Information Security Technology,Guangzhou 510006,China

Abstract Deepfake is a kind of deep network model based on generative adversarial networks(GAN). It uses the source and target faces to generate highly realistic face videos that are difficult to identify. If some malicious person uses this technology to make fake videos and spread rumors on the Internet,it will infringe personal portrait right,cause adverse social impact,or even cause serious judicial disputes. In view of the serious threat brought by deepfake technology,many researchers at home and abroad pay close attention to the study of the deepfake detection technology,and have put forward some effective detection methods. The existing detection methods achieve good detection results in high-quality videos,but most videos in daily applications are usually compressed into low-quality versions through social software. However,most of the existing deepfake detection methods have a significant decline in detecting this kind of low-quality videos. Besides,the detection performances of existing methods are still unsatisfying in the case of cross datasets,limiting their real applications. To address this issue,this paper proposes a dual-flow network structure based on Xception model under multiple attention mechanism. The network structure includes an RGB branch using multiple attention mechanism and a frequency-domain branch for capturing low quality video artifacts. Based on our research,it is found that the tiny difference between real images and fake images tends to concentrates in some local area. The RGB branch under the multiple attention mechanism makes the model focus on different regions of the face,so it can get the global features aggregated by the low-level texture and high-level semantic features under the guidance of the attention map. Com-

基金项目:国家自然科学基金(61702199,62172165,61872152);广东省基础与应用基础研究重大项目(2019B030302008);广东省信息安全技术重点实验室开放基金(2020B1212060078-07);广州市科技计划项目(202102020582,201902010081)
This work was supported by the National Natural Science Foundation of China(61702199,62172165,61872152),Major Program of Guangdong Basic and Applied Research(2019B030302008),Opening Project of Guangdong Province Key Laboratory of Information Security Technology(2020B1212060078-07) and Science and Technology Program of Guangzhou(202102020582,201902010081).
通信作者:边山(bianshan@scau.edu.cn)

bined with the RGB branch, the discrete cosine transform(DCT) is introduced in the frequency domain branch to provide complementary feature representation, which can reflect subtle forgery traces or compression errors. Specifically, the proposed algorithm firstly extracts a large number of face frames from videos by face extractor algorithm, and feeds these face frames into the two-branch network model. The frequency branch decomposes the spectrum of images with three combined filters that provide additional learnable parts. In the RGB branch, the first three layers of the backbone network extract shallow features including texture information etc. Then the attention module makes the model attend to the shallow information from different local areas. The shallow information is then fed to the attention pooling layer to aggregate with the high-level semantic features from the rest layers of the backbone network. Finally, the network merges the feature vectors from both the RGB-branch and the frequency branch to obtain the final discriminant result. The combination of these two branches can significantly improve the detection performance of the model in cross database scenes and low-quality video sets. In order to verify the effectiveness of the proposed network structure, a large number of comparative experiments are conducted on three public datasets, including FaceForensics++, Celeb DF and DFDC. In the low-quality part of FaceForensics++ dataset, the AUC(Area Under the Curve) can reach 0.927 1. In the video level, the detection accuracy of low-quality and high-quality videos can reach 93.84% and 99.69%, respectively. Experimental results show that the proposed algorithm outperforms the existing detection algorithms in low-quality video sets as well as in cross dataset scenes. It verifies that the combination of dual-stream features including the RGB branch and the frequency branch can improve the robustness of the detection method, especially in low-quality video sets and in cross dataset scenes.

Keywords Deepfake, Video forensics, Dual stream network, Attention mechanism, RGB branch, Frequency branch

1 引言

近年来，基于生成对抗网络和自动编码器的深度伪造技术(deepfakes)得到了快速发展。借此攻击者能够对图像中的人脸区域进行篡改并合成伪造人脸图像或者视频。最早将其运用在人脸伪造上的是海外知名论坛上一位名为“deepfakes”的用户，他将影视明星的脸移植到成人电影的演员身上，自此Deepfakes也成为了深度伪造技术的代名词[1-2]。无独有偶，2019年一款名为“ZAO”的交换人脸软件一夜之间风靡了国内整个社交网络。这类虚假视频的广泛传播对多媒体信息安全产生了巨大的威胁，对于个人而言，使用了其人脸的深度伪造视频的恶意传播可能侵犯公民的隐私权和名誉权；对于社会来说，深度伪造技术的滥用将破坏社会的舆情舆论稳定；对于国家而言，虚假视频一旦被用于煽动极端情绪、制造政治矛盾等恶劣途径，将严重威胁国家安全和社会稳定。

2019年，中国互联网信息办印发《网络音视频信息服务管理规定》，要求网络音视频信息服务提供者应健全辟谣机制，采取相应的鉴别措施，对基于深度学习、虚拟现实等技术的虚假音视频信息做到及时的风险管控。2021年，针对涉及“深度伪造”技术的应用，中国互联网信息办约谈了包括小米、网易云音乐等企业，督促其按照法律法规及政策要求，完善风险防控机制和措施。由此可见，针对深度伪造的鉴别技术具有强烈的社会需和国家需求。

深度伪造技术主要是基于生成对抗网络技术[3]，生成对抗网络由生成器和判别器组成，生成器和判别器在训练中互相博弈以习得期望的数据分布。在深度伪造内容生成过程中，生成模型用于生成虚假人脸，通过编码-解码的过程实现人脸替换、人脸重现等内容篡改操作，而判别模型则对生成模型所生成的伪造人脸进行分类，识别生成器所生成的人脸是真实的或是伪造的。在交替训练生成器与判别器的策略之下，生成器将更倾向于输出判别器无法判定的伪造人脸，从而提升伪造人脸的逼真程度。基于深度伪造技术的人脸伪造工具的流行，如FakeApp，Deepfakes，FaceSwap等软件，使得编辑人脸、篡改人脸的成本降低，对互联网中的深度伪造图像/视频进行高效准确的识别成为了当前的热点研究领域。

以Deepfakes为代表的深度伪造技术的取证需要，推动了基于深度学习方法的取证技术的发展，国内外众多研究机构和学者关注到了深度伪造检测技术的研究。然而，由于上传带宽的限制，大多数深度伪造图像/视频上传至社交平台之后，质量都会因为压缩而大大降低。这类低质量的深度伪造产物由于伪造痕迹不明显，且内存占用较小，在现实场景中更易形成广泛的传播，这也就给取证研究带来了困难。

现有的大多数方案在这类高压缩低质量视频上准确度低，同时，由于深度伪造技术的多样性，不同数据集采用的深度伪造技术不同，数据集之间的深度伪造特征差距较大，因此大部分的深度伪造检测算法在跨库场景下性能将显著下降，难以在现实场景中得到应用推广。

跨库性能较低的原因在于，数据驱动的深度伪造检测技术往往能够对数据集中存在的深度伪影进行拟合，而与之不相似的特征则无法很好地识别出来。Zhao等[4]提出，存在于真实区域与篡改区域之间微小而局部的差异与细粒度分类问题有一定的相似。因此，本文引入多重注意力机制促使模型更关心伪造图像中真实区域与伪造区域之间的差异，这类差异广泛存在于各种深度伪造技术输出的伪造图像/视频中，也因此能够作为更加通用的检测线索。

低质量图像/视频的分类难度较高的其中一个原因在于，基于学习的深度伪造技术大多关注真实视频与伪造视频之间的全局特征差异。然而，随着压缩程度的加深，这类特征差异很难在RGB域体现。Qian等[5]提出，尽管在RGB域中伪造特征容易被压缩误差所污染，但这类伪影却能够在频率域中捕捉得到，与真实区域相比，伪造区域将呈现异常的频率分布。

因此，针对现有工作的局限性，本文提出了一种注意力机制下基于XceptionNet模型的双流网络结构。具体地，本文的主要贡献如下：

(1)针对深度伪造检测技术跨库检测精度较低的问题，模型中引入了多重注意力机制，学习人脸伪造图像中局部区域之间的细微不一致性，提升深度伪造检测技术的跨库精度。

(2)针对低质量人脸深度伪造图像，将频率域信息作为

RGB域特征的补充，此分支能够反映细微的伪造痕迹或者压缩误差。

(3)结合上述两点，提出了基于XceptionNet的双流检测网络，并在3个大型数据集上进行了评估。实验结果表明，本文方法在多个压缩质量视频上的表现均优于现有方法，并且本文模型在跨库场景下具有更好的泛化性能。

2　相关工作

2.1　深度伪造技术

深度伪造技术主要通过深度学习方法，对图像中的人脸区域进行篡改操作，从而达到修改图像中人物的身份、表情等目的。现有的两种深度伪造技术主要分为两大类：人脸交换与人脸重现。人脸交换方法使用目标对象的人脸部分替换源对象的面部区域。这些方法通常先将目标图像中的脸部区域通过人脸提取算法提取出来，再将其与源图结合，进行背景混合。人脸重现算法则使用源对象的面部动作，驱动目标对象的运动，这类伪造不会改变对象的身份信息，却能达到通过修改视频中人物的表情、所说的话来歪曲事实的目的。深度伪造技术以生成对抗网络和自动编码器为基础，对源对象首先使用动作提取，将其与源图中的人脸部分形成标签对，生成对抗网络中的生成器将生成更逼真的标签对以欺骗判别器，通过对生成器和判别器进行交替训练，最终使得生成数据与真实数据具有相同的分布，判别器无法识别出图像是否由深度伪造技术生成。

基于生成对抗网络技术的深度伪造技术的基本原理如图1所示，对于希望替换的人脸，通过训练神经网络编码器的方式，来得到编码后的中间状态。每个编码器都对应一个解码器，为了实现替换的过程，编码时应使用统一的编码器，解码时则使用目标人脸的解码器对源人脸的中间状态进行解码。

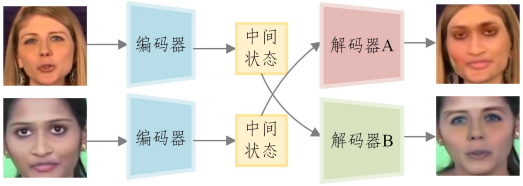


图1　深度伪造技术原理

Fig. 1　Principle of Deepfake technology

2.2　深度伪造检测技术

针对深度伪造带来的各种信息安全风险，国内外众多研究机构和学者关注到了深度伪造检测方法的研究，并提出了许多经典的方法。最早的一些方法将卷积神经网络与手工制作的特征相结合，然而这些方法往往只能针对一种或两种深度伪造技术，在现实场景中不具有推广性。

最新的方法[6-7]则利用深度神经网络从空间域提取深层的信息作为分类线索，对于伪造图像，捕捉其中的伪造线索和图像块的不一致性。为了提高检测技术的泛化性能，更多的研究人员开始探索深度伪造模型由于内部结构、伪造过程在图像中留下的固有特征，如Yu等[8]认为由生成对抗网络得到的伪造图像中将留下对应的模型指纹；McCloskey等[9]则观察到，对抗网络得到的伪造图像将缺乏自然图像中饱和度和曝光度不足的区域。一些生物特征和视觉线索也成为了深度伪造检测技术的基础，如通过不自然的眨眼或异常的头部姿势等来进行识别。

基于注意力机制的模型在目标检测、语义分割等领域收获了很好的效果，最新的一些方法也开始将注意力机制作为深度伪造检测的指导。Gong等[10]提出融合双层注意力的检测模型，Zhao等[4]将深度伪造图像的检测表述为细粒度分类问题，由此提出了基于多注意力图的检测框架，同时增强了从浅层获得的纹理特征之间的差异，聚集低级纹理特征以及高级语义特征作为每个局部区域的表示。

目前基于空间域的伪造检测算法较多，如Bian等[11]提出通过多通道的空洞卷积模块，来提升模型在低质量伪造人脸视频上的检测精度；Li等[12]引入噪音流特征，设计了基于EfficientNet模型的双流检测框架；Bao等[13]提出了改进的ResNet网络，结合数据增强的方式来提升模型在Deepfakes和FaceSwap两类篡改上的检测性能。然而，这些方法大多数只利用了空间域的信息，忽视了在颜色空间中检测不到的细微篡改线索。因此，也有一些研究人员转向了对频率域中篡改特征的研究，文献[5]提出了基于频率域特征的双流网络，利用频率线索来挖掘难以察觉的篡改特征，这在低质量视频的检测上被证明有效。Liu等[14]则认为，上采样操作是人脸伪造中的必要操作，而累积的上采样操作会导致频域中有显著的变化，因此结合了空间图像与相位谱来捕获深度伪造图像中的上采样伪影，通过利用频率域信息和空间信息的多模态方法在跨数据集检测上取得了先进的性能。Wang等[15]结合频率域特征与RGB空间特征，提出了一种多模态多尺度的结构，用于检测不同尺度下的局部不一致性。本文受到现有工作的启发，利用频率域信息和RGB域信息构建了双流检测网络结构，在RGB流引入了多重注意力机制，与频率域流相结合，在空间域习得局部不一致性，同时在频率域提取细微的篡改特征，在低质量图像和跨库的检测性能上寻得一个平衡，在跨库的深度伪造检测上取得了更好的性能。

3　理论基础

3.1　XceptionNet模型

XceptionNet模型是在InceptionNet模型的基础上提出的，过去的卷积神经网络设计离不开模块的堆叠，从而使得为了学到更多的特征就不得不对网络进行加深，网络结构的风格并没有根本性的变化。InceptionNet结构由Szegedy等[16]提出，其最大的特点在于Inception模块的引入。Inception模块对卷积操作进行解耦，将其划分为空间上与通道上的多分支运作。

卷积核实际上是在宽、高两个空间维度以及一个通道维度这样一个三维空间中学习得到的，因此一个卷积核需要同时学习如何表示空间相关性及跨通道相关性。Inception模块希望通过一系列独立的操作将跨通道相关性和空间相关性的学习进行脱钩，极端版本的Inception模块的学习过程如图2所示，对输入数据进行1 * 1卷积以描述跨通道相关性，让其在3或4个更小的不同空间中通过3 * 3或5 * 5的卷积操作独立学习每个输出通道中的空间相关性。

而Chollet等[17]提出，极端版本的Inception模块与深度可分卷积几乎一致，因此使用深度可分卷积来替代InceptionNet家族中的Inception模块，即用深度可分卷积进行堆叠，

构建了一个完全基于深度可分卷积层的网络结构。深度可分卷积结构如图 2 所示，XceptionNet 实质上是深度可分卷积层的线性堆叠，同时采用了残差连接，这也就使得 XceptionNet 作为骨干网络时容易对结构进行定义和修改，引入深度可分卷积也使得网络的复杂度得到了降低，在参数量与 InveptionV3[18] 相同的情况下提高了性能。

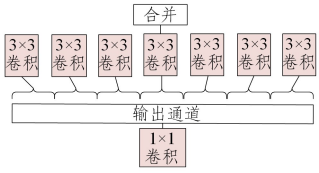


图 2 深度可分卷积

Fig. 2 Depthwise separable convolution

3.2 离散余弦变换

离散余弦变换(Discrete Cosine Transform)与离散傅里叶变换相似，区别于离散傅里叶变换的关键在于离散余弦变换是在实数域中的变换，相比需要进行复数运算的离散傅里叶变换可以提高运算速度，也因此能够在实时使用场景中广泛使用，在数字图像处理中其可以用于图像的压缩，将空域中的信号变换到频域中，同时由于其过程是无损的，因此可以在离散余弦变换之后执行逆变换，恢复原始图像信息。

离散余弦变换具有很好的去相关性，能够聚集重要信息，因此在图像这类相关性很高的自然信号处理上有着广泛的应用。

二维离散余弦变换的计算式如式(1)所示：

$$F(u,v)=\frac{2}{N}\times\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}f(m,n)\cos\frac{(2m+1)u\pi}{2N}\cos\frac{(2n+1)v\pi}{2N} \tag{1}$$

其中，$f(m,n)$为空间域中大小为 $N*N$ 的二维矩阵，$F(u,v)$则表示经二维离散余弦变换后得到的矩阵。处理相关性较强的原始信号时，如图像、音频等，系数能量将集中在左上角，其余大部分系数几乎为零。

在深度伪造检测中，低频信息、中频信息和高频信息对取证而言有着不同的作用，如图 3 所示，因此本文应用离散余弦变换获得不同频带的信息，用于全面捕捉频率域中各个频带中的篡改伪影。

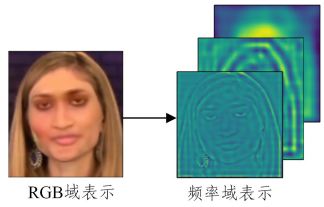


图 3 图像的频率域表示

Fig. 3 Frequency domain representation of image

4 本文模型结构

本文基于频率域和 RGB 域构建了一种新型的双流检测网络结构，在 RGB 流中引入了多重注意力机制，与频率域流相结合，在空间域习得局部不一致性，同时在频率域提取细微的篡改特征。在压缩图像和视频中，频率域信息能够提供一种与 RGB 信息互补的伪影表示，其能够很好地描述在 RGB 图像中因为压缩错误被污染而难以找到的细微伪像[5]。基于此，本文引入了频率域特征，用于提高检测算法在高压缩场景下的检测性能。同时，为了提高检测的鲁棒性，在 RGB 流的学习上引入注意力机制，着重学习篡改区域与真实区域的不一致性，并将两个维度的特征信息融合起来，设计双流结构。由于频率域分支提取得到的特征将着重反映低质量视频中是否存在伪造伪影，且与 RGB 分支最终得到的聚合后的深层特征来自 XceptionNet 模型中同样深度的语义层，因此采用了在决策前融合的策略对特征进行拼接，以充分保留两个分支的深层特征。XceptionNet 模型[17]已被证明在深度伪造检测中作为骨干网络拥有优越的性能，本文采用 XceptionNet 模型作为双流网络结构的骨干网络，检测的主体框架如图 4 所示。其中，骨干网络浅层特征来自 XceptionNet 模型的入口流，浅层特征经重复 8 次的中间流到达出口流，深层特征则来自 XceptionNet 模型的出口流。XceptionNet 模型的入口流包含 4 个模块，中间流包含 8 个模块，由图 4 中中间流的 ReLU 函数与深度可分卷积层搭配实现，出口流包含 2 个模块，其结构与入口流相似，由 ReLU 层、深度可分卷积层、最大池化层组成，并在第一个与最后一个之外的模块之间建立残差连接。

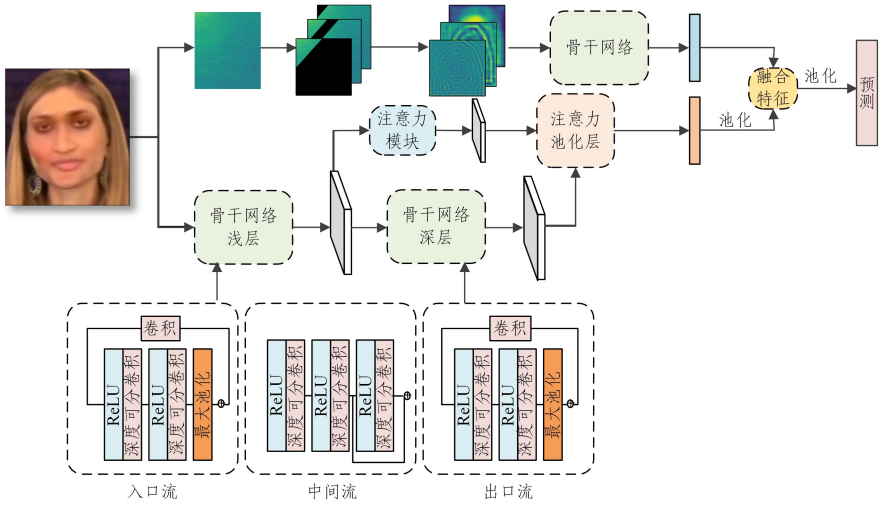


图 4 双流检测框架

Fig. 4 Dual-stream detection framework

4.1　RGB 注意力流

在深度伪造图像/视频中，真实区域和篡改区域之间的差异往往很小，同时由于篡改区域仅仅局限于人脸部分，因此将注意力集中在篡改区域可以收集到更加具有区分度的局部特征作为深度伪造检测的线索。同时，在 RGB 图像中，由于卷积神经网络提取到的浅层特征将比深层特征包含更多有利于寻找篡改线索的纹理信息，取证模型更应该关注的是图像中浅层的特征，而非来自更深层的高级的语义特征[4]。因此，本文将注意力机制应用在骨干网络提取到的浅层特征图上，受文献[4]的启发，加入注意力模块，并使用注意力池化层来代替全局平均池化，注意力模块的结构如图 5 所示。

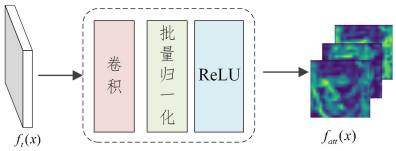


图 5　注意力模块

Fig. 5　Attention module

输入的人脸图像表示为 x，骨干网络为 f，从骨干网络中提取到的特征图为 f_t，其中 t 表示骨干网络中的第 t 层。对于输入图像，通过骨干网络提取到浅层特征图之后，送入到注意力模块得到注意力图。注意力模块是卷积层、批量归一化层和非线性激活层 ReLU 的组合，用于产生本质为加权图的注意力图。注意力模块的输出如式(2)所示：

$$f_{att}(x)=\mathrm{ReLU}(BN(Conv(f_t(x)))) \tag{2}$$

其中，ReLU 表示 ReLU 激活函数，BN 表示批量归一化，Conv 表示卷积层。

骨干网络提取到的浅层特征图通过注意力模块之后，将得到多个注意力图，注意力图关注不同区域的篡改特征，分别对应特定的辨别区域，如眼睛、嘴巴。文献[6]指出，人脸区域一些特殊的区域上的篡改伪影将更为明显，如嘴巴附近以及前额的模糊区域，这也是本文引入多注意力图的目的所在，使得更多区域的篡改伪影能够被注意力模块捕捉到。注意力图的数目 k 根据经验确定为 4。

本文对模型的注意力图进行可视化，图 6 给出了部分注意力图的可视化结果，第一行为真实图像及其对应的 4 个注意力图，每个注意力图中反应最强烈的关注区域存在差别，在图中用红色矩形标注。例如第一个注意力图将更加关注下巴区域，第二个注意力图将更加关注眼睛及鼻梁部分，第二行则为伪造图像及其对应的注意力图。

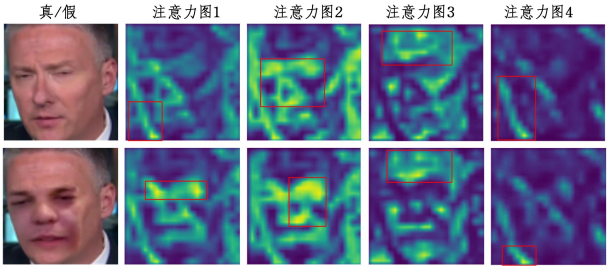


图 6　多重注意力图(电子版为彩图)

Fig. 6　Multiple attetion maps

注意力图可视化展现出了多重注意力对人脸不同区域的反映程度，同时，对于伪造图像，注意力图的反映相比真实图像要稍显强烈。

尽管本文为了减少模型训练难度与时长并没有采用与文献[4]一致的独立损失及相同的数据增强策略，使得多个注意力图能够完全分散在互补的人脸区域，然而从图中可以看出，无约束的注意力机制也能够一定程度地帮助模型关注到人脸中的不同区域。

4.2　注意力池化

本文使用注意力池化层替代全局平均池化层，对 RGB 流中的浅层特征与深层特征通过注意力图进行融合。注意力池化层的输出表示如式(3)所示：

$$f_{\mathrm{RGB}}=A_s\cdot f_{\mathrm{depth}} \tag{3}$$

其中，A_s 表示对于浅层特征图提取得到的注意力图，f_{depth} 表示经由入口流、中间流及出口流得到的深层特征图，最终得到 RGB 流输出的全局特征 f_{RGB}。具体做法是，通过双线性插值，将注意力图与深层特征图调整到同样大小，同时将多重注意力图拼接成单通道注意力图 A_s，并对两者作乘法运算，相当于对深层特征图进行提取，得到了全局的深层特征图，在降低深层特征中的特征冗余的同时，对浅层特征中的重要信息进行了增强。

4.3　频率域特征流

现有工作已经证明，采用压缩方法，如 JPEG 压缩之后的虚假图像和视频中的篡改伪影难以察觉。本文采用与 F3Net[5]一致的方法，从频域中计算特征来对 RGB 进行特征补充，在频率域流中得到的特征将与 RGB 流特征在决策前进行融合。

输入的人脸图像表示为 x，对 x 进行离散余弦变换，将其从 RGB 域变换到频率域，离散余弦变换之后的频谱图、低频信息将集中在左上角，而高频信息则位于右下角。本文将频域分为 3 个频带，分别为低频、中频以及高频。高频信息通常和伪造检测中感兴趣的边缘和纹理相关联，而低频分量能够将全局图像保留下来，将低、中、高频的信息连接在一起，能够让模型在频率域捕捉到更加完整的线索，甚至是在压缩过程中丢失从而在 RGB 域中不可见的信息。

具体做法是，设计 3 个基本的二进制滤波器，能够将频域划分为低、中、高 3 个频带，同时对这 3 个基本滤波器添加可学习的滤波器，可学习的滤波器与基本滤波器的结合得到的组合滤波器将自适应地选择频率，从而在频域中对图像进行分割，表达式如式(4)所示：

$$F_i=b_i+\sigma(l_i) \tag{4}$$

其中，b_i 表示基本滤波器，l_i 表示可学习的滤波器，$\sigma(x)=\dfrac{1-\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}}$，目的是将可学习的滤波器限制在(−1，+1)之间。

组合滤波器将对频率响应进行分解，得到一系列频率图像分量，表达式如式(5)所示：

$$Y_i=DCT(x)\odot F_i \tag{5}$$

其中，$\odot$ 表示点乘运算，F 表示不同频带对应的组合滤波器，组合滤波器中，3 个基本滤波器分别为低频带滤波器、中频带滤波器和高频带滤波器，其中低频带滤波器是整个频谱的前 1/16，中频带滤波器是频谱的 1/16 到 1/8 之间，高频带滤波器则为频谱最后的 7/8。为了在引入频率域线索的同时不丢失自然图像的局部一致性以及平移不变性，对进行了离散余弦变换和分解的图像 x，经过离散余弦变换的逆变换回到 RGB 域，得到新的 RGB 表示，具体表达式如式(6)所示：

$$Z_i=DCT^{-1}(Y_i) \tag{6}$$

其中，Z_i 表示应用了第 i 个频带滤波器得到的频率分量经逆变换之后的 RGB 表示，对 Z_i 沿着通道方向进行重组，最终得到频率域信息图。将频率域信息作为频率域分支的输入，

用于提取频率域中的特征。

骨干网络得到的频率域特征将与池化后的 RGB 流特征进行拼接，在模型决策阶段之前进行融合，最终一起送入到分类器中，得到输入图像是否为深度伪造图像的预测结果。

4.4 损失函数

对来自 RGB 分支和频率域分支的特征向量进行拼接之后，即得到了全局的双流特征，通过交叉熵损失函数，使用这些特征来预测输入图像的真假分类结果，交叉熵损失函数如式(7)所示：

$$L=y\log y+(1-y)\log(1-\hat{y}) \tag{7}$$

若输入图像为经过篡改的深度伪造图像，则 y 被设置为 1，反之设置为 0。$\hat{y}$ 则表示本文模型根据输入图像预测得到的标签。

5 实验结果及分析

本节将提出的网络模型在 3 个大规模伪造人脸数据集中进行测试，包括单数据集内部测试、跨数据集测试、消融测试等。通过多个检测指标的对比，证明了本文模型的有效性。

5.1 参数设置及实验环境

为了提取图像中的重要信息，提高检测的精度与效率，本文使用 Blaze 模型对输入图像进行人脸区域的提取，大小设置为 512 * 512。将双流结构中用于判断正样本的阈值设置为 0.5。骨干网络采用了在 ImageNet 数据集上进行预训练的 XceptionNet 模型，学习率为 1×10^{-5}，学习率每步衰减一次，批量设置为 8，10 个 epoch 训练完整的网络，实验运行环境如表 1 所列。

表 1 实验运行环境

Table 1 Experimental environment

类别	配置
电脑类型	台式电脑
显卡	Nvidia GeForce RTX 2080Ti
CPU	Intel Core i9-9900K
内存大小/GB	32
操作系统	Ubuntu 18.04 LTS
深度学习框架	Pytorch
CUDA 版本	CUDA 10.1
cuDNN 版本	Cudnn 7.6.03
编程语言	Python 3.6.9

5.2 评估指标及数据集

5.2.1 评估指标

本文采用准确度分数 ACC(Accuracy)以及 ROC 曲线下面积 AUC(Area Under Curve)作为本文的评估指标，这两个指标在分类任务中是普遍的评价标准，现有的研究工作均采用了这两种方式。

5.2.2 数据集

本文在 FaceForensics++(下称 FF++)、Celeb-DF[19] 以及更大型的 DFDC[20] 数据集上进行了实验。FF++数据集由 1000 个真实视频及对应的虚假视频组成，其中 720 个视频用于训练，140 个视频用于验证，140 个视频用于测试。每一个真实视频都由 4 种深度伪造技术生成对应的虚假视频，即 Deepfakes，FaceSwap，Face2Face 以及 NeuralTextures。同时每个视频根据压缩程度有 3 种版本，分别为原始质量、高质量版本以及低质量版本。Celeb-DF[19] 则由 890 个真实视频和 5639 个 DeepFake 视频组成，用于做跨数据集的测试。DFDC 数据集[20] 是 2020 年由 FacekBook 发布的大规模的深度伪造检测数据集，对于 DFDC[20] 数据集，本文按照 7:1:2 的比例划分训练集、验证集和测试集，对每个视频抽取 32 帧图像。对于 FF++ 数据集[21]，与文献[4-5]保持一致，本文通过重复采样及数据增强(如随机裁剪、随机翻转等)的方式将真实图像的数量增加 4 倍，平衡数据集中真实样本和虚假样本的数量，对每个视频抽取 270 帧图像。

5.3 FF++数据集的实验结果

FF++[21] 是许多深度伪造检测中广泛使用的数据集，因此本文将双流网络结构与目前最先进的深度伪造检测方法进行比较。对比实验中分别测试了网络在原始质量、高质量以及低质量上的性能。表 2 列出了与最新的方法进行的帧级对比的 ACC 以及 AUC 结果，部分数据来自文献[4,14]，其中包括 MADD[4]，LD-CNN[22] 等方法。

表 2 FF++数据集的帧级检测结果

Table 2 Frame level detection results of FF++ dataset

(单位：%)

模型	LQ		HQ	
	ACC	AUC	ACC	AUC
LD-CNN[22]	58.69	—	78.45	—
MesoNet[6]	70.47	—	83.10	—
DSP-FWA[23]	—	59.15	—	56.89
X-ray[24]	—	61.60	—	87.35
Xception[21]	—	83.93	—	92.30
SPSL[14]	81.57	82.62	91.60	95.32
TB[25]	—	86.59	—	98.70
MADD[4]	**88.69**	90.40	**97.60**	**99.29**
本文模型	87.99	**92.71**	**97.60**	98.80

Face X-ray[24] 是以深度伪造图像中的混合边界为新的图像表示，同时以此为线索进行检测。F3-Net[5] 则由两个分支组成，分别对图像中的频率域信息进行分解和统计，MADD[4] 则提出了多注意力网络结构，来捕获深度伪造图像中的局部不一致性。LQ 列表示在低质量数据集上进行训练，且在低质量数据集上进行测试得到的结果。HQ 列表示在高质量数据集上进行训练，且在高质量数据集上进行测试得到的结果。

本文选择的两个评价指标中，准确度分数与选择的分类阈值有关，AUC 分数在大多数场景下更能够反映模型分离正负样本的性能。从表 2 中可以观察到，相比同样使用了频率域线索的 SPSL[14] 和 TB[25]，本文采用的双流结构在低质量数据集上的 AUC 值分别提升了 9.89% 和 6.12%；相比同样采用了注意力机制的 MADD，本文采用的双流结构在低质量数据集上的 AUC 值提升了 2.31%，高质量数据集上的测试结果相近，这代表本文模型在提高低质量视频检测的精度的同时，在高质量视频的检测上也能够保持同等的精度水平。本文模型的 RGB 分支设计参考了 MADD 中骨干网络分支的结构，但存在 3 点明显差异：1) MADD 模型[4] 为使注意力图更好地分散开，采用了区域独立损失，而本文为了平衡双流结构的分类性能与计算效率，仅使用交叉熵损失对模型进行训练；2) 浅层特征的选择不同，MADD[4] 选择 EfficientNet 中的第二层和第五层分别作为浅层特征层及注意力层，本文选择 XceptionNet 入口流的最后一个块作为浅层特征层及注意力层，对其输出特征图产生注意力图；3) 本文模型针对低质量视频检测精度低的问题，将在低质量数据集上可能造成过度纹理提取的纹理增强模块更换成为了频率域分支，以期改善低

质量场景下的模型精度。

除了帧级的实验，视频级的测试结果更符合实际的检测场景，本文模型与不同模型视频级的实验结果对比如表 3 所列，由于部分算法未开源且未报告视频级检测结果，因此表格中仅与已在论文中报告视频级结果的方法进行对比。对于每个视频，本文采取对视频所有帧计算平均分数的方式得到视频的分类分数。LQ 列表示在低质量数据集上进行训练，且在低质量数据集上进行测试得到的结果。HQ 列表示在高质量数据集上进行训练，且在高质量数据集上进行测试得到的结果。从表中数据可以看到，本文模型在两个压缩质量上都要优于 DSP-FWA，Xception，F3-Net 以及 TB。本文模型与 TB 模型在特征选择上相似，但存在 3 点明显差异：1）TB 由两个独立的 dense block 构成两个分支，分别处理频率域和颜色空间的信息，而本文模型采用 XceptionNet 作为两个分支的骨干网络，XceptionNet 结构使用深度可分卷积作为基础模块，利用通道与空间操作解耦的思想，简单有效且减少了参数量；2）TB 在融合双分支特征之后，采用全局平均池化，本文模型在注意力模块之后使用注意力池化代替全局平均池化，增强了浅层特征信息，提高了篡改区域的检测精度；3）TB 在融合层将双分支的特征相结合，而本文模型采用分类器决策前融合双分支深层特征的策略。

表 3　FF＋＋数据集视频级检测结果

Table 3　Video level detection results of FF＋＋ dataset

（单位：%）

模型	LQ		HQ	
	ACC	AUC	ACC	AUC
DSP-FWA[23]	—	62.34	—	57.49
Xception[21]	—	86.75	—	92.50
F3-Net[5]	93.02	95.80	**98.95**	99.30
TB[25]	—	91.10	—	99.12
本文模型	**93.84**	**98.02**	98.29	**99.69**

以上差异使得本文模型在 3 个压缩质量的数据集上的表现更加出色。图 7 给出了使用类激活映射（Class Activation Mapping，CAM）时本文模型与 XceptionNet[26] 及 EfficientNet[26] 的类激活热力图对比，其中红色部分则为模型最为关注的区域。

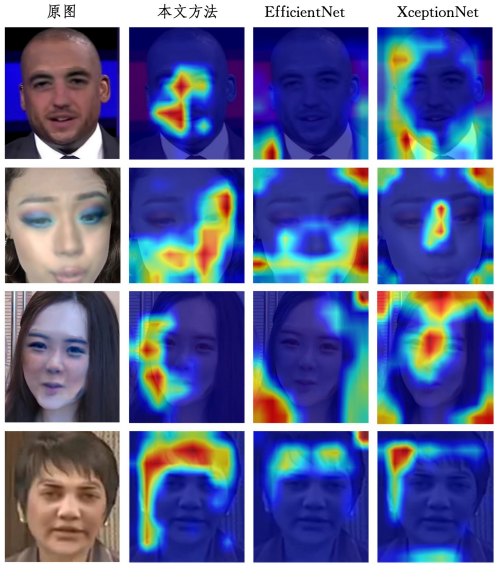


图 7　不同方法的类激活图对比（电子版为彩图）

Fig. 7　Comparison of class activation maps of different methods

本文模型的类激活热力图能够很好地捕捉到深度伪造图像中的伪造边缘，而 XceptionNet 及 EfficientNet 则更加侧重于关注图像的背景信息，因此难以学习到人脸区域中的伪造特征。在频率域分支及注意力机制的协调作用下，本文模型能够将注意力集中在人脸区域及边缘因区域间压缩特征不一致导致的差异，从而使得检测准确率得到提升。

5.4　Celeb-DF 数据集上的实验结果

不同的数据集互不交叉，使用的深度伪造技术不同，数据集中内在的伪造特征也不相同。因此，现有大多数的检测方法虽然在数据集内的检测效果很好，但跨数据集测试的效果却下降明显。为了评估本文方法在跨库场景下的性能，本文在 FF＋＋[21] 数据集（HQ）上对模型进行训练，同时在 Celeb-DF[19] 上进行测试，用于测试本文模型在跨库场景下的性能，以便评估模型的泛化能力。

表 4 列出了本文模型在 FF＋＋[21] 高质量数据集上训练，然后在 FF＋＋[21] 和 Celeb-DF[19] 数据集上分别进行测试的实验结果，所有数字均来自于文献[4,14]。表 4 中第一列数据为 FF＋＋ 库内的测试结果，库内结果均使用的是帧级结果。需要注意的是，现实中的检测场景更多的是面向视频级别，因此帧级结果的库内差距在视频级别的检测上将变得微乎其微。

表 4　跨库测试 AUC 的结果对比

Table 4　AUC results of cross dataset test

（单位：%）

模型	FF＋＋	Celeb-DF
Xception-c40[19]	95.5	65.5
Multi-task[27]	76.3	54.3
Capsule[28]	96.6	57.5
DSW-FPA[23]	93.0	64.6
F3-Net[5]	98.1	65.2
MADD[4]	**99.8**	67.4
DCViT[29]	98.3	60.8
TB[25]	93.20	73.40
SPSL[14]	96.91	**76.88**
本文模型	98.80	72.98

表 4 中的第二列数据则为在 Celeb-DF[19] 上的跨库测试结果。从第二列数据可观察到，相比同样使用了频率域线索的 F3-Net[5]，本文模型的跨库 AUC 提高了 7.78%。F3-Net[5] 同样为双分支检测结构，然而其只利用了频率域特征，数据集之间的频率域特征往往相差较大。而 MADD[4] 对纹理特征进行了增强，同时引入了注意力机制，因此这也使得它相比 F3-Net[5] 有着更好的跨库精度，说明了注意力机制的有效性。纹理特征在高质量数据上有着很好的作用，然而在低质量数据集上反而会影响检测精度，因为低质量的视频中包含着许多无效的纹理线索。本文选择使用频率域特征对注意力机制进行补充，相比同样使用了注意力机制的 MADD[4]，提升了 5.58%。与特征选择上相似的 TB 及 SPSL 相比，本文模型在 FF＋＋库内的检测精度提升了 5.60% 和 1.89%，同时保留了不俗的跨库检测精度，这意味着本文模型能够在低质量场景及跨库场景的检测之间寻得一个平衡，在两种场景下都能达到先进的检测性能。

5.5　DFDC 数据集的实验结果

DFDC 数据集包含了超过 10 万个人脸视频，是目前最具挑战性的深度伪造检测数据集之一。然而，现有的方法很少

报道在此数据集上的实验结果，为了公平起见，本文采取相同的实验设置对开源模型进行重新训练，由于 DFDC 数据集包含的视频更多，受限于硬件条件与耗时，对 3 个模型都采用迭代 60000 次的训练设置，并且与本文模型进行结果的对比。表 5 列出了复现的开源模型 XceptionNet[26]、EfficientNet[26]以及本文模型在 DFDC 数据集上的实验结果。从表中可以观察到，与同类算法相比，本文模型在 DFDC 数据集上，无论是帧级还是视频级的 AUC 指标都提高了超过 1%。这表明，本文模型在最具挑战性的数据集上的表现依然令人满意。

表 5 DFDC 数据集上的 AUC 结果对比

Table 5 Comparison of AUC results on DFDC dataset

（单位：%）

模型	帧级	视频级
XceptionNet[26]	89.11	93.73
EfficientNet[26]	90.13	94.90
本文模型	**91.27**	**95.95**

5.6 消融实验

本文在 FF++[21]的低质量数据集上进行消融实验，用于证明注意力模块与频率域线索结合的有效性。为了进行不同网络模块的对比实验，本文划分了 4 种不同的网络结构，其中 #1 Freq. 对应频率域单分支网络结构的检测性能，#2 RGB 代表 RGB 域单分支网络的检测性能，#3 RGB-Att. 表示在注意力机制下 RGB 域单分支网络结构，#4 FRGB-Att. 表示本文提出的完整的双流结构。表 6 列出了消融实验中的 4 种网络结构以及对应的实验结果对比。

表 6 消融实验结果对比

Table 6 Comparison of ablation experimental results

（单位：%）

模型	网络结构			ACC	AUC
	Freq	RGB	Att		
Freq.	√			85.89	90.39
RGB		√		85.60	89.77
RGB-A		√	√	87.33	91.60
FRGB-A	√	√	√	**87.99**	**92.71**

通过表 6 中的单流模型和双流模型的对比可以看出，频率域单流（Freq）在低质量数据集上的检测性能略高于 RGB 单流（RGB），这也就证明了频率域线索在低质量数据集上的有效性。而当 RGB 流引入注意力机制后（RGB-A），同样能够提高检测的精度。在对注意力机制下的 RGB 分支和频率域分支进行结合（FRGB-A）构建双流网络后，检测性能相比两个单独的分支（Freq 和 RGB-A）分别在 AUC 值上增加了 2.32%和 1.11%。

本文提出的双流网络在性能上优于每个单流网络，分析其原因如下：1）深度伪造技术得到的虚假人脸图像在经过压缩之后，图像中的篡改痕迹会被压缩误差所污染，仅彩色图像通道捕获到的特征信息会显著减少，难以成为后续分类的线索；2）伪造人脸图像往往同时存在着篡改区域和真实区域，因此对伪造图像引入注意力机制，使得网络模型将更多的注意力分配给篡改区域，能够显著地提升检测的性能。同时注意力模块对浅层特征图中的重要信息进行了加强，通过注意力池化层对深层特征做进一步的特征提取，联合多重注意力图得到全局的深层特征表示，给分类提供了更加有力的特征线索。

图 8 为 4 种模型在低质量数据集上对应的 ROC 曲线图，同样可以看到本文模型（FRGB-Att）的性能超过了其他 3 种消融实验中的网络结构。此外，图 9 给出了本文模型在低质量数据集上的视频级检测例子（FF++[21]-Deepfake-285-136 及其原始真实视频 285），其中纵坐标为帧预测分数，横坐标为视频帧。每一帧的分数越接近 1，说明图像经过伪造的概率越大。从图 9 可以看出，本文网络模型能以较高的概率检测出伪造视频的每一帧，即具有良好的视频级的检测性能。

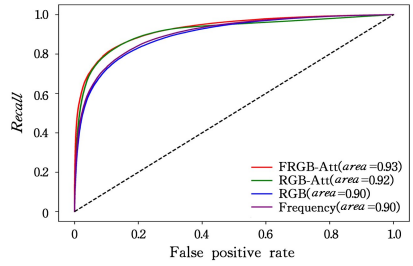
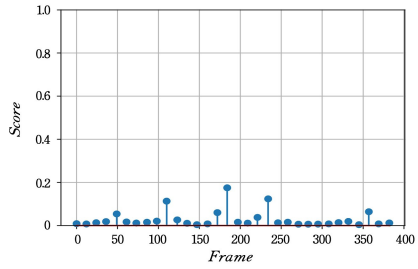
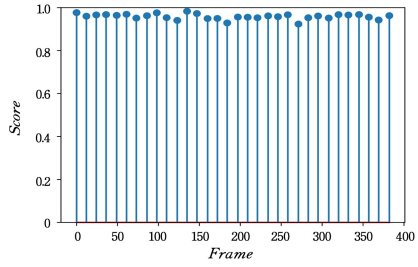


图 8 4 种模型在 FF++低质量数据集上的 ROC 曲线对比

Fig. 8 Comparison of ROC curves of four models on FF++ low-quality datasets



（a）REAL



（b）FAKE

图 9 视频级检测示例

Fig. 9 Video level detection example

结束语 深度伪造技术的迅速发展严重威胁了视频媒体的可信度，如何高效准确地检测深度伪造视频成为了当下迫切需要解决的问题。然而，现有的深度伪造检测方法普遍存在着在低质量视频集、跨库场景下检测精度较低的问题。针对上述问题，本文提出了一种双流结构的深度伪造检测方法，利用频率域特征与 RGB 空间特征的双流分支网络结构，提高了在低质量视频以及跨库场景下的检测精度。在多个数据集中的大量对比实验表明，本文提出的网络模型具有良好的检测性能和泛化能力，优于现有的检测方法。在下一步的工作中，我们将探索利用深度伪造视频中的编码特性与序列特征，设计一种轻量级的序列特征检测网络模型，实现模型检测精度与效率的进一步提升。

参 考 文 献

[1] BRANDON J. Terrifying High-Tech Porn: Creepy《deepfake》

Videos Are on the Rise[EB/OL]. Fox News，2018.(2018-02-16)[2021-06-27]. https://www. foxnews. com/tech/terrifying-high-tech-porn-creepy-deepfake-videos-are-on-the-rise.

[2] ROETTGERS J，ROETTGERS J. Porn Producers Offer to Help Hollywood Take Down Deepfake Videos[EB/OL]. (2018-02-21)[2021-06-27]. https://variety. com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/.

[3] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Networks[J/OL]. arXiv：1406. 2661，2014.

[4] 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.

[5] 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. Cham：Springer，2020：86-103.

[6] 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). IEEE，2018：1-7.

[7] ZHOU P，HAN X，MORARIU V I，et al. Two-stream neural networks for tampered face detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE，2017：1831-1839.

[8] YU N，DAVIS L，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.

[9] MCCLOSKEY S，ALBRIGHT M. Detecting GAN-Generated Imagery Using Color Cues[J]. arXiv：1812. 08247，2018.

[10] XIAO J，GONG L Y，HUANG T Q，et al. Deepfake swapped face detection based on double attention[J]. Chinese Journal of Netword and Information Security，2021，7(2)：151.

[11] BIAN M Y，PENG B，WANG W，et al. Detection of low- quality facial deepfake image based on void convolution[J]. Modern Electronics Technique，2021，44(6)：133-138.

[12] LI X R，YU K. A Deepfakes detection technique based on two-stream network[J]. Journal of Cyber Security，2020，5(2)：84-91.

[13] BAO Y X，LU T L，DU Y H，et al. Deepfake VideosDetection Method Basedoni_ResNet34 Model and Data Augmentation[J]. Comptuter Science，2021，48(7)：77-85.

[14] 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.

[15] WANG J，WU Z，CHEN J，et al. M2TR：Multi-Modal Multi-Scale Transformers for Deepfake Detection[J]. arXiv：2104. 09770，2021.

[16] SZEGEDY C，LIU W，JIA Y Q，et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：1-9.

[17] CHOLLET F. Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：1251-1258.

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

[19] 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.

[20] DOLHANSKY B，BITTON J，PFLAUM B，et al. The deepfake detection challenge(dfdc) dataset[J]. arXiv：2006. 07397，2020.

[21] 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.

[22] COZZOLINO D，POGGI G，VERDOLIVA L. Recasting residual-based local descriptors as convolutional neural networks：an application to image forgery detection[C]// Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017：159-164.

[23] LI Y，LYU S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv：1811. 00656，2018.

[24] LI L，BAO J，ZHANG T，et al. Face x-ray for more general face forgery detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：5001-5010.

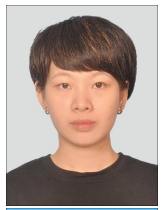
[25] MASI I，KILLEKAR A，MARIAN R，et al. Two-branch recurrent network for isolating deepfakes in videos[C]// European Conference on Computer Vision. Cham：Springer，2020：667-684.

[26] BONETTINI N，CANNAS E D，MANDELLI S，et al. Video face manipulation detection through ensemble of cnns[C]// 2020 25th International Conference on Pattern Recognition(ICPR). IEEE，2021：5012-5019.

[27] NGUYEN H H，FANG F，YAMAGISHI J，et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[J]. arXiv：1906. 06876，2019.

[28] NGUYEN H H，YAMAGISHI J，ECHIZEN I. Use of a capsule network to detect fake images and videos[J]. arXiv：1910. 12467，2019.

[29] WODAJO D，ATNAFU S. Deepfake video detection using convolutional vision transformer[J]. arXiv：2102. 11126，2021.



LI Ying，born in 1998，postgraduate，is a student member of China Computer Federation. Her main research interests include video forensics and so on.

BIAN Shan，born in 1986，Ph. D，associate professor，is a member of China Computer Federation. Her main research interests include video forensics and tampering detection.