文章编号：2096-1618(2022)05-0508-07

基于频域信息的深度伪造检测算法

蒲文博， 胡 靖
(成都信息工程大学计算机学院，四川 成都 610225)

摘要：深度伪造技术作为人脸窜改技术的一种，由它合成的换脸视频已经对隐私安全带来了巨大的隐患。现存的深度伪造检测方法通常基于传统的卷积神经网络提取合成视频中空间域的不连续信息，以判断是否为深度伪造视频。随着深度伪造技术的迭代，传统检测方法精度难以取得显著提升。与传统方法不同，文本将合成视频帧进行离散余弦变换，获得视频帧图像的频域表示，使用残差卷积网络学习频域特征，并通过双向 LSTM 提取帧间不连续信息，从而检测视频帧是否伪造。此外，针对深度伪造数据提出了一种新的数据增强方法 Xray-blur，降低换脸视频的空间域不连续性，从而提升训练难度，加强模型对不连续信息的捕获能力。实验表明，该方法在公开数据集 Celeb-DF 和 FaceForensics++上取得了优秀的准确率(ACC)和 ROC 曲线下面积(AUC)，且在面对低质量视频时，具有更好的鲁棒性。

关 键 词：深度伪造检测；频域学习；时序学习
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.05.004

0 引言

深度伪造(Deepfake)是最近广泛流行的人脸图像窜改技术。与其他使用传统图像处理手段的人脸窜改技术不同，该方法使用深度神经网络合成指定目标人脸图像并对原视频或图像进行换脸操作。具体步骤为：(1)Deepfake 使用自编码器(autoencoders)或生成对抗网络(GAN)合成目标人脸图像，该图像会保留原人脸的非身份信息，例如原人脸的动作表情、人脸位置、环境光照等；(2)通过传统的图像处理手段例如仿射变换替换原视频或图像中的人脸。由 Deepfake 技术得到的换脸视频能模仿原视频人物的行为动作。然而，人脸信息是身份认证的关键信息，在现如今的公民网络生活中已经扮演了举足轻重的角色，Deepfake 技术生成的换脸视频对于公民的个人隐私甚至社会安全构成了严重的威胁。

为应对 Deepfake 技术造成的安全威胁，大量针对 Deepfake 换脸视频的检测算法便由此提出。这些算法的检测基本思路为：由于 Deepfake 技术仍然使用了传统的图像处理技术，将合成人脸变换到原视频人脸位置，这种变换会在图像的空间域上产生不一致性(例如合成人脸与原人脸周围环境的不一致)，从而成为检测视频或图像是否为伪造的重要依据。现存的 Deepfake 检测方法大多便是使用传统的卷积神经网络(CNN)来提取图像空间域的不连续信息。然而随着 Deepfake 技术的迭代，空间域中的不连续信息变得难以捕获，传统的 Deepfake 检测方法的精度受到很大影响。

鉴于 Deepfake 技术的特殊合成方式，会在合成人脸与周围像素间产生不连续的特征，这种边缘的不连续特征往往在频域中属于高频信息。尽管随着 Deepfake 技术的迭代，这种边缘不连续特征在空间域上难以分别，但却可以在频域中被神经网络更有效地提取出来。本文通过对 Deepfake 视频帧进行离散余弦变换(DCT)，将视频映射到频域中，使用残差卷积网络提取伪造帧的频域特征，以提升模型对空间域不连续信息的捕捉能力。此外，引入了双向 LSTM 模块提取 Deepfake 视频的帧间时序信息，以提升模型对帧间不连续信息的提取能力。在此基础上，提出了一种专门针对 Deepfake 视频的数据增强方法 Xray-blur。该方法通过对人脸周围像素进行高斯模糊，提高模型对图像不连续信息的捕捉难度，以此生成难度较高的训练样本，针对性地训练模型对于不连续区域特征的提取能力。实验表明，该数据增强方法能有效提升模型的性能。

1 相关工作

为了应对 Deepfake 技术的威胁，科研工作者提出众多用于检测 Deepfake 换脸视频和图像的方法，这些方法以深度学习方法为主。Rossler 等[1]首次提出使

收稿日期：2022-03-08
基金项目：国家自然基金重点资助项目(42130608)；国家自然基金资助项目(61602065)；四川省科技厅重点研发资助项目(2021YFG0038)

用 XceptionNet[2] 进行 Deepfake 检测,但这种直接使用现存的 CNN 网络的方法由于其未对 Deepfake 数据进行针对性优化而效果欠佳。Afchar 等[3] 设计了更专注于图像的介观特性的 CNN 网络:Meso4 和 MesoInception4。Li 等[4] 提出了一种名为 FWA 的网络,首次提出通过检测图像中人脸与其周围区域的不一致性来判断该图像是否为换脸图像。在此基础上,Li 随后提出了 DSP-FWA 网络,通过引入空间金字塔池化(SPP)[5],来解决检测过程遇到的图像输入尺寸不同的问题。而后,Nguyen 等[6] 提出了 Capsule-Forensics 网络,其使用基于 VGG19[7] 的胶囊结构网络(CapsuleNet)[8] 检测换脸图像。最近,Luo 等[9] 提出了一种能捕捉图像高频噪声的检测网络来提升模型面对不同换脸数据的泛化能力。近期也有部分工作将检测重点放到图像或视频中的生物学特征上:例如 Li 等[10] 提出通过估计视频中人的眨眼频率来判断其是否为换脸视频;Javier 等[11] 设计了一种能通过 rPPG 技术估计人的心率的网络来判断视频是否由 Deepfake 技术合成。此外,近期提出的方法也更加重视视频帧间信息的提取,这些方法能通过检测视频帧间的不连续信息而判断视频是否为 Deepfake 视频。例如,Güera 等[12] 提出了一种包含 CNN 和长短期记忆(LSTM)[13] 的两阶段检测网络,以捕获帧之间的不一致信息;韩语晨等[14] 提出一种基于 Inception[15] 模块的 3D 卷积的网络,该网络则是通过 3D 卷积的方式来提取帧间时序信息,从而检测换脸视频。

2　提出的方法

本文提出方法的流程如图 1 所示。对于一个输入视频,首先使用 Dlib[16] 的人脸检测库(Dlib face detector)逐帧提取出视频中的人脸图像,随后通过 DCT 变换将人脸图像帧转换到频域;之后每帧的频域信息被送入频域学习残差 CNN 中,提取高维频域特征图;这些特征图接着被送入双向 LSTM 网络中,以提取帧间的时序信息;得出的特征图融合了频域信息特征和时序信息特征,通过网络的全连接层进行逐帧判断。

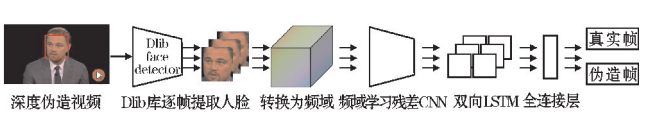

图 1　提出方法的流程

2.1　频域学习残差卷积网络

2.1.1　图像频域转换

受 JPEG 压缩算法的启发,本文提出了将 RGB 图像转换为频域图像的方法。JPEG 压缩算法通过对原始图像划分为的矩阵,再将每个矩阵中的图像通过余弦变换(DCT)转换为频域信息进行保存,这种存储方式相比传统方式有效节省了存储空间。图像频域转换流程如图 2 所示。具体步骤为:(1)人脸提取。对于输入的 RGB 视频帧,首先使用 Dlib 人脸检测库提取人脸图像,提取的人脸图像随后会由 RGB 色彩空间转换到 YCbCr 色彩空间。(2)DCT 转换。Y、Cb、Cr 3 个通道的图像会以 8×8 的矩阵块进行 DCT 变换,分别形成 3 个二维 DCT 系数矩阵。该 8×8 的矩阵块对应图 2 中 2×2 的同色方块。8×8 矩阵中保存了来自不同频域分量的 DCT 系数。位于高频分量的系数存放于矩阵右下角,而低频系数存放于矩阵左上角。(3)DCT 矩阵变维。将各个 8×8 的矩阵块中相同分量的 DCT 系数组合到同一通道中。例如,每个同色 2×2 方块的左上角小块会组合到一个通道里。这个组合过程会按照原相对位置进行,以保证相对位置的统一。这样组合会形成通道数为 $8\times8=64$ 的三维 DCT 立方。DCT 立方中每个通道中保存了位于同一频域分量的 DCT 系数。由于人脸图像由 Y、Cb、Cr 3 个通道构成,因此每个通道都会形成一个 DCT 立方矩阵。(4)DCT 矩阵连接。将每个通道形成的 DCT 立方体做连接操作,最终形成 $8\times8\times3=192$ 个通道的三维张量。(5)归一化。这个张量在经过归一化后作为网络的最终输入。设输入的 RGB 图像大小为 $H\times W\times C$,H、W 为图像的高和宽,C 为图像的通道数且 $C=3$,则该图像经过转换到频域后得到的张量大小为 $H/8\times W/8\times64\times C$。

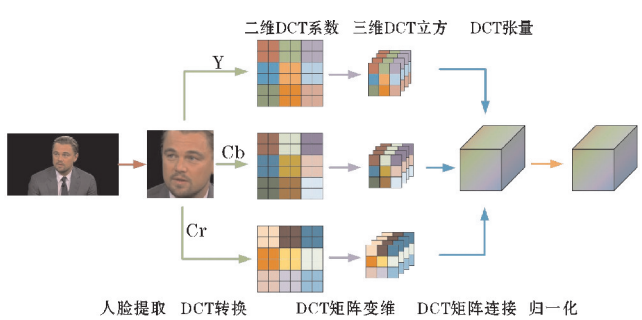

图 2　图像的频域转换

2.1.2　频域学习的 CNN

由于最终输入张量依然保持三维,该三维张量相比于传统输入的 RGB 图像,其本质上只是通道数的不同。因此只需要调整 CNN 网络的第一层卷积层的输入通道数,便可以使其处理频域信息。本文使用简单修改 ResNet-50[17] 作为频域学习的 CNN。具体修改如下:(1)由于 ResNet-50 的第一层卷积层和随后的一层最大池化层的步长为 2,为了不损失频域信息,故将这两层移除。(2)将第二层的卷积层通道数设置为和频

域张量相同的通道数,即 192,使网络能接收频域三维张量。(3)移除原 ResNet-50 的最后一层全连接层,其输出的特征图直接输入到下层 LSTM 中。这样的修改能使原 CNN 模型的结构变化最小,从而快速移植到各种 CNN 模型中。如图 3 所示,虚线框为原 Resnet-50 中被修改的部分。输入的图像经过 DCT 转换到频域后会跳过原 ResNet-50 的第一层中的 7×7 卷积和 3×3 最大池化层(Max Pool)直接输入到 1×1 卷积层中,该层的输入通道设为 192,即与频域张量通道数相等,其他部分保持不变即可。

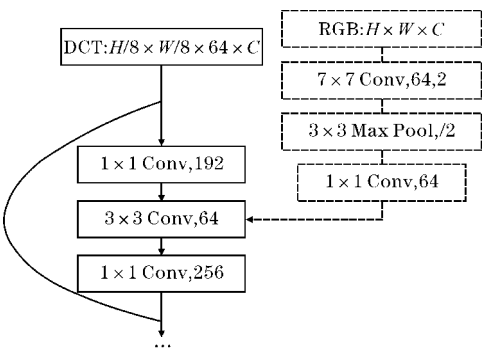

图 3 ResNet-50 的修改

2.2 双向 LSTM

由于多数 Deepfake 算法在合成换脸视频时未考虑帧间信息,从而导致合成的换脸视频在时域中会有一定的不连续现象,例如帧间的异常抖动。和空间域的不连续信息相同,捕获帧间不连续信息也能提升模型对换脸视频的检测能力。本文方法使用双向 LSTM 提取帧间不连续信息。如图 4 所示,相比传统的单向 LSTM,双向 LSTM 使用了两层 LSTM 网络能处理正向和反向传播两个路径。这种设计使得双向 LSTM 不但能考虑视频帧的历史信息,也能考虑视频帧的预测信息,有助于模型更好地提取帧间的不连续信息,从而做到更准确的判断。

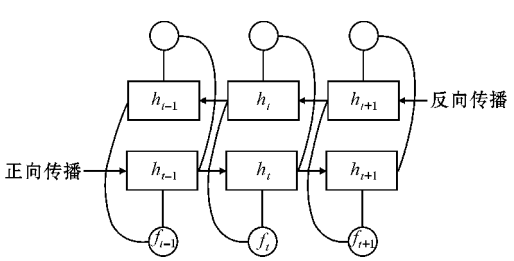

图 4 双向 LSTM

在提取到图像帧的高维频域特征图后,特征图首先被展平成一维特征向量,随后所有帧的特征向量会被堆栈成一个二维特征图,该特征图会经过采用 256 个神经元的双向 LSTM 模块对帧间的时序信息进行提取。经过双向 LSTM 提取后,得到每帧的特征向量,特征向量会被随即传入一个共享的全连接层中,该全连接层则用于输出模型对每帧是否伪造的最终判断。

2.3 数据增强 Xray-blur

鉴于主要的 Deepfake 检测方法以合成人脸与周围区域的不连续性信息作为检测的关键信息。因此如果有一种数据增强方法能针对区域进行模糊,降低此处不连续特征,以困难样本训练该模型,便能提升模型对该不连续区域的提取能力。由该观点出发,本文便提出了一种针对 Deepfake 的视频数据增强的方法 Xray-blur,该方法能对合成人脸的周围边界区域进行模糊处理。受 Face X-ray[18] 中将合成人脸的邻域以光圈的形式暴露的启发,本文将与光圈像素对应的原图像像素进行高斯模糊处理。Face X-ray 光圈生成过程如图 5 所示。首先,给定一个输入的 Deepfake 人脸图像,使用 Dlib 检测人脸的 68 个特征点,如图 5(a)所示。将这些特征点连接形成的凸包做白色填充,形成一个初始 mask,如图 5(b)所示。再经过 5×5 高斯核模糊运算形成最终 mask,记为 M,如图 5(c)所示。最后通过下列运算即可得到图 5(d)的 Face X-ray光圈图像。

$$I_{\text{Xray}} = 4 \cdot M \cdot (1-M)$$

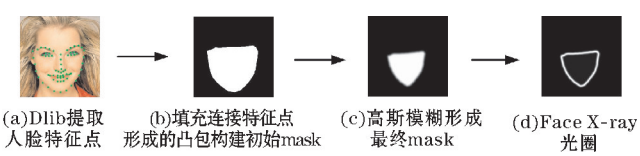

图 5 Face X-ray 光圈的生成过程

通过上述方法生成好 Xray 光圈后,将 Xray 光圈区域对应的原图像区域进行高斯模糊处理,该模糊处理过程使用了 5×5 的高斯核。原图和经过 Xray-blur 增强的图像都会作为模型的训练集进行训练。图 6 为 Xray-blur 细节展示:

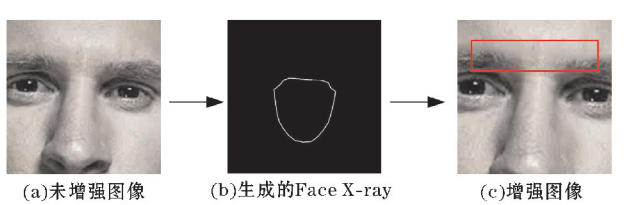

图 6 Xray-blur 增强效果

从图 6 可以明显看出,增强的合成人脸周围有显著的模糊效果,这种模糊效果会减弱合成人脸与周围像素的不连续信息,使得 Deepfake 检测模型难以提取该图像帧的不连续信息。因此,在训练中加入这些复杂样本,将会进一步增强模型对不连续信息的提取能

力。

3　实验及结果

3.1　实验设置

3.1.1　数据集和数据准备

本文将测试提出的方法在公开数据集 Celeb-DF[19] 和 FaceForensics++[1] 上的检测性能。

Celeb-DF(v2)是最近提出的具有挑战性的大规模数据集,用于评估换脸检测方法。其包含 590 个真实视频和 5639 个换脸视频。换脸视频由 59 位不同性别、年龄和种族的名人的公开可用的 YouTube 视频生成。

FaceForensics++包含从 YouTube 抓取的 1000 个真实视频和使用 4 种换脸算法生成的 4000 个换脸视频,每个算法生成的换脸视频数量为 1000。本文使用 Deepfake 算法合成的版本,故该数据为平衡数据集。除此以外,FaceForensics++包含了 3 个视频质量从高到底的 3 个版本,分为称为:raw、c23、c40。

数据准备工作如下:首先,使用 Dlib 库中的人脸检测器对数据集的每个视频中的每一帧中进行人脸检测并提取;然后将提取的人脸调整为 64×64 像素,并使用 ImageNet 的均值和标准差对人脸图像进行归一化。实验中的输入视频的帧长设置为 300,如果某个视频少于 300 帧,则重复其最后一帧以达到 300 帧。此外,将生成的每个视频帧再采用 Xray-blur 增强作为新增数据,与原视频数据一起训练模型。

对于训练集和测试集的划分,本文遵循各数据集原有划分方式。其中 Celeb-DF 训练集包含了 890 个真实视频和 5639 个换脸视频;测试集包含了 178 个真实视频和 340 个换脸视频;FaceForensics++中训练集分布包含 360 个真实视频和换脸视频;测试集和验证集分别包含 70 个真实视频和换脸视频。

3.1.2　对比方法

实验比较了 6 种 Deepfake 检测方法:

(1) DSP-FWA[4]:在 FWA 的基础上加入了空间金子塔池化(SPP)以应对不同输入尺寸换脸图像。

(2) Meso4[3]:Meso4 通过捕获深换脸图像的介观特征以判断图像是否为 Deepfake 合成。

(3) MesoInception4[3]:MesoInception4 为在 Meso4 基础上结合 Inception[15] 模块改进网络。

(4) Xception[1]:Xception 使用常用的卷积神经网络 XceptionNet 提取 Deepfake 图像的空间域信息

(5) Capsule[6]:Capsule 以 VGG19 为基础,基于胶囊网络结构检测换脸视频帧。

(6) Inception3D[14]:Inception3D 通过 3D 卷积网络同时提取伪造视频的空间信息和时序信息,以判断换脸视频是否伪造。

3.1.3　评价指标

实验中使用准确率 ACC(Accuracy)和 ROC 曲线下面积(AUC)作为评价指标,ACC 和 AUC 计算方式分别为:

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{2}$$

其中,TP 为正例预测正确的个数,FP 为负例预测错误的个数,TN 为负例预测正确的个数,FN 为正例预测错误的个数。

$$\mathrm{AUC}=\frac{\sum \mathrm{pred}_{\mathrm{pos}}>\mathrm{pred}_{\mathrm{neg}}}{\mathrm{posNum}\cdot\mathrm{negNum}} \tag{3}$$

其中,posNum 为正样本数,negNum 为负样本数,则分母表示为正负样本总的组合数。$\mathrm{pred}_{\mathrm{pos}}$表示模型对正样的预测结果,$\mathrm{pred}_{\mathrm{neg}}$为模型对负样本的预测结果,分子则表示是正样本大于负样本的组合数。AUC 的含义为分别随机从数据集中抽取一个样本,正样本的预测值大于负样本的概率。

3.1.4　参数设置及实验环境

实验在使用 NVIDIA Tesla P100 GPU 的服务器中完成,本文方法使用 PyTorch 实现。训练迭代 20 次,批处理大小为 16。训练时,使用 Adam 优化器,并使用交叉熵作为损失函数,学习率设置为 1×10^{-4}。此外,为了保证对比实验的公平,其他对比方法使用尽可能相同的设置 。

3.2　实验结果

3.2.1　公开数据集上的对比

实验对比了不同方法在两个公开数据集 Celeb-DF 和 FaceForensics++上的 ACC 和 AUC 性能。从表 1 中可以得出,本文方法优于其他方法:本文方法在 ACC 上达到了0.96,AUC 上达到了0.95。Celeb-DF 作为一个不平衡的数据集,其中换脸视频与真实视频比例为 7∶1。方法 Meso4,由于其只使用了传统的卷积网络提取空间域信息,而未考虑帧间信息,因此在面对不平衡数据时会存在偏向预测,导致 AUC 结果不理想。而 Inception3D 利用 3D 卷积网络,提取帧间不连续特征,表现效果较好。但由于其使用 3D 卷积而致使网络参数较大,加之其仍使用传统 RBG 图像作为输入,空间域不连续信息未能被有效提取,因此其性能仍弱于本文方法。本文所提方法利用频域学习 CNN 和双向 LSTM,在不平衡数据集上仍然取得了较好的结果。这

是由于其能将图像转换为频域信息而放大不连续细节。且得益于其双向 LSTM,模型能捕捉 Deepfake 视频中的帧间不连续抖动,从而进行更准确的判断。

与在 Celeb-DF 上相同,本文方法在 FaceForensics++上仍取得了优越的性能,其中 ACC 达到了0.95,AUC 达到了0.94。本文方法相比于 Inception3D,ACC 领先了 5%,AUC 领先了 3%。实验充分展示了频域学习和时序学习在 Deepfake 检测中扮演的重要角色。

表 1 在公开数据集上的性能对比

数据集	方法	ACC	AUC
Celeb-DF	DSP-FWA[4]	0.65	0.50
	Meso4[3]	0.72	0.83
	MesoInception4[3]	0.87	0.92
	Xception[1]	0.88	0.83
	Capsule[6]	0.92	0.90
	Inception3D[14]	0.93	0.92
	本文	0.96	0.95
FaceForensics++c23	DSP-FWA[4]	0.51	0.51
	Meso4[3]	0.62	0.74
	MesoInception4[3]	0.80	0.86
	Xception[1]	0.83	0.91
	Capsule[6]	0.84	0.87
	Inception3D[14]	0.90	0.91
	本文	0.95	0.94

3.2.2 消融实验

为了研究本文模型各个模块的作用,在 Celeb-DF 数据集上进行消融实验。具体步骤如下:(1)为了验证 Xray-blur 增强方法带给模型的性能提升,实验去除了 Xray-blur 增强的数据,仅使用原数据进行训练,记为“-Xb”;(2)为了考察将图像转换为频域信息从而对模型带来的提升,使用传统的 ResNet-50 代替频域学习的 CNN,记为“-Xb,-DCT”;(3)为了验证双向 LSTM 模块和帧间不连续信息在深度换脸检测中的重要作用,实验在“-Xb,-DCT”基础上移除了双向 LSTM,只使用 ResNet-50 检测深度伪造视频帧,记为“-XB,-DCT,-Bi”。

表 2 是各方法的检测结果。对比本文方法和“-Xb”可以得出,Xray-blur 增强方法带来了 2% 的 ACC 和 1% 的 AUC 提升。Xray-blur 增强方法从数据集入手,模糊对检测的关键信息,即合成人脸及其周围的边界不一致信息。通过增强数据训练的模型在面对正常数据时能更有效地提取这些信息,从而针对性地提升模型对伪造图像的判断能力。模型“-Xb,-DCT”取得了0.91的 ACC 和0.92的 AUC。相比“-Xb”,ACC 下降了 3%,AUC 下降了 2%。实验表明,经过频域转换后,模型性能有了显著提升。这种提升和 Deepfake 检测的注意目标相关,空间的不一致信息通过频域的转换而进行了放大,模型能更好地提取该信息。模型“-XB,-DCT,-Bi”取得了0.75的 ACC 和0.68的 AUC。相较于“-Xb,-DCT”,双向 LSTM 的加入使得模型提高了0.16的 ACC 和0.24的 AUC,模型性能提升显著,展示了帧间信息在换脸检测任务中的至关重要的作用。双向 LSTM 通过前后的帧间信息传播,强化了帧间信息的提取。

表 2 消融实验检测结果

方法	ACC	AUC
本文	0.96	0.95
-Xb	0.94	0.94
-Xb,-DCT	0.91	0.92
-Xb,-DCT,-Bi	0.75	0.68

3.2.3 低质量视频鲁棒性分析

Deepfake 检测算法需要对不同质量的视频具有良好的检测效果,以应对现实应用场景。尤其是 Deepfake 视频在网络上传播会受压缩算法的影响,使其质量明显下降。这种视频会对模型的检测性能有着显著的影响。为验证所提模型在不同视频质量下的鲁棒性,本节测试了模型在 FaceForensics++的 c40 数据集的性能。数据集 c40 使用了 H.264 编码器对原始视频进行低质量压缩,以模拟网络中视频的真实压缩情况。本文方法和其他方法在 c40 上的测试结果见表 3。对比表 1 中 FaceForensics++ c23 的结果,本文方法在视频检测任务中 ACC 性能下降了 3%,AUC 性能下降了 2%,但远高于其他模型在 c40 上的测试性能。该实验充分证明了本文方法面对低压缩率视频时表现出良好的鲁棒性。

表 3 在 FaceForensics++上的面对低质量视频 c40 的测试结果

方法	ACC	AUC
DSP-FWA[4]	0.50	0.50
Meso4[3]	0.65	0.73
MesoInception4[3]	0.73	0.83
Xception[1]	0.80	0.88
Capsule[6]	0.78	0.84
Inception3D[14]	0.82	0.89
本文	0.92	0.92

此外,为了展示低质量视频对模型检测能力的影响,本文方法、Capsule、MesoInception4 以及 Xception 分别在 c23 和 c40 上的预测结果见图 7。图 7 中红色框为换脸视频帧,绿色框为真实视频帧。预测结果大于

0.5则模型判断为换脸图像,小于 0.5 则模型判断为真实图像。从图 7 可以得出,相较于对 c23 的预测结果,对比模型对 c40 换脸视频帧的预测概率有一定下降;在对 c40 真实视频帧预测中出现了错误判别(红色)。相较于其他方法,本文方法做到了正确预测的同时,其输出概率更准确。

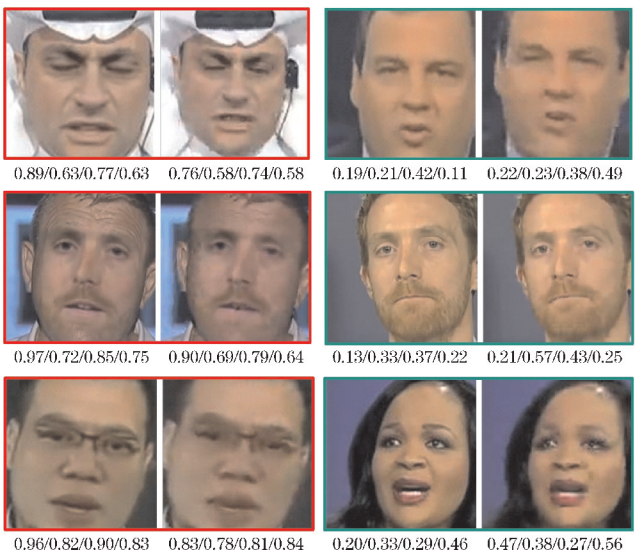

图 7　在中等质量 c23 和低质量 c40 视频帧的预测对比

4　结束语

提出了一种基于视频帧频域信息的 Deepfake 检测方法,该方法能够更好地捕获 Deepfake 合成中产生的不连续信息。此外,该方法通过引入双向 LSTM 以提取 Deepfake 视频的帧间信息,进一步提升模型对 Deepfake 视频的检测能力。针对深度换脸视频的合成特点,提出了 Xray-blur 数据增强方法,其能模糊合成人脸与周围边界区域,使得模型在增强数据训练下能更好捕捉不连续区域。相比于基于传统的卷积网络的检测方法,本方法在公开数据集上取得了优秀的检测效果,且拥有良好的应对低质量视频的能力。本文的方法仍有些不足,例如,模型虽然能捕获频域信息,但缺乏对关键频域信息的注意能力,后续研究会考虑将通道注意力模块引入模型中,使模型能在众多频域中选择有效的信息进行 Deepfake 检测。

参考文献:

[1] 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. Long Beach: IEEE, 2019: 1-11.

[2] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]. Proceedings of the 25th international conference on Machine learning. New York: Association for Computing Machinery, 2008: 160-167.

[3] 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 (WIFS). Hong Kong: IEEE, 2018: 1-7.

[4] Li Yuezun, Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts[EB/OL]. (2019-05-22) [2022-05-18]. https://arxiv.org/abs/1811.00656.

[5] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE trans on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[6] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 2307-2311.

[7] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [EB/OL]. (2017-11-08) [2022-04-18]. https://arxiv.org/abs/1710.09829.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-11)[2022-05-18]. https://arxiv.org/abs/1409.1556.

[9] 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.

[10] 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.

[11] Hernandez-Ortega J, Tolosana R, Fierrez J, et al. Deepfakeson-phys: Deepfakes detection based on heart rate estimation[EB/OL]. (2020-05-14)[2022-05-18]. https://arxiv.org/abs/2010.00400.

[12] Güera D, Delp E J. Deepfake video detection using recurrent neural networks[C]. Proceedings of 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). Auckland: IEEE, 2018: 1-6.

[13] Donahue J, Hendricks LA, Rohrbach M, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 677-691.

[14] 韩语晨, 华光, 张海剑. 基于 Inception3D 网络的眼部与口部区域协同视频换脸伪造检测[J]. 信号处理, 2021, 37(4): 567-577.

[15] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[16] King D E. Dlib-ml: A machine learning toolkit[J]. The Journal of Machine Learning Research, 2009, 10: 1755-1758.

[17] 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. Las Vegas: IEEE, 2016: 770-778.

[18] 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. Seattle: IEEE, 2020: 5001-5010.

[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. Seattle: IEEE, 2020: 3207-3216.

A Deepfake Detection Method based on Frequency Domain Information

PU Wenbo, HU Jing

(College of Computer, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: As a kind of face manipulation technology, the widespread popularity of Deepfake has brought huge hidden dangers to privacy and security. Existing Deepfake detection methods are based on a traditional convolutional neural network to extract spatial discontinuous information in the spatial domain in synthetic videos to judge whether a video is a deepfake. With the iteration of Deepfake, the accuracy of traditional detection methods cannot be significantly improved. Different from these methods, this paper performs discrete cosine transform (DCT) on Deepfake frames to obtain the frequency domain representation, uses a modified residual convolutional network to learn the frequency domain features, and extracts temporal information between Deepfake frames through bidirectional LSTM. In addition, this paper proposes a new data augmentation method called Xray-blur for Deepfake data, which reduces the spatial discontinuity of Deepfake data and enhances the model's ability to capture discontinuous information. Experiments show that this method achieves excellent accuracy and AUC on public datasets of Celeb-DF and FaceForensics++ and has better robustness against low-quality videos.

Keywords: deepfake detection; frequency learning; temporal learning