

## 空域频域相结合的唇型篡改检测方法

林佳滢<sup>1,2</sup>, 周文柏<sup>1,2</sup>, 张卫明<sup>1,2</sup>, 俞能海<sup>1,2</sup>

(1. 中国科学院电磁空间信息重点实验室, 安徽 合肥 230027; 2. 中国科学技术大学网络空间安全学院, 安徽 合肥 230027)

**摘要:** 近年来, 社交网络中的“换脸”视频层出不穷, 对说话者进行唇型篡改是其中的视频代表之一, 这给大众生活增添娱乐的同时, 对于网络空间中的个人隐私、财产安全也带来了不小隐患。大多数唇型篡改检测方法在无损条件下取得了较好的表现, 但广泛存在于社交媒体平台、人脸识别等场景中的压缩操作, 在节约像素和时间冗余的同时, 会对视频质量造成影响, 破坏空域上像素与像素、帧与帧之间的连贯完整性, 导致其检测性能的下降, 从而引发对真实视频的错判情况。当空域信息无法提供足够有效的特征时, 能够抵抗压缩干扰的频域信息就自然而然地成为重点研究对象。针对这一问题, 通过分析频率信息在图像结构和梯度反馈上的优势, 提出了空域频域相结合的唇型篡改检测方法, 有效利用空域、频域信息的各自特点。对于空域上的唇型特征, 设计了自适应提取网络和轻量级的注意力模块; 对于频域上的频率特征, 设计了不同分量的分离提取与融合模块。随后, 通过对空域上的唇型特征和频域上的频率特征进行有侧重的融合, 保留更多关键纹理信息。此外, 在训练中设计细粒度约束, 分开真假唇型特征类间距离的同时, 拉近类内距离。实验结果表明, 得益于频率信息, 所提方法能有效改善压缩情况下的检测准确性, 并具备一定的迁移性。另外, 在对核心模块开展的消融实验中, 相关结果验证了频率分量对于抗压缩的有效性, 以及双重损失函数在训练中的约束作用。

**关键词:** 人脸伪造; 人脸伪造检测防御; 唇型篡改检测; 抗压缩; 深度学习

**中图分类号:** TP309.2

**文献标志码:** A

**DOI:** 10.11959/j.issn.2096-109x.2022075

## Lip forgery detection via spatial-frequency domain combination

LIN Jiaying<sup>1,2</sup>, ZHOU Wenbo<sup>1,2</sup>, ZHANG Weiming<sup>1,2</sup>, YU Nenghai<sup>1,2</sup>

1. Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

2. School of Cyber Science, University of Science and Technology of China, Hefei 230027, China

**收稿日期:** 2022-04-06; **修回日期:** 2022-07-09

**通信作者:** 周文柏, welbeckz@ustc.edu.cn

**基金项目:** 国家自然科学基金 (U20B2047, 62072421, 62002334, 62102386, 62121002); 中国科技大学探索基金项目 (YD3480002001); 中央高校基础研究基金 (WK2100000011)

**Foundation Items:** The National Natural Science Foundation of China (U20B2047, 62072421, 62002334, 62102386, 62121002), Exploration Fund Project of University of Science and Technology of China (YD3480002001), Fundamental Research Funds for the Central Universities (WK2100000011)

**引用格式:** 林佳滢, 周文柏, 张卫明, 等. 空域频域相结合的唇型篡改检测方法[J]. 网络与信息安全学报, 2022, 8(6): 146-155.

**Citation Format:** LIN J Y, ZHOU W B, ZHANG W M, et al. Lip forgery detection via spatial-frequency domain combination[J]. Chinese Journal of Network and Information Security, 2022, 8(6): 146-155.

**Abstract:** In recent years, numerous “face-swapping” videos have emerged in social networks, one of the representatives is the lip forgery with speakers. While making life more entertaining for the public, it poses a significant crisis for personal privacy and property security in cyberspace. Currently, under non-destructive conditions, most of the lip forgery detection methods achieve good performance. However, the compression operations are widely used in practice especially in social media platforms, face recognition and other scenarios. While saving pixel and time redundancy, the compression operations affect the video quality and destroy the coherent integrity of pixel-to-pixel and frame-to-frame in the spatial domain, and then the degradation of its detection performance and even misjudgment of the real video will be caused. When the information in the spatial domain cannot provide sufficiently effective features, the information in the frequency domain naturally becomes a priority research object because it can resist compression interference. Aiming at this problem, the advantages of frequency information in image structure and gradient feedback were analyzed. Then the lip forgery detection via spatial-frequency domain combination was proposed, which effectively utilized the corresponding characteristics of information in spatial and frequency domains. For lip features in the spatial domain, an adaptive extraction network and a light-weight attention module were designed. For frequency features in the frequency domain, separate extraction and fusion modules for different components were designed. Subsequently, by conducting a weighted fusion of lip features in spatial domain and frequency features in frequency domain, more texture information was preserved. In addition, fine-grained constraints were designed during the training to separate the inter-class distance of real and fake lip features while closing the intra-class distance. Experimental results show that, benefiting from the frequency information, the proposed method can enhance the detection accuracy under compression situation with certain transferability. On the other hand, in the ablation study conducted on the core modules, the results verify the effectiveness of the frequency component for anti-compression and the constraint of the dual loss function in training.

**Keywords:** DeepFake forgery, DeepFake detection and defense, lip forgery detection, anti-compression, deep learning

## 0 引言

2017 年, 第一个针对名人的换脸视频在社交平台上发布, 从此人脸伪造技术开始走进公众视野。根据篡改区域的不同, 人脸伪造技术分为全脸替换和局部篡改。前者通常将整个源脸替换成目标脸; 后者修改部分脸部区域, 如篡改唇型来匹配音频内容, 并达到视觉上的同步。前者篡改区域大, 且伴随着身份属性的改变; 后者篡改区域小, 身份属性不变但视频内容会发生更改, 因而后者产生的社会危害更大。随着攻防一体化的发展, 针对人脸伪造的检测技术应运而生, 同样按照篡改区域划分为全脸检测技术和局部检测技术两大类。无损情况下, 唇型篡改视频中浅层纹理特征能够得到完整保留且具有连贯性, 现有的局部检测模型依靠该空域信息即可提取到具有区分性的特征, 在真实、篡改视频的鉴别任务中取得较好的表现。

然而实际生活中, 压缩作为一种基本的数据

处理方式, 普遍存在于各个数字平台中。从像素空间冗余压缩、时间冗余压缩以及编码冗余压缩 3 个维度对视频的分辨率和尺寸进行改变, 能够有效节约带宽资源, 防止自身数据冗余造成不必要的消耗。倘若外界施加的压缩强度过大, 在包括信道传输损失等多路干扰的叠加下, 视频画面极易产生大幅度的缺陷, 甚至出现人眼都无法分辨清楚视频内容的情况。常见的压缩导致的视频受损情况有高斯噪声、像素腐蚀以及高斯模糊等。

毫无疑问, 对视频采取的压缩强度越大, 浅层纹理特征就越容易被破坏, 像素与像素、帧与帧之间的连贯完整性因此被削弱。原本未经篡改的真实视频此时在基于深度神经网络的检测模型看来, 由于噪声的影响其极有可能被认为是经过篡改的假视频而发生错判。作为一种广泛使用的后处理操作, 压缩对当前的局部唇型篡改检测提出了新的挑战。

当空域信息所剩无几, 面对此类困难场景, 本质上需要解答一个问题, 即什么信息对于高压

缩视频的检测是有效的，这也是人脸伪造检测发展至今，学者们不断从各个领域借鉴新思路试图解决的问题。传统图像处理领域除了研究空域信息外，另一个重点研究对象是频域。频域信息能很好地反映图像的不同结构，表征梯度的变化，这一特性使得某些操作在频域上的效果会优于空域，因而将其引入同属于图像处理细分下的人脸伪造检测任务中，具备充分的合理性。

综合上述研究背景，本文提出了空域频域相结合的唇型篡改检测方法。从空域频域两个方面入手，分别提取各自特征并采取有侧重的融合方式，进一步突出纹理信息；另外，为了更好地区分不同真假唇型特征，拉近相同特征，在训练中引入双重损失函数，对模型构成细粒度约束。面对压缩环境下的人脸伪造场景，从多个角度对检测模型开展积极探索，促进实际人脸安全防御体系的构建。

## 1 相关工作

### 1.1 局部唇型篡改方法

早期的局部唇型篡改主要基于单幅图像或者纯视频，现阶段为了进一步营造出自然逼真的效果，通常会与音频进行结合，生成可以说话的伪造人脸视频。其核心思想是篡改目标人物的唇型以匹配当前音频的说话内容，因而在生成过程中会涉及音频、视频等多模态的特征信息。通过音频结合的局部唇型篡改技术，伪造方可以达到操控目标人物说出他们希望说的话的目的。

通过音频生成的唇型在完成张开闭合动作的同时，还要尽可能形状准确，符合人类发音规律。为了实现这一任务，深度学习领域的神经网络模型成为首选。对音频进行编码后作为特征输入生成模型中，得到对应的唇型关键点或者 3D 重建参数，再将这些唇型特征送到解码器中还原出当前唇型。

相关唇型篡改方法包括 Obama lip-sync<sup>[1]</sup>、First order motion<sup>[2]</sup>、Audio-driven<sup>[3]</sup>和 Wav2Lip<sup>[4]</sup>等。其中，Obama lip-sync 通过奥巴马的每周总统演讲视频，学习其特定的说话方式和表情姿势；Audio-driven 利用 3D 重建分别提取音频、表情参数，将属性和身份特征分开；First order motion

则通过视频驱动单幅原始人脸图像。Wav2Lip 通过输入一段动态视频和一段音频，即可实现任意人的唇型篡改。Wav2Lip 框架如图 1 所示，由 3 个模块组成：音频驱动的人脸唇型生成器、生成人脸视觉质量判别器、音频唇型同步判别器。不同于先前方法在每帧视频上进行篡改，该方法一次性输入连续 5 帧，经过编解码器结构得到相应的人脸输出帧，使用 L1 重建损失约束生成人脸与真实人脸间的距离，同时视觉质量判别器会对生成人脸进行真假判断，形成对抗式训练。

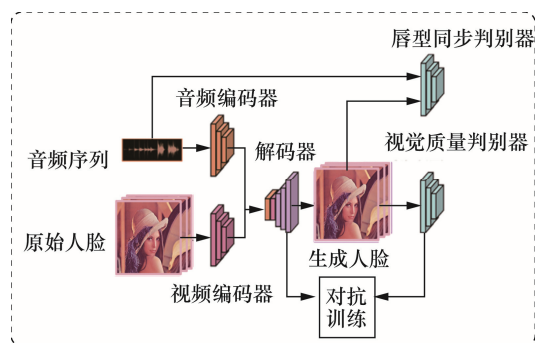


图1 Wav2Lip 框架

Figure 1 The framework of Wav2Lip

### 1.2 局部唇型检测方法

对局部唇型检测方法的发展过程进行大致追溯可以发现，其主要从语音识别、唇型识别等真人音视频任务上迁移而来。利用在真实大规模数据集上经过严密训练和测试的语音、唇读等预训练模型，可以很好地对当前伪造人脸特征进行初步筛选过滤。固定网络的前层模块，损失函数在迭代训练中更新调整最后一个全连接层的权重，实现预训练模型在人脸伪造数据集上的微调。

受到以上研究视角的启发，来自 Facebook 的团队率先提出了 Lip Forensics 算法<sup>[5]</sup>，其框架如图 2 所示，核心思想是挖掘高级语义在嘴部运动中存在的不规则性，如超出正常人嘴唇的开合幅度，或者相反地，不能自如地完成嘴唇的闭合动作。唇读预训练模型的优势在于空间上提取 3D 特征，时间上使用 MS-TCN 网络<sup>[6]</sup>描述时序特征，学习真实自然场景下嘴部运动特有的高级语义表征。因此，能有效避免一些过拟合的情况，如检测网络过于依赖低层次语义表征或者某种生成方法产生的特定伪影。

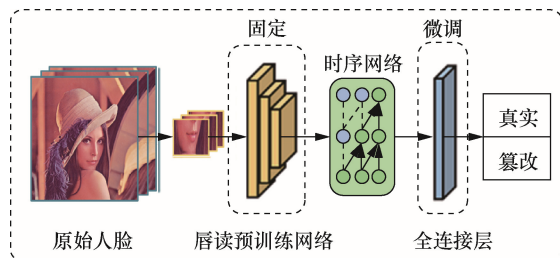


图2 Lip Forensics 框架  
Figure 2 The framework of Lip Forensics

值得一提的是,虽然预训练模型在库内和跨数据库间的实验上展现了优异的性能,但其庞大的模型结构导致适用范围受到局限,在大企业的大平台上更具备施展空间。日常生活中,面向计算机和手机等灵活终端,往往要求模型轻量化、易部署,且对于压缩场景具备一定检测能力。

### 1.3 频率特征变换

按照常规的几种频率变换方法如离散傅里叶变换、离散余弦变换,可以将图像从空域转到频域。在此基础上对频率特征进行划分,进而得到低频、中频、高频3个基本分量。低频信号往往占比最高,主要描述了图像中梯度变化平缓的区域,这意味着在颜色和内容上的波动幅度不大,与空域中的低维纹理特征具有相同性质。高频信号则与此不同,其刻画了梯度变化剧烈的部分。在图像内容出现明显转换的地方,高频信号所蕴含的能量越多,因而大多对应空域中的高维细节特征如边缘、轮廓等,甚至少部分的噪声。中频信号的梯度变化介于两者之间,一方面,减少对

图像内容的补充;另一方面,增加对细节的描绘。

实际上围绕频率信息,已经有相关的人脸伪造检测工作展开。例如,文献[7]和文献[8]均指出空域上低维浅层纹理特征的重要性,另外,文献[9]探讨了不同频率分量对检测模型的有效性。

因此,不仅要引入频率特征与空域上的唇型特征相结合,还要对频率特征进行划分得到不同分量。在频率分量的基础上进行特征融合,起到引导和增强的作用,有利于模型在压缩情况下的检测判断。

## 2 空频域唇型篡改检测

压缩操作会引发图像空域出现一定缺陷,频率信号却因自身在结构、梯度上的特性,能很好地抵抗此类影响。为了尽可能捕捉到更多有效特征减小干扰,可将空域、频域二者结合,相互取长补短,共同用于压缩场景下的唇型篡改检测任务。检测方法的框架如图3所示,采用空域频域相结合的方式,在空域上提取唇型特征,在频域上提取频率特征,经过卷积神经网络将两者进行特征融合,最终得到真实、篡改的判断。

其中,在空域的唇型特征提取模块,使用区域生成网络(RPN, region proposal network)<sup>[10]</sup>实现唇型候选区域的自适应提取,在随后的特征编码阶段引入轻量级的注意力模块分别作用于通道和空间维度,调整每个通道的权重,关注感受野中激励更强的区域。在频率特征提取模块中,

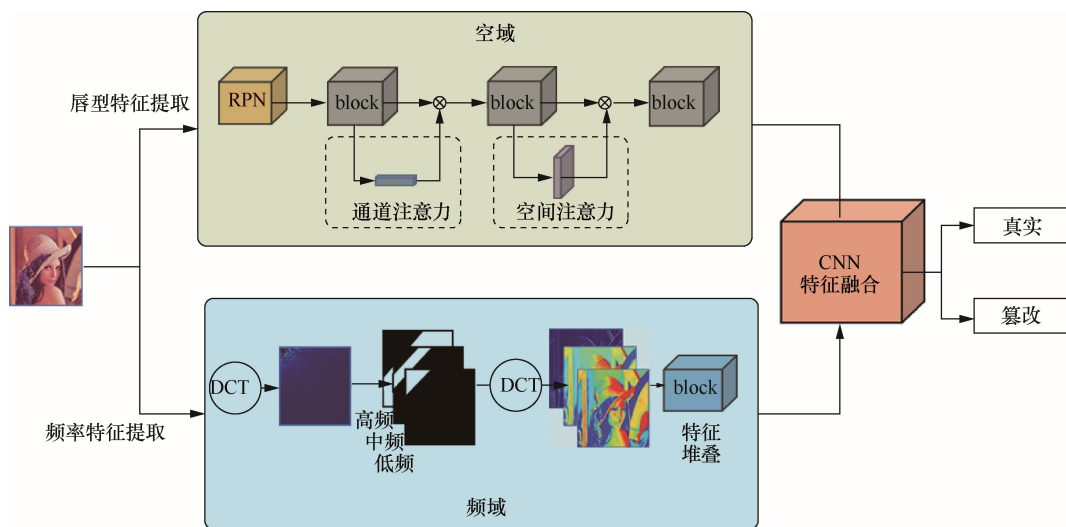


图3 空域频域相结合的唇型篡改检测方法的框架  
Figure 3 The framework of lip forgery detection via spatial-frequency domain combination

利用离散余弦变换去相关性得到频域信号, 滤波器滤波后提取低、中、高 3 个频带信息, 再利用离散余弦反变换回空域, 得到 3 个频带各自对应的图像信号。经过特征堆叠网络将 3 个图像信号合并, 形成最终的频率特征。在特征融合阶段, 将两路分支提取的特征进行结合, 经过 CNN 中的卷积、池化和全连接层, 输出最终的判决结果。

## 2.1 唇型特征提取

唇型特征提取模块的目的在于直接从空域获取显著的低维纹理特征, 核心由自适应提取模块和轻量级注意力模块组成。

RPN 自适应提取模块如图 4 所示, 目标人脸图像经过卷积提取特征图后, 进入 RPN 特征提取网络。对该特征图进行卷积得到两路分支, 上支路进行 Softmax 和 Reshape 操作生成初步的候选框, 下支路保留该特征图并与上支路结合, 得到最终的唇型候选框。在特征响应较大的人脸下半区域, 唇型候选框经过多次迭代优化, 具有一定动态性, 当遇到脸部出现较大转动或外界遮挡的情况, 候选框的大小与位置会做出灵活调整。与先前基于人脸关键点进行唇型特征提取的方法相比, 本文方法减少了人工干预和尺寸固定的限制。

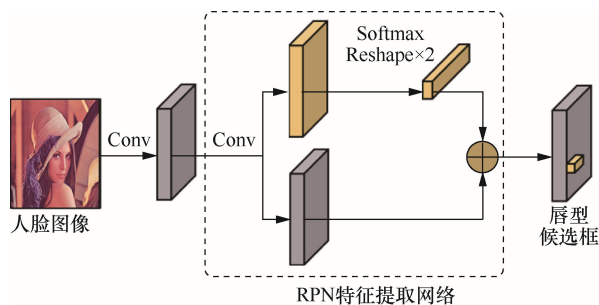


图 4 RPN 自适应提取模块  
Figure 4 RPN adaptive extraction module

轻量级注意力模块如图 5 所示, 本文分别设计了通道注意力模块和空间注意力模块, 前者注重内容理解找到关键信息, 后者确定关键信息所在的具体位置。为了使特征更好地聚合并减少参数量, 在注意力图的生成过程中添加了两池化操作, 分别为平均池化 (AvgPool) 和最大值池化 (MaxPool), 前者能保留全局范围内的反馈, 后者能突出强调局部响应大的反馈, 相互补充。

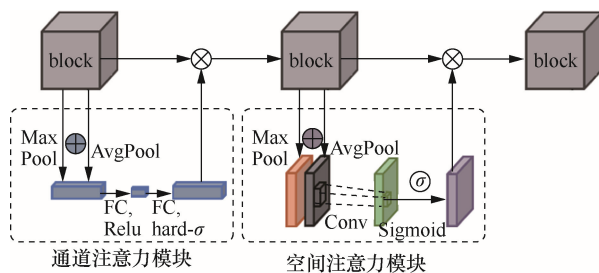


图 5 轻量级注意力模块  
Figure 5 The light-weighted attention module

设  $F$  为输入的特征图,  $\sigma$  表示激活函数, 此处使用 Relu 函数。AP 和 MP 分别表示平均池化和最大值池化, 生成的通道注意力图  $m_c(F)$  的计算如下:

$$m_c(F) = \sigma(\text{mlp}(\text{AP}(F)) + \text{mlp}(\text{MP}(F))) = \sigma(w_1(w_0(F_{\text{avg}}^c)) + w_1(w_0(F_{\text{max}}^c))) \quad (1)$$

其中, mlp 是多层感知机, 此处只包含一个隐层, 生成  $1 \times 1 \times c$  的通道注意力图, 用于表征关联程度。 $w_0$  和  $w_1$  分别表示隐层和输出层的权重。生成的空间注意力图  $m_s(F)$  的计算如下:

$$m_s(F) = \sigma(f([\text{AP}(F); \text{MP}(F)])) = \sigma(f([F_{\text{avg}}^s; F_{\text{max}}^s])) \quad (2)$$

不同于通道注意力图的相加操作, 池化后的特征图  $F_{\text{avg}}^s$  和  $F_{\text{max}}^s$  在通道维度上进行拼接<sup>[11]</sup>再输入 Sigmoid 激活函数中。

## 2.2 频率特征提取

为了更好地发挥频率特征的优势, 频率特征提取模块采取先分离后融合的策略, 如图 6 所示, 本文设计了频率特征分离模块和频率特征融合模块。

首先进行频率特征分离操作。选用离散余弦变换得到人脸频谱图。其中, 低频分量密度大, 集中在左上角; 中频分量信号靠近中间呈带状分布; 高频分量的密度较低, 占据整个右下角。设计 3 种二分类滤波器将低频、中频、高频分量各自提取出来, 其本质是由 0、1 构成的二进制掩码。

频率信号不具备空间位置和内容表达的能力, 不能直接使用在以卷积神经网络为主干的模型中, 因此采用离散余弦反变换操作将分离出的 3 种频率信号还原回空域。显然, 在还原图像中,

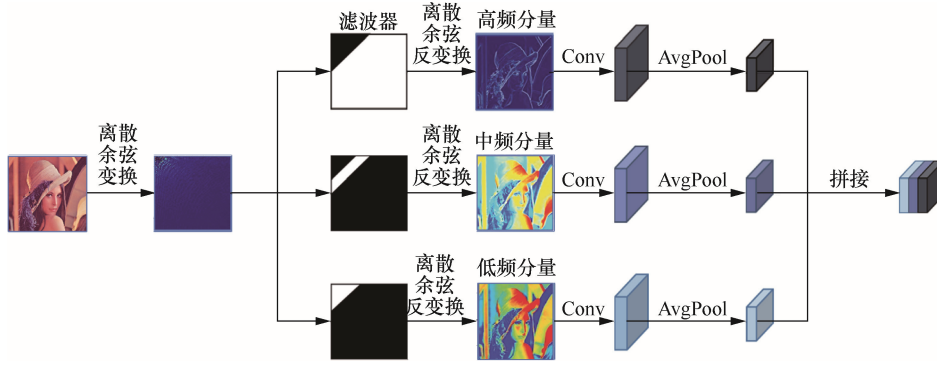


图 6 频率特征提取模块  
Figure 6 Frequency feature extraction module

低频图像响应最强，包含了大部分纹理信息；轮廓、边缘等细节主要反映在高频图像上；中频图像则介于两者之间，但三者均与原始目标人物图像有较大区别。 $x$  为输入的人脸图像， $D(x)$  为离散余弦变换， $D^{-1}(x)$  为离散余弦反变换， $f^i$  为滤波器， $y_i$  为还原图像，频率特征提取分离过程如下：

$$y_i = D^{-1}(D(x) \odot f^i) \quad i=1,2,\dots,N \quad (3)$$

其中， $N$  表示滤波器个数，在该模块中  $N=3$ 。

在频率分量的特征融合阶段，依次经过卷积和平均池化得到尺寸相同的编码特征。随后，在通道方向上按照低、中、高的顺序进行拼接，保持各频率分量独立的同时得到融合后的频率特征。

### 2.3 特征融合

特征融合是特征处理的关键模块，空域频域融合后的总体特征将对模型的最终决策起到指导作用。

特征融合模块如图 7 所示，唇型特征经过平均池化和  $1 \times 1$  卷积后得到归一化的权重，频率特征经过  $3 \times 3$  卷积与该权重相乘，产生新的频率特征。再与唇型特征相加，经过  $3 \times 3$  卷积和 Sigmoid 激活函数得到融合后的特征。

在权重的指导下，相乘后的新频率特征能呈现更多检测所需的纹理信息。相比起均匀融合，以唇型特征为主的有侧重融合，能突出有效特征，增强网络的学习能力。

### 2.4 损失函数

针对压缩场景下的唇型篡改检测任务，本文设计双重损失函数，改善原先方法中普遍采取的

基于二分类任务的粗粒度约束，进一步优化模型的检测性能。损失函数模块如图 8 所示，输入的融合特征经过卷积、平均池化后，再通过全连接层进行矩阵变换，与样本标记空间建立映射得到当前的输出结果。随后，由 Softmax Loss<sup>[12]</sup> 和 Center Loss<sup>[13]</sup> 构成的双重损失函数模块对输出结果进行误差计算。

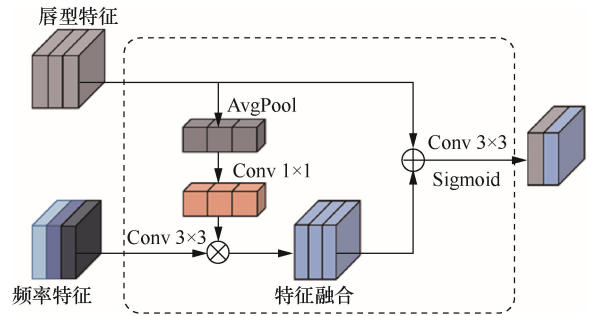


图 7 特征融合模块  
Figure 7 Feature fusion module

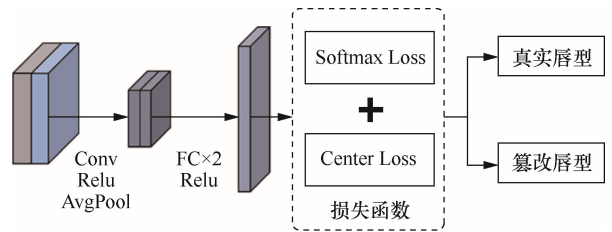


图 8 损失函数模块  
Figure 8 Loss function module

Softmax Loss 计算公式如下：

$$L_{\text{Softmax}} = -\sum_{i=1}^m \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{w_j^T x_i + b_j}} \quad (4)$$

其中， $x_i$  表示进入 Softmax Loss 前全连接层的特

征,  $y_i$  是模型当前的预测结果,  $w_j$  和  $b_j$  分别表示第  $j$  列的权重与偏置参数。训练中的每轮批次大小为  $m$ , 总的类别数为  $n$ , 在唇型篡改检测任务中取  $n=2$ , 代表真实、篡改两类。

Center loss 计算公式如下:

$$L_{\text{Center}} = \frac{1}{2} \sum_{i=1}^m \|x_i - C_{y_i}\|_2^2 \quad (5)$$

将当前批次的每个预测值  $y_i$  包含成一类, 计算得到类别中心  $C_{y_i}$ 。每批次训练完毕, 根据对应的  $L_{\text{Center}}$  对  $C_{y_i}$  进行更新。用  $\frac{\partial L_C}{\partial x_i}$  表示梯度,  $\Delta C_j^t$  表示偏导, 此时梯度为:

$$\frac{\partial L_C}{\partial x_i} = x_i - C_{y_i} \quad (6)$$

进一步, 求得偏导为:

$$\Delta C_j = - \frac{\sum_{i=1}^m y_i \cdot \left( \frac{\partial L_C}{\partial x_i} \right)}{1 + \sum_{i=1}^m y_i} \quad (7)$$

当  $y_i=1$ , 表示当前预测为真实唇型; 当  $y_i=0$ , 则表示预测为篡改唇型。分母项设置1防止分数出现除0异常。用  $C_j^{t+1}$  表示  $t+1$  时的类别中心, 设置  $\alpha$  控制  $t$  时刻  $C_j^t$  的更新速度, 由此得到的更新为:

$$C_j^{t+1} = C_j^t - \alpha \cdot \Delta C_j^t \quad (8)$$

通常取  $\alpha=0.5$ 。

至此, 将两个损失函数结合, 得到最终总的训练损失函数  $L_{\text{total}}$  为:

$$L_{\text{total}} = L_{\text{Softmax}} + \lambda L_{\text{Center}} = - \sum_{i=1}^m \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{w_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^m \|x_i - C_{y_i}\|_2^2 \quad (9)$$

其中,  $\lambda$  为 Center loss 的权重用于调节平衡, 取  $\lambda=0.003$ 。

显然, 两个损失函数的作用有所不同。在特征空间中, Softmax Loss 将不同类别间的各点分开, Center Loss 将同一类别中的各点向中心方向收缩, 共同对唇型篡改检测任务形成细粒度的约

束。模型在上述迭代优化中, 逐渐学习真实、篡改唇型间更具有区分性的特征, 提升检测性能。而大多数方法采用的二分类交叉熵则属于粗粒度约束, 仅从输出结果与标签之间的距离进行衡量, 导致类内特征不够紧凑, 在唇型篡改的关键痕迹获取上更难把握。

### 3 实验

本节首先介绍实验设置, 包括数据集和评价指标等, 随后对提出的唇型篡改检测方法进行实验, 包括库内的真假分类评估、跨方法评估, 跨数据库的迁移性评估以及关键模块的消融实验。

#### 3.1 实验设置

本文实验采用人脸伪造数据集 FaceForensics++<sup>[14]</sup> 和 Celeb-DF<sup>[15]</sup>。

FaceForensics++数据集使用4种人脸伪造方法, 其中, FaceSwap<sup>[16]</sup>和 Face2Face<sup>[17]</sup>是基于计算机图形学的方法, 而 DeepFakes<sup>[18]</sup>和 Neural Textures<sup>[19]</sup>则是基于学习的方法, 在不同方法间可进行检测评估。此外, 采用 H.264 编解码器对视频进行不同程度的压缩, 得到相应的高质量视频 (HQ,C23) 和低质量视频 (LQ,C40)。该数据集是目前首个引入压缩操作的大型公开数据集。

Celeb-DF 数据集则以视频制作精良, 检测难度较大为显著特点, 是评估模型迁移性较为理想的测试数据集之一。

评价指标方面, 采用准确率 (ACC, accuracy)<sup>[23]</sup> 和 ROC 曲线面积 (AUC, area under the curve)<sup>[20]</sup> 对本文实验进行评估。

训练中, 使用 XceptionNet<sup>[24]</sup>作为主干网络。输入的人脸图像尺寸为  $299 \times 299$ , 与 FaceForensics++数据集对齐。模型以批次大小为32的规模训练100轮, 使用学习率为0.001的 Adam 优化器<sup>[21]</sup>。

#### 3.2 真假分类评估

为了评估不同压缩强度下模型的检测性能, 在 FaceForensics++数据集的 C23 和 C40 上分别进行训练和测试, 并与现有方法在 ACC 和 AUC 指标上进行对比, 得到的实验结果如表1所示。

表 1 真假分类评估结果  
Table 1 Evaluation of real/fake classification

| 数据集 | Steg. Features <sup>[22]</sup> |     | Cozzolino et al <sup>[23]</sup> |     | Rahmouni et al <sup>[24]</sup> |     | Bayer & Stamm <sup>[25]</sup> |     | MesoNet <sup>[26]</sup> |     | XceptionNet <sup>[27]</sup> |        | 本文方法          |               |
|-----|--------------------------------|-----|---------------------------------|-----|--------------------------------|-----|-------------------------------|-----|-------------------------|-----|-----------------------------|--------|---------------|---------------|
|     | ACC                            | AUC | ACC                             | AUC | ACC                            | AUC | ACC                           | AUC | ACC                     | AUC | ACC                         | AUC    | ACC           | AUC           |
| C23 | 70.97%                         | —   | 78.45%                          | —   | 79.08%                         | —   | 82.97%                        | —   | 83.10%                  | —   | 92.39%                      | 94.86% | <b>92.88%</b> | <b>95.07%</b> |
| C40 | 55.98%                         | —   | 58.69%                          | —   | 61.18%                         | —   | 66.84%                        | —   | 70.47%                  | —   | 80.32%                      | 81.76% | <b>81.50%</b> | <b>82.81%</b> |

在比较的方法中, Steg. Features<sup>[22]</sup>是基于隐写分析特征的人工检测方法, 从 Cozzolino 等到 XceptionNet 的 5 种方法<sup>[23-27]</sup>则是基于卷积神经网络的特征学习检测方法。从表 1 中结果可以看出, 本文方法在两种压缩场景下的性能均优于先前方法。在高压压缩的 C40 上, 通过结合频率特征捕捉到关键篡改痕迹, 取得了更大的增幅。

### 3.3 库内跨方法评估

在 FaceForensics++ 数据集的不同方法间使用留一法进行评估, 即选择 3 类作为训练集, 剩下一类作为测试集, 总共产生 4 类测试结果。为了保障公平客观性, 本文实验统一使用 C40 高压压缩下的视频进行训练和测试, ACC 作为评价指标, 并与真假分类评估中的检测方法进行比较, 结果如表 2 所示。

表 2 库内跨方法评估结果  
Table 2 Evaluation across methods within FaceForensics++

| 检测方法                            | ACC           |               |               |                 |
|---------------------------------|---------------|---------------|---------------|-----------------|
|                                 | DeepFakes     | Face2Face     | FaceSwap      | Neural Textures |
| Steg. Features <sup>[22]</sup>  | 73.64%        | 73.72%        | 68.93%        | 63.33%          |
| Cozzolino et al <sup>[23]</sup> | 85.45%        | 67.88%        | 73.79%        | 78.00%          |
| Rahmouni et al <sup>[24]</sup>  | 85.45%        | 64.23%        | 56.31%        | 60.07%          |
| Bayer & Stamm <sup>[25]</sup>   | 84.55%        | 73.72%        | 82.52%        | 70.67%          |
| MesoNet <sup>[26]</sup>         | 87.27%        | 56.20%        | 61.17%        | 40.67%          |
| XceptionNet <sup>[27]</sup>     | 95.15%        | 83.48%        | 92.09%        | 77.89%          |
| 本文方法                            | <b>95.36%</b> | <b>85.47%</b> | <b>92.18%</b> | <b>78.19%</b>   |

在 4 种篡改方法中, DeepFakes 上的检测准确率普遍较高, 而在 Neural Textures 的检测上则呈现一定差距, 这可能与篡改方法自身的原理有关。DeepFakes 和 FaceSwap 属于早期的篡改方法, Neural Textures 和 Face2Face 则在前两者的基础上改进, 提高了生成视频的质量。虽然不同方法间的检测难度有所不同, 本文方法依然在库内跨方法评估上取得了更好的表现效果。

### 3.4 跨数据库迁移性评估

为了更好地模拟真实场景, 本文进行了跨数据库迁移性评估。本文实验采用 FaceForensics++ 的 C23 视频作为训练集, Celeb-DF 作为测试数据集, AUC 作为评价指标, 并与主打迁移性的 4 种相关检测方法进行比较, 具体的实验设置如表 3 所示。其中, Two-stream<sup>[28]</sup>采用双流网络结构; Multi-task<sup>[29]</sup>不仅判断人脸图像是否经过篡改, 还定位分割出篡改区域; VA-LogReg<sup>[30]</sup>使用逻辑回归模型, 聚焦于篡改方法在眼睛、牙齿上留下的视觉伪影; FWA<sup>[31]</sup>则捕捉插值和尺寸缩放引起的形变痕迹。上述方法分别从不同的检测视角出发。

从同时列出的 FaceForensics++ (C23) 和 Celeb-DF 的实验结果来看, 两个数据集的特征分布存在明显差异, 模型在后者上的表现出现普遍下降。虽然 Celeb-DF 在检测上具有相当的挑战难度, 本文方法依然取得了 AUC 上的提升, 达到 66.24%, 在跨数据库间展现出一定的迁移性。

表 3 跨数据集迁移性评估结果  
Table 3 Evaluation of transferability across datasets

| 检测方法                        | AUC             |               |
|-----------------------------|-----------------|---------------|
|                             | FaceForensics++ | Celeb-DF      |
| Two-stream <sup>[28]</sup>  | 70.10%          | 53.80%        |
| Multi-task <sup>[29]</sup>  | 76.30%          | 54.30%        |
| VA-LogReg <sup>[30]</sup>   | 78.00%          | 55.10%        |
| FWA <sup>[31]</sup>         | 80.10%          | 56.90%        |
| MesoNet <sup>[26]</sup>     | 84.70%          | 54.80%        |
| XceptionNet <sup>[27]</sup> | 94.86%          | 65.50%        |
| 本文方法                        | <b>95.07%</b>   | <b>66.24%</b> |

### 3.5 消融实验

#### 3.5.1 频率分量选择实验

低频、中频、高频分量分别包含不同的图像特征, 在压缩场景下的唇型篡改检测中产生的作用也不尽相同。设计实验对不同频率信号下的模型检测性能进行评估, 结果如表 4 所示。

表4 频率分量选择实验结果  
Table 4 Ablation study of frequency components

| 数据集 | XceptionNet |        | XceptionNet+低频 |        | XceptionNet+低频+中频 |        | XceptionNet+低频+中频+高频 |               |
|-----|-------------|--------|----------------|--------|-------------------|--------|----------------------|---------------|
|     | ACC         | AUC    | ACC            | AUC    | ACC               | AUC    | ACC                  | AUC           |
| C23 | 91.79%      | 94.08% | 91.85%         | 94.10% | 91.92%            | 94.27% | <b>91.94%</b>        | <b>94.29%</b> |
| C40 | 80.77%      | 82.18% | 80.89%         | 82.26% | 81.06%            | 82.44% | <b>81.10%</b>        | <b>82.48%</b> |

表4中第一行基线模型 XceptionNet 作为对照组,没有频率分量输入。从压缩程度来看,在高压缩的 C40 上,性能提升幅度更大;从频率信息来看,相比高频信息,加入低频和中频信息的提升作用更大,但3个分量的加入均对模型的检测起到正向作用。

### 3.5.2 损失函数选择实验

本文方法的损失函数由 Softmax Loss 和 Center Loss 两部分组成,实验如表5所示。与第一行仅使用 Softmax Loss 的粗粒度约束场景相比,加上 Center Loss 的约束项后,在 C23 和 C40 的场景中模型的性能均有所提高。结果表明,改进后的损失函数在模型训练中进行了细粒度约束,在分开不同特征的同时,引导同一类特征向中心方向更加靠拢。

表5 损失函数选择结果  
Table 5 Ablation study of loss function

| 数据集 | Softmax Loss |        | Softmax Loss+Center Loss |               |
|-----|--------------|--------|--------------------------|---------------|
|     | ACC          | AUC    | ACC                      | AUC           |
| C23 | 91.94%       | 94.29% | <b>92.88%</b>            | <b>95.07%</b> |
| C40 | 81.10%       | 82.48% | <b>81.50%</b>            | <b>82.81%</b> |

## 4 结束语

针对网络空间中广泛存在的压缩操作,对当前唇型篡改检测带来的挑战,本文充分挖掘频率信号抗干扰的特性,提出了空域频域相结合的唇型篡改检测方法。空域上,采用 RPN 自适应提取与轻量级注意力两个模块,关注局部重点区域,实现唇型特征的灵活提取;频域上,使用离散余弦变换与反变化,提取低频、中频、高频率分量再进行通道上的堆叠,保持各分量独立性。随后,在唇型特征指导下对两路特征进行有侧重的融合。训练阶段,为了对模型形成细粒度约束,采用由 Softmax Loss 和 Center Loss 构成的双重损失函数。实验结果表明,与现有方法相比,本文

方法在检测准确性与迁移性上取得更好表现。此外,消融实验的结果证明了频率分量和细粒度约束的有效性。

随着人脸伪造方法的精细化发展,未来的检测工作除了考虑现有的空域、频域外,还可能拓展至时域,从3种特征维度上挖掘出更多潜在篡改痕迹,突破先前框架,进一步提升面向实际的检测性能。

### 参考文献:

- [1] SUWAJANAKORN S, SEITZ S M, KEMELMACHER- SHLIZERMAN I. Synthesizing Obama: learning lip sync from audio[J]. ACM Transactions on Graphics (TOG), 2017, 36: 1-13.
- [2] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[J]. ArXiv, 2019, abs/2003.00196.
- [3] YI R, YE Z, ZHANG J, et al. Audio-driven talking face video generation with learning-based personalized head pose[J]. arXiv: 2002.10137v2, 2020.
- [4] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 484-492.
- [5] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: a generalisable and robust approach to face forgery detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 5037-5047.
- [6] FARHA Y A, GALL J. MS-TCN: multi-stage temporal convolutional network for action segmentation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3570-3579.
- [7] QIAN Y Y, YIN G J, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//Proceedings of Computer Vision – ECCV 2020. 2020: 86-103.
- [8] LI J M, XIE H T, LI J H, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 6454-6463.
- [9] CHEN S, YAO T P, CHEN Y, et al. Local relation learning for face forgery detection[J]. arXiv:2105.02577, 2021.
- [10] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

- [11] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2261-2269.
- [12] SUN Y, WANG X G, TANG X O. Deep learning face representation from predicting 10, 000 classes[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1891-1898.
- [13] WEN Y D, ZHANG K P, LI Z F, et al. A discriminative feature learning approach for deep face recognition[C]//Proceedings of Computer Vision – ECCV 2016. 2016: 499-515.
- [14] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics++: learning to detect manipulated facial images[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 1-11.
- [15] LI Y Z, YANG X, SUN P, et al. Celeb-DF: a large-scale challenging dataset for DeepFake forensics[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 3204-3213.
- [16] Faceswap. Faceswap github[EB].
- [17] THIES J, ZOLLHÖFER M, STAMMINGER M, et al. Face 2 face: real-time face capture and reenactment of RGB videos[J]. ArXiv, 2019, abs/2007.14808.
- [18] DeepFakes. Deepfakes github[EB].
- [19] THIES J, ZOLLHÖFER M, NIEBNER M, et al. Real-time expression transfer for facial reenactment[J]. ACM Transactions on Graphics, 2015, 34(6): 1-14.
- [20] LI L Z, BAO J M, ZHANG T, et al. Face X-ray for more general face forgery detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 5000-5009.
- [21] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. CoRR, 2015, abs/1412.6980.
- [22] FRIDRICH J J, KODOVSKÝ J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
- [23] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[J]. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018: 1-7.
- [24] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1800-1807.
- [25] 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.
- [26] BAYAR B, STAMM M C. A deep learning approach to universal image manipulation detection using a new convolutional layer[C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. 2016: 5-10.
- [27] RAHMOUNI N, NOZICK V, YAMAGISHI J, et al. Distinguishing computer graphics from natural images using convolution neural networks[J]. 2017 IEEE Workshop on Information Forensics and Security (WIFS), 2017: 1-6.
- [28] ZHOU P, HAN X T, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 1831-1839.
- [29] NGUYEN H H, FANG F M, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[C]//Proceedings of 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems. 2019: 1-8.
- [30] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[J]. 2019 IEEE Winter Applications of Computer Vision Workshops(WACVW), 2019: 83-92.
- [31] LI Y, LYU S. Exposing DeepFake videos by detecting face warping artifacts[J]. ArXiv, 2019, abs/1811.00656.

#### [作者简介]



林佳滢（1997—），女，江西赣州人，中国科学技术大学硕士生，主要研究方向为人工智能安全、信息隐藏。



周文柏（1992—），男，安徽合肥人，中国科学技术大学特任副研究员，主要研究方向为信息隐藏、人工智能安全。



张卫明（1976—），男，河北定州人，中国科学技术大学教授、博士生导师，主要研究方向为信息隐藏、多媒体内容安全、人工智能安全。



俞能海（1964—），男，安徽无为，中国科学技术大学教授、博士生导师，主要研究方向为多媒体信息检索、图像处理与视频通信、数字媒体内容安全。