基于局部纹理差异特征增强的 Deepfake 检测方法

韦争争
安徽理工大学 计算机科学与工程学院,安徽 淮南 232001

摘 要:目的 针对当前 Deepfake 检测侧重全局伪造特征,而局部纹理差异特征利用不足导致模型泛化性能差的问题,提出一种基于局部纹理差异特征增强的 Deepfake 检测模型,通过挖掘伪造图像内在的空间伪造模式,提高检测的准确性和泛化性。**方法** 模型首先通过中心差分卷积操作捕捉像素强度和像素梯度两种信息,从而获得更精确的局部纹理差异信息,提高对伪造图像的敏感性。其次,构建双层注意力模块,旨在利用空间注意力学习位置敏感的权重信息,并通过通道注意力自适应调整通道重要性,定位重要纹理差异特征的位置,增强纹理差异特征的表示。**结果** 在高质量和低质量的 FaceForensics++数据集上的实验,平均准确率分别达到了 97.36%和 92.37%,而 Celeb-DF 数据集上的跨数据集实验获得了比当前先进的检测模型更好的泛化性,大量的消融实验表明了方法的有效性。**结论** 实验表明:引入中心差分和双层注意力模块后模型能够更好地捕捉图像的纹理差异信息,适应不同场景和压缩率的伪造检测,有效提高了 Deepfake 检测的准确性和泛化性。

关键词: Deepfake 检测;纹理差异;中心差分卷积;空间注意力;通道注意力
中图分类号:TP391 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0002.011

Deepfake Detection Based on Local Texture Difference Feature Enhancement

WEI Zhengzheng
School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective Current Deepfake detection methods primarily focus on global forgery features, leading to poor generalization performance of the model due to insufficient utilization of local texture contrast features. To address this issue, a Deepfake detection model based on local texture difference feature enhancement was proposed, aiming to improve detection accuracy and generalization by exploring intrinsic spatial forgery patterns in forged images. **Methods** Firstly, the model captured both pixel intensity and pixel gradient by center difference convolution operation, to obtain more accurate local texture difference information and improve the sensitivity to forged images. Secondly, a dual-layer attention module was constructed, aiming to use spatial attention to learn location-sensitive weighting information and adaptively adjust the channel importance through channel attention to locate the position of important texture disparity features and enhance the representation of texture disparity features. **Results** Experiments on high-quality and low-quality FaceForensics++ datasets obtained average accuracies of 97.36% and 92.37%, respectively, while cross-dataset experiments on the Celeb-DF dataset obtained better generalization performance than current state-of-the-art detection models. Extensive ablation studies validate the effectiveness of the proposed method. **Conclusion** Experiments show that integrating center difference convolution and a dual-layer attention module enables the model to better capture texture difference information in images, adapt to different scenarios and compression rates in forgery detection, and effectively improve the accuracy and generalization of Deepfake detection.

Keywords: Deepfake detection; texture difference; center difference convolution; spatial attention; channel attention

收稿日期:2023-09-25 **修回日期:**2023-11-18 **文章编号:**1672-058X(2025)02-0078-08
作者简介:韦争争(1999—),男,安徽阜阳人,硕士研究生,从事计算机视觉研究.
引用格式:韦争争. 基于局部纹理差异特征增强的 Deepfake 检测方法[J]. 重庆工商大学学报(自然科学版),2025,42(2):78-85.
WEI Zhengzheng. Deepfake detection based on local texture difference feature enhancement[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(2): 78-85.

1 引　言

深度学习技术为社会带来便利的同时其负面影响也不可忽视,Deepfake 正是其中的代表之一。Deepfake 是指通过深度学习模型篡改人脸图像和视频,主要包括人脸交换和面部重演两种伪造[1]。近年来,生成对抗网络的发展使 Deepfake 的制作门槛和成本越来越低,非专业人士也能制作虚假图像和视频[2]。这些虚假内容的广泛传播严重破坏了个人名誉、损害了公众对媒体的信任、对社会构成潜在巨大威胁。因此,迫切需要有效的 Deepfake 检测方法。

对于早期的拼接、复制和移动等传统人脸图像伪造技术,研究人员提出了各种检测方法。例如,通过照明颜色[3]、颜色滤波器阵列模式[4]、模糊类型不一致[5]等手工制作的特征区分图像中的特定区域是否被篡改。这些手工特征放大了真实图像和虚假图像之间细微的差异,对特定的伪造技术有很高的检测准确率。然而,随着基于深度学习的 Deepfake 的发展,虚假人脸图像变得更加多样和复杂,这些早期的基于简单的图像处理和特征提取技术的检测往往难以捕捉到这些复杂的特征,导致检测性能下降。因此,为了捕捉到虚假人脸更高级的伪造特征和模式,研究人员们转向了基于深度学习的方法。

基于深度学习的检测方法通常是利用 Deepfake 制作过程中生成的特定伪影识别人脸图像的真或假。伪造人脸图像生成时通常需要将生成人脸与原始人脸背景融合,而由于伪造算法的缺陷,虚假人脸图像的头部姿势、眼睛颜色、牙齿等面部特征难以完美匹配原始人脸,导致 RGB 统计特征、频域信息、面部纹理等不一致特征的产生。Li 等[6]分析了生成图像与真实图像在色彩空间中相邻像素之间的相关性,发现两者在 HSV 和 YCbCr 色彩空间的色度分量上的统计特性是不同的,在残差域的差异更为明显。基于这个发现,提出利用共生矩阵提取多个颜色分量残差图像的特征,进而构建用于虚假图像识别的有效特征集,经过简单的训练即可取得良好的检测效果。刘贤刚等[7]设计了一种基于人脸特征点对齐的检测框架,通过制订人脸检测、定点、对齐、特征提取、假脸识别等检测流程,并引入特征点对齐保障假脸检测效果,在几种主流框架上均取得较好的检测结果。这些手工制作特征与 CNNs 结合的检测方法计算量小、训练难度低,但严重依赖数据集的质量、类型、选择的特征以及训练集和测试集之间的对齐,无法适用低质量和未知数据集的检测。

为了解决上述问题,Qian 等[8]提出频率线索可以很好地挖掘不同质量的 Deepfake 图像共有的伪造模式,提出了 F3-Net (Frequency in Face Forgery Network)学习模型用于提取频率特征,模型由频率感知图像分解和局部频率统计两个频率感知分支组成,通过 CNN 提取细微的伪造模式和高级语义特征,结合交叉注意模块协同学习。方法在低质量人脸伪造检测方面效果突出,但频率信息不能完整表征人脸图像特征,从而在一定程度上影响检测性能。Luo 等[9]在此基础上提出结合图像的空间纹理特征和频域信息增强特征表示。通过双流结构分别提取 RGB 图像的残差引导空间纹理特征和多尺度高频特征,结合跨模态的注意力模块建模两种模式的特征之间的交互,显著提升了 Deepfake 检测的准确率和泛化性。

然而真实图像和虚假图像之间的差异通常是微妙和局部的,通过骨干网络提取全局特征并不是最优的。因此,Zhao 等[10]引入多头注意力机制使网络关注不同的局部区域,并通过纹理特征增强模块分离低频分量以放大浅层特征中的纹理信息,经过放大的局部纹理特征显著改善了网络性能。通过分析上述深度学习网络在 Deepfake 检测任务中的行为,发现局部人脸纹理信息在伪造检测中至关重要。然而之前的研究中局部纹理信息仅作为辅助信息改善特征表示,缺乏了更深入的研究。

针对上述问题,聚焦于人脸图像的局部区域纹理信息,与以往将局部纹理视为辅助信息不同,明确了局部纹理的重要性,旨在更好地捕捉局部纹理信息。其次,引入了局部纹理差异特征,提出使用局纹理差异特征来建模局部区域附加信息与纹理特征之间的关系,挖掘伪造人脸图像更一般性的伪造模式。同时,为了提取高质量的纹理差异特征,提出了中心差分卷积(Center Difference Convolution,CDC)模块与双层注意力(Dual-layer Attention,DA)结合的网络(CDCDANet)。CDC 允许网络在处理特征时考虑中心与局部区域像素信息的关系,能够同时利用像素强度级和像素梯度级信息,因此,利用中心差分卷积使模型倾向于学习更具泛化性的局部纹理差异特征。然而,由于中心差分卷积需要对所有相邻区域特征进行差分操作,导致差异特征存在较大的冗余。为了精炼特征,增强局部纹理差异特征的表示,引入了 DA 模块。通过双层注意力,使网络更专注重要的纹理差异特征,从而增强网络特征表示能力,改善检测性能。

2 模型设计细节

基于中心差分卷积和注意力机制的检测方法由基于 ResNet-50 架构的骨干网络,计算局部纹理差异特征的中心差分卷积及纹理差异特征增强的双层注意力模块组成。

2.1 整体框架

提出的模型总体框架如图 1 所示。由于图像中的人脸区域是主要研究对象,而图像背景会造成信息冗余,因此准确地检测和定位图像中的人脸至关重要。多任务卷积神经网络(Multi-Task Convolutional Neural Networks,MTCNN)引入了多任务学习的优势,能够同时处理人脸检测、人脸关键点定位等多个相关任务,高效地检测图像中的人脸。因此,首先通过 MTCNN 定位人脸,截取并保存为图像,图像尺寸为 224×224,作为网络的输入;中心差分卷积是一种特殊的卷积操作,用于捕获图像中局部纹理差异特征。通常涉及在局部窗口内对像素梯度信息进行卷积运算,同时结合像素强度信息,建模纹理差异。模块有助于检测人脸图像局部区域之间的微弱纹理差异,提高对伪造图像的敏感性;随后,使用空间注意力和通道注意力级联组成的双层注意力模块。其中,空间注意力能够关注图像的不同区域,使模型更有能力识别不同位置的特征,而通道注意力有助于选择模型中最重要的特征通道,从而减少了冗余信息。通过双层注意力强化模型在局部区域内的特征表示,更好地捕获真伪人脸图像之间的差异。

在 ResNet-50 网络的基础上,方法使用中心差分卷积模块替换原始 Bottleneck 的 3×3 卷积,同时在 1×1 卷积后添加双层注意力模块,从而构成新的 Bottleneck。将新的 Bottleneck 嵌入到原始 ResNet-50 结构的每一个 Layer 中,输入输出尺寸保持不变。最后,经过平均池化和全连接层后进行二分类,从而预测给定人脸图像的真或假。

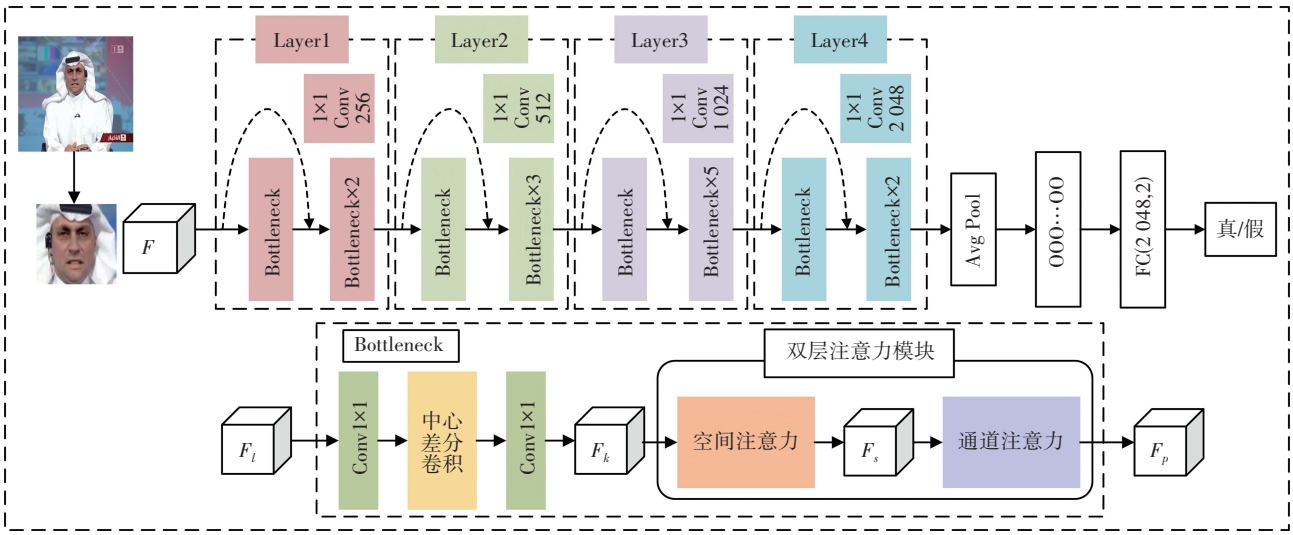


图 1 模型总体设计

Fig. 1 General design of the model

2.2 中心差分卷积

中心差分卷积的目的是提取像素强度和梯度信息,从而利用局部纹理的判别信息。对于像素强度信息使用 ResNet-50 原始 Bottleneck 中的 3×3 卷积提取。设输出特征图 $F_l \in \mathbf{R}^{C\times H\times W}$,其中 C,H 和 W 分别代表通道数,高度和宽度,从特征图 F_l 提取下一层像素强度特征图的过程如式(1)所示:

$$F_{l+1}(p_n) = \sum_{p_c \in c} w(p_c) F_l(p_n + p_c) \tag{1}$$

式(1)中,$F_{l+1} \in R^{C\times H\times W}$,$p_n$ 表示特征图 F_l 和特征图 F_{l+1} 的当前位置,c 则是卷积运算的局部感受野区域,而 p_c 则列举了 c 中的位置。例如,在 3×3 的卷积核中,$p_c \in \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$,$w(p_c)$ 表示卷积核的权重。

对于像素梯度信息,使用卷积运算与差分思想相结合的中心差分卷积提取,具体过程如图 2 所示。设从输入特征图 $F_l \in \mathbf{R}^{C\times H\times W}$ 中经中心差分卷积提取下一层像素梯度特征图 F_{l+1} 的计算过程如式(2)所示:

$$F_{l+1}(p_n) = \sum_{p_c \in c} w(p_c)(F_l(p_n + p_c) - F_l(p_n)) \tag{2}$$

式(2)中,$F_{l+1} \in \mathbf{R}^{C\times H\times W}$。

输入特征图经过 3×3 卷积和中心差分卷积得到像素强度信息和像素梯度信息描述。将像素强度特征图和像素梯度特征图使用参数 α 结合得输入特征图 F_l 完

整的伪造图像纹理差异信息描述,如式(3)所示:

$$F_{l+1}(p_n)=(1-\alpha)\sum_{p_c\in c}w(p_c)F_l(p_n+p_c)+\alpha\sum_{p_c\in c}w(p_c)(F_l(p_n+p_c)-F_l(p_n)) \tag{3}$$

式(3)中,经过分解合并后,最终的表示如下:

$$F_{l+1}(p_n)=\sum_{p_c\in c}w(p_c)F_l(p_n+p_c)-\alpha\sum_{p_c\in c}w(p_c)F_l(p_n) \tag{4}$$

式(4)中,描述的是完整的中心差分卷积,其中 α 是超参数且 $\alpha\in[0,1]$,目的是用于平衡两种像素信息。

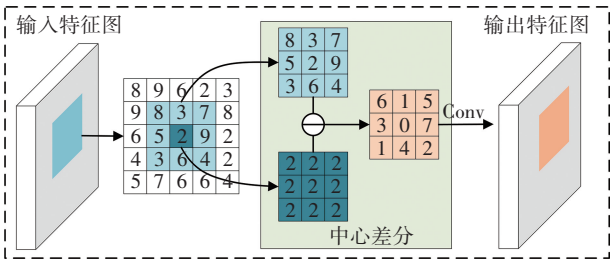


图 2　中心差分卷积的过程

Fig. 2　Process of center difference convolution

2.3　双层注意力模块

通过中心差分卷积获得了纹理异常特征的表示,但纹理差异特征的提取不仅需要关注纹理差异,还要考虑异常特征的位置信息,加强异常特征的表达能力,进一步决策。注意力机制可以结合输入数据生成权重分布,从而引导模型的关注重心,因此通过注意力机制指导特征提取的方向,关注纹理特征的局部区域以加强表征。基于文献[11]的工作,构建了双层注意力模块。如图 3 所示,模块由空间注意力模块和通道注意力模块串联组成。设输入特征图 $F_{l+1}\in\mathbf{R}^{C\times H\times W}$,对输入特征图进行通道维度的平均池化操作平均池化,得到维度为 $1\times H\times W$ 的平均池化特征图,从而获取整个区域的平均信息,在分配权重时增强全局信息表达。同时使用最大池化操作得到 $1\times H\times W$ 的最大池化特征图,强调局部区域内显著的差异信息。将两个特征图在通道维度连接,经过一个 7×7 的卷积和 Sigmoid 激活后得到位置敏感权重 M_I,从而形成对空间位置的有效描述。将权重 M_I 与输入特征图 $F_{l+1}\in\mathbf{R}^{C\times H\times W}$ 对应像素相乘得到纹理差异的空间特征图 $F_s\in\mathbf{R}^{C\times H\times W}$,具体的计算步骤如式(5)、式(6)、式(7)所示:

$$\mathrm{Cat}_{\mathrm{out}}=\mathrm{Cat}([\mathrm{AvgPool}(F_k),\mathrm{MaxPool}(F_k)],dim=1) \tag{5}$$

$$M_I=\sigma(\mathrm{Conv7\times7}(\mathrm{Cat}_{\mathrm{out}})) \tag{6}$$

$$F_s=M_I\otimes F_k \tag{7}$$

式(5)中,Cat 是连接函数,AvgPool 表示平均池化操作,MaxPool 是最大池化操作,$dim=1$ 表示沿通道维度的连接,式(6)中的 σ 是 Sigmoid 函数,式(7)中的⊗表示元素乘法。使用通道注意力模块增强重要纹理差异特征通道,提高网络捕获关键特征的能力。在模块中,对经过空间注意力模块增强得到的特征图 $F_s\in\mathbf{R}^{C\times H\times W}$ 进行平均池化,生成维度为 $C\times1\times1$ 的平均池化特征图,捕获通道之间的整体信息,从而调整通道的重要性。最大池化通过压缩空间信息,得到 $C\times1\times1$ 大小的最大池化特征图,更好地区分通道之间的不同特性。将最大池化特征图和平均池化特征图输入到同一个多层感知机,经过 Sigmoid 函数激活获得输入特征图的权重 M_O,与输入特征图 F_s 对应元素相乘得到通道特征图 F_p,计算过程如式(8)、式(9)所示:

$$M_O=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(F_s))+\mathrm{MLP}(\mathrm{MaxPool}(F_s))) \tag{8}$$

$$F_p=M_O(F_s)\otimes F_s \tag{9}$$

式(8)中,MLP 指的是多层感知机,F_s 为经过空间注意力得到的空间特征图。模块通过空间注意力捕获位置敏感信息,利用通道注意力自适应调整通道重要性,使网络更关注重要特征,有效增强了纹理差异特征表示。

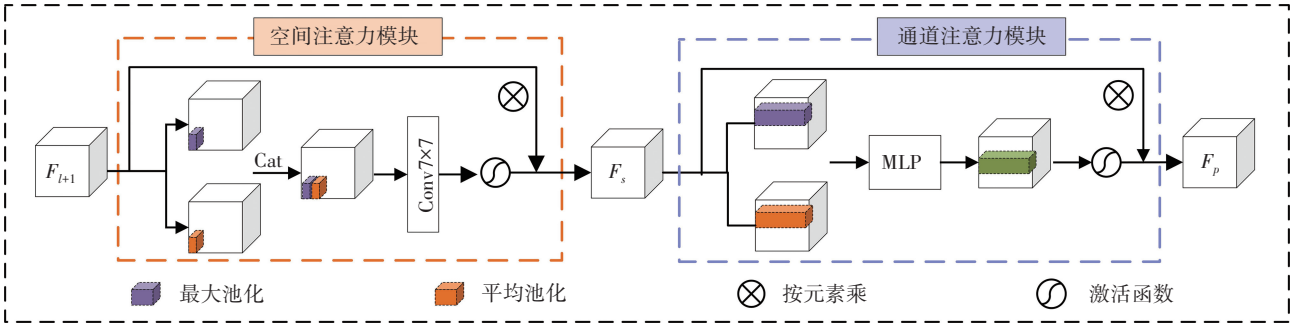


图 3　双层注意力模块的结构

Fig. 3　Architecture of the bilayer attention module

3 实验结果及分析

3.1 数据集介绍

FaceForensics++(FF++)数据集[12]和Celeb-DF数据集[13]是Deepfake检测领域广泛使用的开源数据集。FF++数据集由1 000个原始视频以及4个不同伪造算法生成的视频子数据集Face2Face(F2F)、FaceSwap(FS)、DeepFakes(DF)和NeuralTextures(NT)组成,每个子数据集包含1 000个伪造视频。同时针对不同分辨率的数据需求,采用H.264编码中的C0,C23以及C40进行压缩。而Celeb-DF则是一个以人脸交换为中心的大规模数据集,包括从YouTube上收集的590个真实视频和生成的5 369个假视频。为了验证所提方法的有效性,选择在检测难度高的中等压缩的FF++(C23)及强压缩的FF++(C40)和Celeb-DF数据集上进行训练和验证。在实验中,参考之前方法对数据集的处理,以6∶2∶2的比例划分训练集,验证集和测试集。

3.2 实验设置和评估指标

整个算法采用PyTorch框架实现。根据之前的工作经验,超参α设置为0.7,模型训练期间,batchsize设置为64,epoch为150,初始学习率为0.000 4,并使用带余弦下降的AdamW优化器进行梯度更新。实验中使用的硬件平台包括:CPU为Intel Xeon Platinum 8255C,GPU为RTX 3090,内存43 G,显存24 G。软件平台由PyTorch 1.7,CUDA 11.0,cuDNN 8.0.5和Conda环境组成。

采用两种指标评估模型性能。第一种是分类模型通常使用的准确率(Accuracy,P_{ACC}),指标越高,模型的效果越好。为了更全面地评估模型,除了P_{ACC},还引入了受试者工作特征曲线下面积(Area Under the Receiver Operating Characteristic Curve,P_{AUC}),指标可以很好地描述模型的泛化性能,P_{AUC}值越大,则可说明模型的性能越好。

3.3 实验结果

3.3.1 数据集内实验

为了验证模型对于不同质量人脸图像的有效性,选择在FF++(C23)、FF++(C40)各个子数据集和完整的C23和出C40上进行实验,计算测试时所获得的P_{ACC}来评估模型,并与目前先进的算法进行比较,实验结果如表1所示。

表1 不同方法在不同质量FF++数据集上的性能

Table 1 Performance of different methods on different quality FF++ datasets

方法	P_{ACC}[FF++(C23)]					P_{ACC}[FF++(C40)]					平均
	DF	F2F	FS	NT	C23	DF	F2F	FS	NT	C40	P_{ACC}
MesoNet[14]	95.26	95.84	93.43	85.96	91.74	89.52	84.44	83.56	75.74	74.31	86.98
Xception[12]	98.85	98.36	98.23	94.50	95.28	94.28	91.56	93.70	82.11	88.86	93.57
Gram-Net[15]	97.63	96.31	97.96	92.07	94.16	92.39	90.67	91.99	84.69	86.43	92.41
F3-Net[8]	98.43	98.51	98.34	93.22	96.35	96.01	93.62	94.33	86.37	92.18	94.73
DMGTNet[16]	98.31	98.46	97.67	92.46	96.46	95.67	92.74	93.27	85.05	90.47	94.02
Ours	98.52	98.35	98.40	94.26	97.29	96.48	92.46	93.61	86.23	93.05	94.86

从表1结果发现,提出的模型性能在不同质量的不同子数据的准确率均达到了检测的较高水准,显著超过了早期的MesoNet[14]检测模型。对比Liu等[15]、Qian等[8]和Liang等[16]提出的Gram-Net、F3-Net和DMGTNet,方法在C23的DF、FS、NT子数据集上的性能均超过了这些模型,获得了与先进的Xception[14]可比的检测结果。此外,在F2F的准确率上也达到了与先进的方法可比的98.35%。而Xception[14]由于引入了深度可分离卷积,降低了过拟合的风险,在C23的各个子数据集上都表现优异,这也侧面反映了单一伪造技术生成的伪造人脸的检测相对简单。

在对低质量的C40子数据集进行分析时发现,相比在C23上的检测结果,所有检测方法的性能明显下降,这是由于C40数据集的图像经过强压缩处理,导致

大量原始篡改痕迹的丢失,进而使伪造图像的检测变得更具有挑战性。然而,所提出的方法在各个子数据集上均表现出与先进检测方法相媲美的性能,尤其在 DF 子数据集上,其检测准确率高达 96.48%,显著超过了现有方法的检测表现。这一优越性可以归因于 DF 子数据集的特性,它包含了由深度学习生成的人脸交换图像。方法巧妙地应用了中心差分卷积,更好地挖掘了这些伪造图像的内在伪造模式,从而提高了检测性能。

完整的 C23 和 C40 是涉及多种篡改技术的复杂数据集,其中伪造图像的人物身份、年龄、性别等存在更大的差异和多样性。为了评估模型检测复杂伪造模式人脸图像的能力,分别在完整的 C23 和 C40 数据集上进行了实验。表 1 中的结果表明,方法在 C23 上获得了 97.29%的检测准确率,显著超越了现有的检测方法。而在强压缩的 C40 数据集上,方法也取得了 93.05%的准确率,比先进的 F3-Net[6] 高 0.87 个百分点,表明方法在面对伪造模式复杂的数据时也表现出显著的判别能力。

通过在数据集内进行实验,方法展现了在不同条件下(无论伪造模式是简单还是复杂,图像质量高还是低),对伪造人脸数据的有效识别。证明了人脸的局部纹理差异特征在伪造图像中的广泛存在。通过充分挖掘真伪人脸图像的纹理差异,方法成功地降低了预测误差,提高了检测性能。

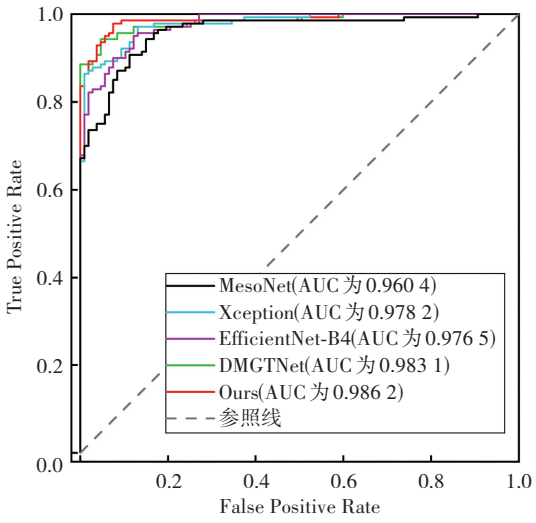
3.3.2　跨数据集实验

不同的伪造技术导致生成的伪造数据集之间存在差异,目前先进的检测技术训练和测试在同一数据集上表现出色,但当应用到不同数据集上进行跨数据集测试时,检测性能会急剧下降。为了验证提出模型的泛化能力,在 FF++(C23)和 Celeb-DF 数据集上进行了跨数据集性能测试,同时与先进的检测模型进行对比,结果如表 2 所示。从表 2 发现,当训练和测试在相同数据集上时,各种方法通常都能获得出色的性能,均表现出超过 0.96 的 P_{AUC} 值,其中方法在 C23 上达到了当前最优性能检测性能,显示了强大的检测能力。在图 4 展示的 ROC 曲线和 P_{AUC} 值中,方法的 ROC 曲线几乎覆盖了整个区域,进一步说明了方法的有效性。

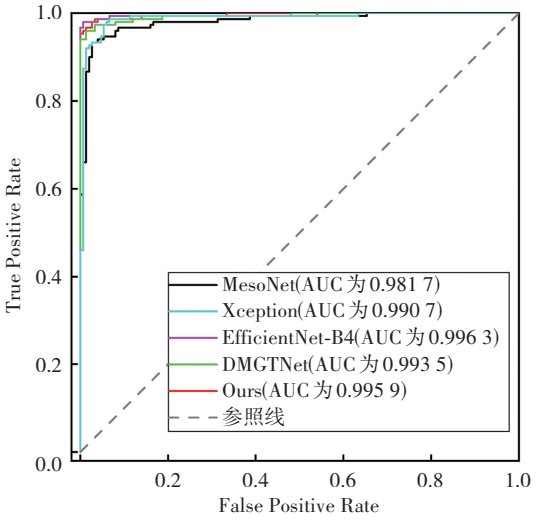
表 2　不同方法的泛化性能 P_{AUC} 值比较

Table 2　Comparison of P_{AUC} values for generalization performance of different methods

方　法	P_{AUC}[FF++(C23)]		P_{AUC}[Celeb DF]	
	FF++(C23)	Celeb DF	FF++(C23)	Celeb DF
MesoNet[14]	0.960 4	0.648 8	0.612 4	0.981 7
Xception[12]	0.978 2	0.671 2	0.582 4	0.990 7
EfficientNet-B4[17]	0.976 5	0.711 4	0.638 5	0.996 3
DMGTNet[16]	0.983 1	0.723 0	0.634 6	0.993 5
Ours	0.986 2	0.743 2	0.641 7	0.995 9



(a) FF++



(b) Celeb-DF

图 4　不同方法在不同测试集上的 ROC 曲线

Fig. 4　ROC curves of different methods on different test sets

当模型在不同数据分布的数据集上进行跨数据集

测试时,P_{AUC} 值急剧下降,揭示了伪造检测模型泛化难的问题。其中,提出的方法在跨数据集检测方面表现出色,相较于先进的 EfficientNet-B4[17] 和 DMGTNet[16],其性能高出 1~3 个百分点,表现出良好的泛化性。进一步观察发现,用 C23 训练的模型在其他数据集的测试集上表现比用 Celeb-DF 训练的模型更有效。这启示了在具有更多伪造类别分布的数据上进行模型训练能够提升模型的泛化性能。尽管如此,跨数据集的模型检测 P_{AUC} 值仍然只有 0.692 5,表明存在大量的错检。这也侧面强调了不同数据分布的检测仍是 Deepfake 检测的难点和重点。

总结而言,不同伪造技术导致不同的数据集存在显著差异,对检测模型的泛化性能构成了挑战,而提出的方法在跨数据集测试中表现出良好的泛化性,超越了其他先进模型。

3.3.3 消融实验

为了验证提出方法中 CDC 模块和 DA 模块对模型性能的重要性,对两个模块设置了消融实验,实验以 ResNet-50 为骨干网络。通过添加 CDC 模块、DA 模块和两个模块的组合使用分别验证各个模块的有效性。实验在 FF++(C23)数据集和 Celeb-DF 数据集上进行,结果如表 3 所示。由表 3 可知,添加 CDC 模块和 DA 模块后,在 FF++数据集上的实验性能分别提升了 1.95%和 1.73%,而在 Celeb-DF 数据集上性能也分别提升了 2.73%和 1.44%,当这两个模块共同作用时,P_{AUC} 值提升了 0.035 2 和 0.035 8。

表 3 不同模块消融实验的 P_{AUC} 值

Table 3 P_{AUC} values for different module ablation experiments

模 型	P_{AUC}[FF++(C23)]		P_{AUC}[Celeb-DF]	
	FF++(C23)	Celeb-DF	FF++(C23)	Celeb-DF
ResNet-50	0.951 0	0.643 6	0.604 5	0.960 1
ResNet-50+CDC	0.970 5	0.713 2	0.634 1	0.987 4
ResNet-50+DA	0.968 3	0.697 6	0.624 9	0.974 5
ResNet-50+CDC+DA	0.986 2	0.743 2	0.641 7	0.995 9

针对跨数据集的实验,可以发现添加 CDC 模块后,性能平均提高了近 5 个百分点。说明 CDC 模块使网络能够学习到 Deepfake 图像中的共有不一致特征,显著改善了检测性能。此外,添加 DA 模块后,性能同样提高了近 4 个百分点,表明 DA 模块能够有效地捕捉位置敏感信息,用于定位异常特征。值得一提的是,实验结果还显示,当这两个模块结合使用时,性能比单独使用任何一个模块都要好。说明位置敏感信息对于捕获纹理差异特征具有积极作用,模块组合的使用能够更有效地提升网络的泛化性。

此外,为了评估模型对计算资源的需求,还分析了模型的计算复杂度。模型的计算复杂度主要包括浮点运算数(FLOPs)和参数数量,其中参数数量以百万(1 M)为单位。原始的 ResNet-50 网络有 36 亿 FLOPs 和 25 M 参数,而提出的模型在采用了 CDC 模块和 DA 模块后,计算复杂度分别提升到 49 亿 FLOPs 和 29 M 参数。可以发现:所提出的模型在引入 CDC 和 DA 模块后,显著提高了泛化性能,但同时也要求更多的计算资源。因此,平衡模型的性能和资源需求也将成为进一步研究的重点。

综上所述,消融实验表明了 CDC 模块和 DA 模块在提出方法中的重要性。CDC 模块有助于学习到 Deepfake 图像中的共有特征,而 DA 模块则能够捕捉位置敏感信息,有助于定位异常特征。此外,两个模块的组合使用效果最佳,能够显著提高网络的泛化性能。

4 结 论

提出了一种基于局部纹理差异特征增强的 Deepfake 检测方法用以提高检测的准确性和泛化性。该方法明确了局部纹理差异是伪造人脸图像广泛存在的不一致特征。使用中心差分卷积模块,加强模型对伪造信息的捕捉,学习伪造人脸过程中更具泛化性的局部纹理差异信息,提高对局部细微差异信息的检测。使用双层注意力模块精炼特征,空间注意力定位局部纹理差异特征的位置信息,帮助模型集中关注重要人脸区域,通道注意力保留与任务相关的差异特征,同时抑制差异特征中的冗余信息,增强模型的特征表示,提高判别能力。

通过上述设计,方法在保持良好数据集内检测性能的同时,有效提升了模型的泛化性能。与其他同类方法相比,在应对低质量图像和跨数据集检测等复杂任务方面表现出色。方法的有效性在 FF++和 Celeb-DF 数据集上得到了验证。然而,尽管已取得显著进

展,目前仍需要为不同数据分布的人脸训练不同的模型,这是一个具有挑战性的问题。未来的研究将致力于寻找通用方法,以促进人脸伪造检测的广泛应用。这一方向的研究有望进一步提升该领域的检测性能和适用范围。

参考文献(References):

[1] 李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述[J]. 软件学报, 2021, 32(2): 496-518.
LI Xu-rong, JI Shou-ling, WU Chun-ming, et al. An overview of deep forgery and detection techniques[J]. Journal of Software, 2021, 32(2): 496-518.

[2] 李泽宇, 张旭鸿, 蒲誉文, 等. 多模态深度伪造及检测技术综述[J]. 计算机研究与发展, 2023, 60(6): 1396-1416.
LI Ze-yu, ZHANG Xu-hong, PU Yu-wen, et al. A survey on multimodal Deepfake and detection techniques[J]. Jisuanji Yanjiu yu Fazhan/Computer Research and Development, 2023, 60(6): 1396-1416.

[3] DE C T J, RIESS C, ANGELOPOULO E, et al. Exposing digital image forgeries by illumination color classification[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(7): 1182-1194.

[4] FERRARA P, BIANCHI T, DE R A, et al. Image forgery localization via fine-grained analysis of CFA artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566-1577.

[5] BAHRAMI K, KOT A C, LI L, et al. Blurred image splicing localization by exposing blur type inconsistency[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(5): 999-1009.

[6] LI H D, LI B, TAN S Q, et al. Identification of deep network generated images using disparities in color components[J]. Signal Processing, 2020, 174(9): 1-12.

[7] 刘贤刚, 范博, 郝春亮. 一种基于特征点对齐的假脸检测框架[J]. 通信技术, 2020, 53(5): 1133-1137.
LIU Xian-gang, FAN Bo, HAO Chun-liang. A fake face detection framework based on feature point alignment[J]. Communications Technology, 2020, 53(5): 1133-1137.

[8] QIAN Y Y, YIN G J, LU S, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]// European conference on computer vision. Berlin: Springer International Publishing, 2020: 86-103.

[9] LUO Y C, ZHANG Y, YAN J C, et al. Generalizing face forgery detection with high-frequency features[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York: IEEE, 2021: 16317-16326.

[10] ZHAO H Q, ZHOU W B, CHEN D D, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York: IEEE, 2021: 2185-2194.

[11] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). Berlin: Springer Science, 2018: 3-19.

[12] 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. New York: IEEE, 2019: 1-11.

[13] LI Y Z, 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. New York: IEEE, 2020: 3207-3216.

[14] 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). New York: IEEE, 2018: 1-7.

[15] LIU Z Z, QI X J, 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. New York: IEEE, 2020: 8060-8069.

[16] LIANG B Y, WANG Z Y, HUANG B J, et al. Depth map guided triplet network for deepfake face detection[J]. Neural Networks, 2023, 159(3): 34-42.

[17] 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). New York: IEEE, 2021: 5012-5019.

责任编辑:代小红