基于操纵痕迹融合的人脸伪造检测方法

黄继胜,杨高明
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 针对目前缺乏一种能够在复杂场景中暴露假脸图像的强大的假人脸检测模型,提出了一种新的网络,称为双流操纵痕迹网络(Two-stream Manipulation Trace Network, TSMTN),用于学习假图像面部区域上的细微操纵痕迹。**方法** 该方法不同以往直接从图像中学习特征,而是先从图像中提取操纵痕迹,之后通过操纵痕迹检测人脸是否被操纵。该网络由于 3 个关键模块组成:空间域操纵痕迹提取(Spatial Domain Manipulation Trace Extraction, SDMTE)、频域操纵痕迹提取(Frequency Domain Manipulation Trace Extraction, FDMTE)以及基于自注意力机制的特征融合模块(Feature Fusion Module, FFM)。SDMTE 使用卷积神经网络(CNNs)来学习图像空间域中的细微操纵痕迹。FDMTE 学习图像频域中高频信息的操纵痕迹。FFM 融合空间域和频域中的操纵痕迹,以生成用于分类的最终特征。**结果** 实验结果表明:该模型具有良好的性能,在常用检测数据集上到达了先进的水平。**结论** 该方法表现出较好的鲁棒性和泛化能力,取得了一些进步,具有重要意义。

关键词:人脸检测;操纵痕迹;空间域;频域;自注意力机制
中图分类号:TP183 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0004.010

Face Forgery Detection Method Based on Manipulation Trace Fusion

HUANG Jisheng, YANG Gaoming
School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective In view of the current lack of a powerful fake face detection model that can expose fake face images in complex scenes, a new network called two-stream manipulation trace network (TSMTN) is proposed for learning subtle manipulation traces on facial regions in fake images. **Methods** This method is different from the previous methods of directly learning features from images. Instead, it first extracts manipulation traces from the image and then uses the manipulation traces to detect whether the face has been manipulated. The network consists of three key modules: spatial domain manipulation trace extraction (SDMTE), frequency domain manipulation trace extraction (FDMTE) and feature fusion module (FFM) based on self-attention mechanism. SDMTE uses convolutional neural networks (CNNs) to learn subtle manipulation traces in the image spatial domain. FDMTE learns the manipulation traces of high-frequency information in the frequency domain of images. FFM fuses manipulation traces in the spatial and frequency domains to generate final features for classification. **Results** The experimental results show that the model has good performance and has reached an advanced level on commonly used detection datasets. **Conclusion** This method shows good robustness and generalization ability and has made some progress, which is of great significance.

Keywords: face detection; manipulation traces; spatial domain; frequency domain; self-attention mechanism

收稿日期:2024-01-03 **修回日期:**2024-02-27 **文章编号:**1672-058X(2025)04-0080-08
基金项目:安徽省自然科学基金(2008085MF220).
作者简介:黄继胜(1998—),男,硕士研究生,从事计算机视觉研究.
通信作者:杨高明(1974—),男,博士,教授,从事信息安全研究. Email:gmyang@aust.edu.cn.
引用格式:黄继胜,杨高明. 基于操纵痕迹融合的人脸伪造检测方法[J]. 重庆工商大学学报(自然科学版),2025,42(4):80-87.
HUANG Jisheng, YANG Gaoming. Face forgery detection method based on manipulation trace fusion[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(4): 80-87.

1 引　言

Deepfake 技术的快速发展和广泛应用引发了人们对全球数字内容可信度和安全性的严重担忧。Deepfake 是指使用深度学习和人工智能技术,通过合成、修改或替换真实数据来创建虚假图像和视频内容。随着卷积神经网络(Convolutional Neural Networks, CNN)[1]和生成对抗性网络(Generative Adversarial Networks, GANs)[2]的发展,面部操纵的实现变得更加容易。不幸的是,这项技术对社会各个领域构成了威胁,包括新闻业、社交媒体传播、法律诉讼等。因此,必须开发一种能够有效区分真实人脸和伪造人脸的方法。

早期的人脸伪造检测算法主要基于生理信号,随着伪造技术的进步,这种简单的算法已不再适用。随着深度学习的发展,检测方法开始训练深度卷积神经网络来学习人脸伪造数据集中的伪造特征。这些方法在很大程度上依赖于特定的训练数据集,并根据其特性进行了优化,在该数据集中产生了良好的性能。然而,当应用于看不见的数据时,它们的有效性会显著下降,这表明它们的鲁棒性和泛化能力较差。这就要求找到人脸伪造最具代表性的特征。

Li 等[3]提出了一种称为面部 X 射线的图像表示,该图像表示可以揭示输入图像是否是两个不同源图像的混合,从而识别假图像。Liu[4]观察到,上采样是大多数人脸伪造技术的必要步骤,累积上采样会导致频域发生变化,特别是相位谱的明显变化。自然图像的相位谱保留了大量的频率特征,这些特征提供了额外的信息,减少了幅度谱的损失。为此,他们提出一种结合空间图像和相位谱来捕捉人脸伪造的上采样伪像的浅学习方法,提高了人脸伪造检测的可转移性。瞿远近等[5]改进了高斯滤波对地标点进行去噪和深度网络模型,将高斯滤波应用到空域和值域上,尽可能保留高频噪声,提高了标记点的精度,增加了鉴别虚假视频的效率,针对不同的人脸特征采用不同深度的神经网络,使得更高效地判别真假人脸。Bappy 等[6]开发一种高置信度检测框架,该框架可以定位图像中的操纵区域。图像处理的定位关注可能被篡改的区域,使用长短期记忆网络(Long Short-Term Memory, LSTM)和 CNN 结合的模型来捕捉操纵区域和非操纵区域之间的区别特征。操纵区域的一个关键特性是,它们在与相邻的未操纵像素共享的边界中表现出判别特征。通过 LSTM 和卷积层的组合来学习操纵区域和非操纵区域之间的边界差异。随着注意力机制的发展,Zhao 等[7]将深度伪造检测定义为一个细粒度的分类问题,并提出了一种新的多注意力深度伪造检测网络,该方法使用多个注意力头是网络关注不同的局部部分,利用纹理特征增强块放大细微伪影,聚合了低级文本特征和高级语义特征,实现了先进的性能。Heo 等[8]提出了一种用于 Deepfake 检测的有效视觉 Transformer 模型,以提取局部和全局特征,将矢量连接的 CNN 特征和基于补丁的定位相结合,与所有位置进行交互,以指定伪影区域。Nguyen 等[9]使用胶囊网络来检测各种 Deepfake。实验证明该方法对各种伪造的图像和视频攻击是有效的。

尽管目前的方法在泛化能力方面取得了一定进展,但普遍存在泛化能力差的问题,仍值得研究。原因在于上述方法把输入图像作为学习对象,直接从图像中学习伪造信息,不免会导致学习到图像中与图像真假无关的冗余信息,造成过拟合,进而导致泛化能力下降。为了解决这一问题,一些文献[10-11]采用了操纵痕迹提取网络来抑制图像内容并强调操纵痕迹。这些方法在一定程度上避免了图像冗余信息的过拟合。尽管如此,它们只突出了伪造图像的单一痕迹,并且可能无法有效地提取伪造图像的一般特征。在仔细研究了现有检测方法的泛化能力问题和局限性后,本文提出了一种称为双流操纵痕迹人脸伪造检测方法。该方法在两个分支上提取两种不同的操纵痕迹特征:使用操纵痕迹提取网络来抑制图像内容并突出操纵痕迹;探索图像频域中的操纵痕迹信息。还设计了一个基于注意力机制的特征融合模块,以提取两个操纵痕迹的共同特征,从而更好地表示伪造图像的一般属性。使用两种操纵痕迹信息,并精心设计了一个特征融合模块。

2 模型设计

TSMTN 包括 3 个主要组件:SDMTE 模块、FDMTE 模块以及基于注意力的特征融合模块 FFM。TSMTN 的总体模型框架如图 1 所示。首先,利用先进的人脸检测方法 MTCNN[12]来提取人脸地标。计算的面部标志用于裁剪、调整大小和对齐每帧中的面部区域。然后,分别使用 SDMTE 和 FDMTE 模块提取图像的操纵痕迹,并采用注意力机制融合这两种类型的操纵痕迹。最后,对融合后的特征进行分类。人脸伪造检测过程如算法 1 所示。

算法 1:人脸伪造检测过程

输入:人脸视频
输出:视频中人脸的预测结果(真或假)
Step1:从人脸视频中均匀提取 N 个视频帧
Step2:使用 MTCNN 模型检测每一个视频帧的人脸区域
Step3:将人脸区域放大 10%
Step4:从每个视频帧裁剪出大小为 299×299 的人脸图像
Step5:将人脸送入到 SDMTE 和 FDMTE 中分别提取空间域操纵痕迹和频域操纵痕迹
Step6:将提取到的两种操纵痕迹分别送入各自的骨干网络 CNN 中以学习操纵痕迹特征
Step7:将两种学习到的特征送入 FFM 融合模块中融合出更具代表性的痕迹特征
Step8:使用全连接层分类器对融合后的代表性痕迹特征进行检测分类

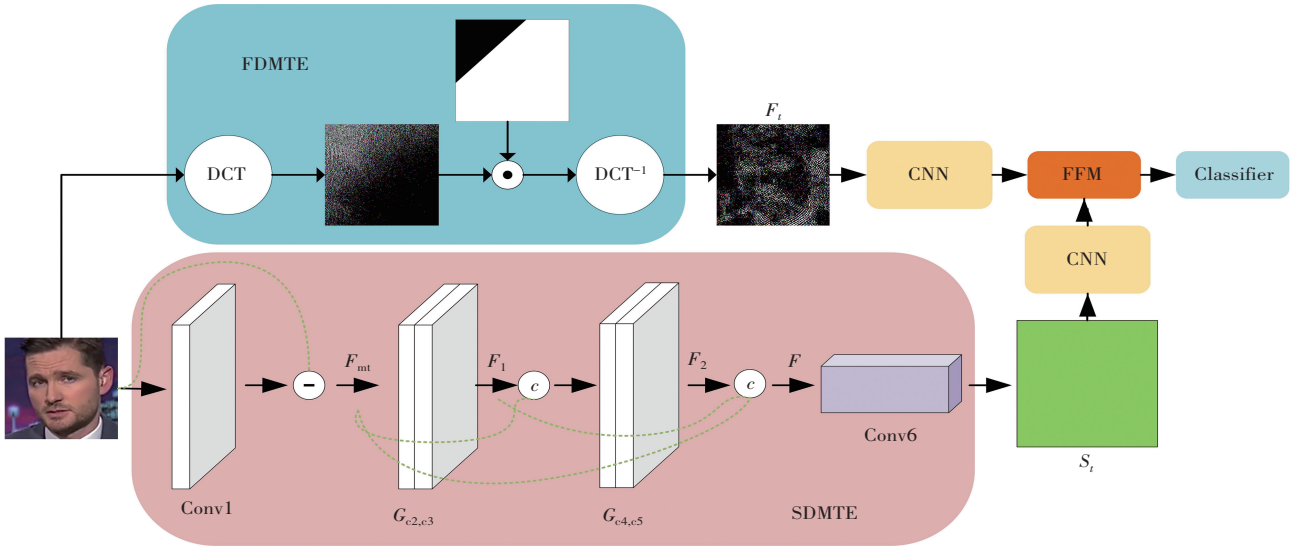


图 1 TSMTN 的总体模型框架

Fig. 1 The overall model framework of TSMTN

2.1 空间域操纵痕迹的提取

一些现有的工作,如 SRM[13] 和 SPAM[14],从预测的操纵痕迹中学习特征。它们首先通过固定预测器 $\boldsymbol{f}$ 来生成一组预测像素值。然后,通过从预测的像素值中减去原始像素值来获得操纵痕迹 $\boldsymbol{P}$。可以表示为

$$\boldsymbol{P}=\boldsymbol{f}(\boldsymbol{I})-\boldsymbol{I} \tag{1}$$

其中,$\boldsymbol{I}$ 是输入图像。通过预测器生成一组预测像素值。然后,通过从预测像素值中减去原始像素值来获得操纵痕迹。通过简单地使用减法运算获得的操纵痕迹是不稳定的。本文的模型不是这样简单地得到操纵痕迹,该模型借鉴了特征复用的思想,使用多个卷积神经网络层并重复使用每层获得的操纵痕迹来合成最后更稳定的操纵痕迹。这些卷积神经网络层的结构如表 1 所示。一次 CNN 计算可以定义为

$$\boldsymbol{F}_i = \sum_{i=1}^{m} \boldsymbol{I}_i * \boldsymbol{w}_{ij} + \boldsymbol{b}_j \tag{2}$$

首先,Conv1 用于提取初始操纵痕迹 $\boldsymbol{F}_{\mathrm{mt}}$,并且所获得的痕迹 $\boldsymbol{F}_{\mathrm{mt}}$ 依次通过 Conv2 和 Conv3 以获得中间特征图 $\boldsymbol{F}_1$。这个过程可以描述为

$$\boldsymbol{F}_1=\boldsymbol{G}_{c2,c3}(\boldsymbol{F}_{\mathrm{mt}})=\boldsymbol{G}_{c2,c3}(\mathrm{Conv1}(\boldsymbol{I})-\boldsymbol{I}) \tag{3}$$

然后,将通过连接 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_{\mathrm{mt}}$ 而获得的特征图发送到 Conv4 和 Conv5 以获得中间特征图 $\boldsymbol{F}_2$。这个过程可以描述为

$$\boldsymbol{F}_2=\boldsymbol{G}_{c4,c5}([\boldsymbol{F}_1,\boldsymbol{F}_{\mathrm{mt}}]) \tag{4}$$

最后,$\boldsymbol{F}_{\mathrm{mt}}$、$\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 被连接起来作为最终的操纵痕迹。这个过程可以描述为:

$$\boldsymbol{F}=[\boldsymbol{F}_{\mathrm{mt}},\boldsymbol{F}_1,\boldsymbol{F}_2] \tag{5}$$

将稳定的操纵痕迹 $\boldsymbol{F}$ 传递到 Conv6 层,得到具有三个通道的特征图,即空间域中的操纵痕迹 $\boldsymbol{S}_t$。

表 1 提取操纵痕迹的卷积模块

Table 1 Convolution module for extracting manipulation traces

Layers	Kernel sizes	Kernel quantities	Strides	Paddings	Output sizes
Conv1	3×3	3	1	1	299×299
Conv2	3×3	3	1	1	299×299
Conv3	3×3	3	1	1	299×299
Conv4	3×3	6	1	1	299×299
Conv5	3×3	6	1	1	299×299
Conv6	1×1	3	1	0	299×299

2.2 频域操纵痕迹的提取

在频域中已经有一些关于深度伪造检测的工作。Frank 等[15] 首先分析了频域中的深度伪造图像。由于频域中的操纵痕迹可以提供空间域中不存在的信息,因此可以提取频域中的操纵痕迹。观察发现,人脸图像中的操纵痕迹经常出现在图像的细节和边缘,这些细节和边缘对应于图像频域中的高频信息。因此,设计了这个模块。给定输入图像,$\boldsymbol{I}\in\mathbf{R}^{C\times H\times W}$,其中 $\boldsymbol{C}$、$\boldsymbol{H}$ 和 $\boldsymbol{W}$ 分别表示通道的数量、高度和宽度。首先使用离散余弦变换(Discrete Cosine Transform, DCT)将图像转换为频域表示,表示为 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$。使用二维 DCT 对图像进变换,二维 DCT 计算如式(6)所示:

$$\boldsymbol{F}(\boldsymbol{u},\boldsymbol{v})= \boldsymbol{c}(\boldsymbol{u})\boldsymbol{c}(\boldsymbol{v})\sum_{i}^{N-1}\sum_{j}^{N-1}\boldsymbol{f}(\boldsymbol{i},\boldsymbol{j})\cos\left(\frac{(\boldsymbol{j}+0.5)\pi}{\boldsymbol{N}}\boldsymbol{u}\right)\cos\left(\frac{(\boldsymbol{i}+0.5)\pi}{\boldsymbol{N}}\boldsymbol{v}\right) \tag{6}$$

其中,$\boldsymbol{x}=0$ 时,$\boldsymbol{c}(\boldsymbol{x})=\sqrt{\frac{1}{N}}$,$\boldsymbol{x}\neq 0$ 时,$\boldsymbol{c}(\boldsymbol{x})=\sqrt{\frac{2}{N}}$。下文式(7)中的 $\boldsymbol{c}$ 和此处一样。由于 DCT 的高度对称性,在进行相关运算时可以使用更直接的矩阵处理方法,计算矩阵表示为 $\boldsymbol{A}\in\mathbf{R}^{H\times W}$。矩阵 $\boldsymbol{A}$ 计算方式如下:

$$\boldsymbol{A}(i,j)=\boldsymbol{c}(i)\cos\left(\frac{(j+0.5)\pi}{\boldsymbol{N}}i\right) \tag{7}$$

其中,i,j 是图像像素的坐标。将输入图像 $\boldsymbol{I}$ 的两侧分别乘以矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{A}^{\mathrm{T}}$,以获得图像的频域表示。构造了一个与 $\boldsymbol{F}$ 大小相同的滤波器 $\boldsymbol{M}$。滤波器 $\boldsymbol{M}$ 像素值分布如图 2 所示。通过计算图像频域表示和 $\boldsymbol{M}$ 的逐元素乘积来提取图像的高频信息 $\hat{F}$。这个该过程可以表示为

$$\hat{\boldsymbol{F}}=\boldsymbol{M}\cdot(\boldsymbol{A}\times\boldsymbol{I}\times\boldsymbol{A}^{\mathrm{T}}) \tag{8}$$

然后,将高频信息 $\hat{F}$ 逆变换到空间域表示中,以获得频域中的最终操纵痕迹 F_t。这个该过程可以表示为

$$\boldsymbol{F}_t=\boldsymbol{A}^{\mathrm{T}}\times\hat{\boldsymbol{F}}\times\boldsymbol{A} \tag{9}$$

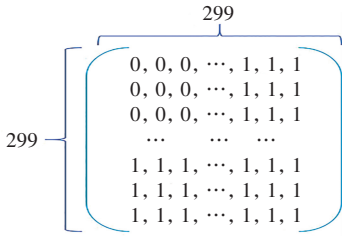


图 2　滤波器 $\boldsymbol{M}$ 像素值分布

Fig. 2　Distribution of pixel values in filter $\boldsymbol{M}$

2.3　基于注意力机制的特征融合模块

Vaswani 等[16] 首先提出了一种基于注意力的 Transformer 模型,并展示了捕捉上下文信息的能力。在这种方法的基础上,开发了一个特征融合模块,该模块利用注意力机制来关注两种痕迹之间的关系。这种方法有效地融合了来自空间域和频域的操纵痕迹。将获得的空间域操纵痕迹信息和频域操纵痕迹信息发送到它们各自的骨干网络,以生成它们的操纵痕迹特征图 $\boldsymbol{S}_m\in\mathbf{R}^{C\times H\times W}$ 和 $\boldsymbol{F}_m\in\mathbf{R}^{C\times H\times W}$。如图 3 所示,卷积 $\boldsymbol{W}_Q$ 和 $\boldsymbol{W}_V$ 分别将空间域中的操纵痕迹特征图编码为查询矩阵 $\boldsymbol{Q}\in\mathbf{R}^{N\times C}$ 和值矩阵 $\boldsymbol{V}\in\mathbf{R}^{N\times C}$,卷积 $\boldsymbol{W}_K$ 用于将频域中的操纵痕迹特征图编码为关键矩阵 $\boldsymbol{K}\in\mathbf{R}^{N\times C}$,$\boldsymbol{N}=H\times W$ 是特征图中的像素数。这个过程可以表示为

$$\boldsymbol{Q}=\boldsymbol{S}_m\times\boldsymbol{W}_Q\quad\boldsymbol{K}=\boldsymbol{F}_m\times\boldsymbol{W}_K\quad\boldsymbol{V}=\boldsymbol{S}_m\times\boldsymbol{W}_V \tag{10}$$

使用自注意机制将这两个特征融合,以获得用于分类的最终特征 $\boldsymbol{G}$。这个过程涉及矩阵乘法、特征缩放以及 softmax 回归,可以表示为

$$\boldsymbol{G}=\boldsymbol{S}_m+\mathrm{Softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{K})^{\mathrm{T}}}{\sqrt{C}}\right)\boldsymbol{V} \tag{11}$$

其中,C 是缩放量。

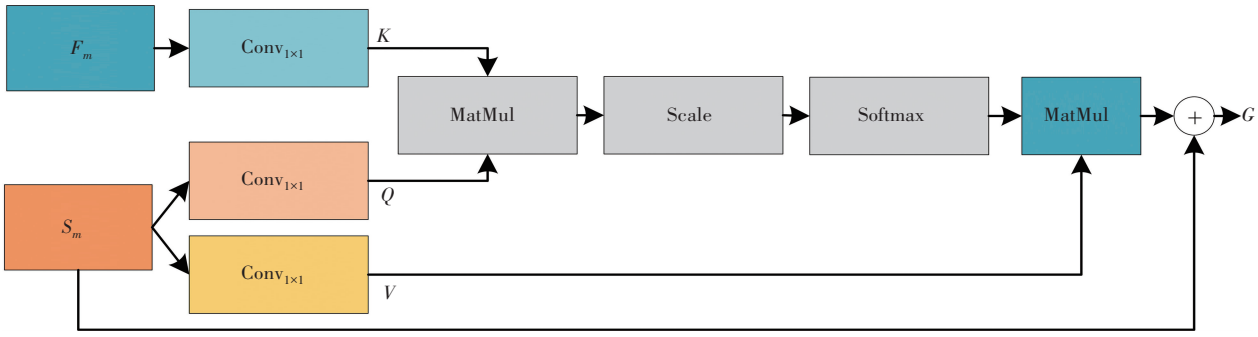


图 3　FFM 结构组成

Fig. 3　Composition of FFM structure

3　仿真实验与结果分析

3.1　实验环境

(1) 数据集。为了评估方法的有效性,在以下数据集上进行了实验:FF++、Celeb-DF 和 DFDC。

FF++包含 1 000 个原始视频,其中 720 个视频用于训练,140 个视频用于验证,140 个用于测试。每个视频都使用 4 种深度伪造方法进行处理,即 DeepFakes、FaceSwap、Face2Face 和 NeuralTextures。因此,该数据集共有 5 000 个视频(1 000 个真实视频和 4 000 个虚假视频)。真实图像和伪造图像在不同级别被压缩,以生成高质量(High Quality, HQ)和低质量(Low Quality, LQ)版本的 FF++。

Celeb-DF 由 890 个真实视频和 5 639 个操纵视频组成,总计超过 230 万帧。真实视频是从公共 YouTube 视频中收集的,590 个真实视频对应 59 个不同领域的名人,在 V1 版本中增加了 300 个真实视频。6 011 个视频用于训练,518 个视频用于测试。

DFDC 的容量高达 472 GB,包括 119 197 个视频,每个视频 10 个秒长,但帧速率在 15~30 fps 之间,分辨率也在 320~240 到 3 840~2 160 之间。在训练视频中,19 197 个视频是实际拍摄的约 430 名演员的片段,其余 10 万个视频是由真实视频生成的假脸视频。假脸生成使用各种主流的假脸生成算法,如 DeepFakes、Face2Face 等,因此该数据集包含大量的假脸视频。只选择部分数据作为测试集。

(2) 交叉熵损失。为了对图像进行分类,将特征 $\boldsymbol{G}$

输入到一系列神经网络层中,包括卷积层、池化层、线性层和 Softmax 层。最后,使用交叉熵损失函数 $\boldsymbol{L}$ 对模型进行优化。损失函数表示如下:

$$\boldsymbol{L}=\boldsymbol{y}\log\hat{\boldsymbol{y}}+(1-\boldsymbol{y})\log(1-\hat{\boldsymbol{y}}) \tag{12}$$

(3) 评估指标。应用分类正确率 $\boldsymbol{P}_{\mathrm{ACC}}$(Accuracy, ACC)和受试者工作特征曲线下面积 $\boldsymbol{P}_{\mathrm{AUC}}$(Area Under Curve, AUC)作为评估指标,这些指标通常用于 Deepfake 检测任务。

(4) 实施细节。在本文的方法中,使用 MTCNN 来裁剪大小为 299×299 的人脸区域(检测框放大了 1.3 倍)作为输入。使用在 ImageNet 上预训练的 Xception 作为骨干网络。新引入的层和块被随机初始化。使用 Adam 优化模型,学习率为 0.000 1,每 50 个时期衰减 10 倍。动量设置为 0.9,批量大小设置为 16。整个网络被训练了 100 个时期。使用 Albumentations 库对输入图像进行数据增强,应用常见的变换,如引入模糊、高斯噪声、换位、旋转等。

3.2 在 FF++数据集上的实验结果

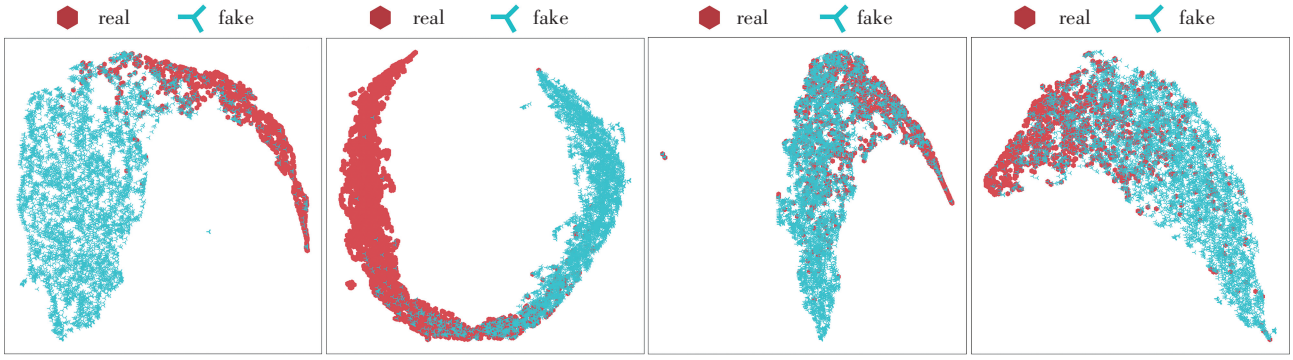
在 FF++数据集的 HQ 和 LQ 版本上进行了实验。在实验中使用 100 个时期来训练模型。结果如表 2 所示。所提出的 TSMTN 在 HQ 版本上优于所有参考方法。所提出的 TSMTN 实现了 95.49% 的准确率和 0.996 的 $\boldsymbol{P}_{\mathrm{AUC}}$,与最先进的方法相比,这是一个显著的改进,与性能最佳的参考方法相比,$\boldsymbol{P}_{\mathrm{AUC}}$ 得分增加了约 0.033(Xception 的 $\boldsymbol{P}_{\mathrm{AUC}}$ 为 0.963),与性能最佳参考方法相比精度性能仅降低了0.0024,但远高于其他参考方法的得分。在 LQ 版本上,观察到由于图像压缩,性能显著下降。然而,能够通过利用频率信息来解决这个问题。TSMTN 在 LQ 条件下的 $\boldsymbol{P}_{\mathrm{AUC}}$ 得分达到 0.959,比 Xception 方法高 6.6%,检测精度 $\boldsymbol{P}_{\mathrm{ACC}}$ 得分达到了 85.33%,比 Xception 方法只降低了 1.53%。降低了一点点精度,但取得了非常好的鲁棒性,是不错的进步。这得益于模型不是直接从图像中学习特征,而是从操纵痕迹中学习特征,避免学习低质量图像压缩而产生的冗余信息。

表 2 在 FF++数据集 HQ 和 LQ 版本上的帧级检测结果

Table 2 The frame-level detection results on the FF++dataset for both HQ and LQ versions /%

Method	HQ		LQ	
	$\boldsymbol{P}_{\mathrm{ACC}}$	$\boldsymbol{P}_{\mathrm{AUC}}$	$\boldsymbol{P}_{\mathrm{ACC}}$	$\boldsymbol{P}_{\mathrm{AUC}}$
Steg. Features[13]	70.97	—	55.98	—
LD-CNN[17]	78.45	—	58.69	—
MesoNet[18]	83.10	—	70.47	—
Face X-ray[3]	—	87.4	—	61.6
Xception[19]	**95.73**	96.3	**86.86**	89.3
Ours	95.49	**99.6**	85.33	**95.9**

性能的提高主要是由于从频域和空间域都提取了操纵痕迹信息,并使用注意力机制融合这两种类型的信息。这有助于所提出的 TSMTN 比常规的网络更好地检测细微的操纵伪影,并且对严重的压缩误差具有鲁棒性。为了更好地理解所提出方法的有效性,将 TSMTN 提取的特征图可视化,如图 4(a)所示。



(a) Visualization on FF++ (b) Visualization on Celeb-DF (c) Feature distribution on model 1 (d) Feature distribution on model 3

图 4 特征图可视化

Fig. 4 Visualization of feature maps

3.3 在 Celeb-DF 数据集上的实验结果

在本节中,在 Celeb-DF 数据集上评估了 TSMTN 的帧级检测精度,模型 TSMTN 在 Celeb-DF 上的 AUC 达到 0.989。具体结果如表 3 所示。本文的方法在性能方面显著优于参考方法,这要归功于模型准确捕捉操纵痕迹的能力,使本文的方法能够在使用其他类型的伪造图像

进行训练时保持其有效性。具体而言,与 Xception 相比,本文的方法将 P_{AUC} 值提高了 0.013,将 P_{ACC} 值提高了 1.15%。也优于相比较的所有方法。这表明该方法在不同的数据集上也优于其他深度伪造检测方法。

一般的检测方法通常在一种数据集上学习得很好,却不能够学习新的数据集,这是因为以图像作为学习对象的方法,会学习到冗余的信息,不同数据集冗余信息会有所不同。因此有些方法在不同的数据集上不能都学习得很好。但是本文方法是从根本上学习操纵痕迹,将空间域与频域操纵痕迹相结合学习。因此可以挖掘出不同类型数据集的代表性伪造特征。所以本文方法表现出强大的伪造特征捕获能力和鲁棒性。提供了数据可视化结果(图 4(b)),以说明该模型的有效性及其区分真实和虚假人脸的能力。

表 3　在 Celeb-DF 数据集上的帧级检测结果

Table 3　The frame-level detection results on the Celeb-DF dataset　/%

Method	P_{ACC}	P_{AUC}
Xception[19]	94.77	97.6
Multi-task[20]	—	90.5
Capsule[9]	—	93.2
DSW-FP[21]	—	94.8
Ours	**95.92**	**98.9**

3.4　泛化能力的评估

泛化能力是 deepfake 检测的核心。本文使用 FF++(HQ)、Celeb-DF 和 DFDC 来评估 TSMTN 的泛化能力,对每个测试集视频采样 30 帧,并计算帧级 P_{AUC} 分数。

对不可见操纵技术的泛化。为了评估本文方法的通用性,对看不见的操纵技术进行了实验。FF++数据集有 4 种不同的操纵方法:DeepFakes、FaceSwap、Face2Face 和 NeuralTextures。在 FF++数据集中执行交叉评估。从表 4 中可以看出,由于 Xception 过度依赖于纹理信息,其性能在不可见操纵技术中急剧下降。但是本文方法在数据集中表现良好,对看不见的操纵技术实现了较好的泛化。该方法对训练集中没有的 4 种操纵技术生成的图像取得了良好的检测结果,实现了显著的泛化性能。在 DeepFakes 和 FaceSwap 上训练的模型的平均泛化能力超过了最先进的方法。在 Face2Face 和 NeuralTextures 上训练的模型也非常接近最先进的方法。这得益于基于自注意力的特征融合机制,它提取了空间域操纵痕迹和频域操纵痕迹中更一般的伪造信息,从而对看不见的操纵技术也能捕捉到相关伪造信息。

对不可见数据集的泛化。为了进一步证明泛化能力,测试和训练分别在两个不同类型的数据集上进行,进行跨数据集评估。表 5 显示:与其他最先进的方法相比,本文方法在跨数据集场景中可以实现更好的泛化能力。还可以观察到,大多数基于图像的方法,如 Xception 和 F3-Net,在跨数据集场景中检测性能显著下降。这是因为特定操纵方法留下的静态纹理伪影相对均匀,因此基于这些伪影的和基于图像的方法容易受到过拟合的影响。FTCN 从空间域全局角度学习特征表示,因此对定位在关键面部区域(如口腔)的一些痕迹不太敏感。相比之下,本文模型旨在探索两个域的全局操纵痕迹,从而表现出更好的泛化能力。

表 4　对不可见操纵技术的模型泛化能力 P_{AUC} 值评估

Table 4　P_{AUC} value evaluation of model generalization ability for unseen manipulation techniques　/%

Training Set	Methods	DF	F2F	FS	NT	Avg
DF	Xception[19]	99.7	67.7	37.7	71.7	69.2
	Efficientnet-b4[22]	99.2	59.0	41.4	74.8	68.6
	Ours	99.8	70.4	38.3	71.8	70.1
F2F	Xception[19]	82.6	99.7	49.3	56.7	72.1
	Efficientnet-b4[22]	77.5	99.5	46.6	58.3	70.5
	Ours	80.6	99.8	56.1	60.1	74.2
FS	Xception[19]	60.1	61.2	99.7	46.6	66.9
	Efficientnet-b4[22]	53.1	59.6	99.6	43.3	63.9
	Ours	51.3	69.0	99.8	44.5	66.2
NT	Xception[19]	89.7	62.3	41.1	97.1	72.6
	Efficientnet-b4[22]	89.0	60.0	39.8	95.8	71.2
	Ours	83.5	61.6	42.1	97.1	71.1

表 5 对不可见数据集的模型泛化能力 P_{AUC} 值评估

Table 5 P_{AUC} value evaluation of model generalization ability for unseen datasets /%

Training Set	Methods	FF++	Cele-DF	DFDC	Avg
FF++	Xception[19]	96.3	48.2	69.0	71.2
	F3-Net[23]	98.1	65.2	70.1	77.8
	Face X-ray[3]	87.4	79.5	65.5	77.5
	FTCN[24]	97.9	86.9	74.0	86.3
	Ours	99.6	83.6	91.6	91.6
Cele-DF	Xception[19]	41.0	97.6	66.8	68.5
	F3-Net[23]	62.8	99.0	71.0	77.6
	Face X-ray[3]	51.3	99.1	69.1	73.2
	FTCN[24]	72.1	99.3	74.7	82.0
	Ours	88.5	99.3	89.9	92.6

3.5 异常区域可视化

为了直观地理解模型的决策,基于整个人脸的分类结果来可视化模型决策所依赖的空间异常区域。它使用流入最终层的权重参数来生成高亮显示决策区域的热图。可视化结果来自全局分支,在全局分支中,所设计的融合模块 FFM 提供了可视化的全局特征。不同数据集的可视化结果如图 5 所示。全局特征集中在不同的操纵区域,如眼睛、眉毛、前额、鼻子、混合边界等。这与本文的方法的目标一致,即全局分支保持感知全局操纵痕迹的能力。此外,当操纵痕迹在口腔区域不明显时,本文的方法仍然可以基于鼻子和眼睛区域中存在的其他操纵痕迹来检测 Deepfake,而不是强迫全局特征仅集中在嘴部区域。可视化结果进一步证明了该方法的有效性和泛化能力。

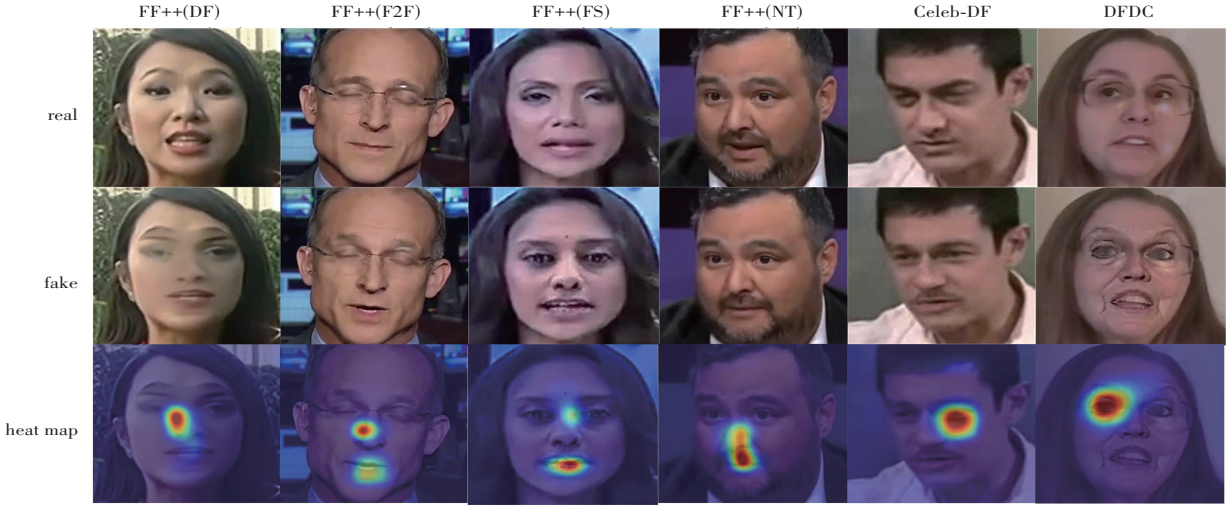


图 5 在不同数据集上的异常区域可视化

Fig. 5 Visualization of anomalous regions on different datasets

3.6 消融研究

FDMTE 和 FFM 的有效性。为评估所提出的 FDMTE 和 FFM 的有效性,定量评估了 TSMTN 及其变体:不含有 FDMTE 模块和 FFM 模块的 TSMTN(基线);不含有 FFM 的 TSMTN。

在 FF++(LQ)上进行了训练和验证。定量结果如表 6 所示,通过比较模型 1(基线)和模型 2,所提出的 FDMTE 稳步提高了 P_{ACC} 和 P_{AUC} 得分。在模型 2 的基础上加上 FFM(模型 3),P_{ACC} 和 P_{AUC} 得分更高。P_{ACC} 和 P_{AUC} 得分分别为 85.40%和 0.960 6,获得了最佳性能。

将模型 1 和模型 3 提取的特征可视化,如图 4(c)和 4(d)所示。模型 1(基线网络)不能很好地将真属性和假属性分离开来,而模型 3 可以充分地分离它们。这再次证明了添加的模块提供了更好的分类特征。这些逐渐改进的性能验证了所提出的 FDMTE 和 FFM 模块有助于 Deepfake 检测并相互补充。FFM 可以更有效地利用 FDMTE 学习到的特征来获得额外的收益。

表 6 模型消融研究帧级检测结果

Table 6 Frame-level detection results of model ablation studies /%

ID	SDMTE	FDMTE	FFM	P_{ACC}	P_{AUC}
1	✓	—	—	82.66	94.73
2	✓	✓	—	83.99	95.76
3	✓	✓	✓	85.40	96.06

4 结 论

在这项工作中,提出了一种基于深度学习的方法来检测使用 Deepfake 技术生成的伪造人脸。该方法使用了操纵痕迹提取和特征融合模块。操纵痕迹提取模块可以提取假脸中的细微操纵痕迹,如伪影、混合边缘和不合理的细节,分别在空间域和频域提取了两种操纵痕迹。特征融合模块借鉴注意力机制有选择地将两种操纵痕迹信息相融合,提取出更能代表人脸真假的一般性特征。在 3 个常用的公共伪造数据集的实验

上,本文方法的性能优于这里对比的先进方法。为了进一步验证该方法在复杂环境(如低光照和面部遮挡)以及更具挑战性的场景中的有效性,还需要额外的验证。探索 Xception 模型之外的其他卷积神经网络模型作为骨干网络值得尝试。在未来的研究中,将继续完善和增强该模型在各种现实世界环境中检测 Deepfake 的能力。

参考文献(References):

[1] GU J, WANG Z, KUEN J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[2] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[3] 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. Piscataway: IEEE Press, 2020: 5000-5009.

[4] 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. Piscataway: IEEE Press, 2021: 772-781.

[5] 瞿远近, 吴起. 基于改进高斯滤波网络的深度伪造检测方法[J]. 重庆工商大学学报(自然科学版), 2023, 40(4): 41-47.
QU Yuan-jin, WU Qi. Depth forgery detection method based on improved Gaussian filter network[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 41-47.

[6] BAPPY J H, ROY-CHOWDHURY A K, BUNK J, et al. Exploiting spatial structure for localizing manipulated image regions[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4980-4989.

[7] ZHAO H, WEI T, ZHOU W, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2185-2194.

[8] HEO Y J, YEO W H, KIM B G. DeepFake detection algorithm based on improved vision transformer[J]. Applied Intelligence, 2023, 53(7): 7512-7527.

[9] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: using capsule networks to detect forged images and videos[C]//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 2307-2311.

[10] BAYAR B, STAMM M C. Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2691-2706.

[11] GUO Z, YANG G, CHEN J, et al. Fake face detection *via* adaptive manipulation traces extraction network[J]. Computer Vision and Image Understanding, 2021, 204: 103170.

[12] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[13] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[14] PEVNY T, BAS P, FRIDRICH J. Steganalysis by subtractive pixel adjacency matrix[J]. IEEE Transactions on Information Forensics and Security, 2010, 5(2): 215-224.

[15] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//Proceedings of the Proceedings of the 37th International Conference on Machine Learning. 2020: 3247-3258.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. 2017: arXiv: 1706.03762. http://arxiv.org/abs/1706.03762.pdf.

[17] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[18] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network[C]//Proceedings of the IEEE International Workshop on Information Forensics and Security. Piscataway: IEEE Press, 2018: 1-7.

[19] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1800-1807.

[20] NGUYEN H H, FANG F, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[C]//Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems. Piscataway: IEEE Press, 2019: 1-8.

[21] LI Y, LYU S. Exposing deepfake videos by detecting face warping artifacts[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshop. Long Beach, CA, USA, 2019: 46-52.

[22] TAN M, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. 2019: 1905.11946. https://arxiv.org/abs/1905.11946v5.

[23] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 86-103.

[24] ZHENG Y, BAO J, CHEN D, et al. Exploring temporal coherence for more general video face forgery detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 15024-15034.

责任编辑:陈　芳