基于时空一致性的视频篡改检测方法

程 健[1],杨高明[2],杨新露[2]
1. 安徽理工大学 人工智能学院,安徽 淮南 232001
2. 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 随着视频篡改技术的快速发展,原始视频与篡改视频之间的差距愈发缩小,现有检测方法在处理篡改视频数据时,仍需提高检测准确度以及泛化性能,为此,提出一种基于时空一致性的视频篡改检测方法。**方法** 首先通过不同的视频采样步幅预处理视频数据,利用时序卷积核在高采样率视频帧数据中侧重提取帧间时序特征信息,空间卷积核在低采样率视频数据中侧重提取帧内空间特征信息,并在高采样视频帧数据与低采样视频帧数据间建立横向连接,从而获得更有效的视频时空特征;同时结合 Transformer 模型在时空特征序列中提取时空特征的不一致性,实现对篡改视频的判定。**结果** 改进的方法在高质量和低质量 FaceForensics++数据集上进行性能测试,AUC 数值分别达到 99.47%和 93.05%,此外在 FaceForensics++数据集上的域内跨伪造方式实验以及 Celeb-DF 数据集上的跨数据集实验中,测试结果相较于目前主流检测算法同样表现出竞争性,消融实验结果验证了方法中每个单一模块的有效性。**结论** 各项实验结果验证,所提方法在域内性能测试中有着优于现有算法的检测精度,并且在跨域性能测试中具有更好的泛化性能,即验证了联合时空卷积 Transformer 模型可以提高模型泛化性能。

关键词:视频篡改检测;时空一致性;时序卷积核;空间卷积核;Transformer 模型
中图分类号:TP391.41 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0004.014

Tampering Video Detection Methods Based on Spatiotemporal Consistency

CHENG Jian[1], YANG Gaoming[2], YANG Xinlu[2]
1. School of Artificial Intelligence, Anhui University of Science and Technology, Anhui Huainan 232001, China
2. School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective With the rapid development of video tampering technology, the gap between original videos and tampered videos is narrowing. Existing detection methods need to improve accuracy and generalization performance in detecting video tampering. Therefore, a video tampering detection method based on spatiotemporal consistency was proposed. **Methods** Firstly, the video was processed with different sampling strides. Temporal convolutional kernels were used to extract temporal features in high-sampling-rate data, while spatial convolutional kernels focused on extracting spatial features in low-sampling-rate data. A lateral connection was established between high-sampling-rate and low-sampling-rate video data to obtain a better representation of spatiotemporal features. Additionally, a Transformer model was used to extract inconsistencies in the spatiotemporal feature sequence to detect tampered videos. **Results** The improved method was tested on the high-quality and low-quality datasets of FaceForensics++, achieving AUC values of 99.47% and 93.05%, respectively. Furthermore, in-domain cross-forgery experiments on the FaceForensics++ dataset and cross-dataset experiments on the Celeb-DF dataset showed competitive performance compared with current mainstream

收稿日期:2023-12-18 **修回日期:**2024-01-30 **文章编号:**1672-058X(2025)04-0109-07
基金项目:科技部高端境外人才项目(G2021019006L);安徽省自然科学基金(2008085MF220).
作者简介:程健(1997—),男,安徽安庆人,硕士研究生,从事计算机视觉、隐私保护研究.
通信作者:杨高明(1974—),男,安徽阜阳人,教授,博士,从事隐私保护、机器学习研究. Email: gyang@aust.edu.cn.
引用格式:程健,杨高明,杨新露. 基于时空一致性的视频篡改检测方法[J]. 重庆工商大学学报(自然科学版),2025,42(4):109-115.
CHENG Jian, YANG Gaoming, YANG Xinlu. Tampering video detection methods based on spatiotemporal consistency[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(4): 109-115.

detection algorithms. Ablation experiments validated the effectiveness of each module. **Conclusion** Based on the experimental results of various groups, the proposed method demonstrates superior detection accuracy in domain-specific performance testing compared with existing algorithms. Additionally, it exhibits better generalization performance in cross-domain testing, verifying that the combined spatiotemporal convolution and Transformer model can enhance model generalization performance.

Keywords: video tampering detection; spatiotemporal consistency; temporal convolution kernel; temporal convolution kernel; Transformer model

1 引 言

随着科技的发展进步,深度神经网络在计算机视觉的各个领域被广泛应用,视频篡改技术因此发展迅速,篡改视频与原始视频间差距愈发缩小。视频篡改通过生成对抗网络[1]和自编码器对视频中的人像数据进行修改以及人脸交换[2]。虽然该技术可以应用于虚拟现实、教育以及动画制作等有利方面,但是篡改方法也可能会用于恶意的目的,由于其简单易用的特点,通过封装好的 APP 一键便能完成视频篡改过程。现如今的互联网时代,一些名人政客的肖像数据可以被很轻易地收集,不良分子利用篡改技术,散步不实消息,传播虚假视频图像,在社会舆论、司法刑侦等方面造成诸多不良影响。因此,需要研究行之有效的伪造检测方法。

早期的检测思路将注意力集中在寻找方法缺陷上,篡改之后的图像会存在一些特殊纹理、空间域、频域的细微波动。比如 Zhang 等[3],借助人工设计的规则鉴别伪造数据,通过简单的训练便可取得较好的分类效果。但随着深度伪造技术的发展,肉眼可见的方法缺陷逐渐被隐藏,人工设定的判别指标逐渐失效。之后的一些研究中,人们从图像本身出发,深入挖掘图像信息,Liu 等[4]提出一种融入相位谱信息的伪造检测方法 SPSL,发现视频篡改过程中图像会进行多次的上采样。该操作会导致明显的伪像或锯齿状噪声,这种缺陷在相位谱中可以更为显著地被观测。并且因为上采样过程在伪造流程中存在的普遍性,SPSL 方法在跨数据集测试中表现出更好的泛化性能。瞿远近[5]提出的 IGFNet 方法改进了高斯滤波方式,提高了篡改视频的检测精度,但对图像多个层级不同频率信息利用不够充分。Zhao 等[6]提出一种基于多注意力的深度伪造检测方法,将深度伪造检测问题转换为细粒度分类问题,关注图像中存在的细微结构特征,并利用多注意力模型将检测重心集中到篡改区域,以降低特征冗余,增强分类特征信息。这些检测方法在当时取得了良好的结果,图像本身所包含的信息被很好地发掘利用。但随着篡改技术的快速发展,篡改痕迹被隐藏,在图像内提取的判别信息渐渐不能满足需求,研究方向开始发生变化。

虽然一些篡改方法可以在空间维度上消除篡改痕迹,但是视频相邻帧之间存在着紧密的关联和连贯性,篡改方法可以篡改帧内数据,但是无法完美复现视频帧之间存在的时序相关性。基于此,Zheng 等[7]提出一个端到端的框架,限制了空间相关卷积核的大小,鼓励模型学习时间上的不一致性特征。这种方法在多个数据集上有着优异的性能表现。此外,Coccomini 等[8]提出一种具有左右分支的卷积神经网络结构,将 ViT[9](Vision Transformer)与卷积网络 EfficientNet B0 相结合,提取人脸特征,基于简单投票的方案,用于处理同一视频镜头中的多个不同人脸。在时间上和跨多个人脸上聚合推断出视频片段的真伪,在多个实验数据集中表现优异。

但是现有的一些检测方法在时间维度中提取存在的时序信息时,会忽视空间信息对模型检测性能的影响。利用篡改视频与真实视频间存在的时空不一致性鉴别视频真伪,如何在保留时序信息的同时,增强空间特征表示成为研究的重点。

由此,提出一种基于时空一致性的视频篡改检测方法,在关注数据时间维度的同时,保留空间特征表示。通过不同的视频采样步幅预处理视频数据,结合时序卷积核以及空间卷积核在高低采样率视频数据中分别侧重提取时空特征的不同方向,对提取到的空间语义特征和时序不一致特征进行逐阶段的特征连接,从而获得更完备的时空特征表示,并结合 ViT 在时空特征序列中捕捉数据中存在的长距离依赖关系,增强时空特征表示,提高模型检测性能。

2 方法原理

本文提出一种基于时空一致性的视频篡改检测方法,检测模型整体架构如图 1 所示。设置不同输入视频采样间隔 t,获得不同的输入视频帧序列,同时为了加速模型训练过程,采用 RetinaFace[10]定位人脸面部矩形区域,去除人物背景,降低模型计算量。低采样率视频帧序列更好地保留了空间特征,高采样率视频帧序列放大了帧之间时序不一致的存在痕迹,使得模型可以充分探索视频帧内空间伪造痕迹以及视频帧间存在的时空不一致性。

深度卷积网络有着优异的特征提取能力,选择 ResNet50 作为主干网络,利用时序卷积核(Temporal Convolution Kernel, TCK)以及空间卷积核(Spatial Convolution Kernel,SCK)替换原有卷积结构,增强模型对于时空特征信息的提取能力。同时在时序特征与空间特征之间建立横向连接,在捕获视频帧间时序相关性的同时强化空间特征表示。此外,利用 ViT 模型提取长距离依赖关系,在图像全局理解信息。ViT 通过自注意力机制在图像的不同区域之间建立全局关联,这使得模型能够捕获图像中远距离的上下文信息,进一步聚合增强存在的时空不一致性,从而实现对视频数据是否经过篡改问题的判定。

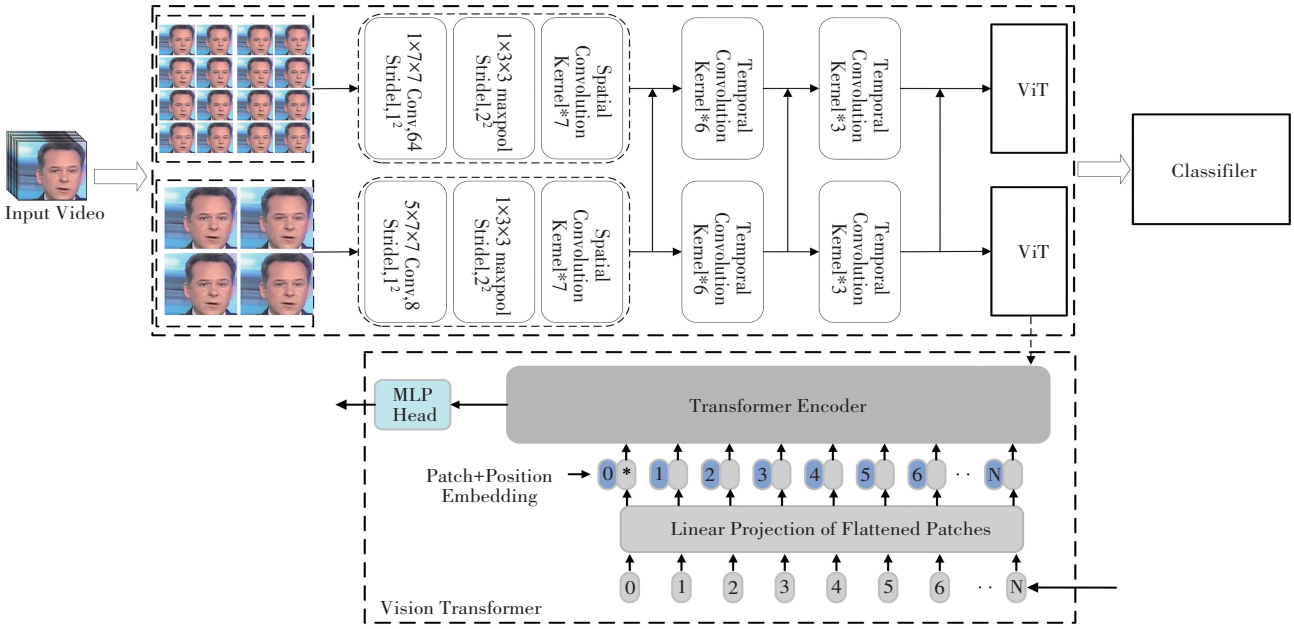


图 1　基于 Transformer 的时空一致性视频篡改检测

Fig. 1　Video tampering detection based on Transformer for spatiotemporal consistency

2.1　数据预处理

视频数据输入模型前,通过不同的时间步长 t_s 以及 t_f 对输入数据进行采样。将输入视频视为连续的图像序列 T,图像序列 T 被划分为 T/t 个单独的图像集合,t 设置为 16,则模型仅处理包含在 16 帧中的一帧画面。对于不同的采样间隔 t,以较大的时间步长 t_s($t_s/8=t_f$)对输入视频进行采样时,可以更好地利用视频帧内存在的空间特征信息。相似地,通过缩小采样步长 t,提高视频采样率,能够更好地提取视频帧存在的时序不一致。

此外,深度伪造技术对人像数据进行篡改时,一般会集中在人的面部区域,通过对背景数据的裁剪,可以降低模型计算量,同时避免模型对背景数据的拟合程度。因此首先采用 RetinaFace 预处理输入数据,根据锚点定位人脸面部区域,设置人脸定位框。此外,将每个帧上的面部区域放大 1.5 倍后进行裁剪,大小调整为 224×224。

2.2　时空特征提取

虽然一些视频篡改方法可以消除空间维度上存在的篡改痕迹,但是在时间维度上,篡改技术的发展稍显落后,伪造痕迹并不能完全被隐藏。通过以不同的时间间隔 t_s 以及 t_b 预处理视频数据,引入时序卷积核 TCK 以及空间卷积核 SCK,提取不同采样率数据中存在的篡改痕迹,可以充分探索空间和时间维度上存在的不一致性特征。SCK 以及 TCK 结构如图 2 所示,空间卷积核本质上与二维卷积类似,目标在于提取存在的空间伪造痕迹,时序卷积核融合了 3D CNN 的相关概念,在时间维度捕获特征。

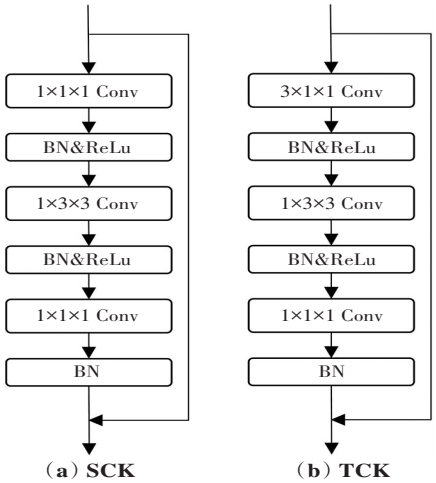


图 2　空间卷积核以及时间卷积核结构

Fig. 2　Structures of spatial convolution kernel and temporal convolution kernel

低采样率数据层中采样步幅较大,时间感受野几乎没有相关性,过早使用 TCK 会降低模型的整体性能。因此,如图 1 所示,将 ResNet 第 3 个块中的卷积核用 SCK 取代,以便于捕获更详细的伪造指纹。模型的前三块可以视为空间层面的 2D 卷积过程,模型末端两个模块的卷积核用 TCK 取代。在高采样率数据层中,为了更好地提取帧间时序信息,将池化层之后的所有层卷积核替换为 TCK,增强提取视频帧间时序信息的能力。

同时,每次经过时序卷积前后在高低采样率视频数据之间建立横向连接,实现这两个层级之间的信息通信,补充高采样率层中的时序信息,输出如式(1)所示:

$$T_i = f_t(T_{i-1}) \tag{1}$$

其中,$f_t(\cdot)$ 表示时序层的卷积函数,$T_{i-1} \in \mathbf{R}^{T\times W\times H\times C}$ 表示上一层的时序特征输出,T 指帧样本的数量,W 以及 H 表示图像尺寸,C 代表通道数。

另外,由于采样步幅不一致,会导致高低采样率数据提取的特征尺寸不同,将 t_s 与 t_b 间比值表示为 α,且 $\beta = 1/\alpha$。低采样率层特征即空间特征尺寸表示为 $T\times W\times H\times C$,由此可得高采样率层特征即时序特征尺寸表示为 $\alpha T\times W\times H\times \beta C$。为了建立空间特征与时序特征之间的信息通道,对高采样率层特征进行形状匹配,将时间维度打包到通道维度,即得到各维度相同尺寸的表示 $\{T, W, H, \alpha\beta C\}$,经过匹配后的特征形状,与空间特征尺寸一致。由此可得空间特征层的特征表示,空间特征层输出如式(2)所示:

$$S_i = \{f_t(T_{i-1}) + D[f_s(S_{i-1})]\}_{i=1}^{i=N} \tag{2}$$

其中,$D(\cdot)$ 表示特征匹配函数,$S_{i-1} \in \mathbf{R}^{T\times W\times H\times \alpha\beta C}$ 表示上一层的空间特征输出,T 指帧样本的数量,W、H、C 分别代表单个帧的宽度、高度以及通道数,α 设置为 8,β 设置为 1/8。

2.3 Transformer

相比于传统的卷积神经网络(CNN),ViT 更能够理解整体语境,在长距离关系的图像数据处理任务中有着优异的表现,通过自注意力机制在不同区域之间建立全局关联,捕获全局依赖关系。不同采样率的输入视频帧,经过主干网络特征的提取,得到低采样率特征表示 $S \in \mathbf{R}^{C\times T\times H\times W}$ 以及高采样率特征表示 $T \in \mathbf{R}^{T\times W\times H\times C}$。为了匹配 Transformer 模型的输入序列表示,将空间特征 S 与时间 T 再拆分成一个类似的特征序列,并在特征序列中添加图像的编码特征 I_{class} 以及位置编码 E_{pos}。将 Transformer 模型中的输入序列 embbedding 表示为 Z_S,得到输入序列 Z_S 如式(3)所示:

$$Z_S = [I_{\text{class}}; S_1E, S_2E, \cdots, S_NE] + E_{\text{pos}} \tag{3}$$

其中,S_i 表示特征序列 S 的第 i 个特征序列,E 表示可训练的线性映射,同理可得 Z_T 的序列表示,如式(4)所示:

$$Z_T = [I_{\text{class}}; T_1E, T_2E, \cdots, T_NE] + E_{\text{pos}} \tag{4}$$

得到标准的输入序列 Z_S 以及 Z_T 后,将其送入 Transformer 模型进行训练,Transformer 编码器模型结构如图 3 所示。

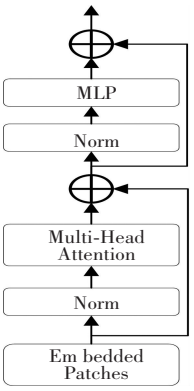


图 3 Transformer 编码器

Fig. 3 Transformer encoder

多头注意力模块(MSA)对输入 $\boldsymbol{X}$ 执行 3 个可学习的线性投影 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$,分别生成查询嵌入 $\boldsymbol{Q}$(Query)、键嵌入 $\boldsymbol{K}$(Key)和值嵌入 $\boldsymbol{V}$(Value),计算过程如式(5)所示:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \boldsymbol{V} = XW_V \tag{5}$$

对输出的 $\boldsymbol{Q}$、$\boldsymbol{K}$ 以及 $\boldsymbol{V}$ 矩阵进行自注意力计算,过程如式(6)所示:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{6}$$

其中,d_K 是 $\boldsymbol{K}$ 中每个输入向量的维度。

如图 3 所示,Transformer 模型训练过程中在计算不同的注意力嵌入和分类标签之前进行归一化,由此可得中间层 z_l 输出结果,计算过程如式(7)、式(8)所示:

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{7}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z_l \tag{8}$$

其中,LN(·)表示归一化,z_l 表示第 l 层的编码图像,$l \in [1, 2, \cdots, L]$,最终层输出表示 Z_{TL} 以及 Z_{SL}。将 Z_{TL} 以及 Z_{SL} 的第一个元素作为特征表示,转发到分类器进行结果预测。

3 实验验证

3.1 数据集设置

选择两个广泛使用的基准数据集进行实验,包括 FaceForensics++[12](FF++)以及 Celeb-DF[13] 数据集。FaceForensics++是目前最流行的数据集之一,其中包含 1 000 个原始真实视频以及相应的 4 000 个篡改视频,该数据集包含的篡改视频通过 Deepfakes(DF)、FaceSwap(FS)、Face2Face(F2F)以及 NeuralTextures(NT)4 种方式对真实视频进行处理后生成,同时采用 H. 264 编码进行压缩获得不同分辨率数据。Celeb-DF 数据集是一个高质量伪造视频数据集,其中包括 590 个标注不同年龄、种族和性别受试者的原始视频,以及 5 639 个相应的 DeepFake 视频,总帧数超过 230 万。对于大多数现有检测方法而言,具有一定的挑战性。

3.2　实验设置及评估指标

为了统一实验结果,实验数据集采用统一的划分比例。其中 60%的数据作为训练集,验证集与测试集各占据 20%。对于不同的数据集采用相同的处理方式,按照不同采样率在视频中抽取视频帧数据,保留视频帧的连贯性以确保能够提取到正确的时序特征信息。人像定位工具选择 RetinaFace,将所有视频帧数据保存为对应的面部信息。

模型基于 pytorch 深度学习框架实现,优化器为 AadmW,初始学习率 α 设置为 0.000 2,学习率随训练过程逐步衰减,batch size 设为 32,平衡因子设为 0.1,最大训练轮次 150 轮。

评估指标选择检测准确率 f_{Acc}(Accuracy)以及受试者工作特征曲线下面积 f_{AUC}(Area Under the Receiver Operating Characteristic Curve),计算方式如式(9)所示:

$$f_{\mathrm{AUC}}=\frac{1}{2}\sum_{i=1}^{n-1}(R_{\mathrm{FP}}^{(i+1)}-R_{\mathrm{FP}}^{(i)})\times(R_{\mathrm{TP}}^{(i+1)}+R_{\mathrm{TP}}^{(i)}) \tag{9}$$

其中,R_{FP} 表示在所有实际为伪造的样本中,被错误地判断为真实的比例即假正例率;R_{TP} 表示在所有实际为真实的样本中,被正确地判断为真实的比率即真正例率;ROC 曲线的横坐标为 R_{TP},纵坐标为 R_{FP},曲线下面积即为 f_{AUC} 值。

3.3　实验结果及分析

为了评估不同质量视频数据对于模型性能的影响,将所提方法及其对比方法在 FaceForensics++数据集不同压缩倍率的高质量(C23)以及低质量(C40)视频上进行训练与测试,评估指标为 f_{Acc} 以及 f_{AUC},实验结果如表 1 所示。所提方法在低质量视频上,f_{AUC} 数值达到了 93.05%,相较于 ResNet 骨干网络,数值上提高了 11.29%。在低质量分列其他各组对比实验中,实验结果均优于其他对比方法。高质量数据分组的对比实验中,仅 f_{AUC} 数值略微落后于先进的 M2TR 方法,高质量分列其他各组对比实验均优于其他对比方法。所提方法在检测精度以及泛化性能方面均表现出竞争性。

表 1　FF++数据集评估结果

Table 1　FF++dataset evaluation results　/%

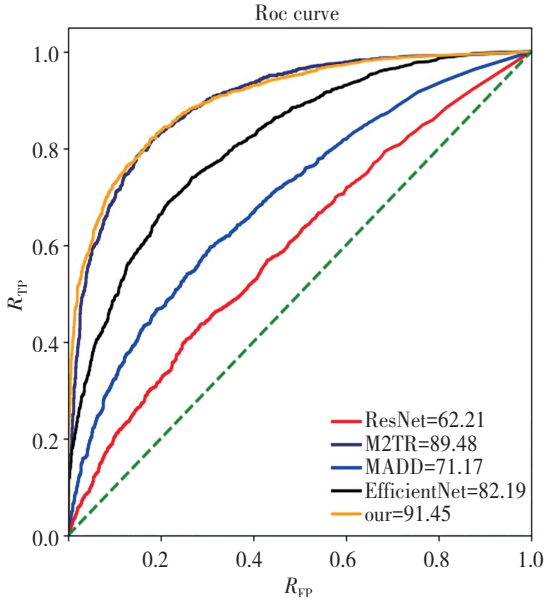
方法	C23		C40	
	f_{Acc}	f_{AUC}	f_{Acc}	f_{AUC}
ResNet[14]	93.71	94.86	80.32	81.76
EfficientNet[11]	91.67	93.21	82.19	82.77
M2TR[15]	96.23	99.51	81.57	90.47
F^3-Net[16]	93.71	98.88	87.43	92.18
MADDl[6]	95.74	99.13	83.27	91.31
Ours	96.14	99.47	88.91	93.05

不同伪造方式生成的图像之间存在差异,对于未经拟合的伪造方法,会对模型性能造成不同程度的影响,甚至检测精度大幅度下降。为进一步评估所提方法,在 FF++数据集的不同伪造方法生成的伪造数据中进行评估,模拟缺乏对模拟数据来源了解的情况,对模型性能进行跨伪造方式测试。FF++(HQ)数据集包含 Deepfakes(DF)、FaceSwap(FS)、Face2Face(F2F)以及 NeuralTextures(NT)4 种伪造方式,选择 3 种伪造方式用于模型训练并在剩下的一种伪造方式上进行测试,测试数据如表 2 所示。本文方法每一组对比实验中,f_{AUC} 数值结果均超过 90%,相较于同组对比方法,略微落后于 M2TR 方法,在其他组的对比实验中,f_{AUC} 数值均有不同幅度提高,尤其在 FS 伪造方式中,大幅提高了模型性能,f_{AUC} 数值达到 92.31%。为直观感受模型性能表现,如图 4 所示,绘制 NT 以及 FS 数据子集测试 ROC 曲线。依据跨伪造方式的性能测试结果,在相同图像质量设定下,虽然测试集数据对于检测模型而言是不可见的,但是本文方法依旧有着优异的性能表现,均领先组内其他对比方法,验证了检测模型能够很好地处理未知伪造方式的图像数据,反映出模型具备良好的泛化性能。

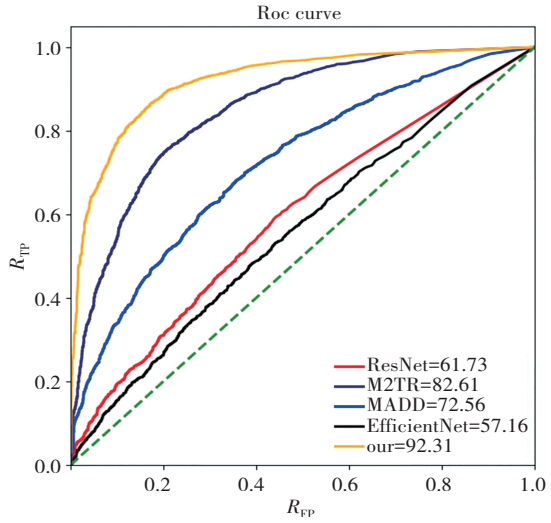
表 2　跨伪造方式的 f_{AUC} 评估结果

Table 2　Evaluation results across forgery methods f_{AUC}　/%

模型	DF	FS	F2F	NT
Resnet	93.92	61.73	87.92	62.21
M2TR	99.57	82.61	93.96	89.48
MADD	97.82	72.56	91.23	71.17
EfficientNet	97.69	57.16	87.83	82.19
Ours	97.91	92.31	95.53	91.45



(a) NT



(b) FS

图 4 ROC 曲线

Fig. 4 ROC curve

不同伪造数据集包含的伪造手段存在差异,图像质量也不相同,一些模型在数据集内测试有着优异的性能表现,但在跨数据集性能测试中却表现出明显的性能下降。除了篡改方式未知的原因外,图像质量的波动也会导致模型性能的下降。因此,为了进一步验证模型泛化性能,对所提方法以及一些主流检测方法进行跨数据集性能测试。训练集选择 FF++数据集,验证集选择 Celeb-DF 数据集。所提方法与对比方法测试结果如表 3 所示,表 3 中数值代表检测模型 f_{AUC} 的数值。依据表中结果,域内测试结果以及跨域测试结果均优于对比方法,包括先进的 MADD 方法,跨数据集验证 f_{AUC} 数值提高了 7.99%。在与 M2TR 方法的比较中,训练集测试数据略高于所提方法,但在跨数据集测试中,f_{AUC} 数值提高 5.44%。综合本文方法与对比方法的性能表现,依据跨域泛化性实验中的测试结果,当输入视频质量存在波动,并且在篡改技术未知的实验设置下,本文方法能够提取到更完备的时空特征表示,结合 ViT 模型在时空特征序列中捕捉全局依赖关系,所提取的伪造特征更加具备泛化性。

表 3 跨数据集的 f_{AUC} 评估结果

Table 3 Evaluation results across datasets f_{AUC} /%

方法	FF++	Celeb-DF
Resnet	94.86	57.30
MADD	99.13	61.74
M2TR	99.51	64.29
EfficientNet	93.21	59.17
Ours	99.47	69.73

3.4 消融实验

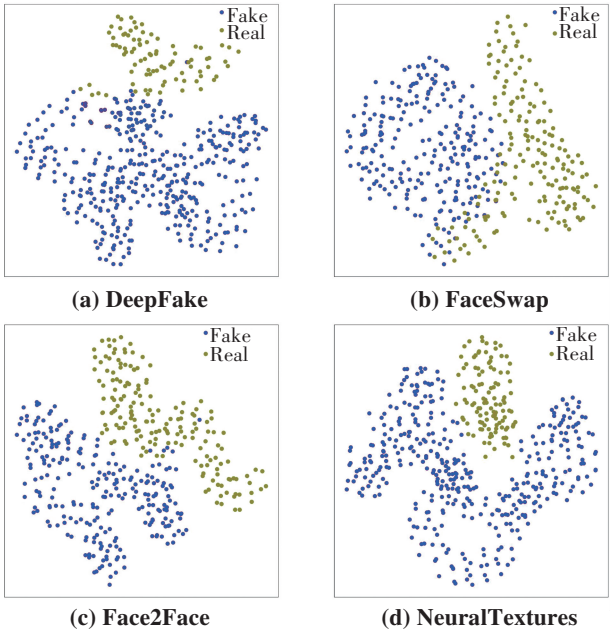
本文方法可以分为引入时空卷积的 ResNet 网络以及 ViT 特征提取模块。为验证改进的 Resnet 网络结构以及 ViT 对模型性能的影响,设计 3 组消融模型分别为 ViT 模型、原始 ResNet 与 ViT 联合模型以及引入时空卷积核的 ResNet 与 ViT 联合模型。各组消融模型在 FaceForensics++高质量数据子集上进行性能研究,评估指标选择 f_{AUC},实验结果如表 4 所示。在与单一 ViT 模型比较中,f_{AUC} 提高了 1.84%,虽然 ViT 模型得益于强大的长距离上下文依赖关系捕捉能力,能够获得优异的性能测试结果,f_{AUC} 达到了 97.63%,但在联合 Resnet 后,f_{AUC} 数值提高了 0.28%,与 Resnet 的联合补充了空间信息的不足,对模型性能产生了正向增益。在同时引入时空卷积机制以及 ViT 特征提取后,模型性能得到了最大幅度的提升,在保留时序信息不损失的前提下,进一步在时空特征序列中补充了空间信息。依据表 4 内数据结果,表明了本文方法中各个模块的有效性。

表 4 FF++(C23)数据集上的消融实验结果

Table 4 Ablation experimental results on FF++ (C23) dataset /%

模型	ViT	ResNet	时空卷积	f_{AUC}
1	✓	—	—	97.63
2	✓	✓	—	97.91
3	✓	✓	✓	99.47

t-SNE[17] 呈现了数据在低维空间中的分布,通过观察 t-SNE 图,可以发现数据点之间的相似性和聚类关系,因此使用 t-SNE 来可视化时空特征。图 5 展示了时空特征序列的可视化结果,模型分别在 FF++4 种篡改方式下进行训练,可以显著观察到,时空一致性特征在伪特征和真特征之间具有明显的可区分性。



(a) DeepFake (b) FaceSwap

(c) Face2Face (d) NeuralTextures

图 5 时空特征可视化

Fig. 5 Visualization of spatiotemporal features

4 结 论

针对现有检测算法出现的性能不足问题,提出了一种基于时空一致性的视频篡改检测方法。不同采样率视频数据与改进的卷积核结构加强了对时空特征的提取能力,通过 ViT 模型对全局关系的建模,进一步补充了时空特征信息。相较于传统 CNN 方法,ViT 通过自注意力机制更擅长全局信息的建模。结合两者可以更好地平衡计算效率,提高模型对图像信息的全面理解。不同采样率的数据预处理与时空卷积核的协同作业加强了卷积模型捕获时空不一致性特征的能力,有效增强了模型检测精度以及泛化性能。针对模型分别进行域内以及域外测试,模拟检测未知伪造方式以及输入数据质量存在波动情况下测试模型的性能表现,定性定量分析模型数据。实验结果表明:所提方法能够有效鉴别视频是否篡改,提升了模型性能。然而,随着篡改技术的不断发展,需要研究更加有效的检测方法,未来考虑从不同检测角度出发,利用多模态信息特征协同研究视频篡改问题。

参考文献(References):

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

[2] SHI Y, YANG X, WAN Y, et al. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 11254-11264.

[3] ZHANG Y, ZHENG L, THING V L. Automated face swapping and its detection[C]//2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP). IEEE, 2017: 15-19.

[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. IEEE, 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): 40-47.

[6] 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. IEEE, 2021: 2185-2194.

[7] 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. IEEE, 2021: 15044-15054.

[8] COCCOMINI D A, MESSINA N, GENNARO C, et al. Combining efficientnet and vision transformers for video deepfake detection[C]//International conference on Image Analysis and Processing. Cham: Springer International Press, 2022: 219-229.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30(1): 6000-6010.

[10] DENG J, GUO J, VERVERAS E, et al. Retinaface: Single-shot multi-level face localisation in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 5203-5212.

[11] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. ACM Press, 2019: 6105-6114.

[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. IEEE, 2019: 1-11.

[13] 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. IEEE, 2020: 3207-3216.

[14] 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. IEEE, 2016: 770-778.

[15] WANG J, WU Z, OUYANG W, et al. M2TR: Multi-modal multi-scale transformers for deepfake detection[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. New York: ACM, 2022: 615-623.

[16] 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 International Publishing, 2020: 86-103.

[17] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

责任编辑:李翠薇