计算机科学
COMPUTER SCIENCE

基于多尺度Transformer融合多域信息的伪造人脸检测

马欣, 吉立新, 李邵梅

引用本文

马欣, 吉立新, 李邵梅. 基于多尺度Transformer融合多域信息的伪造人脸检测[J]. 计算机科学, 2023, 50(10): 112-118.

MA Xin, JI Lixin, LI Shaomei. Forgery Face Detection Based on Multi-scale Transformer Fusing Multi-domain Information [J]. Computer Science, 2023, 50(10): 112-118.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900048

基于多尺度 Transformer 融合多域信息的伪造人脸检测

马　欣[1,2]　吉立新[2]　李邵梅[2]
1 郑州大学网络空间安全学院　郑州 450001
2 战略支援部队信息工程大学信息技术研究所　郑州 450002
(15543782756@163.com)

摘　要　当前,基于 Deepfakes 等深度伪造技术生成的“换脸”类伪造视频泛滥,给公民个人隐私和国家政治安全带来巨大威胁,为此,研究视频中深度伪造人脸检测技术具有重要意义。针对已有伪造人脸检测方法存在的面部特征提取不充分、泛化能力弱等不足,提出一种基于多尺度 Transformer 对多域信息进行融合的伪造人脸检测方法。基于多域特征融合的思路,同时从视频帧的频域与 RGB 域进行特征提取,提高模型的泛化性;联合 EfficientNet 和多尺度 Transformer,设计多层级的特征提取网络以提取更精细的伪造特征。在开源数据集上的测试结果表明,相比已有方法,所提方法具有更好的检测效果;同时在跨数据集上的实验结果证明了所提模型具有较好的泛化性能。

关键词:伪造人脸检测;多尺度 Transformer;EfficientNet;频域特征;特征融合
中图法分类号　TP391

Forgery Face Detection Based on Multi-scale Transformer Fusing Multi-domain Information

MA Xin[1,2],JI Lixin[2] and LI Shaomei[2]
1 School of Cyber Science and Engineering,Zhengzhou University,Zhengzhou 450001,China
2 Institute of Information Technology,PLA Strategic Support Force Information Engineering University,Zhengzhou 450002,China

Abstract　At present,the proliferation of “face-changing” fake videos generated based on deep forgery technologies such as Deepfakes poses a considerable threat to citizens' privacy and national political security. Therefore,it is of great significance to study deep-faked face detection technology in videos. Aiming at the problems of insufficient extraction of facial features and weak generalization ability of existing forged face detection methods,this paper proposes a fake face detection method based on multi-scale Transformer for the fusion of multi-domain information. First,based on the idea of multi-domain feature fusion,feature extraction from the frequency domain and RGB domain of video frames improves the generalization of the model. Second,the EfficientNet and multi-scale Transformer are combined to design a multi-level feature extraction network to extract more elaborate forged features. The test results on open-source datasets show that the proposed method has better detection performance than the existing methods. At the same time,experimental results on cross-datasets prove that the proposed model has better generalization performance.

Keywords　Forgery face detection,Multi-scale Transformer,EfficientNet,Frequency domain features,Feature fusion

1　引言

随着深度学习技术的不断发展,基于 Deepfakes 等深度伪造技术生成的“换脸”类伪造视频在网络上广泛传播。此类技术可将视频中的人脸替换成目标人物,从而制作出目标人物做特定动作的假视频[1]。该技术已经被用于丑化名人政客、操纵选举,甚至在俄乌战争中也被应用[2],潜在危害巨大。因此,研究此类深度伪造视频的自动检测技术具有重要意义。

针对视频中伪造人脸的检测,研究人员提出了大量的伪造检测方法,主要分为基于帧内特征和基于帧间特征两类。其中基于帧内特征的研究更广泛,其原理是将视频处理成帧,使用分类器进行逐帧检测以判别真实与伪造人脸。早期用于帧内检测的分类器主要都是基于卷积神经网络(Convolutional Neural Networks,CNN)实现的,如 Zhou 等[3]使用双流网络分别捕获的面部篡改伪影特征和局部噪声残差特征训练分类器;Nguyen 等[4]设计了一个包含编码器和 Y 型解码器的网络,同时识别被伪造的面部并通过多任务学习定位被伪造的区域;Hsu 等[5]提出了一种基于嵌入对比损失的方法,该方法可以捕捉到不同伪造图像的联合特征,最后通过连接一个分类器完成分类任务;Tariq 等[6]通过集成多种 CNN 网络

到稿日期:2022-09-06　返修日期:2022-12-10
基金项目:国家自然科学基金创新研究群体科学基金(61521003)
This work was supported by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China(61521003).
通信作者:李邵梅(lishaomei_may@126.com)

对伪造人脸视频进行检测判别。

近年来，Transformer 在自然语言处理领域成功应用，它也被引入到计算机视觉领域。已有研究证明，在图像处理中，相较于卷积操作，Transformer 的动态注意力机制可以捕获全局上下文信息，具有强大的长期依赖建模能力[7]。因此，Transformer 模型被引入到计算机视觉的各类任务中，并且表现出具有竞争力的甚至更好的性能[8]。文献[9]首次把视觉 Transformer 用于视频中的伪造人脸检测，取得了不错的效果。随后，研究人员对基于 Transformer 的伪造视频人脸检测方法进行了改进。文献[10]使用两个不同的 CNN 分支分别提取输入图像的局部特征和全局特征，基于注意力机制，利用 Transformer 网络将两类特征相结合，证明了局部和全局特征相结合的方法可以获得更好的结果。文献[11]提出了一种具有蒸馏方法的视觉 Transformer 模型，结合 CNN 定位伪造视频中的篡改区域，在 DFDC 数据集上实现了 SOTA (State of the Art)的性能。上述研究成果表明，CNN 和 Transformer 相结合的网络结构应用于视频伪造人脸检测任务具有出色的表现，因此本文也将基于此设计用于视频中伪造人脸特征提取的网络。

此外，针对现有的伪造检测方法大多主要从图像 RGB 域中提取特征，对在色彩空间中被巧妙操纵的伪造视频的检测存在一定局限性[12]的不足，本文利用大多数人脸伪造技术都会导致图像频域出现明显变化的特点，融合图像频域特征和 RGB 域特征对伪造人脸进行检测，以补充在图像 RGB 域中不可感知的伪影信息[13]。

综上，本文提出了一种基于多尺度 Transformer 对多域信息进行融合的伪造人脸检测方法。首先，基于空域和频域特征融合的思路，从原始图像和基于相位谱重构的图像中同时进行特征提取；然后，设计基于 EfficientNet[14] 和多尺度 Transformer[12] 相结合的多层级特征提取网络。EfficientNet 作为特征提取网络捕获人脸图像的底层和局部信息，多尺度 Transformer 基于 EfficientNet 提取的特征，进一步实现局部特征到全局特征的融合，充分提取不同粒度的面部特征并实现统一表征，进而提高检测的精度。

2　基于多尺度 Transformer 融合多域信息的伪造人脸检测

本文基于 EfficientNet 和多尺度 Transformer，提出了融合多域特征的人脸特征提取网络，并基于此设计了视频伪造人脸检测方法。图 1 给出了对每个视频帧中伪造人脸的检测流程。首先，对于待检测视频，从每帧中检测并裁剪人脸图像区域，将人脸 RGB 图像和基于相位谱重构的人脸图像融合，一并输入轻量化网络 EfficientNet 中进行特征提取；然后，将上述特征图送入多尺度 Transformer 模型中进一步进行特征提取，实现局部特征到全局特征的融合以及多尺度特征的融合；最后，将多级融合后的特征送入分类层进行分类，得到视频中人脸图像的检测结果。下面将分别对图 1 中的各主要功能模块进行介绍。

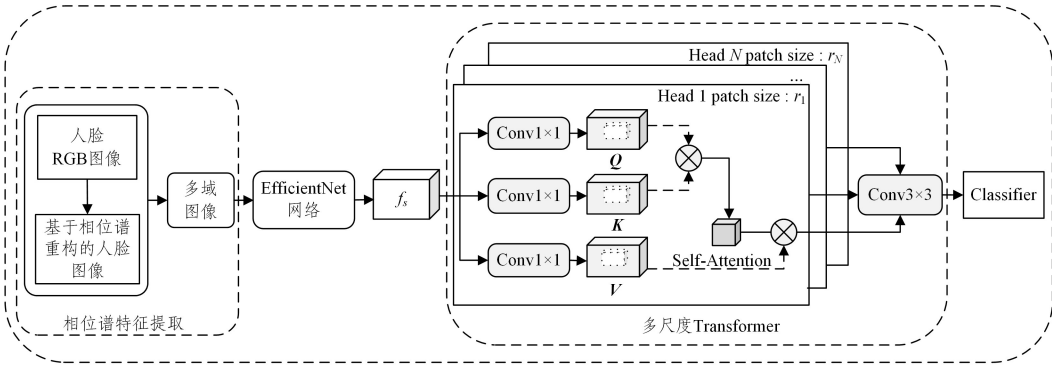

图 1　基于多尺度 Transformer 融合多域信息的检测原理图

Fig. 1　Detection principle diagram based on multi-scale Transformer fusing multi-domain information

2.1　基于相位谱的人脸图像重构

根据文献[13]的研究成果可知，对于基于自动编码器[15]和生成对抗网络[16]生成的伪造视频，上采样是解码目标人脸的关键步骤，累积上采样操作会在图像上引入频率伪影，而相位谱有助于捕获视频中伪造人脸区域由上采样操作引入的伪影。为此，本文拟提取人脸区域的相位谱特征作为检测的依据。为了将相位信息应用于图 1 所示的 EfficientNet 和多尺度 Transformer 融合的模型，对于待检测的人脸图像，首先，将其变换到频域；其次，将人脸图像频域的相位谱重构成图像，得到人脸相位谱的空域表示 P；再次，将相位谱的空域表示 P 与原始人脸的 RGB 图像进行融合，生成四通道人脸图像 RGB-P；最后，将上述 RGB-P 四通道图像输入神经网络模型中进行特征提取并分类。

具体提取过程如下：

首先，对每一帧视频中的人脸图像进行离散傅里叶变换得到：

$$X(u)=\frac{1}{N}\sum_{n=0}^{N-1}x(n)\left(\cos\frac{2\pi un}{N}-j\sin\frac{2\pi un}{N}\right)=Re(u)+jIm(u) \tag{1}$$

其中，$X(u)$表示图像频谱在频率 u 处的值，$x(n)$表示图像像素点的值，N 表示像素总个数，j 表示复数。

基于式(1)计算得到的频谱 $X(u)$，通过式(2)计算得到其相位谱：

$$P(u)=\arctan\frac{\mathrm{Im}(u)}{\mathrm{Re}(u)} \tag{2}$$

其中，$Im(u)$和 $Re(u)$分别为频谱 $X(u)$的虚部和实部。

然后，对相位谱做傅里叶逆变换得到相位谱的空域表示，公式如下：

$$P(n)=\frac{1}{N}\sum_{n=0}^{N-1}p(u)\left(\cos\frac{2\pi un}{N}+j\sin\frac{2\pi un}{N}\right) \tag{3}$$

其中，$P(n)$表示相位谱空域像素点的值，$P(u)$表示相位谱在频率 u 处的值，N 表示像素总个数，j 表示复数。

经过一系列傅里叶变换和逆变换后，得到如图 2 所示的基于相位谱重构的人脸图像表示。对于 FaceForensics++数据集中的 4 种伪造方法，其基于相位谱重构的人脸图像与原始真实人脸图像相位谱存在明显差异，同时它们之间也是可区分的。特别地，基于 NeuralTextures 伪造方法生成的图像只是略微篡改了嘴部区域，与原始人脸图像非常相似，导致在图像 RGB 域中几乎无法分辨，但在基于相位谱重构的人脸图像表示中仍然是可区分的。

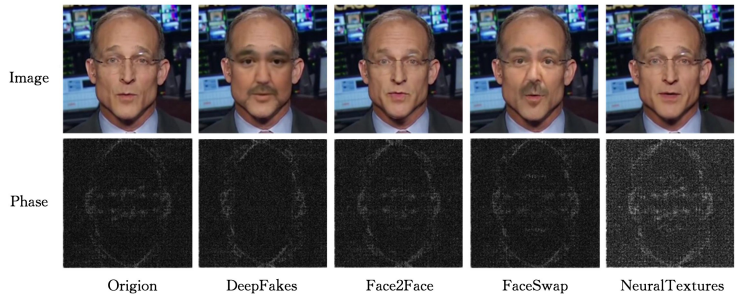

图 2 基于相位谱重构的人脸图像表示

Fig. 2 Face image representation based on phase spectrum reconstruction

2.2 基于 EfficientNet 网络的初始特征提取

EfficientNet 是 Google 于 2019 年提出的经典卷积神经网络模型，它利用一种复合模型缩放的方法，同时对网络的深度、宽度以及分辨率进行均匀缩放，综合 3 个维度提升网络指标。已有研究表明，EfficientNet 能够在大大减少模型参数量和计算量的情况下，在图像分类上取得较高的准确率[14]，是当前计算机视觉领域应用较广的模型。本文在综合考虑检测精度和计算成本的基础上，选用 EfficientNet-B4 网络对输入人脸图像进行初始的特征提取。其网络结构组成如表 1 所列。

表 1 EfficientNet-B4 网络结构

Table 1 EfficientNet-B4 network structure

阶段 i	操作 F_i	分辨率 $H_i\times F_i$	通道 C_i	层数 L_i
1	Conv3×3	380×380	48	1
2	MBConv1, k3×3	190×190	24	2
3	MBConv6, k3×3	190×190	32	4
4	MBConv6, k5×5	95×95	56	4
5	MBConv6, k3×3	48×48	112	6
6	MBConv6, k5×5	24×24	160	6
7	MBConv6, k5×5	24×24	272	8
8	MBConv6, k3×3	12×12	448	2
9	Conv1×1&Pooling&FC	12×12	1 792	1

EfficientNet-B4 网络主要由 3 部分组成：第一部分为 1 个卷积核大小为 3×3、步长为 2 的卷积层；第二部分共 7 层，每层均由多个移动倒置瓶颈卷积模块(Mobile Inverted Bottleneck Convolution，MBConv)组成，该模块来源于 MobileNetV2 中的移动倒置瓶颈模块，激活函数为 Swish，并引入了压缩和激励网络(Squeeze-and-Excitation，SE)模块进行优化，MBConv 模块重复的次数对应 EfficientNet-B4 网络结构中的层数；第三部分由 1 个 1×1 的卷积层、1 个全局平均池化层和 1 个全连接层构成。

如图 1 所示，在本文所提的检测方法中，输入人脸图像表示为 $X\in R^{H\times W\times C}$，其中 H，W 和 C 分别表示图像的高度、宽度和通道数。与传统的图像特征提取应用不同，EfficientNet-B4 特征提取网络的输入为融合原始 RGB 图像和基于相位谱重构人脸图像的 RGB-P 四通道图像，即这里的 $C=4$；输出为从该网络提取的浅层特征图 f_s。如图 1 所示，将 f_s 进行维度尺寸变换后送入多尺度 Transformer 网络中进行更高层次的特征提取和融合。

2.3 多尺度 Transformer 网络的特征融合

伪造人脸大部分的操作可能作用于面部不同的五官区域，且伪造区域的大小各不相同。为更好地捕获伪造特征，需要对图像中人脸各部位的远程关系进行建模，即不仅要计算局部邻域中区域的相似性，还要计算相距较远的区域的相似性。受 Transformer 模型在捕获全局上下文信息方面取得巨大成功的启发，本文在 Wodajo 等[9]工作的基础上采用多尺度 Transformer 结构[12]，对图像块划分不同尺度进行多次分析，覆盖不同大小的区域，以捕获更精细的伪造特征。整体结构如图 3 所示。

为了捕获多尺度的伪造模式，我们将特征图分割成不同大小的空间图像块，并计算不同注意力头的图像块之间的自注意力。如 2.2 节所述，将从 EfficientNet-B4 网络提取的特征图 $f_s\in R^{(H/4)\times(W/4)\times C}$ 输入到多尺度 Transformer 后，Transformer 从中提取形状为 $r_h\times r_h\times C$ 的空间图像块 $\{\boldsymbol{I}^h\}_{h=1}^N$，其中 $N=(H/r_h)\times(W/r_h)$，并将它们规整为第 h 个注意力头的一维输入向量。之后，使用全连接层将扁平化的图像块 $\{I^h\}_{h=1}^N$ 嵌入到查询嵌入 $\{q^h\}_{h=1}^N$ 中，按照类似的方法分别获得相应的键嵌入 $\{k^h\}_{h=1}^N$ 和值嵌入 $\{v^h\}_{h=1}^N$。然后，通过矩阵乘法和 softmax 函数计算得到不同图像块的相似度 $a_{i,j}^h$，公式如下：

$$a_{i,j}^h=\mathrm{softmax}\left(\frac{\boldsymbol{q}_i^h\times(\boldsymbol{k}_j^h)^{\mathrm{T}}}{\sqrt{r_h\times r_h\times C}}\right),1\leqslant i,j\leqslant N \tag{4}$$

其中，$\boldsymbol{q}_i^h$ 表示第 i 个查询嵌入，$\boldsymbol{k}_j^h$ 表示第 j 个键嵌入。

之后通过对相关图像块的相似度值进行加权求和，得到查询图像块的输出：

$$\boldsymbol{o}_i^h=\sum_{j=1}^N a_{i,j}^h\boldsymbol{v}_j^h \tag{5}$$

其中，$\boldsymbol{v}_j^h$ 是第 h 个注意力头中第 j 个图像块对应的值嵌入，在得到所有图像块的输出后，将它们拼接在一起并规整为原始

空间分辨率。然后，将来自不同注意力头的特征拼接起来，通过一个 2D 残差模块，处理后得到多域多尺度融合特征 $f_{m\_t}\in R^{(H/4)\times(W/4)\times C}$。最后，通过 softmax 分类层进行分类处理，输出检测结果，实现对视频中的伪造人脸的检测。

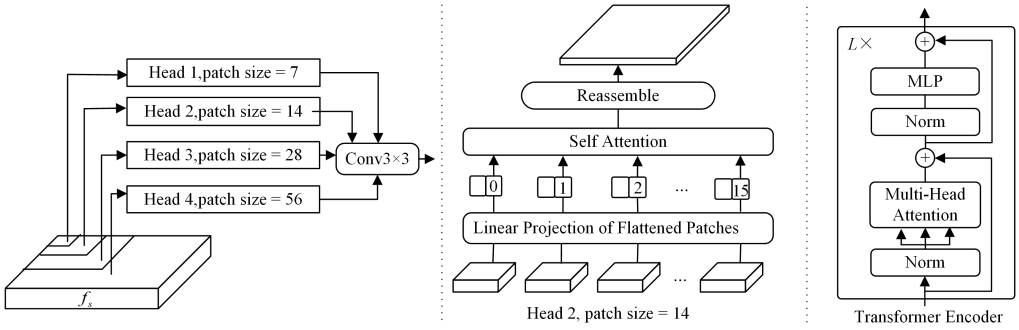

图 3　多尺度 Transformer 模块

Fig. 3　Multi-scale Transformer module

3　实验与分析

3.1　实验数据

为了评估所提方法的有效性，本文在公开数据集 FaceForensics＋＋(FF＋＋)[17]和 Celeb-DF[18]上进行实验。FF＋＋数据集是伪造人脸检测中使用最广泛的数据集，它包含来自互联网的 1000 个真实人脸视频，每个真实视频对应 4 个伪造视频，分别由 DeepFakes，Face2Face，FaceSwap，NeuralTextures 方法生成。同时为了模拟真实场景，采用 H.264 编码对真假视频进行不同程度的压缩，每个视频生成 Raw(c0)，HQ(c23)，LQ(c40)这 3 种压缩方式的视频[1]，分别表示原始视频、低压缩率下的高质量视频和高压缩率下的低质量视频。Celeb-DF 数据集由取自 YouTube 的 590 个真实视频和生成的 5639 个深度伪造视频组成，每条视频时长约为 13 s[16]。其中，深度伪造视频由改进的 DeepFakes 生成，欺骗性更强，是更具挑战性的数据集。

在本文实验中，在预处理环节采用人脸检测算法 RetinaFace[19]对视频逐帧进行人脸区域检测，将人脸区域裁剪并缩放至 224×224 大小再输入到检测模型中进行检测。在训练模型的过程中，使用 Albumentations 数据增强库对其进行常见的变换，如引入模糊、高斯噪声、转置、旋转及随机裁剪操作，目的是增加训练样本的多样性，避免模型过度关注图像的部分区域，以提高模型的泛化能力和鲁棒性。

3.2　实验环境及训练参数配置

本文实验平台为 Ubuntu 16.04 操作系统，所有代码均在 PyTorch 深度学习框架下实现，在 Nvidia TitanX 的 GPU 显卡上运行。为提高检测性能，本文使用在 ImageNet 上预训练的 EfficentNet-B4 网络。在模型训练过程中使用 Adam 优化器动态调整学习率，初始学习率为 0.0001，当损失值在 5 个完整 epoch 后没有下降时，学习率降为原来的一半。Batch size 设置为 32，共训练 40 个 epoch。

3.3　评价指标

本文使用准确率(Accuracy，Acc)和 ROC 曲线下的面积 AUC(Area Under Curve)作为评价指标。Acc 指分类正确的样本数占总样本数的比例，是对总体准确率的度量。计算公式如式(6)所示：

$$Acc=\frac{TP+TN}{TP+FP+FN+TN} \tag{6}$$

其中，TP(True Positive)表示将伪造视频检测为伪造视频的数量，FP(False Positive)表示将真实视频检测为伪造视频的数量，FN(False Negative)表示将伪造视频检测为真实视频的数量，TN(True Negative)表示将真实视频检测为真实视频的数量。

AUC 评价指标不依赖于具体阈值的设定，在正负样本分布不均衡的数据集中，可以很好地评价模型的分类能力。该指标和以下两个变量相关：

$$FPR=\frac{FP}{FP+TN} \tag{7}$$

$$TPR=\frac{TP}{TP+FN} \tag{8}$$

其中，FPR(False Positive Rate)为 ROC 曲线的横坐标，表示将真实视频检测为伪造视频的比率；TPR(True Positive Rate)为 ROC 曲线的纵坐标，表示将伪造视频检测为伪造视频的比率。ROC 曲线一般都位于 $y=x$ 这条直线的上方，AUC 值越接近 1.0，模型分类效果越好。

3.4　实验结果及分析

3.4.1　与其他方法的对比分析

为了分析本文提出的模型的有效性，我们在 FF＋＋数据集上与其他经典算法进行了对比实验。因为这些算法的实验结果主要是在 FF＋＋数据集中 HQ(c23)版本和 LQ(c40)版本上获得的，为了能够引用这些算法原始论文中的实验结果进行对比，本文也在 FF＋＋数据集的这两个版本的数据上进行对比实验。分别将 FF＋＋数据集中 HQ(c23)版本和 LQ(c40)版本的视频按照 8∶1∶1 的比例划分为训练集、验证集和测试集进行实验，实验结果如表 2 所列。从实验结果可以看出，大多数方法在高质量视频集上，即 HQ 版本上的性能尚可，但是在低质量的视频集，即 LQ 版本上的性能显著下降。这是因为图像压缩导致色彩信息丢失，模型不能提取到充分的 RGB 图像特征。而本文方法通过引入频域特征为解决这个问题提供了思路。本文方法在 LQ 版本上的检测性能最优，相比目前性能最优的 MaDD 方法，精度高出 1.41%，AUC 值高出 2.21%。另外，本文方法在 HQ 版本上的检测效果和性能最好的 MaDD 基本相当，证明了本文方法在视频伪造

人脸检测任务中的有效性。

表 2 各方法在 FaceForensics++数据集上的检测结果

Table 2 Detection results of each method on FaceForensics++ dataset

(单位:%)

Methods	LQ(c40)		HQ(c23)	
	ACC	*AUC*	*ACC*	*AUC*
Face X-ray[20]	—	61.60	—	87.35
Xception[21]	86.86	89.30	95.73	96.30
EfficientNet-B4[22]	86.67	88.20	96.63	99.18
Two Branch[23]	—	86.59	—	98.70
SPSL[13]	81.57	82.82	92.39	95.32
MaDD[24]	88.69	90.40	**97.60**	**99.29**
Ours	**90.10**	**92.61**	97.39	99.20

3.4.2 跨数据集的结果及分析

为了评估本文提出的模型的泛化性,本小节将在 FF++(HQ)数据集和 Celeb-DF 数据集进行测试。与其他经典算法进行对比,不同模型的跨数据集检测结果如表 3 所列。

表 3 不同模型的跨数据集 AUC 结果

Table 3 Cross-dataset AUC results of different models

(单位:%)

Methods	FF++(HQ)	Celeb-DF
Two-stream[3]	70.10	53.80
Meso4[25]	84.70	54.80
DSP-FWA[26]	93.00	64.60
Capsule[4]	96.60	57.50
Two Branch[23]	93.18	73.41
SMIL[27]	96.80	56.30
SPSL[13]	96.91	**76.88**
MaDD[24]	99.29	67.44
Ours	99.20	**83.28**

从实验结果可以看出,现有的检测方法在应用于跨数据集检测时,性能都有显著的下降。跨数据集检测效果最好的是 SPSL 和本文方法,这是因为这两种方法都采用了频域特征,证明了该特征在跨数据集迁移方面的有效性;另外,本文方法的 *AUC* 比 SPSL 的 *AUC* 高出了 6.4%,说明了本文基于多尺度 Transformer 提出的空域和频域特征融合架构的合理性。

3.4.3 消融实验

如前所述,本文的主要贡献主要有两部分:(1)基于相位谱引入了频域特征,对传统空域特征进行了补充;(2)引入了多尺度 Transformer,和 EfficientNet 联合设计了人脸伪造特征提取网络。为更好地分析这两处改进的效果,本节设计了消融实验,以 EfficientNet-B4 网络和 Vision Transformer 网络组成的检测模型作为基线模型,记为 EfficientNet-ViT。其中,EfficientNet-B4 网络的初始化参数是基于 ImageNet 预训练好的,Vision Transformer 网络使用的是与 Transformer 相同的组件,只对输入进行轻微修改,然后基于公开的深度伪造的数据集进行训练。在 EfficientNet-ViT 的基础上增加频域特征提取模块记为 EfficientNet-ViT-P 模型;在 EfficientNet-ViT 的基础上增加多尺度 Transformer 结构记为 EfficientNet-MViT 模型;在 EfficientNet-ViT 的基础上融合上述两处改进,即本文提出的模型记为 EfficientNet-MViT-P。分别在 FF++数据集 HQ 版本上进行训练及测试,对比分析本文两处改进对模型检测性能的影响,实验结果如表 4 所列。

表 4 消融实验 AUC 结果

Table 4 Ablation experiment AUC results

模型	频域特征	多尺度 Transformer	FF++(HQ)上的 *AUC*/%
EfficientNet-ViT	—	—	98.15
EfficientNet-ViT-P	√	—	98.61
EfficientNet-MViT	—	√	98.83
EfficientNet-MViT-P	√	√	99.20

从实验结果可以看出,采用频域特征和多尺度 Transformer 网络对模型检测性能的提升均有贡献,其中多尺度 Transformer 网络对模型检测性能的提升更加明显。综合采用两种改进后,本文提出的模型的 *AUC* 值比改进前的基线模型的 *AUC* 值提高了 1.05%。证明本文提出的检测方法是合理、有效的。

此外,图 4 给出了上述 4 个模型在 FF++(HQ)数据集上的 Loss 曲线对比。其中,横坐标为训练轮数(Epoch),纵坐标为损失值(Loss)。由实验结果可见,随着训练轮数增加,4 个模型均具有较好的收敛性。在添加频域特征提取模块与多尺度 Transformer 结构后,能加快模型的收敛,提升模型的检测性能。

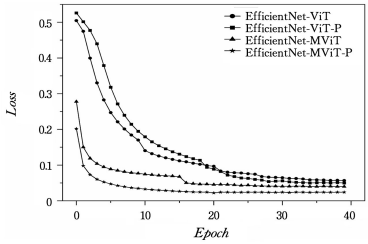

图 4 4 种模型在 FF++数据集上的 Loss 曲线对比

Fig. 4 Comparison of Loss curves of four models on FF++ dataset

3.5 本文方法可视化分析

3.5.1 决策区域可视化分析

为了分析本文提出的方法在鉴别伪造人脸时是否关注了人脸图像的不同伪造区域并提取到有效的面部特征,本节使用 Grad-CAM[28](Gradient-weighted Class Activation Mapping)算法对 FF++数据集中多种伪造方法生成的样本进行可视化分析。

对于 FF++的每种伪造方法,分别使用 3.4.3 节中训练好的基线模型 EfficientNet-ViT 和本文提出的模型通过 Grad-CAM 算法生成热力图,结果如图 5 所示,其中第二行是基线模型在不同伪造样例上的可视化结果,第三行是本文模型的结果。

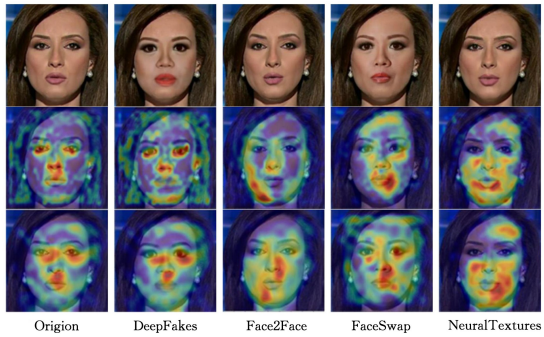

图 5 基线模型与本文模型在 FF++数据集上的热力图对比

Fig. 5 Heat map comparison between baseline and the proposed model on FF++ dataset

由可视化结果可以看出，本文提出的模型可以激活更多的面部伪造区域并且激活区域更加准确，这有助于学习到更具有鉴别性的特征，同时也可以在一定程度上解释为何本文方法在 FF++数据集上表现出较好的分类性能。

3.5.2　模型输出可视化分析

为了进一步分析本文方法能否有效区分视频中的真伪人脸，使用 T-SNE[29]（T-distributed Stochastic Neighborhood Embedding）算法将模型最后一个全连接层输出的 1 792 维度的人脸图像特征数据降至 2 个维度，进行可视化分析。

本节对本文模型在 FF++数据集 HQ(c23)版本测试集上的测试数据进行 T-SNE 可视化，结果如图 6 所示。图中蓝色圆形表示真实人脸降维后的特征在空间的分布，红色圆形表示伪造人脸降维后的特征在空间的分布。由可视化结果可以看出，大部分原始真实视频中的人脸特征集中分布在左边区域，伪造视频中的人脸特征则分布在右边区域，可见本文提出的基于多尺度 Transformer 融合多域信息的伪造人脸检测方法能够有效检测视频中的伪造人脸。

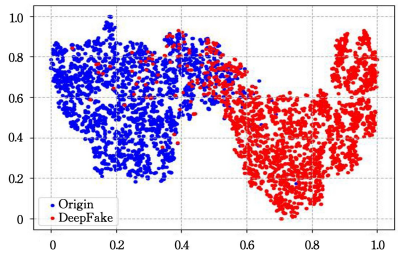

图 6　模型输出的 T-SNE 可视化结果（电子版为彩图）

Fig. 6　T-SNE visualization results of model output

结束语　为了提高视频中伪造人脸检测的精度及模型的泛化性能，本文提出了一种基于多尺度 Transformer 融合多域信息的伪造人脸检测方法，通过在 RGB 空域信息的基础上融合基于相位谱的频域特征提高了检测的泛化性，同时基于 EfficientNet 和多尺度 Transformer 设计了合理的伪造特征提取网络，充分发挥了 Transformer 和 CNN 的联合优势。多个数据集上的实验结果表明本文提出的模型对视频中伪造人脸具有较好的检测性能和良好的泛化性。后续工作将在现有工作基础上继续探索更加通用的检测模型以检测现实世界中未知伪造方式的视频，进一步提高检测模型的泛化性。

参 考 文 献

[1] LI X R, JI S L, WU C M, et al. Survey on deepfakes and detection techniques[J]. Journal of Software, 2021, 32(2): 496-518.

[2] Big Data Digest. Deepfake 'involved in the war' for the first time: The Ukrainian president was faked to surrender the video, and the rumor was dispelled on Twitter[EB/OL]. https://www.thepaper.cn/newsDetail_forward_17262083.

[3] ZHOU P, HAN X, MORARIU V I, et al. Two-Stream Neural Networks for Tampered Face Detection[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2017: 1831-1839.

[4] NGUYEN H, FANG F, YAMAGISHI J, et al. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos[C]// Proceedings of the 2019 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway: IEEE Press, 2019: 2307-2311.

[5] HSU C C, ZHUANG Y X, LEE C Y. Deep Fake Image Detection Based on Pairwise Learning[J/OL]. Applied Sciences, 2020, 10(1): 370. http://doi.org/10.33901app10010370.

[6] TARIQ S, LEE S, KIM H, et al. Detecting Both Machine and Human Created Fake Face Images In the Wild[C]// Proceedings of the 2nd International Workshop on Multimedia Privacy and Security. Canada: CCS Press, 2018: 81-87.

[7] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. 2019: 2978-2988.

[8] HAN K, WANG Y, CHEN H, et al. A Survey on Visual Transformer[J]. arXiv. 2012. 12556, 2020.

[9] WODAJO D, ATNAFU S. Deepfake Video Detection Using Convolutional Vision Transformer[J]. arXiv: 2102. 11126, 2021.

[10] COCCOMINI D A, MESSINA N, GENNARO C, et al. Combining EfficientNet and Vision Transformers for Video Deepfake Detection[C]// Proceedings of the 21st International Conference on Image Analysis and Processing. Cham: Springer, 2022: 219-229.

[11] HEO Y J, CHOI Y J, LEE Y W, et al. Deepfake Detection Scheme Based on Vision Transformer and Distillation[J]. arXiv: 2104. 01353, 2021.

[12] WANG J, WU Z, CHEN J, et al. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection[C]// Proceedings of the 2022 International Conference on Multimedia Retrieval. New York: ACM Press, 2022: 615-623.

[13] LIU H, LI X, ZHOU W, et al. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 772-781.

[14] TAN M, LE Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]// Proceedings of the 2019 International Conference on Machine Learning. Piscataway: IEEE Press, 2019: 6105-6114.

[15] PU Y, GAN Z, HENAO R, et al. Variational Autoencoder for Deep Learning of Images, Labels and Captions[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 2360-2368.

[16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[17] ROSSLER 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. New York: IEEE Press, 2019: 1-11.

[18] LI Y, YANG X, SUN P, et al. Celeb-DF: A Large-scale Challenging Dataset for Deepfake Forensics[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern

Recognition. New York: IEEE Press, 2020: 3204-3213.
[19] DENG J, GUO J, ZHOU Y, et al. RetinaFace: Single-stage Dense Face Localisation in the Wild[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 5202-5211.
[20] LI L, BAO J, ZHANG T, et al. Face X-Ray for More General Face Forgery Detection[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 5000-5009.
[21] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 1800-1807.
[22] TAN M, LE Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]// Proceedings of the 36th International Conference on Machine Learning. New York: PMLR Press, 2019: 6105-6114.
[23] MASI I, KILLEKAR A, RM MASCAREN, et al. Two Branch Recurrent Network for Isolating Deepfakes in Videos[C]// Proceedings of the 2020 European Conference on Computer Vision. Cham: Springer, 2020: 667-684.
[24] ZHAO H, ZHOU W, CHEN D, et al. Multi-attentional Deepfake Detection[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 2185-2194.
[25] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a Compact Facial Video Forgery Detection Network[C]// Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security. New York: IEEE Press, 2018: 1-7.
[26] LI Y, LYU S. Exposing DeepFake Videos By Detecting Face Warping Artifacts[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. California: CVPR workshop, 2019: 46-52.
[27] LI X, LANG Y, CHEN Y, et al. Sharp Multiple Instance Learning for DeepFake Video Detection[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1864-1872.
[28] SELVARAJU R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. USA: IEEE Press, 2017: 618-626.
[29] MAATEN L V D. Accelerating t-SNE using tree-based algorithms[J]. Journal of Machine Learning Research, 2014, 15(1): 3221-3245.

MA Xin, born in 1997, postgraduate. Her main research interests include deep learning and computer vision.

LI Shaomei, born in 1982, Ph. D, associate professor. Her main research interests include computer vision and so on.

（责任编辑：何杨）