E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)03-0804-16

论文引用格式:Li Y, Bian S, Wang C T and Lu W. 2023. CNN and Transformer-coordinated deepfake detection. Journal of Image and Graphics, 28(03):0804-0819(李颖,边山,王春桃,卢伟. 2023. CNN 结合 Transformer 的深度伪造高效检测. 中国图象图形学报,28(03):0804-0819)[DOI:10.11834/jig.220519]

CNN 结合 Transformer 的深度伪造高效检测

李颖[1,2],边山[1,2*],王春桃[1,2],卢伟[3]

1. 华南农业大学数学与信息学院,广州　510642;2. 广州市智慧农业重点实验室,广州　510642;
3. 中山大学计算机学院,广州　510006

摘　要:**目的**　深度伪造视频检测是目前计算机视觉领域的热点研究问题。卷积神经网络和 Vision Transformer(ViT)都是深度伪造检测模型中的基础结构,二者虽各有优势,但都面临训练和测试阶段耗时较长、跨压缩场景精度显著下降问题。针对这两类模型各自的优缺点,以及不同域特征在检测场景下的适用性,提出了一种高效的 CNN(convolutional neural network)结合 Transformer 的联合模型。**方法**　设计基于 EfficientNet 的空间域特征提取分支及频率域特征提取分支,以丰富单分支的特征表示。之后与 Transformer 的编码器结构、交叉注意力结构进行连接,对全局区域间特征相关性进行建模。针对跨压缩、跨库场景下深度伪造检测模型精度下降问题,设计注意力机制及嵌入方式,结合数据增广策略,提高模型在跨压缩率、跨库场景下的鲁棒性。**结果**　在 FaceForensics ++ 的 4 个数据集上与其他 9 种方法进行跨压缩率的精度比较,在交叉压缩率检测实验中,本文方法对 Deepfake、Face2Face 和 Neural Textures 伪造图像的检测准确率分别达到 90.35%、71.79% 和 80.71%,优于对比算法。在跨数据集的实验中,本文模型同样优于其他方法,并且同设备训练耗时大幅缩减。**结论**　本文提出的联合模型综合了卷积神经网络和 Vision Transformer 的优点,利用了不同域特征的检测特性及注意力机制和数据增强机制,改善了深度伪造检测在跨压缩、跨库检测时的效果,使模型更加准确且高效。

关键词:深度伪造检测;卷积神经网络(CNN);Vision Transformer(ViT);空间域;频率域

CNN and Transformer-coordinated deepfake detection

Li Ying[1,2], Bian Shan[1,2*], Wang Chuntao[1,2], Lu Wei[3]

1. College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China;
2. Guangzhou Key Laboratory of Intelligent Agriculture, Guangzhou 510642, China;
3. School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

Abstract: **Objective** The research of deepfake detection methods has become one of the hot topics recently to counter deepfake videos. Its purpose is to identify fake videos synthesized by deep forgery technology on social networks, such as WeChat, Instagram and TikTok. Forged features are extracted on the basis of a convolutional neural network (CNN) and the final classification score is determined in terms of the features-forged classifier. When facing the deep forged video with low quality or high compression, these methods improve the detection performance by extracting deeper spatial domain informa-

收稿日期:2022-06-10;**修回日期**:2022-10-27;**预印本日期**:2022-11-03

*通信作者:边山　bianshan@scau.edu.cn

基金项目:国家自然科学基金项目(62172165,61872152,U2001202,62072480);广东省基础与应用基础研究重大项目(2019B030302008);广州市科技计划项目(202102020582,201902010081)

Supported by: National Natural Science Foundation of China (62172165, 61872152, U2001202, 62072480); Major Program of Guangdong Basic and Applied Research (2019B030302008); Science and Technology Program of Guangzhou (202102020582, 201902010081)

tion. However, the forged features left in the spatial domain decrease with the compression, and the local features tend to be similar, which degrades the performances severely. This also urges us to retain the frequency domain information of forged image artifacts as one of the clues of forensics, which contains less interference caused by JPEG compression. The CNN-based spatial domain feature extraction method can be conducted to capture facial artifacts via stacking convolution. But, its receptive field is limited, so it is better at modelling local information but ignores the relationship between global pixels. Transformer has its potentials at long-term dependency modelling in relevant to natural language processing and computer vision tasks, therefore it is usually employed to model the relationship between pixels of images and make up for the CNN-based deficiency in global information acquisition. However, the transformer can only process sequence information, making it still need the cooperation of convolutional neural network in computer vision tasks. **Method** First, we develop a novel joint detection model, which can leverage the advantages of CNN and transformer, and enriches the feature representation via frequency domain-related information. The EfficientNet-b0 is as the feature extractor. To optimize more forensics features, in the spatial feature extraction stage, the attention module is embedded in the shallow layer and the deep features are multiplied with the activation map obtained by the attention module. In the frequency domain feature extraction stage, to better learn the frequency domain features, we utilize the discrete cosine transform as the frequency domain transform means and an adaptive part is added to the frequency band decomposition. In the training process, to accelerate the memory-efficient training, we adopt the method of mixed precision training. Then, to construct the joint model, we link the feature extraction branches to a modified Transformer structure. The Transformer is used to model inter-region feature correlation using global self-attention feature encoding through an encoder structure. To further realize the information interaction between the dual-domain features, the cross attention is calculated between branches on the basis of the cross-attention structure. Furthermore, we design and implement a random data augmentation strategy, which is coordinated with the attention mechanism to improve the detection accuracy of the model in the scenarios of cross compression rate and cross dataset. **Result** Our joint model is compared to 9 state-of-the-art deepfake detection methods on two datasets called FaceForensics ++ (FF ++) and Celeb-DF. In the experiments of cross compression-rate detection on the FF ++ dataset, our detection accuracy can be reached to 90.35%, 71.79% and 80.71% for Deepfakes, Face2Face and Neural Textures (NT) manipulated images, respectively. In the cross-dataset experiments, i.e., training on FaceForensics ++ and testing on Celeb-DF, our training time is reduced. **Conclusion** The experiments demonstrate that our joint model proposed can improve datasets-crossed and compression-rate acrossed detection accuracy. Our joint model takes advantage of the EfficientNet and the Transformer, and combines the characteristics of different domain features, attention, and data augmentation mechanism, making the model more accurate and efficient.

Key words: deepfake detection; convolutional neural network (CNN); Vision Transformer (ViT); spatial domain; frequency domain

0　引　言

随着基于深度学习的生成技术的发展,深度伪造技术能够生成高度逼真难以辨别的人脸图像。将一个人的脸变换到另一个人的脸上,或是使用一个人的脸部动作驱动另一个人的五官运动,这两类技术分别称为人脸交换和人脸重现(Wang 等,2021),是目前深度伪造应用最广泛的技术。人脸重现能够使画面中的人物说自己从未说过的话,人脸交换则能够使画面中人物主体的脸被替换。由于这两种应用都威胁到公民的名誉权与隐私权及互联网信息安全,研究人员开始寻找有效的方法来对抗日益强大且便利的深度伪造技术。

传统方法更关注图像本身的统计信息和物理特征(Liu 等,2022),现有的大多数方法则都以卷积神经网络为基础,从深度伪造视频中提取人脸图像,用卷积神经网络学习伪造特征并进行真假分类。受限于感受野的大小,基于卷积神经网络的方法提取到的图像伪造特征往往更为局部,难以考虑到整幅图像中全局像素之间的关系(Wang 等,2021)。深度伪造图像中同时存在着包含不同特征的伪造区域和真实区域,这些区域块特征之间的关系对于深度伪造的检测而言有着至关重要的作用,然而卷积神经网络通常难以捕捉到区域间的这种相关性。除此之外,这些基于单一网络的检测方法往往在固定压缩

率的数据集上进行训练并测试。这意味着在某一压缩率数据集上训练得到的模型只能用于该压缩率级别数据的推理,在其他压缩级别的视频检测中将很难得到令人满意的效果。此时,模型检测精度将大幅下降,尤其是其在高质量数据集中学习到了对应的数据分布,而在应用场景中却需要检测低质量图像和视频的情况下。在现实的应用场景中,互联网上的伪造人脸图像和视频的压缩率往往是不固定的,这就要求深度伪造检测模型具备一定的跨压缩率检测能力,即面对压缩应有良好的泛化性能。

同时,深度伪造压缩视频的取证也是一个重要的问题。压缩伪造视频在社交网络上的传播往往更为迅速,这是因为在网络带宽的限制下,用户上传到互联网上的视频往往会经过压缩,体积更小的视频文件将更易形成传播。而压缩视频中,深度伪造伪影往往更难以为人察觉。因此,假若不法分子故意传播经压缩后的低质量深度伪造图像,人们将更难以对其进行分辨。

为了解决以上问题,本文将 Vision Transformer(ViT)与卷积神经网络结合进行深度伪造检测。Vision Transformer 将 Transformer(Vaswani 等,2017)具有的长距离建模能力迁移到计算机视觉领域,在目标检测、图像分类等领域得到了广泛的应用。与现有的利用 Vision Transformer 的深度伪造检测方法不同,本文提出的联合模型将利用 EfficientNet 在 RGB 域和频率域分别提取图像特征,目的是保留卷积神经网络应对深度伪造图像时优秀的局部建模能力与异常挖掘能力,同时利用频率域分支能够提取伪造压缩视频中伪影特征的特性。经过 Transformer 结构中的 Transformer 层编码后执行交叉注意,以对区域块特征之间的关系进行建模,生成类别(classification,CLS)令牌,最后经多层感知器(multi-layer perceptron head,MLP Head)得到分类结果。此外,为了提高模型跨压缩场景下的检测能力,本文在训练阶段结合注意力机制引入了数据增强。主要工作和贡献如下:1)提出了一种卷积神经网络结合 Transformer 的双路深度伪造检测模型,使模型兼具局部提取和全局建模的能力,更全面地学习深度伪造图像人脸空间特征。在 FaceForensics ++(FF ++)和 Celeb-DF 两个不同的数据集上达到了先进的检测性能。2)设计了结合注意力机制的随机数据增强策略,提高模型的跨压缩检测能力,增强模型的泛化能力。3)在 Transformer 结构之前采用卷积神经网络提取双域特征,避免了 Transformer 训练时需要从图像中直接学习自相关性的时间损耗,使训练及推理都变得更加高效。

1 相关工作

1.1 基于卷积神经网络的深度伪造检测技术

深度伪造技术的精进促使研究人员探索更有效的深度伪造检测算法。社交网络上的深度伪造内容泛滥也引起了人们的关注。Fagni 等人(2021)对推特中存在的深度造假现象进行了分析。早期利用传统算法进行深度伪造视频检测的研究更关注与视频成像相关的特征。Li 等人(2020)认为虚假视频中相邻帧的光照方向难以自然吻合,能够作为检测线索。基于机器学习的解决方案往往需要人们对人脸面部可分辨的特征进行提取,这类提取特征的方式受限于视觉层级,即难以提取到像素级别的特征。因此,研究人员利用能够对特征进行自适应提取的卷积神经网络进行深度伪造检测。早期研究大多基于空间域对深度伪造图像进行特征提取。MesoNet(Afchar 等,2018)利用 InceptionNet(Szegedy 等,2015)检测伪造视频。Two-Stream(Zhou 等,2017)则设计双流网络架构,利用两个分支分别捕捉图像中的伪造线索和图像块之间的差异。Li 等人(2020a)提出基于混合边界辨别伪造图像的方法,认为深度伪造检测同时应该关注模型的泛化性能,以对抗不断精进的伪造技术。MADD(multi-attentional deepfake detection)(Zhao 等,2021)则将伪造图像的分类视为细粒度分类任务,因为真实图像与伪造图像之间的差异很小,与细粒度分类所面对的困境相似。MADD 从这一点出发,提出多注意力检测框架,用于捕捉更局部及不易察觉的伪造痕迹。随着研究的不断深入,一些工作尝试利用频率域线索来解决伪造检测中面临的低质量视频难以分类的问题。例如,一些研究(Zhang 等,2019;Wang 等,2020;Durall 等,2019)利用离散傅里叶变换将图像从空间域转换到频域,Durall 等人(2019)将变换后的频域信息中不同频带的振幅取平均,Stuchi 等人(2017)使用频域滤波器提取不同频率范围的信息,最后在全连接层之后得到输出。F3-Net(Qian 等,2020)则结合频率分解与深度学习技术,提出了一种对频率信息自适应分解及局部统计的深度伪造

检测算法。本文假设在不同压缩程度中的频率域特征存在不同，利用卷积神经网络结合空间域和频率域进行双路特征提取及令牌（token）的转换，在训练中引入数据增广策略，从而提升模型在跨压缩场景下的泛化性能。

1.2　基于 Vision Transformer 的深度伪造检测技术

Transformer（Vaswani 等，2017）最初应用在自然语言处理任务中，内部自我关注的结构使 Transformer 擅长对远距离的上下文信息进行建模。在计算机视觉领域，Vision Transformer 作为其变体得到了广泛应用。在 Vision Transformer 中，图像直接分割成图像块并进行投影得到线性嵌入向量。一些研究也将 Vision Transformer 应用在深度伪造检测技术上。DCViT（deepfake video detection using convolutional vision transformer）（Wodajo 和 Atnafu，2021）将卷积神经网络与 ViT 进行结合，在 DFDC（deepFake detection challenge）数据集上取得了具有一定竞争性的结果。CrossViT（Coccomini 等，2022）则设计了具有左右分支的卷积神经网络——Transformer 结构，M2TR（multi-modal multi-scale transformers）（Wang 等，2021）为了捕获不同局部大小中存在的伪影，设计了多尺度的 Transformer 用于集成多尺度信息并进行深度伪造取证。Zhu 等人（2022）将多关键帧的特征信息作为 Transformer 结构的输入，设计了多关键帧能够进行特征交互的卷积 Transformer 结构。

Transformer 在深度伪造检测领域得到应用也得益于其对全局信息强大的建模能力，而随着深度伪造技术的发展，图像中的伪影区域趋于局部甚至微不可察，过去的深度伪造检测技术也根据这一特点在设计上更多地专注于局部特征的捕捉。然而，这往往会使人们忽略真实区域与伪造区域存在差异这一现象，即从全局的角度来看，伪造图像中存在少数区域与多数区域之间的区别特征。本文从这一观察出发，利用卷积神经网络分别在两个分支中提取图像块的空间域特征与频率域特征，并对其进行投影得到线性嵌入向量，利用改进的 Transformer 的自我关注机制及交叉注意力计算对全局信息进行建模，实现深度伪造检测，并提升模型在跨压缩、跨库检测场景下的性能。

2　本文方法

本文通过对双域全局信息建模来捕捉伪造图像中区域间的不一致性。为了利用卷积神经网络良好的特征提取能力，同时缩减自注意力机制在训练时的大量时间损耗，采用卷积神经网络提取空间域及频率域特征并交由 Transformer 结构捕捉特征全局关联性的方式，使模型在训练及测试阶段都更为高效。而且，本文方法设计注意力机制与随机数据增强策略并进行结合，以提高模型在压缩、跨压缩及跨库场景下的准确率。

本文模型的基本结构如图 1 所示。由空间域特征提取分支（RGB branch）、频率域特征提取分支（freq branch）及 Transformer 结构组成。

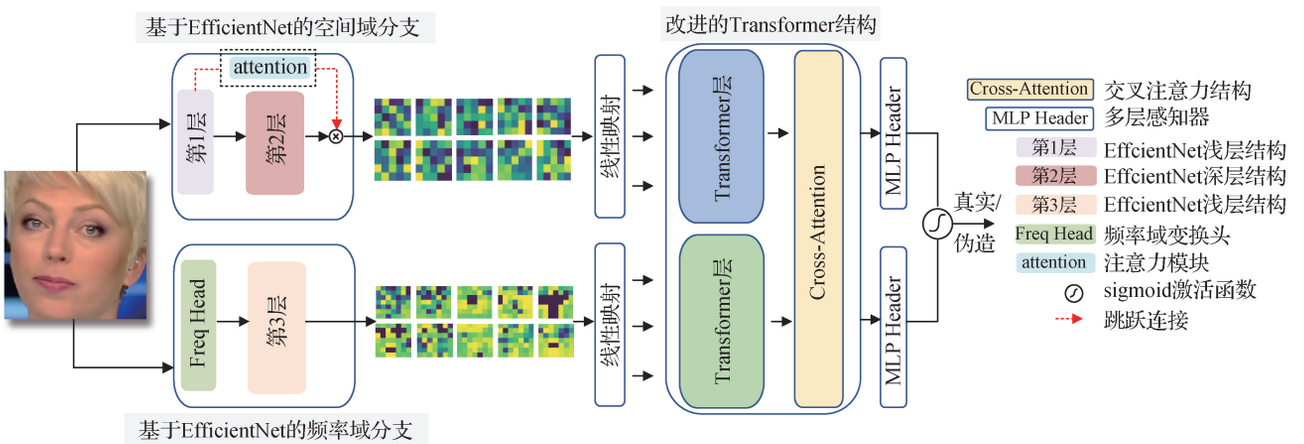


图 1　联合模型检测框架

Fig. 1　Joint model detection framework

2.1 基于 EfficientNet 的空间域分支

由于本文模型为双分支结构,为了提升训练效率,使用 EfficientNet 系列(Tan 等,2019)中最为轻量化的基础网络 EfficientNet-B0 进行特征提取,以使得模型在双分支的情况下训练和测试两个阶段都更为高效,并在空间域特征提取分支中嵌入注意力模块对其进行改进,同时结合注意力模块进行数据增强。

原 EfficientNet-B0 模型在 MBConv(mobile inverted bottleneck conv)内部即深度卷积与点卷积之间添加了 SE(squeeze excitation)模块,目的是为了学习通道之间的相关性,形成了面向通道的注意力机制。然而这将无法学习到像素之间空间上的关联性,对于伪造图像而言,即难以学到区域间(真实区域与伪造区域)的相关性。因此本文从嵌入注意力模块到空间域分支和结合注意力信息进行数据增强两个方面对基于 EfficientNet-B0 的空间域分支进行改进。具体来说,在空间域分支中引入了注意力模块,并将其嵌入到浅层特征提取之后,以在提取人脸区域空间层次的注意力,辅助数据增强策略进行数据增广。

2.1.1 注意力模块的嵌入

注意力模块如图 2 所示,参考 MADD(multi-attentional deepfake detection)(Zhao 等,2021),为 3 × 3 卷积与 1 × 1 卷积交替使用的异构结构,并使用 swish 激活函数替换常用的 ReLU 激活函数。与 ReLU 相比,swish 更加平滑,有助于模块优化。此外,3 × 3 卷积和 1 × 1 卷积交替的异构结构能够保证在注意力机制生效的同时减少参数量、计算成本及内存消耗。

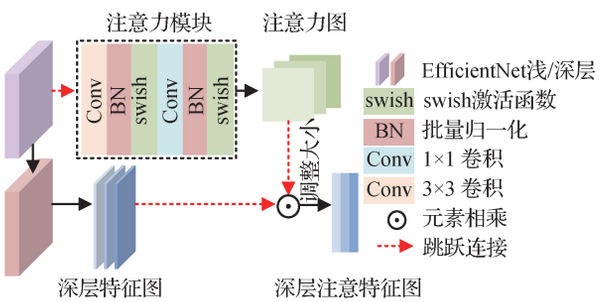


图 2 注意力模块

Fig. 2 Attention module

由于深度伪造检测任务不同于常见的图像分类任务,相比于需要更多语义信息的常见图像分类任务,深度伪造检测更需要保留伪造伪影或边缘的浅层特征信息。本文选择将注意力模块嵌入到 EfficientNetB0 中的浅层特征提取阶段。此外,为了保持注意力机制对后续的深层特征的影响,保留深层特征对浅层特征的记忆,采用跳跃连接的方式融合注意力图与深层特征,即在浅层实现注意力机制生成注意力图之后,与深层特征进行矩阵相乘,最终获得加权后的深层注意力特征。

伪造区域与真实区域往往存在不一致性,且该不一致性分布在多个区域,因此注意力模块只对固定区域做出反应将无法捕捉到全面的伪影信息。此外,卷积神经网络的特征提取过程中,越深层的特征图将包含越多与伪造取证关联较弱的语义信息。为此,本文的注意力模块将对 EfficientNet-B0 的浅层特征生成 k 幅注意力图,k 的数目将经过实验确定。FF ++(FaceForensics ++)全集实验中真假人脸图像及对应注意力图如图 3 所示,O 代表原始图像,DF、F2F、FS 及 NT 分别代表 FaceForensics ++ 中的 4 种篡改类型,每行为对应图像的注意力图。可以看出,注意力区域分布在图像中的不同区域,而同列的注意力区域则趋于相似。跳跃连接的注意力机制将利用浅层特征信息,在空间域分支对人脸图像生成分布更广泛的注意力图。生成的注意力图将指导深层特征的提取,同时与数据增广策略进行结合,以提升模型在跨压缩场景的泛化性能。

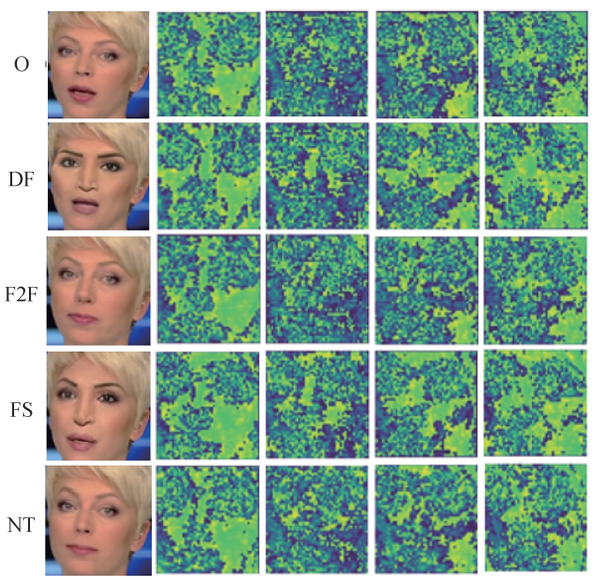


图 3 FF ++ 全集实验中真假人脸图像及对应注意力图

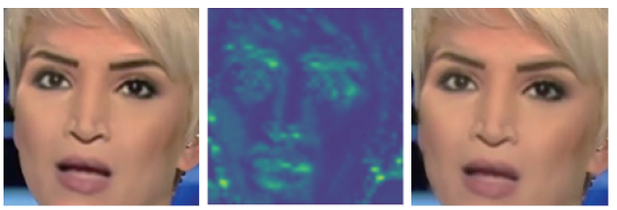
Fig. 3 Real and fake face images and the corresponding attention maps in the experiment on FaceForensics ++

2.1.2　随机数据增广策略

跨压缩率检测的难点在于高质量图像训练得到的模型在压缩伪造图像上的检测精度大幅下降。这是因为压缩与传输过程将不可避免地导致像素信息的损失，表现为图像质量降低，这也将使得深度伪造检测所需的伪影信息在图像中减弱甚至消失。因此，本文受 MADD（Zhao 等，2021）启发，设计数据增广策略，对高质量训练人脸图像进行注意力引导的区域模糊，在高质量图像检测模型的训练过程中，人为引入区域像素的损失，以提升模型在跨压缩率场景下的检测性能。

注意力图结合数据增广策略生成的人脸图像如图 4 所示，首先应用高斯模糊或均值模糊对人脸图像 $\boldsymbol{I}$ 进行随机噪声因子的全脸模糊，定义为 $\boldsymbol{Blur}$，具体为

$$\boldsymbol{I}_d = \boldsymbol{Blur}(\boldsymbol{I}) \times \boldsymbol{A}_k + \boldsymbol{I} \times (\boldsymbol{1} - \boldsymbol{A}_k) \quad (1)$$



(a) 原始图像　(b) 插值后的注意力图　(c) 退化后的增强副本

图 4　注意力图及对应增强图像

Fig. 4　Attention map and the corresponding enhanced image ((a) original image; (b) attension map after bilinear interpolation; (c) sample after degradation)

将模糊后的全脸图像 $\boldsymbol{Blur}(\boldsymbol{I})$ 与随机选择的注意力图 $\boldsymbol{A}_k$ 相乘，得到注意力区域的模糊图像。将原图 $\boldsymbol{I}$ 与注意力图以外的区域部分即 $\boldsymbol{1} - \boldsymbol{A}_k$ 相乘，得到未经模糊处理的原图区域图像。原图区域图像与只包含注意力区域的退化图像进行相加，得到数据增广后的图像，记为 $\boldsymbol{I}_d$。

由于训练数据缺乏注意力的区域标签，因此在训练过程中，注意力机制会趋向对同一个区域作出反应（Zhao 等，2021），同时向整幅图像扩散，最终 k 幅注意力图相加甚至单幅注意力图的区域为整幅图像。因此在训练阶段，数据增广策略需要随机选择一个注意力区域并进行模糊，在达到增强模型对于低质量模型的泛化性能的同时，注意力区域不会进行无效的扩散。

注意力机制在浅层特征提取阶段生成多幅注意力图，并在整个训练阶段对训练集中的样本随机选择注意区域与模糊方式生成增广后的模糊图像。在深层特征提取阶段，注意力图将通过双线性插值调整到与深层特征图一样的大小，并与深层特征图进行相乘，最终得到注意力机制指导下的深层注意力特征，与频率域分支得到的特征图一并送入到 Transformer 结构中进行编码，得到最终的分类结果。

2.2　基于 EfficientNet 的频率域分支

本文模型将 EfficientNet-B0 作为频率域特征提取分支的骨干网络，并在骨干网络之前插入频率域头以将空间域信息转换到频率域。

低质量视频的检测效果不佳一直是深度伪造检测技术期望解决的问题之一。伪造图像生成过程中通常未考虑频域信息的分布，因此生成器通常无法还原与真实图像类似的频域信息（He 等，2022）。而有研究证明频率域伪影在低质量视频中依然能够检测到，因此在许多深度伪造检测方法中频率域线索也作为取证的重要信息（Qian 等，2020；Zhang 等，2019）。本文采取基于离散余弦变换的频率域头连接骨干网络的方式，设计频率域特征提取分支对图像中的频率域特征进行提取，为后续 Transformer 的编码做准备。

频率域变换头的设计与 F3-Net（frequency in face forgery network）（Qian 等，2020）类似，如图 5 所示，$\boldsymbol{I}$ 即人脸图像。本文设计 3 个组合滤波器对经离散余弦变换后的输入图像进行频带分解。

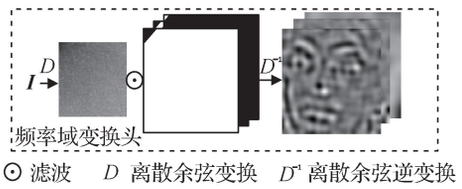


图 5　频率域变换头

Fig. 5　Frequency head

组合滤波器都由可学习的部分与基本部分组成。基本部分分别为低频带滤波器、中频带滤波器及高频带滤波器。低频带为整个频谱的前 1/16，中频带为整个频谱的 1/16 ～ 1/8，高频带为整个频谱剩余的部分。可学习的部分 f_l 与基本滤波器 f_b 结合之后，将对输入图像进行自适应的频域划分。具体为

$$y_i = DCT^{-1}[(f_b^i + \sigma(f_l^i)) \odot DCT(x)] \quad (2)$$

式中，i 表示第 i 个频带，分解得到的图像分量通过

离散余弦逆变换回到图像表示,以适应卷积神经网络的输入宽度。$\sigma(x)=\frac{1-e^{-x}}{1+e^{-x}}$,目的是将 x 限制在(-1,1)中。频率域头得到的图像表示将输入到 EfficientNet-B0 中提取频率域的特征,即频率域表示。通过骨干网络得到深层特征图,并与空间域特征图一同送入 Transformer 结构。

2.3 Transformer 结构

本文采用卷积神经网络与 Transformer 结构结合的方式进行联合模型的设计,在双域分支提取得到相应特征后送入 Transformer 结构中。Transformer 结构中包含了 Transformer 的编码器结构(即 Transformer 层)与交叉注意力结构,以实现分支内特征的自注意力机制与分支间特征的交叉注意力计算,形成双域特征信息的进一步提取与交互。

在原始的 Vision Transformer 中,输入图像将进行分块,展平成序列后,输入到原始 Transformer 结构的 Transformer 层,最终送入全连接层中对输入图像输出分类结果。为了计算区域特征之间的相关性,对空间域分支及频率域分支得到的特征图而非图像进行划分,结合其他研究(Chen 等,2021;Coccomini 等,2022)中的 Transformer 编码器结构进行改进,提出多头特征注意力,由多个特征注意力头组成。

2.3.1 多头特征注意力

多头特征注意力如图 6 所示。每个特征块为 7×7 或 5×5,同时对每个特征块进行位置信息的嵌入,线性投影成特征块序列后,采用 1×1 卷积代替原始 Transformer 编码器结构中的全连接层,分别得到查询嵌入 $\boldsymbol{q}$、键嵌入 $\boldsymbol{k}$ 与值嵌入 $\boldsymbol{v}$,在不破坏基本信息的前提下减少通道,更加方便计算。通过查询嵌入 $\boldsymbol{q}$ 和键嵌入 $\boldsymbol{k}$,计算注意力 $a_{i,j}$。即

$$a_{i,j}=\mathrm{softmax}\left(\frac{\boldsymbol{q}_i\cdot\boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{r\times r\times C}}\right),1\leqslant i,j\leqslant N \tag{3}$$

式中,r 表示特征块序列中 patch 的大小,C 为通道数量,N 即每一行或每一列包含的特征块的数量。对特征块的值进行加权求和,得到查询块对应的输出。在多个头对所有特征块采用相同的计算得到输出之后,将这些输出重新组合在一起,并整形为输入的分辨率。整形后的输出与输入建立一个残差连接,并送入到前向传播部分。

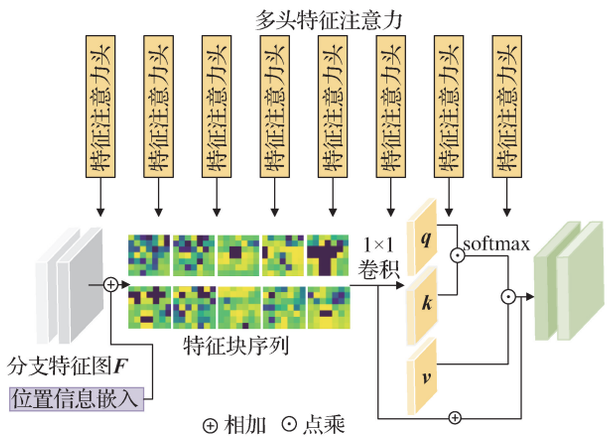


图 6 多头特征注意力

Fig. 6 Multi-head feature attention

前向传播部分由一个前馈神经网络组成。在结束前向传播之后也将与输入建立一个残差连接,随后进行归一化。

2.3.2 分支交叉特征注意力

两个分支的特征图分别通过两个编码器结构后将经过交叉注意力结构计算交叉注意力。对两个分支特征图交叉注意力的计算是为了更好地融合来自两个不同域的特征。而两个编码器结构都将输出 CLS(classification)令牌,分别包含了两个分支学习到的重要信息。利用分支 Transformer 编码器结构输出的 CLS 令牌,与另一个分支的特征块令牌进行交互。

本文将编码器与交叉注意力结构相互交替重复 4 次,形成堆叠结构。交叉注意力模块中输出的 CLS 令牌将会在下一次的 Transformer 的编码器结构中再次与本分支的特征块令牌交互,即完成了其他域特征信息到本域信息的传递,增强了分支中的特征表示。

分支交叉特征注意力如图 7 所示。对于空间域分支 s,从频率域分支 f 中取得特征块令牌,使用线性投影将两者维度对齐后,将本分支的 CLS 令牌与其连接。采用与自注意力计算相类似的方式计算交叉注意力(Chen 等,2021;Coccomini 等,2022)。图 7 中,$\boldsymbol{x}_{\mathrm{cls}}^{s}$ 为空间域分支取得的 CLS 令牌,$\boldsymbol{x}'^{s}$ 由 $\boldsymbol{x}_{\mathrm{cls}}'^{s}$ 和 $\boldsymbol{x}_{\mathrm{patch}}^{f}$ 拼接而成,用于键嵌入与值嵌入,由于其融合了 patch 令牌中的信息,因此不嵌入到查询嵌入 $\boldsymbol{q}$ 中。在计算时,$\boldsymbol{x}_{\mathrm{cls}}^{s}$ 需通过投影对齐调整到与 $\boldsymbol{x}_{\mathrm{patch}}^{f}$ 相同的尺寸,$\boldsymbol{y}_{\mathrm{cls}}'^{s}$ 需调整到与 $\boldsymbol{x}_{\mathrm{patch}}^{s}$ 相同的尺寸。

交叉注意力计算阶段与 Transformer 的编码器

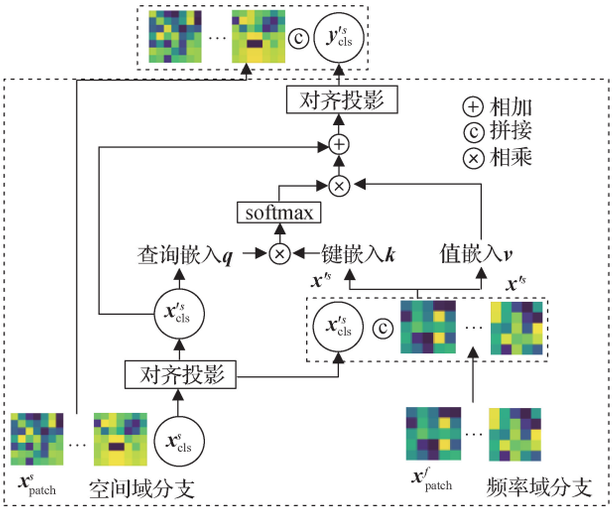


图 7　分支交叉特征注意力

Fig. 7　Cross-feature attention of branches

中有所不同的是不再执行前向传播操作，与 Transformer 的编码器相类似的是同样采用了多头设计，目的是将区域特征之间及双域特征之间的关系映射到多个不同的子空间，增强模型对于区域特征和双域特征之间的相关性表达。

堆叠结构之后两个分支都将输出最终的 CLS 令牌，分别通过多层感知器得到结果，两个分支的结果相加即得最终的分类结果。

3　实验结果与分析

为验证评估模型的性能，在 FaceForensics ++（Rössler 等，2019）的 4 个数据集及 Celeb-DF（Li，2020b）数据集上进行实验。

3.1　参数设置及实验环境

对于人脸图像的获取，使用 Blaze 模型（Bazarevsky 等，2019）提取输入图像中的人脸区域，并将人脸图像尺寸调整为 512 × 512 像素。两个分支中的骨干网络采用在 ImageNet 数据集上经过预训练的 EfficientNet-B0 模型。学习率设置为 0.000 01，学习率每步衰减 1 次，批次设置为 32、40 个 epoch 完成模型训练。实验配置及环境如表 1 所示。

3.2　数据集介绍

FaceForensics ++（Rössler 等，2019）数据集包含 4 个伪造数据集，使用了不同的伪造方法，即 Deepfake、Face2Face、NeuralTextures 和 FaceSwap。所有的视频采用 H. 264 编码方式进行了不同程度的压缩，因此根据压缩率分为 3 个版本，分别为原始版本、C23 版本和 C40 版本。FaceForensics ++ 数据集中的视频来自 Youtube，每个压缩版本包含 1 000 个真实视频，4 000 个伪造视频，即每种伪造方式每个版本分别包括 1 000 个真实视频与 1 000 个伪造视频。本文与其他方法保持了一致的数据集划分，按照 720∶140∶140 的比例将数据集分为了训练集、验证集及测试集，在 4 种伪造方法、2 种压缩版本的 FaceForensics ++ 数据集上进行了实验，并评估了模型在 FaceForensics ++ 数据集上跨压缩率及库内场景下的性能。对于 FaceForensics ++ 数据集，本文对每个视频抽取 270 帧图像，与其他研究保持一致。在训练过程中通过重复采样的方式，平衡真实样本与伪造样本的数量。

表 1　实验配置及环境

Table 1　Experimental configuration and environment

类别	配置
电脑类型	台式电脑
显卡	Nvidia GeForce RTX 2080Ti
CPU	Intel Core i9-9900K
内存大小	32 GB
操作系统	Ubuntu 18.04 LTS
深度学习框架	Pytorch
CUDA 版本	CUDA 10.1
cuDNN 版本	Cudnn 7.6.03
编程语言	Python 3.6.9

Celeb-DF（Li 和 Lyu，2019）数据集由 890 个真实视频与 5 639 个 Deepfake 视频组成，测试集包括 518 个视频。本文在 Celeb-DF 数据集上进行了跨库场景下的算法性能测试。对于 Celeb-DF 数据集，本文同样对每个视频抽取 270 帧图像。

3.3　评价标准

使用准确率（accuracy，ACC）评价模型分类的精度，使用 ROC（receiver operating characteristic curve）曲线下面积（area under curve，AUC）评价模型的分类性能。ACC 及 AUC 均为分类任务中常见的评估标准，现有的深度伪造检测研究大多采用这两项评价指标。

3.4　FaceForensice ++ 数据集上跨压缩率及库内实验结果

FaceForensics ++ 是深度伪造检测方法中广泛

使用的评估数据集。因此,本文将联合模型与目前最新的深度伪造检测技术进行对比。实验比较了提出的联合模型与其他方法在 FaceForensics ++ 的 4 个伪造方法的数据集上跨压缩率场景下的检测性能,即在 C23 高质量数据集上进行训练,在 C40 低质量数据集上进行测试。C40 视频在编码压缩的过程中丢失了大量的纹理细节,因此现有方法在高质量数据集上进行训练后,在低质量数据,如 C40 数据集中的视频上进行测试时将会有明显的精度下降。跨压缩率检测是现实中较为常见的情景之一。

与对比方法的准确率对比如表 2 所示,部分数据取自 Hu 等人(2022)的研究。Deepfake、FaceSwap、Face2Face 和 NeuralTextures 分别为 FaceForensics ++ 数据集中的 4 种伪造类型。C23-C40 表示在高质量数据集上进行训练、在低质量数据集中进行测试。由表 2 可知,本文提出的联合模型在 Deepfake 及 NeuralTextures 这两类伪造视频上都达到了最先进的跨压缩率检测精度。

表 2　FaceForensics ++ 数据集上跨压缩实验结果

Table 2　Experimental results of cross compression on Faceforensics ++ dataset

/%

模型	精确率(C23-C40)			
	Deepfake	FaceSwap	Face2Face	NeuralTextures
Capsule(Nguyen 等,2019)	67.75	54.50	55.75	53.75
MesoNet(Afchar 等,2018)	78.75	59.75	69.75	53.25
CNN-RNN(Güera 和 Delp,2018)	71.93	52.10	50.95	51.76
AR-3D-CNN(Carreira 和 Zisserman,2017)	57.04	50.00	56.82	60.57
3D-CNN(Tran 等,2015)	79.55	54.52	69.10	59.80
Xception(Rössler 等,2019)	63.96	55.63	50.83	55.84
Siamese-ResNet-50(Zhang 等,2020)	82.49	–	71.70	–
Siamese-ResNet-101(Zhang 等,2020)	83.18	–	71.61	–
Siamese-ResNet-152(Zhang 等,2020)	84.99	–	68.93	–
CrossVitEfficientNet(Coccomini 等,2022)	63.81	67.14	50.35	66.40
MADD(Zhao 等,2021)	88.07	**83.57**	62.14	52.85
本文	**90.35**	81.78	**71.79**	**80.71**

注:加粗字体表示各列最优结果,“–”表示无对应实验数据。

与结构相似的 CrossVitEfficientNet(Coccomini 等,2022)相比,本文的联合模型特征来源更加丰富,在 4 种伪造手段上的精确度分别提高了 24.48%、6.78%、21.08% 和 11.81%。MADD(Zhao 等,2021)采用了注意力指导的数据增广策略,该方法中纹理增强模块的设计使其更适合视觉伪影明显的深度伪造视频的库内检测,而 Deepfake 和 FaceSwap 伪造图像无论是高压缩还是低压缩,视觉伪影在空间域都要较其他伪造算法明显,因此只利用了空间域特征的 MADD 在这两类视频上的检测性能较为优秀。而本文模型在检测难度更高(即视觉伪影更为微弱)的 Face2Face 和 NeuralTextures 伪造视频上的检测精度均达到最先进的性能,在 NeuralTextures 伪造类型的视频上,与采用相似增广策略的 MADD 相比,提高了 25.36%。在 4 类伪造视频上的检测性能较为接近,也说明相比于其他模型,本文模型选择的伪造特征更为广泛并具有普遍性,因此更适用于跨压缩率及跨库场景下的检测。

双域分支为了高效提取用于相关性建模的浅层特征,采用了不完整的特征提取结构,如空间域分支只利用了 EfficienNet-B0 的前 15 层 MBConv 结构,频率域分支只利用了 EfficientNet-B0 的前 8 层 MBConv 结构。因此相比于使用了完整结构的 F3-Net、MADD 以及 M2TR,联合模型存在更深层的特征提取不够充分的情况,这导致联合模型在库内更深层

的数据分布的学习上存在一定困难。

表 3 为在 FaceForensics ++ 数据集库内与其他方法对比的准确率及 AUC。其中，LQ（low-quality）是低质量数据集；HQ（high-quality）是高质量数据集。尽管本文方法的主要目的是提高在跨压缩率上的可转移性和跨库场景下的泛化性能，但本文提出的联合模型在高质量数据集上依然达到了与其他方法接近的检测精度与分类能力。此外，在两个基于 EfficientNet-B0 的特征提取分支上，只采用不完整结构以及 Transformer 堆叠结构数量选择为 4 的情况下，结合 Pytorch 的混合精度训练策略，本文大幅缩短了模型训练所需的时间，使模型在深度伪造检测中更加高效，对比结果如表 4 和表 5 所示。可以看出，与性能接近的 MADD 方法相比，本文模型的训练和推理耗时分别缩减了 82% 和 7%，验证了模型的高效性。

表 3　不同模型在 FaceForensics ++ 数据集库内的实验结果

Table 3　Experimental results of different models on Faceforensics ++ dataset

/%

模型	LQ		HQ	
	精确率	AUC	精确率	AUC
Steg. Features（Fridrich 和 Kodovsky，2012）	55.98	–	70.97	–
LD-CNN（Cozzolino 等，2017）	58.69	–	78.45	–
MesoNet（Afchar 等，2018）	70.47	–	83.10	
Face X-ray（Li 等，2020a）	–	61.60	–	87.40
F3-Net（Qian 等，2020）	90.43	93.30	97.52	98.10
MADD（Zhao 等，2021）	88.69	90.40	97.60	99.29
M2TR（Wang 等，2021）	**92.35**	**94.22**	**98.23**	**99.48**
本文	89.71	89.27	97.57	99.17

注：加粗字体表示各列最优结果，“–”表示无对应实验数据。

表 4　不同方法在 FaceForensics ++ 中 Deepfake 篡改类型视频测试集上的消耗时间对比

Table 4　Comparison of consumption time among different methods on Deepface sub-dataset of FaceForensics ++ test dataset

模型	模型训练 1 轮耗时/s	模型测试耗时/ms
CrossVitEfficientNet（Coccomini 等，2022）	1 253.11	4.41
MADD（Zhao 等，2021）	6 802.81	4.42
Siamese-Inception-V1（Zhang 等，2020）	1 938.86	**3.52**
Siamese-ResNet-50（Zhang 等，2020）	2 895.00	4.06
本文	**1 240.47**	4.10

注：加粗字体表示各列最优结果。

表 5　不同方法在 FaceForensics ++ 中 Face2Face 篡改类型视频测试集上的消耗时间对比

Table 5　Comparison of consumption time among different methods on Face2Face sub-dataset of FaceForensics ++ test dataset

模型	模型训练 1 轮耗时/s	模型测试耗时/ms
CrossVitEfficientNet（Coccomini 等，2022）	1 231.05	4.36
MADD（Zhao 等，2021）	6 756.11	3.94
Siamese-Inception-V1（Zhang 等，2020）	1 992.70	**3.70**
Siamese-ResNet-50（Zhang 等，2020）	2 734.70	4.16
本文	**1 229.04**	4.25

注：加粗字体表示各列最优结果。

3.5 Celeb-DF 数据集上跨库实验结果

Celeb-DF(Li 和 Lyu,2019)数据集中的伪造视频来源于 Deepfake 伪造算法,视频的伪造质量相比 FaceForensics ++ 要更高,因此在大多数研究中常作为跨库检测的通用数据集。目前基于深度学习的算法多由数据驱动,即大多数深度伪造模型在推理时依据的是训练阶段从训练集中学习到的特征。在面对跨库场景时,训练后的模型将不再有库内检测时同等的辨别显著性。而现实中用于生成 Deepfake 视频的算法有很多,导致 Deepfake 视频中包含的特征各不相同,因此训练一个通用的具有泛化能力的模型用于应对跨库检测需求具有重要意义。

本文模型在结构设计及特征选取上考虑到跨库检测的有效性,将通过注意力模块及自注意力机制捕捉到的空间域中区域间特征的不一致性和频率域中广泛存在的压缩伪影特征作为检测依据,提升了联合模型在跨库场景下的检测性能。在 Celeb-DF 上本文模型及其他方法的跨库检测 AUC 如表 6 所示。即在 FaceForensics ++ 数据集上进行训练,在 Celeb-DF 上进行测试。可以看出,本文模型在跨库实验上具有最先进的检测性能。与同样应用离散余弦变换的 F3-Net(Qian 等,2020)和 M2TR(Wang 等,2021)相比,AUC 指标分别提升了 3.61% 和 2.11%。F3-Net 中通过双频率域分支学习到的伪造特征使其在库内低压缩视频的检测中有着卓越的成效,然而其对于空间域的线索利用不够充分。

表 6 Celeb-DF 数据集上跨库实验结果

Table 6 Cross database experimental results on Celeb-DF dataset

/%

模型	AUC
Xception(Li 等,2020b)	48.2
Multi-task(Nguyen 等,2019b)	54.3
Capsule(Nguyen 等,2019a)	57.5
DSW-FPA(Li 和 Lyu,2019)	64.6
F3-Net(Qian 等,2020)	65.2
MADD(Zhao 等,2021)	67.4
DCVit(Wodajo 和 Atnafu,2021)	60.8
M2TR(Wang 等,2021)	65.7
本文	**67.81**

注:加粗字体表示最优结果。

M2TR 采用多尺度 Transformer 对伪造图像中不同尺寸大小的图像块进行自注意力的计算,并使用频率域特征作为补充,其结合双域特征的方式与本文模型有着本质的不同。M2TR 对离散余弦变换后的图像表示进行卷积操作得到键嵌入与值嵌入,并用于跨域融合。本文采用对双域特征图进行相似的嵌入操作之后进行交叉注意力的计算,更有利于双域特征之间的交互及补充。相比于应用了多重注意力机制捕捉区域间的细粒度特征的 MADD,本文模型在跨库检测 AUC 分数上提高了 0.41%,这得益于频率域分支的补充以及 Transformer 结构对于全局区域间特征差异信息更好的学习能力。

3.6 消融实验结果

3.6.1 注意力图数量的影响

本文模型在空间域分支加入了注意力模块,用以对骨干网络提取到的浅层特征生成注意力图。伪造区域和真实区域存在不一致性,因此,本文引入注意力机制使模型在空间域特征提取过程中更加关注不一致性显著的区域,即为注意力区域。为了捕捉图像中更全面的局部区域信息,注意力将分散在图像的 k 个局部区域。为了测试注意力机制结合随机数据增广策略的最佳注意力图数量,本文将模型应用于 FaceForensics ++ 数据集上,并改变注意力机制生成的注意力图数量。

注意力图数量消融实验结果如表 7 所示。可以看出,在跨压缩场景的表现下,注意力图数量 M 为 6 时,模型在 FaceForensics ++ 数据集的 4 种深度伪造方法上的表现均为最佳。此外,从图 3 可以看出,本文方法中的注意力模块在全集测试中生成的注意力图对于伪造图像有更强烈的反应,并且每列

表 7 注意力图数量消融实验结果

Table 7 Ablation experiment results of the number of attention maps

/%

注意力图数量	精确率(C23-C40)			
	Deepfake	FaceSwap	Face2Face	NeuralTextures
2	89.64	73.21	67.5	78.21
4	89.29	73.92	71.43	78.21
6	**90.35**	**81.78**	**71.79**	**80.71**
8	88.92	73.57	67.85	77.85

注:加粗字体表示各列最优结果。

注意力图注意区域趋于相似,而每行中注意力图区域又形成互补,这与本文捕捉图像中更全面的局部区域信息的目的一致。

3.6.2　随机数据增广策略的影响

在训练阶段,本文采用了结合注意力机制的随机模糊数据增广策略。具体做法是对于训练阶段的图像通过随机模糊产生一个增广副本,并进行训练,通过对高质量图像进行退化增强模型的泛化性能。随机模糊的方式包括高斯模糊和均值模糊,在训练过程中,随机选择一个注意力图对应的区域进行高斯模糊或均值模糊。为了证明随机数据增广策略对于跨压缩场景的有效性,本文比较了使用不同数据增广策略和不使用数据增广策略训练的模型的性能,对于其他设置则保持一致。

实验对 FaceForensics ++ 数据集中的 4 种伪造类型进行测试,结果如表 8 所示。其中,NONE 表示不采用数据增广策略,DA(data augmentation)表示采用数据增广策略,增广方式为指定某一注意力区域且使用固定核大小的高斯模糊,DA-R(data augmentation-random)则表示对随机注意力区域使用随机核大小高斯模糊或均值模糊。从表 8 可以观察到,随机区域的随机模糊方式的数据增广策略对于跨压缩场景下的检测具有显著贡献。其中,在 FaceSwap 伪造方法上的检测上,随机数据增广策略带来的增益更为明显。

表 8　随机数据增广策略消融实验结果

Table 8　Ablation experiment results of randomized data augmentation strategy

/%

策略	精确率(C23-C40)			
	Deepfake	FaceSwap	Face2Face	NeuralTextures
NONE	86.43	75.36	70.00	74.28
DA	89.64	76.07	69.28	79.29
DA-R	**90.35**	**81.78**	**71.79**	**80.71**

注:加粗字体表示各列最优结果。

3.6.3　分支组合的影响

本文模型主要由空间域特征提取分支、频率域特征提取分支、Transformer 层和交叉注意力结构组成。其中,Transformer 层作为模型的主体结构,用以计算特征间的自注意力,将在所有分支方案中作为基础结构。为了验证模型中分支组合的有效性,本文在 FaceForensics ++ 上对比各种分支组合的跨压缩场景和库内场景的检测性能。

实验结果如表 9 所示。其中,RGB 表示模型由空间域特征提取分支和 Transformer 层组成;Freq. 表示模型由频率域特征提取分支和 Transformer 层组成;RGB + Freq. 表示模型由空间域特征提取分支、频率域特征提取分支及 Transformer 层组成;RGB + Freq. + Cross 表示模型由空间域特征提取分支、频率域特征提取分支、Transformer 层及交叉注意力结构组成。由表 9 可以看出,单独的空间域特征提取分支与 Transformer 层的结合在 Face2Face 上的表现比较好,这说明频率域特征的加入一定程度上影响了 Face2Face 伪造图像的取证。在 NeuralTextures 伪造图像上,单独的频率域特征提取分支与 Transformer 层的结合表现比较好,而与空间域特征提取分支的结合相比于单独的空间域特征提取分支的检测精度有所提高。空间域特征提取分支、频率域特征提取分支、Transformer 层及交叉注意力的结合在 4 种图像伪造方法上的效果都增益显著。

表 9　分支策略消融实验结果

Table 9　Ablation experimental results of branch strategy

/%

方法	精确率(C23-C40)			
	Deepfake	FaceSwap	Face2Face	NeuralTextures
RGB	81.79	70.80	70.35	72.86
Freq.	87.50	70.94	63.92	77.50
RGB + Freq.	88.83	75.71	67.86	73.57
RGB + Freq. + Cross	**90.35**	**81.78**	**71.79**	**80.71**

注:加粗字体表示各列最优结果。

图 8 展示了 4 种分支策略在测试集上的类激活映射(class activation mapping)对比,红色突出部分为模型偏向关注的区域。

从图 8 可以看出,RGB 分支结合 Transformer 的策略(RGB)从表现上来看更加关注图像的背景,频率域特征提取分支与 Transformer 层的结合(Freq.)趋向于学习人脸面部的信息,而两者与 Transformer 层的联合模型(RFreq.)相比于单分支的策略(RGB、Freq.)逐渐将激活区域转移到了脸部或交界区域,但仍会受到背景信息的干扰。深度伪造技术在生成伪造图像时,实际上是对源人脸的面部区域

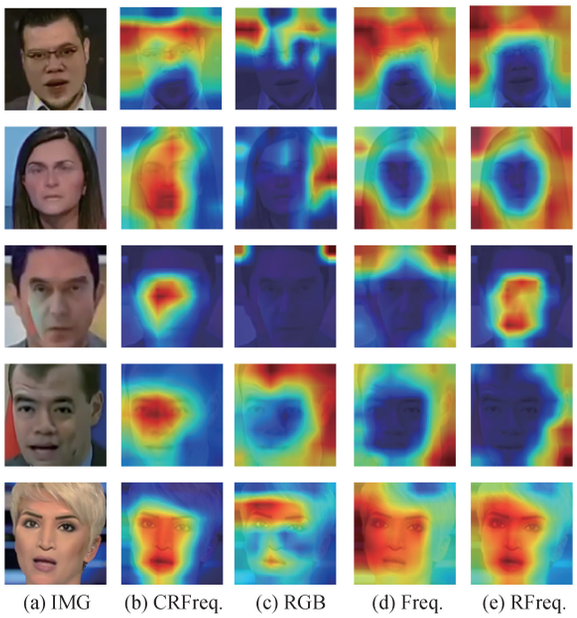


图 8　不同分支策略得到的 CAM 热力图对比

Fig. 8　Comparison of class activation mapping heatmaps obtained by different branching strategies

((a) IMG; (b) CRFreq.; (c) RGB; (d) Freq.; (e) RFreq.)

进行修改，通过对生成网络的训练去拟合源人脸的分布，因此面部区域与其他区域必定存在真实差异，同时在边界中生成局部的伪影。对于联合空间域及频率域的特征，同时利用 Transformer 结构的自注意力计算及交叉注意力计算后，模型对两种区域间的显著区分特征赋予更高的权重，使模型进一步将这些特征作为判别依据。由 RGB 分支、频率域特征提取分支、编码器及交叉注意力结构组成的联合模型（CRFreq.）能够更好地学习到人脸面部的特征，且将其作为区分真实图像和深度伪造图像的线索进行正确分类，从而提升检测准确率。

图 9 展示了不同分支策略在 4 种伪造类型数据集上跨压缩检测的 AUC 曲线。AUC 越接近 1，表示模型的分类性能越好，即对正样本和负样本进行正确划分的能力。图中#1 ～#4 分别代表 4 种分支策略，即单 RGB 分支结合 Transformer 层、单 Freq. 分支结合 Transformer 层、RGB + Freq. 双分支结合 Trans-former 层以及本文提出的 RGB + Freq. 结合 Transformer

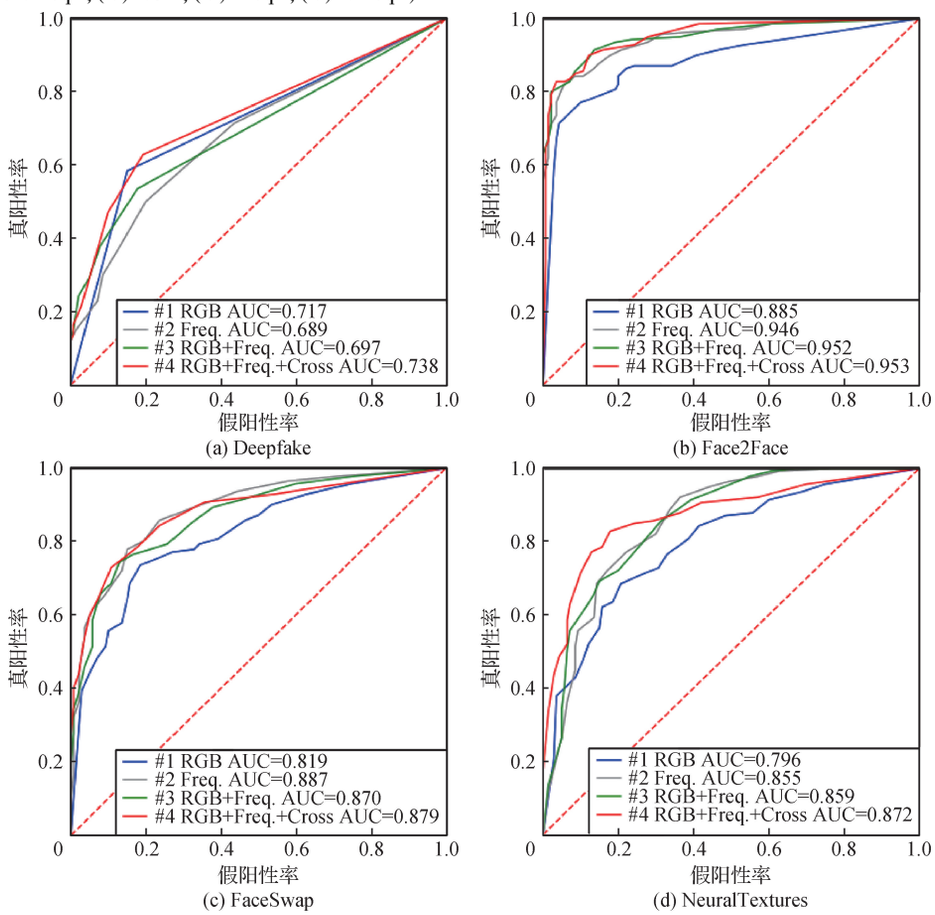


图 9　不同分支策略在 4 种伪造类型数据集的 AUC 对比

Fig. 9　Comparison of AUC of different branching strategies for four tampering types in the FaceForensics++

((a) Deepfake; (b) Face2Face; (c) FaceSwap; (d) NeuralTextures)

层 + 交叉注意力结构组成的联合模型。从图 9 可以观察到，本文方法最终选择的分支策略在多个数据集中具有最优的检测性能。

4 结　论

针对现有深度伪造检测方案在跨压缩率、跨库检测场景下检测精度显著下降且训练时间过长的问题，本文提出了一种基于 Vision Transformer 及 EfficientNet-B0 的双分支联合网络模型，结合随机模糊的数据增广策略，提高了模型在跨压缩率上的检测准确率。首先，对待检测人脸图像使用嵌入注意力模块的 EfficientNet-B0 模型提取空间域特征，并利用注意力引导的模糊方式进行数据增强以适应跨压缩检测场景，同时对经离散余弦变换的频率域表示使用 EfficientNet-B0 提取频率域特征。然后，对双域特征进行线性投影实现嵌入，并进行特征块的划分，送入到编码器与交叉注意力结构堆叠的 Transformer 结构中。最后使用多头特征注意力机制及分支交叉注意力机制在不同域的特征块之间进行信息交互，从而对全局信息进行更好的建模。

本文方法仅采用 EfficientNet-B0 部分结构作为特征提取分支，大幅减少了 Transformer 对图像自相关性进行建模所需的时间消耗，同时保留了卷积神经网络的局部特征提取能力。通过交叉注意力机制，对来自不同域的特征能够更好地实现信息的交互。采用基于随机模糊的数据增强策略，对高质量样本进行退化，以便增加训练数据集样本的多样性。本文方法在主流深度伪造数据集 FaceForensics ++ 和 Celeb-DF 上与现有方法进行对比实验。实验结果表明，本文方法能够更好地应对跨压缩、跨库场景的检测需求，提升了检测性能，同时缩短了训练消耗时间，较现有主流检测算法更易推广至现实应用场景。

在 FaceForensics ++ 数据集上进行了多方面的消融实验，包括注意力图数量的选择、分支组合策略的选择以及数据增强方式的选择。实验结果表明，结合注意力机制下的随机数据增强方式，双分支双特征的模型结构能够更好地捕捉来自不同域的特征信息，同时实现特征交互，并能够更好地适应跨压缩率场景下的检测。此外，时间消耗对比实验结果表明，采用部分结构的卷积神经网络与 4 层改进 Transformer 结构结合的方式，相比于其他方法有着更短的训练时间与推理损耗。基于此，本文框架能够实现对深度伪造视频的高效检测。

虽然跨压缩实验及跨库实验显示本文模型有着良好的检测性能，但在库内实验上可以发现，本文框架的库内检测精度稍低于部分主流模型。这是由于在特征提取分支，本文框架采用了不完整的卷积神经网络结构，因此在层级较浅时网络通常难以学习到数据集的深层分布，从而难以为后续的全局信息建模提供足够的库内特征信息。

在未来工作中，将进一步探索数据增强策略，如进一步加强随机模糊力度，或利用频域注意力引入自适应的无关区域抹除，以便适应不完整的卷积神经网络结构的特征提取，使其在较浅的阶段能获得强有力的判别特征。在结构方面，将继续探索双域特征的提取和增强方式，设计更全面的特征交互融合方法；在此基础上，探究 Transformer 结构与卷积神经网络的结合方式，更完全地利用两种结构的特性和长处，以更好地拟合不同场景下的数据分布，包括库内外、跨库跨压缩等场景。

参考文献 (References)

Afchar D, Nozick V, Yamagishi J and Echizen I. 2018. MesoNet: a compact facial video forgery detection network//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE: 1-7 [DOI: 10.1109/WIFS.2018.8630761]

Bazarevsky V, Kartynnik Y, Vakunov A, Raveendran K and Grundmann M. 2019. Blazeface: sub-millisecond neural face detection on mobile GPUs [EB/OL]. [2022-07-14]. https://arxiv.org/pdf/1907.05047.pdf

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen C F R, Fan Q F and Panda R. 2021. CrossViT: cross-attention multi-scale vision transformer for image classification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 347-356 [DOI: 10.1109/ICCV48922.2021.00041]

Coccomini D A, Messina N, Gennaro C and Falchi F. 2022. Combining efficientnet and vision transformers for video deepfake detection//Proceedings of the 21st International Conference on Image Analysis and Processing. Lecce, Italy: Springer [DOI: 10.1007/978-3-031-

06433-3_19]

Cozzolino D, Poggi G and Verdoliva L. 2017. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Philadelphia, USA: ACM: 159-164 [DOI: 10.1145/3082031.3083247]

Durall R, Keuper M, Pfreundt F J and Keuper J. 2019. Unmasking deepfakes with simple features [EB/OL]. [2022-11-08]. https://arxiv.org/pdf/1911.00686.pdf

Fagni T, Falchi F, Gambini M, Martella A and Tesconi M. 2021. TweepFake: about detecting deepfake tweets. PLoS One, 16(5): #e0251415 [DOI: 10.1371/journal.pone.0251415]

Fridrich J and Kodovsky J. 2012. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 7(3): 868-882 [DOI: 10.1109/TIFS.2012.2190402]

Güera D and Delp E J. 2018. Deepfake video detection using recurrent neural networks//Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand: IEEE: 1-6 [DOI: 10.1109/AVSS.2018.8639163]

He P S, Li W C, Zhang J Y, Wang H X and Jiang X H. 2022. Overview of passive forensics and anti-forensics techniques for GAN-generated image. Journal of Image and Graphics, 27(1): 88-110 (何沛松, 李伟创, 张婧媛, 王宏霞, 蒋兴浩. 2022. 面向GAN生成图像的被动取证及反取证技术综述. 中国图象图形学报, 27(1): 88-110) [DOI: 10.11834/jig.210430]

Hu J, Liao X, Wang W and Qin Z. 2022. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 32(3): 1089-1102 [DOI: 10.1109/TCSVT.2021.3074259]

Li J C, Liu F B, Hu Y J, Wang Y F, Liao G J and Liu G Y. 2020. Deepfake video detection based on consistency of illumination direction. Journal of Nanjing University of Aeronautics and Astronautics, 52(5): 760-767 (李纪成, 刘琲贝, 胡永健, 王宇飞, 廖广军, 刘光尧. 2020. 基于光照方向一致性的换脸视频检测. 南京航空航天大学学报, 52(5): 760-767) [DOI: 10.16356/j.1005-2615.2020.05.012]

Li L Z, Bao J M, Zhang T, Yang H, Chen D, Wen F and Guo B N. 2020a. Face X-ray for more general face forgery detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5001-5009 [DOI: 10.1109/CVPR42600.2020.00505]

Li Y Z and Lyu S W. 2019. Exposing deepFake videos by detecting face warping artifacts//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 46-52

Li Y Z, Yang X, Sun P, Qi H G and Lv S W. 2020b. Celeb-DF: a large-scale challenging dataset for deepfake forensics//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3204-3213 [DOI: 10.1109/CVPR42600.2020.00327]

Liu L Y, Wang J X, Cao S L, Zhao L and Zhang X Q. 2022. U-Net for detecting small forgery region. Journal of Image and Graphics, 27(1): 176-187 (刘丽颖, 王金鑫, 曹少丽, 赵丽, 张笑钦. 2022. 检测小篡改区域的U型网络. 中国图象图形学报, 27(1): 176-187) [DOI: 10.11834/jig.210438]

Nguyen H H, Yamagishi J and Echizen I. 2019a. Use of a capsule network to detect fake images and videos [EB/OL]. [2022-10-29]. https://arxiv.org/pdf/1910.12467.pdf

Nguyen H H, Fang F M, Yamagishi J and Echizen I. 2019b. Multi-task learning for detecting and segmenting manipulated facial images and videos//Proceedings of the 10th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS). Tampa, USA: IEEE: 1-8 [DOI: 10.1109/BTAS46853.2019.9185974]

Qian Y Q, Yin G J, Sheng L, Chen Z X and Shao J. 2020. Thinking in frequency: face forgery detection by mining frequency-aware clues//Proceedings of the 16th European Conference on Computer Vision. Cham, Germany: Springer: 86-103 [DOI: 10.1007/978-3-030-58610-2_6]

Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J and Niessner M. 2019. FaceForensics++: learning to detect manipulated facial images//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 1-11 [DOI: 10.1109/ICCV.2019.00009]

Stuchi J A, Angeloni M A, Pereira R F, Boccato L, Folego G, Prado P V S and Attux R R F. 2017. Improving image classification with frequency domain layers for feature extraction//Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP). Tokyo, Japan: IEEE: 1-6 [DOI: 10.1109/MLSP.2017.8168168]

Szegedy C, Liu W, Yang Q J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tan M X and Le Q. 2019. Efficientnet: rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 6105-6114

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need [EB/OL]. [2022-05-29]. http://arxiv.org/pdf/1706.03762.pdf

Wang J K, Wu Z X, Ouyang W H, Han X T, Chen J J, Jiang Y G and Chen J. 2021. M2TR: multi-modal multi-scale transformers for deepfake detection//Proceedings of ICMR'22: International Conference on Multimedia Retrieval. Newark, USA: ACM: 615-623

Wang S Y, Wang O, Zhang R, Owens A and Efros A A. 2020. CNN-generated images are surprisingly easy to spot... for now//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8695-8704 [DOI: 10.1109/CVPR42600.2020.00872]

Wodajo D and Atnafu S. 2021. Deepfake video detection using convolutional vision transformer [EB/OL]. [2022-08-19]. http://arxiv.org/pdf/2102.11126.pdf

Zhang X, Karaman S and Chang S F. 2019. Detecting and simulating artifacts in gan fake images//Proceedings of 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, the Netherlands: IEEE: 1-6 [DOI: 10.1109/WIFS47025.2019.9035107]

Zhang Y X, Li G, Cao Y and Zhao X F. 2020. A method for detecting human-face-tampered videos based on interframe difference. Journal of Cyber Security, 5(2): 49-72 (张怡暄, 李根, 曹纭, 赵险峰. 2020. 基于帧间差异的人脸篡改视频检测方法. 信息安全学报, 5(2): 49-72) [DOI: 10.19363/J.cnki.cn10-1380/tn.2020.02.05)

Zhao H Q, Zhou W Y, Zhou W B, Zhang W M, Chen D D and Yu N H. 2021. Multi-attentional deepfake detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2185-2194 [DOI: 10.1109/CVPR46437.2021.00222]

Zhou P, Han X T, Morariu V I and Davis L S. 2017. Two-stream neural networks for tampered face detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE: 1831-1839 [DOI: 10.1109/CVPRW.2017.229]

Zhu K M, Xu W B, Lu W and Zhao X F. 2022. Deepfake video detection with feature interaction amongst key frames. Journal of Image and Graphics, 27(1): 188-202 (祝恺蔓, 徐文博, 卢伟, 赵险峰. 2022. 多关键帧特征交互的人脸篡改视频检测. 中国图象图形学报, 27(1): 188-202) [DOI: 10.11834/jig.210408]

作者简介

李颖,女,硕士研究生,主要研究方向为视频取证。
E-mail: spade@stu.scau.edu.cn
边山,通信作者,女,副教授,主要研究方向为视频取证和篡改检测。E-mail: bianshan@scau.edu.cn
王春桃,男,副教授,主要研究方向为多媒体信息安全、信息隐藏和农业人工智能。E-mail: wangct@scau.edu.cn
卢伟,男,教授,主要研究方向为多媒体信息安全、数字取证和多媒体大数据智能。E-mail: luwei3@mail.sysu.edu.cn