doi:10.11887/j.cn.202304004 http://journal.nudt.edu.cn

结合多视角学习与一致性表征的人脸伪造检测*

张 军,于淼淼,杨佳鑫
(国防科技大学 大数据与决策实验室,湖南 长沙 410073)

摘 要:现有的人脸伪造检测方法通常在已知域上表现较好,但面临过拟合的风险,在应对未知场景时无法保持良好的检测能力。为解决此问题,提出一种结合多视角学习与一致性表征的人脸伪造检测框架。为捕获更全面的伪造痕迹,将输入图像转换为两种互补视角并采用双流骨干网络进行多视角特征学习。引入一致性度量,以补丁级监督的方式明确约束不同视角输出的局部特征的相似度。为提高检测精度,采用特征分解策略进一步优化伪造特征,减少不相关因素的干扰,并以伪造特征空间的决策作为最终的预测结果。在基准数据集上进行的大量实验表明,所提出的方法优于现有的主流模型,具备良好的跨域泛化能力。

关键词:人脸伪造;频域特征;多视角学习;一致性度量
中图分类号:TP391 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2023)04-028-09

Combining multi-view learning and consistent representation for face forgery detection

ZHANG Jun, YU Miaomiao, YANG Jiaxin
(Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410073, China)

Abstract: Most of the existing face forgery detection methods usually achieve acceptable detection performance on known attacks, but still face the risk of overfitting and fail to maintain good detection capability when dealing with unknown scenes. To solve this problem, an effective face forgery detection framework based on multi-view learning and consistent representation was proposed. To capture more comprehensive forgery traces, the input image was transformed into two complementary views and a dual-stream backbone network was used for multi-view feature learning. The consistency metric was introduced to explicitly constrain the similarity of local features output from different viewpoints in a patch-level supervised manner. To improve the detection accuracy of the model, the feature decomposition strategy further optimized the forgery-relevant feature to reduce the interference of irrelevant factors, and the decision made from the forgery-relevant feature space was used as the final prediction. Extensive experiments on benchmark datasets show that the proposed method outperforms the existing mainstream approaches with good cross-domain generalization capability.

Keywords: face forgery; frequency features; multi-view learning; consistency metric

近年来,以生成式对抗网络[1](generative adversarial networks, GAN)为代表的深度生成模型得到了业界的广泛关注,进而掀起了一场伪造图像的热潮[2-3],“深度伪造”一词由此而来。深度伪造是指借助深度学习算法,实现对音频和视频的模拟和伪造,主要包括语音模拟、图像生成、换脸、表情重塑等。一个没有任何专业知识储备的普通用户,只需简单操纵电子设备中的应用软件即可实现对一幅图像或一段视频中人物面部区域的篡改和伪造[4-5]。人脸图像中包含着丰富且敏感的个人身份信息,因此,对人脸图像的伪造需要引起格外重视,一旦被不法分子出于恶意目的使用,后果将不堪设想[6]。除对个人隐私的威胁外,深度伪造技术对国家和社会同样会带来极大的安全隐患,例如会损坏企业的公众形象、散播虚假消息并引导政治舆论、促使恐怖主义行动等[7-9]。为了降低这些伪造内容的传播力度和可信度,迫切需要研究有效的检测手段以应对由此带来的种种威胁[10]。

现有的人脸伪造检测方法可大致分为两种类

* **收稿日期**:2023-02-17
基金项目:国家自然科学基金资助项目(62101571);湖南省研究生科研创新资助项目(CX20210058);湖南省自然科学基金资助项目(2021JJ40685)
作者简介:张军(1975—),女,湖南长沙人,教授,博士,博士生导师,E-mail:zhangjun1975@nudt.edu.cn;
于淼淼(通信作者),女,山东青岛人,博士研究生,E-mail:yumiaomiaonudt@nudt.edu.cn

型:一种是采用手工设计的特征描述符结合小型分类器来判断输入图像的真伪[11-13],这类方法适用于特定伪造场景下的检测任务,但在处理复杂场景时通常表现不佳;另一种是基于卷积神经网络(convolutional neural networks, CNNs)的深度学习检测方法[14-19],这类方法凭借 CNNs 卓越的特征学习和数据拟合能力在数据集内部测试中取得了令人满意的检测性能。然而,输入图像中存在大量与标签无关的干扰信息,如背景、身份等,导致有利于当前分类的任意模式和线索都可能被检测器注意并学习到,而关键的伪造特征却被忽略。因此,如何从整个特征空间中挖掘出真实人脸与伪造人脸之间最具判别性的特征是研究的重点[20-22]。对此,多视角学习策略结合伪造特征分解方案是解决该问题的有效手段。

目前大部分工作将人脸伪造检测问题定义为一种二元分类任务,即采用图像级标注对模型输出的预测结果进行监督训练。为了学习全面且泛化的特征表达,仅采用图像级监督信号并不能充分地引导模型进行可靠的特征学习,而引入额外的像素级监督信号(如:伪造区域掩码、人脸深度图等)会不同程度地增加计算开销,进而限制了在现实场景中的应用。基于以上分析,可以采用自监督学习结合二元监督学习的解决方案,进一步提升人脸伪造检测模型的检测性能。

1　相关工作

1.1　面向特定伪造的检测模型

研究帧内图像伪影或不一致的生物信号是判断图像真伪最主要的依据之一。文献[23]利用真实视频和伪造视频中眼睛运动方式的差异来检测输入视频的真伪。Li 等[24]观察到伪造的人脸区域在融合到源视频之前都要经过仿射变换,这一过程会留下独特的伪影,并以此来判断图像是否被篡改过。Jafar 等[25]通过分析嘴巴区域张开时的异常表现来识别虚假视频。Afchar 等[26]采用基于中层语义分析的检测方法,设计了两个具有少量层的深度神经网络模型 Meso-4 和 MesoInception-4,通过关注图像的介观性质来判断图像的真伪。Nguyen 等[27]提出了一种多任务学习框架,既能检测出被伪造的图像,同时又能对伪造区域进行定位。与上述方法不同,Zhao 等[20]认为真假人脸图像之间的差异通常是细微且局部的,因此将这一挑战重新定义为细粒度分类问题,并开发了一个多注意力检测框架,在注意图的指导下将增强的纹理特征和高层次的语义特征结合起来进行最终分类。上述方法大都适用于特定的伪造模式,在域内评估中表现较好,但当检测训练集中未出现的攻击类型时,性能会不同程度地下降。

1.2　可泛化的人脸伪造检测模型

可泛化的人脸伪造检测旨在寻找不同伪造算法遗留下来的共同的伪造痕迹,以实现对任意类型的伪造样本进行准确检测的目的。为了涵盖更全面的伪造表征,Zhou 等[28]提出了一种双流网络框架,人脸分类分支利用真实的和伪造的人脸图像进行训练来捕获高水平的篡改伪迹证据,补丁三元组分支利用隐藏特征提取器捕获局部补丁的低水平的噪声残差证据,最终将两个分支相融合实现了鲁棒的篡改检测。Masi 等[29]同样采用双流检测网络,两个分支分别以 RGB 域和频域信息作为输入,融合模块将两个流的输出结合起来,再经骨干网络和长短期记忆(long short-term memory, LSTM)递归神经网络抽取帧间信息以对视频的真伪进行判断。Liu 等[30]将相位谱的空间域表示与原始 RGB 域合并起来,得到 4 通道的输入图像,并送入 Xception 网络中进行分类。Li 等[31]设计了一种通用的人脸伪造检测算法,只利用真实人脸图像进行简单的融合来自动合成换脸图像以及融合边界图,这两类数据一同输入 HRnet 骨干网络中进行训练,训练好的模型通过预测融合边界来判断图像的真伪。此方法在面对低分辨率的图像时检测性能会显著下降,并且不适用于检测由 GAN 完全生成的伪造图像。

由于整个特征空间中同时包含了与标签相关的信息以及干扰信息,而干扰信息的存在会迷惑检测器做出错误的决策,因此,从整个特征空间中逐步挖掘出真实人脸与伪造人脸之间最具判别性的特征,尽可能消除干扰因素对决策的影响,是本研究的重点。

2　方法

本文提出一种结合多视角学习与一致性表征的人脸伪造检测方法,整体流程如图 1 所示。多视角学习模块旨在从原始图像中提取全面且丰富的互补信息。块间一致性度量模块采用余弦相似性度量促使不同视角的相同位置的局部特征更加接近。特征分解与分类模块旨在将判别性特征空间从整个特征空间中进一步分离出来,减小决策中不相关信息的干扰。

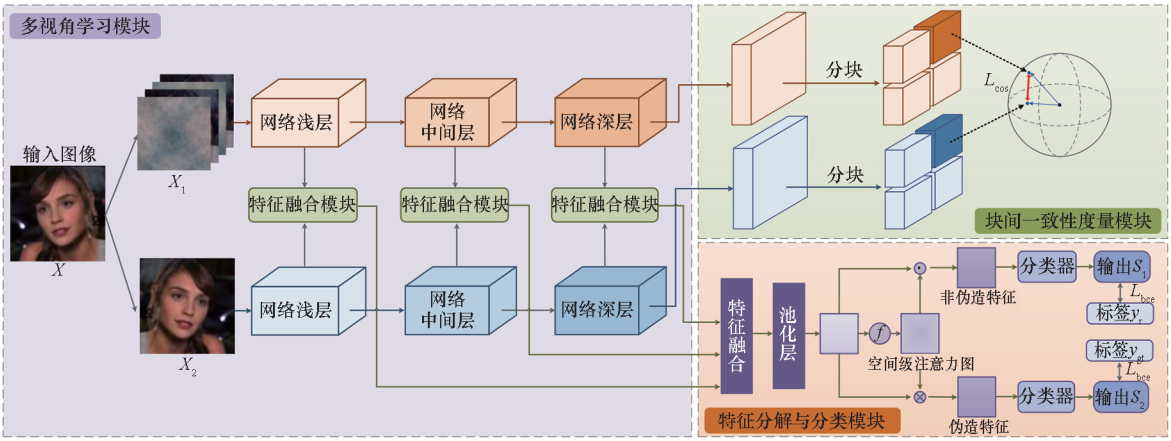


图 1　所提方法整体流程

Fig. 1　Overview of the proposed method

2.1　频域特征

对于人脸伪造检测任务来说，一些关键的判别性线索无法直接从原始 RGB 图像中学习到，而是隐藏在频域中，因此，频域特征和空间特征作为两种互补特征，同时利用它们能获得更全面更泛化的特征表示。在提取频域特征时，采用固定的或手工设计的滤波器很难自适应地充分暴露细微的伪造痕迹，因此选择采用一种自适应频率感知的特征提取方法。首先将输入图像 $\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$ 经离散余弦变换（discrete cosine transform，DCT）由原始 RGB 空间转换到频域中，然后通过将二元基础滤波器 $\{\boldsymbol{t}^i_{\text{base}}\}_{i=1,2,3,4}$ 和可学习频率滤波器 $\{\boldsymbol{t}^i_{\text{learn}}\}_{i=1,2,3,4}$ 相结合来自适应地将其划分为多个频域分量[21]。具体来说，前 3 个基础滤波器将频谱大致分为 3 个子带，分别对应低频（整个频谱的前 1/16）、中频（频谱的 1/16 和 1/8 之间）和高频（频谱的最后 7/8）成分。此外，考虑到分割的频率成分可能不足以挖掘出真假人脸之间全面的伪造痕迹，因此这里又增加了一个额外的基础滤波器 $\boldsymbol{t}^4_{\text{base}}$，用于捕获图像的全频（整个频谱）成分。4 个可学习频率滤波器用于自适应地调整和选择基础滤波器之外的感兴趣的频率响应。最后，利用逆离散余弦变换（inverse discrete cosine transform，IDCT）将划分的频率成分反变换到空间域上。上述过程可以表示为：

$$\boldsymbol{f}^i_{\text{fre}}=D^{-1}\{D(X)\cdot[\boldsymbol{t}^i_{\text{base}}+\sigma(\boldsymbol{t}^i_{\text{learn}})]\},i=\{1,2,3,4\} \tag{1}$$

式中：$D(\cdot)$ 和 $D^{-1}(\cdot)$ 分别表示 DCT 和 IDCT 操作；函数 $\sigma(x)=(1-\exp(-x))/(1+\exp(-x))$ 用于将 x 归一化为 $[-1,1]$。随后，将获得的四个频域分量 $\{\boldsymbol{f}^i_{\text{fre}}\}_{i=1,2,3,4}$ 沿通道维度进行堆叠，最终得到频域特征 $\boldsymbol{X}_1\in\mathbb{R}^{H\times W\times 12}$。

2.2　双流特征提取网络及特征融合模块

作为对频域信息的补充，对原始 RGB 图像采用一般的数据增强技术（如随机翻转）生成增强后的空间图像，记为 $\boldsymbol{X}_2$。接下来，将 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 输入双流特征提取网络中学习全面丰富的特征表示。考虑到 Xception 网络在图像取证方面显示的优越性能，选择其作为双流框架的骨干网络（每个分支不共享参数），主要由三层组成：网络浅层、网络中间层和网络深层。将两个分支中每层输出的特征图经过特别设计的特征融合模块（feature fusion module，FFM）进行融合和增强，得到三个不同尺度的混合特征。

FFM 示意图如图 2 所示。首先采用元素相加运算混合两个输入特征，记为 $\boldsymbol{f}_{\text{ffm}}\in\mathbb{R}^{h\times w\times c}$；随后是一个自注意力机制，其目的是学习通道维度上各个子特征图之间的相关关系，并为它们分配不同的权重。具体来说，将特征 $\boldsymbol{f}_{\text{ffm}}$ 按通道划分为 k 组子特征图 $\boldsymbol{A}^n\in\mathbb{R}^{h\times w\times(c/k)}$ $(n=1,2,\cdots,k)$，然后将其平铺成 1 维特征向量，再分别利用三个嵌入函数 λ、μ 和 γ 生成三个矩阵：

$$\begin{cases}\boldsymbol{K}=\lambda(\boldsymbol{A}^n)\\ \boldsymbol{Q}=\mu(\boldsymbol{A}^n)\\ \boldsymbol{V}=\gamma(\boldsymbol{A}^n)\end{cases} \tag{2}$$

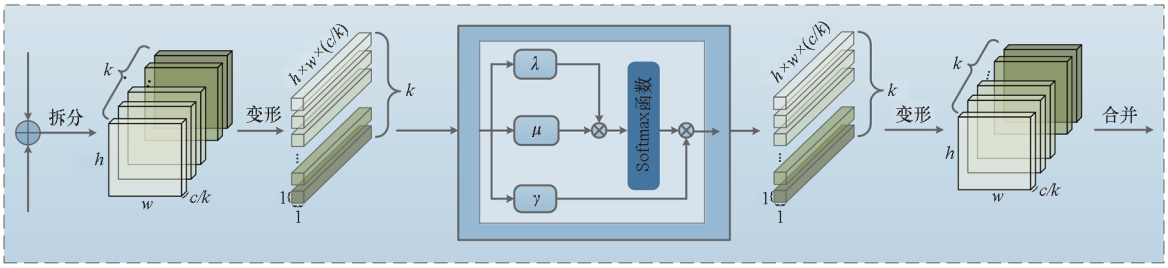


图 2　FFM 示意图

Fig. 2　Schematic diagram of the FFM

接着,将 $\boldsymbol{K}$ 和 $\boldsymbol{Q}$ 组合起来,生成权重矩阵:

$$\boldsymbol{M}^n = \mathrm{Softmax}(\boldsymbol{Q} \otimes \boldsymbol{K}^{\mathrm{T}} / \sqrt{c}) \tag{3}$$

得到的 $\boldsymbol{M}^n$ 即描述了当前的局部特征与其他局部特征之间的相关性。理论上,来自伪造区域的局部特征之间的权重较大,而来自伪造区域的局部特征与来自非伪造区域的局部特征之间的权重较小。再次,将矩阵 $\boldsymbol{V}$ 和 $\boldsymbol{M}^n$ 进行融合:

$$\hat{\boldsymbol{A}}^n = \boldsymbol{M}^n \otimes \boldsymbol{V} \tag{4}$$

最后,将所有生成的 $\hat{\boldsymbol{A}}^n (n = 1, 2, \cdots, k)$ 按通道串联起来,得到增强后的特征图。这里设置 $k = c$,即将每个通道的特征图单独划分为一组。

2.3　块间一致性度量

考虑到仅采用图像级标注对学习鲁棒的和泛化的特征表征方面能力不足,存在过拟合的问题,因此,采用一致性学习策略,在不引入额外的监督信号的情况下,促使同一人脸实例的不同视角的特征更加相似,以自我监督的方式提高输出特征的一致性。将频域分支输出的最终特征记为 $\boldsymbol{f}_{\mathrm{f}} \in \mathbb{R}^{h_{\mathrm{f}} \times w_{\mathrm{f}} \times c_{\mathrm{f}}}$,RGB 分支输出的最终特征记为 $\boldsymbol{f}_{\mathrm{r}} \in \mathbb{R}^{h_{\mathrm{r}} \times w_{\mathrm{r}} \times c_{\mathrm{r}}}$,并将 $\boldsymbol{f}_{\mathrm{f}}$ 和 $\boldsymbol{f}_{\mathrm{r}}$ 空间上划分为 $s \times s$ 大小的补丁,分别记为 $\boldsymbol{u}_i \in \mathbb{R}^{(h_{\mathrm{f}}/s) \times (w_{\mathrm{f}}/s) \times c_{\mathrm{f}}}$ 和 $\boldsymbol{v}_i \in \mathbb{R}^{(h_{\mathrm{r}}/s) \times (w_{\mathrm{r}}/s) \times c_{\mathrm{r}}}$,其中 $i \in \{1, 2, \cdots, s^2\}$。然后,将 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 展平为 1 维向量 $\hat{\boldsymbol{u}}_i$ 和 $\hat{\boldsymbol{v}}_i$,并计算二者之间的余弦相似性:

$$k_i(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{v}}_i) = \left(\frac{<\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{v}}_i>}{\|\hat{\boldsymbol{u}}_i\| \|\hat{\boldsymbol{v}}_i\|} + 1 \right) \Big/ 2 \tag{5}$$

式中,$k_i \in [0, 1]$,值越高表示补丁 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 之间越相似。理想情况下,k_i 应该接近于 1,因此,这里需要构建一个全 1 矩阵 $\hat{\boldsymbol{k}} \in \mathbb{R}^{s \times s}$ 来指导 $\boldsymbol{k}$ 的学习。最终,将块间一致性损失函数表示为:

$$L_{\cos} = \|\boldsymbol{k} - \hat{\boldsymbol{k}}\|_2 \tag{6}$$

2.4　特征分解策略

经过 FFM 处理后会得到三个不同尺度的混合特征,考虑到不同的伪造技术和生成管道会产生不同尺度的伪造痕迹,因此,将这三个特征按通道合并起来以得到输入图像的丰富的多尺度特征表示,记为 $\boldsymbol{f}_{\mathrm{c}} \in \mathbb{R}^{h_{\mathrm{c}} \times w_{\mathrm{c}} \times c_{\mathrm{c}}}$。特征分解旨在为 $\boldsymbol{f}_{\mathrm{c}}$ 每个通道的子特征图分配相应的权重,进一步分离出伪造特征和非伪造特征。首先将 $\boldsymbol{f}_{\mathrm{c}}$ 喂入平均池化层(Avg)压缩其通道信息,再利用多层感知器(multilayer perceptron, MLP)网络和 Sigmoid 激活函数 Sig 得到空间注意力图 $\boldsymbol{g} \in \mathbb{R}^{h_{\mathrm{c}} \times w_{\mathrm{c}}}$:

$$\boldsymbol{g} = Sig(MLP(Avg(\boldsymbol{f}_{\mathrm{c}}))) \tag{7}$$

非伪造特征 $\boldsymbol{f}_{\mathrm{non}}$ 和伪造特征 $\boldsymbol{f}_{\mathrm{forgery}}$ 的分解过程可表示为:

$$\boldsymbol{f}_{\mathrm{non}} = (1 - \boldsymbol{g}) \cdot \boldsymbol{f}_{\mathrm{c}} \tag{8}$$

$$\boldsymbol{f}_{\mathrm{forgery}} = \boldsymbol{g} \cdot \boldsymbol{f}_{\mathrm{c}} \tag{9}$$

将特征 $\boldsymbol{f}_{\mathrm{non}}$ 和 $\boldsymbol{f}_{\mathrm{forgery}}$ 分别输入两个分类器(每个分类器由两个全连接层构成)中,得到两个二元决策 S_1 和 S_2。最终以 S_2 的值作为对输入图像的预测结果。

2.5　混合损失函数

所提出的框架以监督方式进行端到端训练,整体损失函数由分类损失和一致性损失两部分组成:

$$L = \alpha L_{\mathrm{bce}} + \beta L_{\cos} \tag{10}$$

其中,α 和 β 表示权衡参数。分类损失采用二元交叉熵(binary cross entropy, BCE)函数:

$$L_{\mathrm{bce\_1}} = -\frac{1}{N} \sum [y_{\mathrm{r}} \lg S_1 + (1 - y_{\mathrm{r}}) \lg(1 - S_1)] \tag{11}$$

$$L_{\mathrm{bce\_2}} = -\frac{1}{N} \sum [y_{\mathrm{gt}} \lg S_2 + (1 - y_{\mathrm{gt}}) \lg(1 - S_2)] \tag{12}$$

$$L_{\mathrm{bce}} = L_{\mathrm{bce\_1}} + L_{\mathrm{bce\_2}} \tag{13}$$

其中,N 表示训练样本总数,标签 y_{gt} 表示输入图像实际所属的类别,标签 y_{r}($y_{\mathrm{r}} = 1$)表示类别为真。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

为了评估模型性能,在三个公开的人脸取证数据集上进行实验,分别是 TIMIT(DeepFake-TIMIT)[32]、Celeb-DF[33] 以及 FF ++(FaceForensics ++)[34]。TIMIT 包含关于 32 个对象的 320 个真实视频,每个真实视频由 FS(FaceSwap)算法进行篡改,最终生成了包括高质量(HQ)和低质量(LQ)两个版本的共计 640 个伪造视频;Celeb-DF 由 590 个原始视频和 5 639 个通过改进的 DF(DeepFakes)算法生成的高质量伪造视频组成;FF ++ 包含来自 YouTube 的 1 000 个原始视频片段,每个原始视频由 4 种经典的人脸伪造技术进行篡改——DF、F2F(Face2Face)、FS 以及 NT(NeuralTextures),共计 4 000 个伪造视频,所有视频都由三种不同的压缩设置创建而来,分别是 c0(无压缩)、c23(高质量或轻度压缩)和 c40(低质量或重度压缩)。

在随后的实验中,提取 TIMIT 数据集中每个视频的前 20 帧,共计 6 400 个真实帧和 6 400 个伪造帧;从 Celeb-DF 中提取 40 000 个真实帧和 40 000 个伪造帧;随机提取 FF ++ 的 15 000 个真实帧和 60 000 个伪造帧(每种伪造算法 15 000 帧)。每个图像帧通过 OpenCV 的级联分类器 CascadeClassifier 进行人脸区域的检测,检测到的人脸框由中心向外扩展一定的倍数(TIMIT 为 1.1 倍,Celeb-DF 和 FF ++ 为 1.3 倍)并裁切。

3.1.2 实施细节

实验采用 Pytorch 实现,在两个 NVIDIA GPU GeForce RTX 1080 上进行训练。使用的骨干网络 Xception 是在 ImageNet[35] 上预训练的权重进行初始化的。使用 Adam[36] 作为优化器,初始学习率设置为 $2e^{-4}$,每经过 10 个训练周期学习率衰减 0.1。批量大小设置为 16,总的训练周期设置为 50。在评估模型性能时,采用了两个广泛使用的评价指标作为主要衡量标准,即准确率以及接受者操作特征曲线(receiver operating characteristic, ROC)下的面积(area under curve, AUC)。另外,精确率、召回率及 F1 分数被作为辅助评估指标。

在块间一致性度量步骤中,按照经验,将参数 s 设置为 2,即将特征图空间上分割为 $s^2=4$ 个补丁。在混合损失函数中,参数 α 和 β 用于权衡各个项的重要性,为了寻找最佳的设置,表 1 展示了在 FF ++(c23)的 DF 数据集中不同参数设置对模型性能的影响。显然,当 $\alpha=\beta=1.0$ 时,模型获得了最佳的整体性能,后续实验都遵循这一设置。

表 1 损失函数中设置不同参数的性能比较

Tab. 1 Performance comparison when setting different parameters in the loss function

%

(α,β)	准确率	AUC	召回率	精确率	F1 分数
(0.6,0.4)	97.54	98.57	97.20	97.89	97.54
(0.7,0.3)	97.76	99.08	97.60	97.93	97.76
(0.8,0.2)	97.23	98.75	95.35	**99.10**	97.19
(0.9,0.1)	97.80	98.88	96.76	98.84	97.79
(1.0,1.0)	**98.41**	**99.70**	**98.58**	98.25	**98.41**

3.2 数据集内测试

为了验证所提出的方法与现有的其他主流模型相比具有更优越的检测性能,表 2 列出了在 TIMIT、Celeb-DF 和 FF ++/DF 三个数据集上各个方法的 AUC 得分。显而易见,所提方法在所有数

表 2 数据集内部测试的 AUC 结果比较

Tab. 2 Comparison of the AUC results of the in-dataset evaluation

%

模型	TIMIT LQ	TIMIT HQ	Celeb-DF	FF ++/DF
HeadPose[11]	55.1	53.2	54.8	47.3
Multi-task[27]	62.2	55.3	36.5	76.3
VA-MLP[12]	61.4	62.1	48.8	66.4
VA-LogReg	77.0	77.3	46.9	78.0
Capsule[16]	78.4	74.4	57.5	96.6
Two-stream[28]	83.5	73.5	55.7	70.1
Meso-4[26]	87.8	68.4	53.6	84.7
MesoInception-4	80.4	62.7	49.6	83.0
Xception-raw[33]	56.7	54.0	48.2	99.7
Xception-c23	95.9	94.4	65.3	99.7
Xception-c40	75.8	70.5	65.5	95.5
FWA[24]	99.9	93.2	53.8	80.1
DSP-FWA	99.9	99.7	64.6	93.0
FFR_FD[13]	99.9	85.1	78.0	92.3
DFT-MF[25]	98.7	73.1	71.25	
FSSPOTTER[7]	99.5	98.5	77.60	
Tracking eye[23]	99.6	95.5		91.8
Two-branch[29]			73.41	93.18
本文	99.92	100	98.90	99.70

据集上都优于其他模型。具体来说,对于 TIMIT 数据集,无论是低质量还是高质量的伪造版本,本框架都取得了令人满意的结果。由于 Celeb-DF 是一个极具挑战性的数据集,其中的伪造图像用肉眼几乎看不到伪造痕迹,这大大增加了检测难度。本文方法在 Celeb-DF 数据集上以大幅度优势超越了现有模型。另外,本文方法在面对 DF 伪造类型时,同样展现出了卓越的识别能力。

为了评估所提出的框架在面对不同压缩质量的样本时的检测性能,表 3 展示了本方法以及其他几种主流模型在 FF ++ 数据集上的准确率和 AUC 指标,每个指标下最好的结果加粗。从表 3 中可以明显看出,本文方法在高质量数据集 c23 上的检测性能显著优于其他模型。在面对重度压缩样本时,虽然准确率有所下降,但 AUC 指标比排名第二的模型高了 2.65%。本实验证明了所提出的方法在域内评估中的有效性。

表 3　模型在 FF ++ 数据集上的性能比较

Tab. 3　Performance comparison of models on FF ++ dataset

%

模型	c23		c40	
	准确率	AUC	准确率	AUC
Steg. Features + SVM[37]	70.97		55.98	
LD-CNN[19]	78.45		58.69	
Bayar et al.[2]	82.97		66.84	
Rahmouni et al.[3]	79.08		61.18	
DSP-FWA[24]		56.89		59.15
Face X-ray[31]		87.35		61.60
Meso-4[26]	83.10	84.30	70.47	72.62
Multi-task[27]	85.65	85.43	81.30	75.59
SPSL[30]	91.50	95.32	81.57	82.82
Xception[15]	95.73	96.30	86.86	89.30
Xception-ELA[17]	93.86	94.80	79.63	82.90
Xception-PAFilters[18]			**87.16**	90.20
MultiAtt(Xception)[20]	96.37	98.97	86.95	87.26
Two-branch[29]	96.43	98.70	86.34	86.59
本文	**96.68**	**99.18**	84.66	**92.85**

值得一提的是,所提出方法的整体检测性能要略优于 Two-branch 模型。研究发现,Two-branch 同样利用了频域特征,采用了高斯拉普拉斯算子提取图像中的高频信息,并与原始 RGB 域一同作为双流网络的输入信号。对于深度伪造检测任务而言,图像的高频成分作为一种关键判别性特征,能很好地捕获图像因伪造操作而引起的边缘轮廓和纹理细节的改变,同时对引入的噪声敏感。除此之外,中频和低频分量也起到了关键作用,具体来说,中频分量形成了图像的主要边缘结构,低频分量反映了图像中灰度值变化缓慢的区域。通常来说,为了消除伪造产生的抖动,换脸区域需要做进一步的模糊和平滑处理才能与源视频中的背景区域相匹配,这会导致换脸区域的皮肤过于平滑,五官不再清晰锐化,改变了低频和中频成分。因此,与 Two-branch 不同,所提方法还提取了图像的低频和中频特征。另外,考虑到分割的频率成分可能不足以挖掘出真伪人脸之间全面的伪造痕迹,全频分量被用于捕获更大范围的信息。基于以上分析,本文模型性能有所提升的主要原因在于将频谱分成了 4 个不同的子带,而不是只考虑高频分量。

3.3　跨数据集交叉验证

数据集内部测试旨在评估模型对特定伪造的检测能力。与之不同,模型的泛化性或迁移性是检验模型在面对未知伪造类型时能否依然保持良好的性能。在真实场景中,防御模型往往无法获取到任何关于攻击者的先验知识,这就要求模型具备应对未知攻击的能力。在本小节中,通过跨数据集交叉验证来评估所提出的框架的泛化性能。具体来说,首先将模型在 FF ++ (c23)数据集上进行训练,然后在 Celeb-DF 数据集上进行测试,实验结果如表 4 最后一列所示。显然,本文方法的泛化能力比现有的其他模型都更优越。虽然 MultiAtt 模型的域内检测性能稍微优于所提方法,但其泛化性能不足。本实验进一步证明了所提出的模型的可靠性,适用于真实场景下的伪造检测任务。

3.4　*t*-SNE 可视化

对于基于 CNN 的深度学习框架而言,模型的可解释性一直以来都是研究的焦点问题[38],其决定了模型是否真正可靠,可视化是一个重要方式。

如前所述,所提出的特征分解策略能够很好地优化伪造特征空间,进而减少不相关因素的干扰,提升了最终决策的精度。为了证明这一点,以 DF 和 FS 两种伪造为例,通过 *t*-SNE 工具对特征空间优化前后的分布进行可视化,结果如图 3所示,每次实验随机选取 2 000 幅图像进行

表 4 跨数据集交叉验证的 AUC 结果比较

Tab. 4 Comparison of the AUC results of cross-validation across datasets

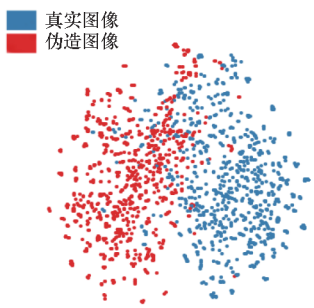
%

模型	FF ++/DF	CDF
HeadPose[11]	47.30	54.60
Multi-task[27]	76.30	54.30
VA-MLP[12]	66.40	55.00
VA-LogReg	78.00	55.10
Capsule[16]	96.60	57.50
Two-stream[28]	70.10	53.80
Meso-4[26]	84.70	54.80
MesoInception-4	83.00	53.60
Xception-raw[33]	99.70	48.20
Xception-c23	99.70	65.30
Xception-c40	95.50	65.50
FWA[24]	80.10	56.90
DSP-FWA	93.00	64.60
EfficientNet-B4[14]	99.70	64.29
MultiAtt[20]	99.80	67.44
F^3-Net[21]	97.97	65.17
MTD-Net[22]		70.12
本文	99.70	70.39

可视化。从图中可以看出,优化前真伪图像的特征空间分布是相互纠缠的,彼此之间没有十分明确的分类边界,这会干扰检测器的判断。当经过特征分解后,两类图像之间的差异进一步被明确,因而证明了所采用的特征分解策略对最终的预测起到了积极作用。

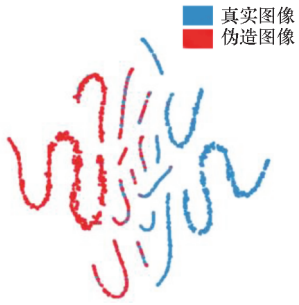
3.5 检测结果示例

为了进一步验证所提出的方法能有效识别出图像的真伪,本小节展示了利用训练好的模型分别对真实人脸图像和四种典型的伪造人脸图像进行检测的结果,如图 4 所示。对于输入的图像帧,由于伪造操作只发生在人脸区域上,因此首先需要对每一帧中的人脸区域进行检测,即图 4 中矩形框框出的区域,然后利用训练好的模型对检测到的人脸区域进行预测。从图中可以看出,模型能够以较高的概率得分预测出不同伪造的目标帧所属的真实类别。



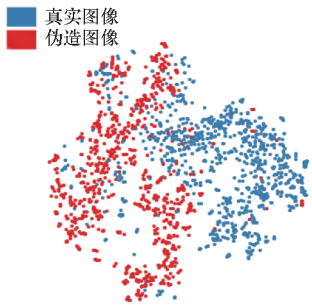
(a) DF 伪造样本采用特征分解策略前的特征分布

(a) Feature distribution of DF forgery samples before adopting feature decomposition strategy



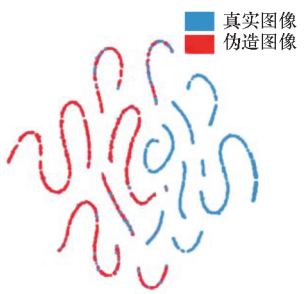
(b) DF 伪造样本采用特征分解策略后的特征分布

(b) Feature distribution of DF forgery samples after adopting feature decomposition strategy



(c) FS 伪造样本采用特征分解策略前的特征分布

(c) Feature distribution of FS forgery samples before adopting feature decomposition strategy



(d) FS 伪造样本采用特征分解策略后的特征分布

(d) Feature distribution of FS forgery samples after adopting feature decomposition strategy

图 3 *t*-SNE 可视化示例

Fig. 3 Examples of *t*-SNE visualization

(a) 真实图像　(b) DF 伪造　(c) F2F 伪造　(d) FS 伪造　(e) NT 伪造
(a) Real images　(b) DF forgery　(c) F2F forgery　(d) FS forgery　(e) NT forgery

图 4　检测结果展示
Fig. 4　Display of detection results

4　结论

本文提出了一种结合多视角学习与一致性表征的人脸伪造检测框架。为了学习全面且泛化的表示以及弱化输入图像中与标签无关的干扰因素对决策的影响,首先采用两个并行工作的骨干网络提取输入图像的两个互补视角中隐藏的伪造痕迹,同时结合特征分解策略进一步优化判别性特征空间。考虑到仅采用图像级标注对学习泛化的特征能力不足,结合一致性度量,通过余弦相似性以自监督方式引导不同视角分支的输出相一致。对比实验(数据集内评估和跨数据集评估)表明,所提出方法的检测性能优于现有的其他主流模型,适用于真实场景下的伪造检测任务。

参考文献(References)

[1] 孙书魁, 范菁, 曲金帅, 等. 生成式对抗网络研究综述[J]. 计算机工程与应用, 2022, 58(18): 90 – 103.
SUN S K, FAN J, QU J S, et al. Survey of generative adversarial networks [J]. Computer Engineering and Applications, 2022, 58(18): 90 – 103. (in Chinese)

[2] BAYAR B, STAMM M C. A deep learning approach to universal image manipulation detection using a new convolutional layer [C]//Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, 2016.

[3] RAHMOUNI N, NOZICK V, YAMAGISHI J, et al. Distinguishing computer graphics from natural images using convolution neural networks[C]//Proceedings of 2017 IEEE Workshop on Information Forensics and Security (WIFS), 2017.

[4] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2Face: real-time face capture and reenactment of RGB videos [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] LI L Z, BAO J M, YANG H, et al. Advancing high fidelity identity swapping for forgery detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] 张远婷. 人工智能时代“深度伪造”滥用行为的法律规制[J]. 理论月刊, 2022(9): 118 – 130.
ZHANG Y T. Legal regulation on abuse of “deep forgery” in the age of artificial intelligence [J]. Theory Monthly, 2022(9): 118 – 130. (in Chinese)

[7] CHEN P, LIU J, LIANG T, et al. FSSPOTTER: spotting face-swapped video by spatial and temporal clues [C]//Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2020.

[8] YU M M, JU S G, ZHANG J, et al. Patch-DFD: patch-based end-to-end DeepFake discriminator [J]. Neurocomputing, 2022, 501: 583 – 595.

[9] 姜文瀚, 田青, 郭小波. 深度伪造技术应用的公共安全挑战与治理[J]. 警察技术, 2023(1): 3 – 9.
JIANG W H, TIAN Q, GUO X B. Public security challenge and governance of deep forgery technology application[J]. Police Technology, 2023(1): 3 – 9. (in Chinese)

[10] 佟昕宇, 陆诗慧, 聂康善, 等. 基于帧内关系建模的人脸深度伪造视频帧间检测模型[J]. 信息与电脑, 2022(24): 56 – 58.
TONG X Y, LU S H, NIE K S, et al. Inter-frame detection model of human-face deepfake video based on intraframe

relationship modeling [J]. China Computer & Communication, 2022(24): 56 - 58. (in Chinese)

[11] YANG X, LI Y Z, LYU S W. Exposing Deep Fakes using inconsistent head poses [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[12] MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//Proceedings of IEEE Winter Applications of Computer Vision Workshops(WACVW), 2019.

[13] WANG G J, JIANG Q, JIN X, et al. FFR_FD: effective and fast detection of DeepFakes via feature point defects[J]. Information Sciences, 2022, 596: 472 - 488.

[14] TAN M, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks [C]//Proceedings of International Conference on Machine Learning, 2019.

[15] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[16] NGUYEN H H, YAMAGISHI J, ECHIZEN I. Capsule-forensics: using capsule networks to detect forged images and videos[EB/OL]. (2018 - 10 - 26)[2023 - 02 - 10]. https://arxiv.org/abs/1810.11215.

[17] GUNAWAN T S, HANAFIAH S A M, KARTIWI M, et al. Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis [J]. Indonesian Journal of Electrical Engineering and Computer Science, 2017, 7(1): 131 - 137.

[18] CHEN M, SEDIGHI V, BOROUMAND M, et al. JPEG-phase-aware convolutional neural network for steganalysis of JPEG images[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017.

[19] COZZOLINO D, POGGI G, VERDOLIVA L. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017.

[20] ZHAO H Q, WEI T Y, ZHOU W B, et al. Multi-attentional deepfake detection [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

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

[22] YANG J C, LI A Y, XIAO S, et al. MTD-Net: learning to detect deepfakes images by multi-scale texture difference[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4234 - 4245.

[23] LI M, LIU B B, HU Y J, et al. Exposing deepfake videos by tracking eye movements [C]//Proceedings of 25th International Conference on Pattern Recognition (ICPR), 2021.

[24] LI Y Z, LYU S W. Exposing DeepFake videos by detecting face warping artifacts [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[25] JAFAR M T, ABABNEH M, AL-ZOUBE M, et al. Forensics and analysis of deepfake videos [C]//Proceedings of 11th International Conference on Information and Communication Systems (ICICS), 2020.

[26] AFCHAR D, NOZICK V, YAMAGISHI J, et al. MesoNet: a compact facial video forgery detection network [C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[27] NGUYEN H H, FANG F M, YAMAGISHI J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos [C]//Proceedings of 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019.

[28] ZHOU P, HAN X T, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[29] MASI I, KILLEKAR A, MASCARENHAS R M, et al. Two-branch recurrent network for isolating deepfakes in videos//Proceedings of the 16th European Conference on Computer Vision ECCV, 2020.

[30] LIU H G, LI X D, ZHOU W B, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[31] LI L Z, BAO J M, ZHANG T, et al. Face X-ray for more general face forgery detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[32] KORSHUNOV P, MARCEL S. DeepFakes: a new threat to face recognition? Assessment and detection [EB/OL]. (2018 - 12 - 20)[2023 - 02 - 10]. https://arxiv.org/abs/1812.08685.

[33] LI Y Z, YANG X, SUN P, et al. Celeb-DF: a large-scale challenging dataset for DeepFake forensics[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[34] RÖSSLER A, COZZOLINO D, VERDOLIVA L, et al. FaceForensics + +: learning to detect manipulated facial images [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

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

[36] KINGMA D P, BA J. Adam: a method for stochastic optimization [C]//Proceedings of the 3rd International Conference for Learning Representations, 2014.

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

[38] 曾春艳, 严康, 王志锋, 等. 深度学习模型可解释性研究综述[J]. 计算机工程与应用, 2021, 57(8): 1 - 9.
ZENG C Y, YAN K, WANG Z F, et al. Survey of interpretability research on deep learning models [J]. Computer Engineering and Applications, 2021, 57(8): 1 - 9. (in Chinese)