计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009701]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

有监督多视图对比学习和两阶段双线性特征融合的人脸活体检测①

孙文赟[1], 李　进[2], 金　忠[3]

[1](南京信息工程大学 人工智能学院, 南京 210044)
[2](南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
[3](南京理工大学 计算机科学与工程学院, 南京 210094)
通信作者: 孙文赟, E-mail: wenyunsun@nuist.edu.cn

摘　要: 本文提出了一种将多尺度频率特征和生成对抗网络 (GAN) 训练的深度图特征融合的多分支网络. 具体地, 高频特征中的边缘纹理信息有利于捕捉摩尔纹. 低频特征对色彩失真更为敏感. 作为辅助信息, 深度图在视觉层面上比 RGB 图像更具辨别力. 有监督多视图对比学习的应用进一步增强了多视图特征的学习. 此外, 还提出了两阶段双线性特征融合方法, 以融合来自不同视图的多分支特征. 为了评估该模型, 我们在 4 个广泛使用的公共数据集 (CASIA-FASD、Replay-Attack、MSU-MFSD 和 OULU-NPU) 上进行了消融实验, 特征融合对比实验, 单一数据集实验和跨数据集实验. 跨数据集实验结果表明, 本文模型在 4 种测试协议上的平均 HTER 比只使用 RGB 图转换为深度图 (DFA) 的方法好 5% (20.3% 减至 15.0%).

关键词: 人脸活体检测; 对比学习; 特征融合; 生成对抗网络; 深度学习

引用格式: 孙文赟,李进,金忠.有监督多视图对比学习和两阶段双线性特征融合的人脸活体检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9701.html

Face Anti-spoofing Based on Supervised Multi-view Contrastive Learning and Two-stage Bilinear Feature Fusion

SUN Wen-Yun[1], LI Jin[2], JIN Zhong[3]

[1](School of Artificial Intelligence, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[3](School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: In this study, a multi-branch network that integrates multi-scale frequency features and depth map features trained by generative adversarial network (GAN) is proposed. Specifically, edge texture information in high-frequency features is beneficial to capturing moire patterns. Low-frequency features are more sensitive to color distortion. Depth maps are more discriminative than RGB images from the visual level as auxiliary information. Supervised multi-view contrastive learning is employed to further enhance multi-view feature learning. Moreover, a two-stage bilinear feature fusion method is proposed to effectively integrate multi-branch features from different views. To evaluate the model, ablation experiments, feature fusion comparison experiments, intra-set experiments and inter-set experiments are conducted on four widely used public datasets, namely CASIA-FASD, Replay-Attack, MSU-MFSD, and OULU-NPU. The experiment result shows that the average HTER of the proposed model on the four tested protocols is 5% (20.3% to 15.0%) better than the DFA method in the inter-set evaluation.

Key words: face anti-spoofing (FAS); contrastive learning; feature fusion; generative adversarial network (GAN); deep learning

① 收稿时间: 2024-05-08; 修改时间: 2024-05-29, 2024-06-17; 采用时间: 2024-07-04; csa 在线出版时间: 2024-09-27

由于其便利性和准确性, 人脸识别已被应用于一些交互式智能应用, 如签到和移动支付. 然而, 现有的人脸识别系统容易受到各种表现形式的攻击, 包括打印、重放、化妆、3D 面具等[1]. 因此, 学术界和工业界都广泛关注开发用于保护人脸识别系统的人脸活体检测技术[2].

由于手工特征已被证明在区分真实样本和攻击样本方面具有辨别力[2], 一些最近的混合方法将手工特征与深度特征结合用于人脸活体检测[3−5]. 关于真实样本和攻击样本之间的区别, 高频特征强调了高频干扰的存在, 例如在打印和视频攻击中出现的莫尔纹. 另一方面, 低频特征突出了颜色失真的问题[5]. 采用像素级监督可以提供更精确和与任务相关的上下文线索, 从而改善内在特征学习. 因此, 最近的研究探索了伪深度标签来指导模型训练[6−8]. 这些方法涉及训练模型来预测真实样本的实际深度, 同时攻击样本置零值. 深度图被用于通过像素级监督估计面部深度信息[6]. Wang 等人[7]通过主成分分析表明, 从深度图领域提取的特征在相同网络结构中比从 RGB 领域提取的特征表现出更大的鲁棒性.

Tian 等人[9]发现, 从更多的视图学习, 得到的表示能更好地捕捉底层场景语义. Khosla 等人[10]将自监督对比方法扩展到全监督, 有效地利用标签信息. 在特征归一化超平面中, 属于同一类的点簇被拉在一起, 同时将不同类别的样本簇推开. 本文将监督对比学习[10]与多视图对比学习[9]相结合. 此外, 本文将非对称三元组损失[11]加入特征学习中, 进一步减小了正样本之间的距离, 从而得到了更紧凑的正样本分布.

常用的多尺度特征融合方法包括拼接和加权求和等[12−14]. 然而, 在处理多视图差异时, 这些方法并不总是最优选择. 因此, 本文提出了两阶段双线性特征融合方法. 在第 1 阶段, 将 3 个频率特征融合视为多尺度特征融合, 本文采用可学习的独立权重的加权求和方法来融合频率特征. 在第 2 阶段, 由于视图之间存在较大尺度, 语义等差异性, 将频率特征和深度图特征融合视为常规的特征融合可能会导致特征信息冗余, 降低多特征学习的多信息优势. 本文引入了双线性模型[15]来融合这些特征. 总结起来, 本文的主要贡献如下.

1) 本文提出了一种结合了频率图像和深度图的多分支网络. 该网络结合了高频和低频编码器, 以提取经常被忽视的高频干扰信息和凸显色彩失真的低频信息, 以及由生成对抗网络生成的深度图提高人脸活体检测的泛化能力.

2) 将有监督多视图对比学习应用到多视图学习中, 最大化了相同样本不同视图之间的相似性, 更好地捕捉人脸潜在的语义信息.

3) 提出了一种两阶段双线性特征融合方法, 以解决高频、低频、深度图中不同尺度和不兼容特征的难题.

1 相关工作

1.1 人脸活体检测

一些基于卷积神经网络的人脸活体检测模型在单一数据集实验中取得了良好的结果. 然而, 在面对跨数据集实验时, 这些只由二元交叉熵损失监督的模型表现出较差的泛化能力. 而且, 端到端的模型很难解释. 此外, 混合算法存在手工设计特征与深度特征之间不兼容的明显缺点, 这可能限制模型的性能[2]. 频率特征和深度图像的融合同样面临着这一挑战. 为了缓解这个问题, 本文确定了不同特征的独立权重, 并采用双线性模型进行特征融合.

生成对抗网络是人脸活体检测算法中较为流行的网络, 特别是在风格迁移和域转换等任务中. 编码器-解码器生成对抗网络[16]能够生成具有连续姿态变化的逼真面部图像. 许多当前的域自适应算法利用 GAN 来减少源域和目标域之间的差异. 例如, Feng 等人[17]采用 GAN 将特征从源域映射到目标域. Wang 等人[18]利用 GAN 混合不同域中的内容特征来学习域无关的内容特征.

在人脸活体检测 (face anti-spoofing, FAS) 中利用面部深度图是一种常见的做法. 最近的几项研究[6,19]已经将像素级伪深度标签纳入到辅助 FAS 模型中. 这些标签强制对真实样本进行准确的深度预测, 而对欺诈样本则采用零值映射. Liu 等人[6]将深度图像引入作为 FAS 模型的辅助信息. Wang 等人[7]采用了神经网络的域对抗训练[20], 用于训练编码器将 RGB 图像转换为深度图像.

1.2 对比学习

对比学习是一种判别方法, 旨在将相似样本聚集在一起, 将不同样本远离彼此[21]. Chen 等人[22]提出了一种用于对比学习视觉表示的简单框架. 标签作为掩码矩阵, 用于区分正负样本[10]. 在 FAS 的背景下, Sun 等人[23]采用了监督对比学习来实现不同域之间的数据

分离. 这是通过将分类标签与域标签相结合来实现的. 域信息也被引入到掩码矩阵中. Tian 等人[9]认为, 人们通过多个感官通道来观察世界. 每个视角都是嘈杂且不完整的, 但它们是重要的因素. 因此, 他们尝试学习能够捕捉多个感官视角之间共享信息的表示. 本文通过利用不同视图的两两排列作为样本对, 并使用类标签作为正负样本分类的指示器来实现这一目标.

1.3 特征融合

特征融合是多分支网络中的重要且常见的操作, 它增强了多个编码器提取的特征. 基本的融合操作包括逐元素相加, 相乘, 拼接等. He 等人[24]采用基于注意力的方法来融合多视图特征. Kim 等人[15]利用点积近似双线性池化方法来融合多视图特征. 双线性池化涉及对特征进行元素级乘积操作, 有效地保留了每个视图的信息.

2 方法

2.1 总体架构

如图 1 所示, 本文所提出的网络可以分为 3 个模块.

(1) 模块 1: RGB 图像被分解为高频图、低频图和 RGB 原图像. 其次, 使用生成对抗网络生成面部深度图. 深度图编码器结构采用了 DepthNet[6].

(2) 模块 2: 使用多个编码器提取深度图特征和多尺度频率特征. 提取频率特征的编码器采用 ResNet18[25]. 提取深度图特征的编码器为一个具有 3 个 3×3 卷积核的 3 层卷积神经网络. 之后, 这些特征通过全连接和 L2 归一化映射到超平面上. 接着, 有监督多视图对比学习方法被用来训练多视图特征.

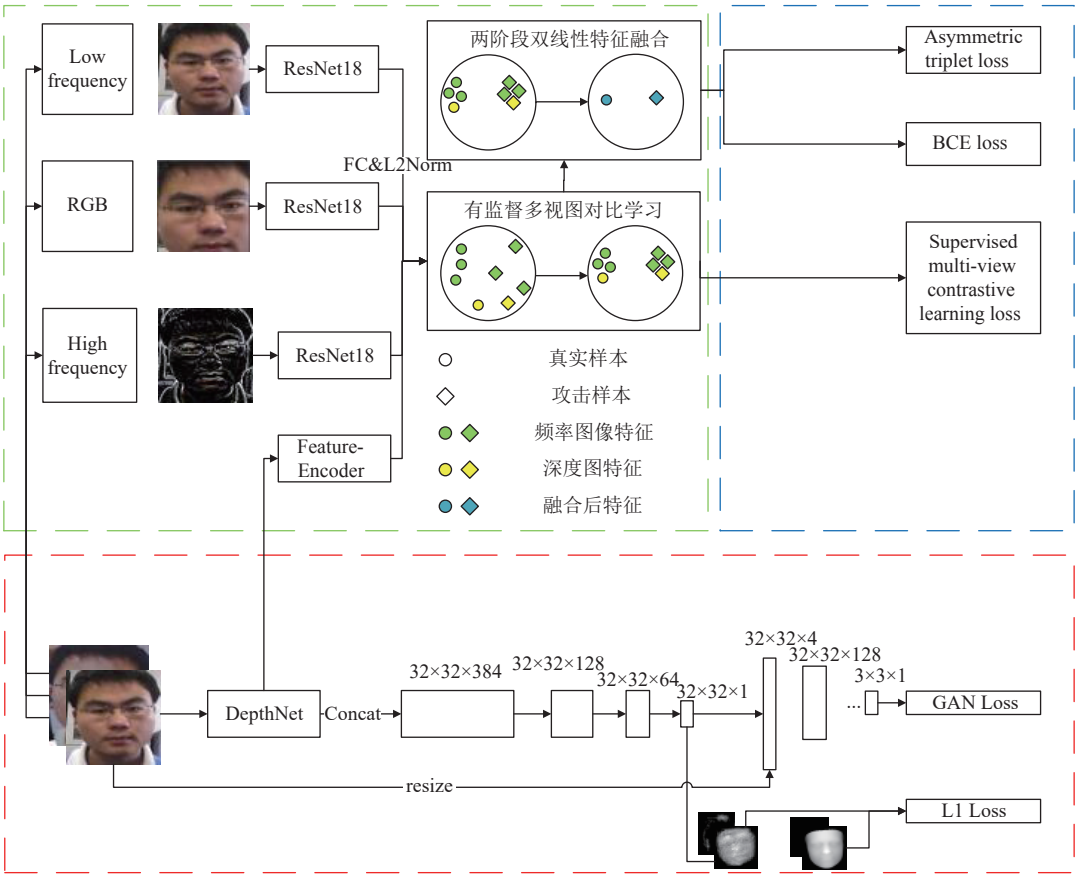

图 1 网络总体架构

(3) 模块 3: 通过两阶段双线性特征融合方法融合频率特征和深度图特征. 我们还使用非对称三元组损失来学习紧凑的正样本特征分布. 最后, 通过二元交叉熵 (BCE) 损失进行特征分类.

综上, 本文的模型共有两个同步的训练流程, 流程 1 中通过生成对抗网络将 RGB 图像转换为深度图, 为流程 2 提供生成深度图作为输入之一. 在流程 1 中, 本文使用 DepthNet 提取原 RGB 图像, 为生成对抗网络中

的深度图生成器提供特征输入.

流程 2 对各个视图进行特征提取, 对比学习, 特征融合以及最后的分类. 流程 2 中, 本文使用一个 3 层卷积神经网络提取已生成的深度图特征, 使用 ResNet18 提取预处理后的频率图像特征. 在提取了 4 种视图特征后, 本文使用多视图有监督对比学习来训练各个视图特征, 使用 full graph 方法, 将所有的视图进行完全的两两配对, 配对的视图作为对比学习的输入, 使得不同视图相同类型的特征拉近, 不同类型的特征拉远. 最后, 本文通过两阶段双线性特征融合方法融合多视图特征并送入分类器分类.

在两个流程之前, 本文通过 MTCNN 来对人脸视频数据进行人脸检测并以 15 帧每次的速率截取人脸图像.

2.2 图像分解与真实深度图生成

本文按照方法[8]来对 RGB 图像进行分解. 首先, 使用平均池化层对原始图像进行下采样. 原始图像大小设置为 256×256, 下采样后的图像大小为 64×64. 然后, 使用双线性插值将图像上采样回原始图像大小. 通过将 RGB 图像从再次上采样的图像中减去, 得到高频图像. 再次上采样的图像即为低频图像. 具体公式如下.

$$\begin{cases} I_H = I - U_{256}(D_{64}(I)) \\ I_L = U_{256}(D_{64}(I)) \end{cases} \tag{1}$$

其中, I表示原始图像, I_H表示高频图像, I_L表示低频图像. $D_{64}()$表示将图像下采样到 64 的大小, $U_{256}()$表示将图像上采样到 256 的大小. 我们使用 3DDFA-V2[26]模型生成真实的深度图作为深度估计的真实值. 在图 2 中, 从上到下的图像依次表示 RGB 图像、高频图像、低频图像和 Ground Truth 深度图像. 从左到右的图像依次表示原始图像、印刷攻击、重放攻击和面具攻击.

2.3 有监督多视图对比学习

给定一个大小为K的 mini batch size $\{x_i, y_i\}_{i=1,\cdots,K}$, 用于训练的 mini batch size 对应由$2K$个配对$\{\tilde{x}_i, \tilde{y}_i\}_{i=1,\cdots,2K}$组成. 其中$\tilde{x}_i$和$\tilde{x}_{2i-1}$是两个不同的视图, 且$\tilde{y}_{2i} = \tilde{y}_{2i-1} = y_i$, 也就是对应的标签相同.

在一个多视角 batch 中, $i \in I \equiv \{1, \cdots, 2K\}$, i是任意视图的索引. $p \in P(i)$, p是具有相同标签的其他视图的索引. $|P(i)|$表示相同标签的不同视图的基数. 全监督对比学习[10]的定义为:

$$L_{\text{Sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \left(\frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \right) \tag{2}$$

多个编码器的输入经过编码、全连接和 L2 归一化后, 这些输入被投影到一个超平面上, 分别表示为z_i, z_p, z_a. τ是温度参数, 它影响学习效果和特征分布. $A(i)$表示除了自身之外的所有样本. 全监督对比学习方法用来训练两个不同的视角, 本文在此基础上, 引入了 full graph[9]方法来完成多视图的对比学习, 方法描述为, 假设有M个视角, 从V_1到V_M. 考虑所有的配对$(i, j), i \neq j$, 可以得到$M \times (M-1)$个配对. 通过涉及所有视图的配对, 监督多视角对比学习的损失为:

$$L_{\text{SMCL}} = \sum_{1 \leqslant i < j \leqslant M} L_{\text{Sup}}(V_i, V_j) \tag{3}$$

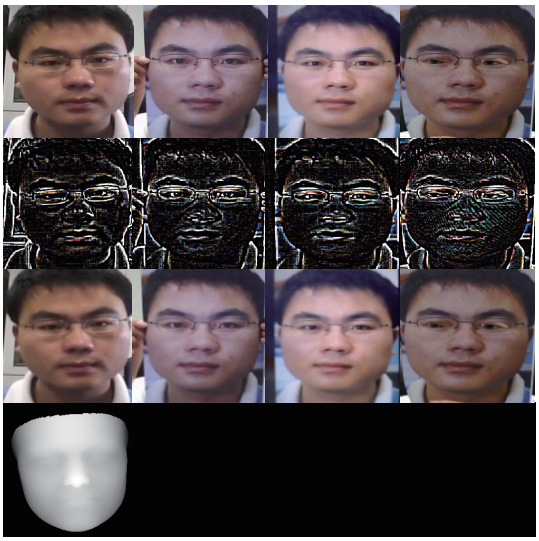

图 2 频率图像与 Ground Truth 深度图

2.4 两阶段双线性特征融合

如图 3 所示, 特征融合可以分为两个阶段. 第一个阶段涉及融合频率特征. 尽管高频图像和低频图像来自同一个 RGB 领域, 但它们具有不同的输入, 针对具有相同输入的场景设计的传统多视角融合方法并不适用于从不同输入融合特征的任务. 因此, 频率特征融合, 求和的权重被初始化为独立的可学习参数. 然后, 这些权重参数与它们的频率特征相乘, 最后将多个频率特征相加. 如式 (4) 所示:

$$F_{\text{fre}} = \lambda_h F_h + \lambda_l F_l + \lambda_i I \tag{4}$$

其中, F_{fre}表示频率特征. λ_h是高频特征的权重, λ_l是低

频特征的权重, λ_i 是 RGB 图像特征的权重. 考虑到频率图像和 RGB 图像来自同一领域, 为了补充图像信息而不引入过多冗余, 将 λ_i 的值设置为 0.1, λ_h 和 λ_l 的值均为 1.0.

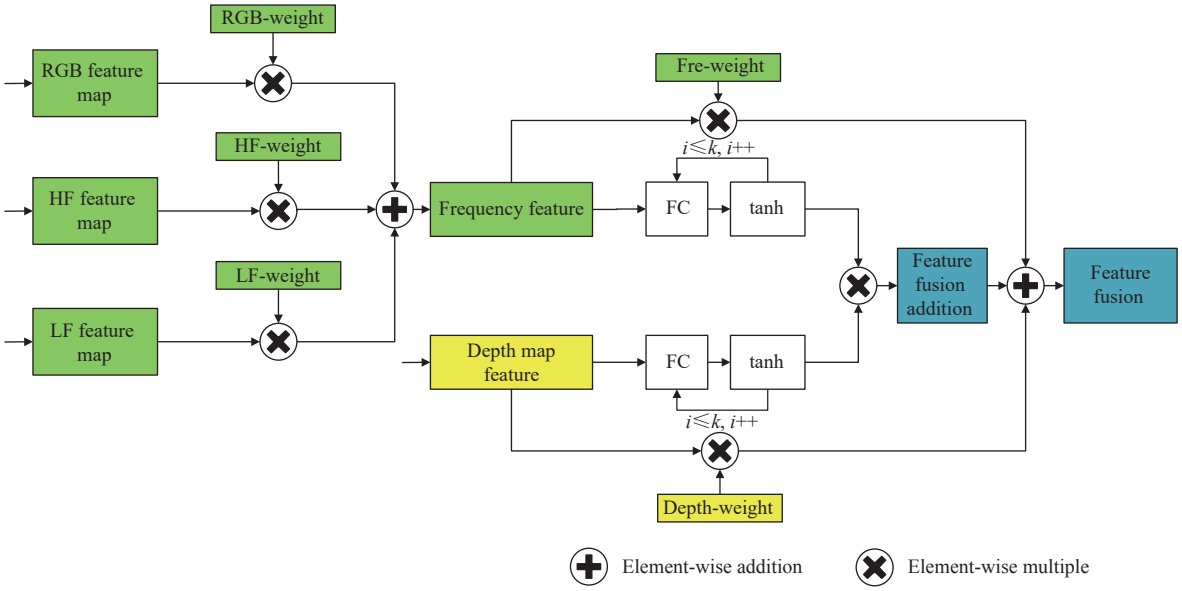

图 3 两阶段双线性特征融合

第 2 阶段是融合频率特征和深度图特征. 频率特征和深度图特征来自不同的领域. 深度图特征的编码器结构与频率特征的编码器结构不同, 因此这些特征之间存在兼容性问题. 然而, 双线性池化融合方法可以缓解或忽略这些问题. 通过利用特征的互相点积, 双线性池化可以实现特征融合, 而不依赖于显式的权重分配. 此外, 双线性池化不受特征维度的限制, 可以应用于任意维度的特征. 这种特性使其特别适用于融合特征, 包括深度图特征和融合的频率特征. Kim 等人[15]通过使用矩阵之间的 Hamada 乘积来近似双线性池化. 这有效地减少了涉及的参数数量. 本文按照他的方法在双线性池化中进行操作[15]. 式 (5) 描述了融合过程, 如下所示.

$$\begin{cases} F_{\text{bp}} = U_{\text{fre}} \circ V_{\text{dep}} \\ F = \lambda_{\text{fre}} F_{\text{fre}} + \lambda_{\text{dep}} F_{\text{dep}} + \lambda_{\text{bp}} F_{\text{bp}} \end{cases} \tag{5}$$

其中, F 是融合后的特征, U 和 V 是权重矩阵, F_{fre} 和 F_{dep} 是特征向量. F_{bp} 表示通过双线性池化融合得到的特征. 类似于跳跃操作, 通过双线性池化融合得到的特征被视为残差. 然后, 将原始的融合频率特征、深度图特征和残差进行加权求和. 在第 2 阶段, 融合频率特征、深度图特征和双线性池化特征的权重分别为 0.5、1.0 和 1.0.

2.5 损失与训练

通过生成对抗网络生成深度图的损失如式 (6)[27].

$$\lambda_{\text{GAN}} \arg\min_G \max_D L_{\text{GAN}}(G,D) + \lambda_{L_1} L_1(G) \tag{6}$$

其中, λ_{GAN} 的值设为 1.0, λ_{L_1} 的值设为 200.0. L_{GAN} 是生成对抗网络损失, L_1 是 L1 重构损失, 用于辅助生成器生成面部深度图.

分类的总损失如式 (7) 所示.

$$L = \lambda_{\text{SMCL}} L_{\text{SMCL}} + \lambda_{\text{tri}} L_{\text{tri}} + L_{\text{cls}} \tag{7}$$

其中, L_{SMCL} 是有监督多视图对比学习 (supervised multi-view contrastive learning, SMCL) 的损失函数. L_{tri} 是非对称三元组损失[11], 用于进一步拉进正样本分布距离. L_{cls} 是二元交叉熵损失. λ_{SMCL} 和 λ_{tri} 是权重参数, 用于在反向传播过程中避免由于不同值的大小而产生偏差. 在本实验中, λ_{SMCL} 为 0.1[23], λ_{tri} 为 2[11].

本文具体的方法流程为, 首先, 对原 RGB 图像进行频率图像预处理, 生成高频图像和低频图像. 使用 3DDFA-V2 模型将原 RGB 图像作为输入, 生成 Ground Truth 深度图. 通过生成对抗网络生成深度图, 输入为原 RGB 图像和 Ground Truth 深度图. 接着, 频率图像和生成的深度图作为输入, 使用 ResNet18 和 3 层卷积神经网络提取多视图特征. 将多视图特征全连接并归

一化, 使用 full graph 方法两两配对, 使用有监督多视图对比学习训练配对的多视图特征, 使用两阶段双线性特征融合方法融合多视图特征. 之后, 将多视图特征送入分类器分类. 最后, 根据式 (6) 计算生成对抗损失, 根据式 (7) 计算对比学习损失, 有监督多视图三元组损失和二元交叉熵分类损失.

3 实验

3.1 数据集

本次实验使用了 CASIA-FASD (CASIA face anti-spoofing dataset) 数据集[28], Idiap Replay-Attack 数据集[29], MSU-MFSD 数据集[30]及 OULU-NPU 数据集[31].

MSU-MFSD 数据集包含来自 35 个受试者的 280 个视频, 分别使用笔记本电脑摄像头和智能手机摄像头进行拍摄, 分辨率分别为 640×480 和 720×480. 主要包括两种不同的欺骗攻击, 例如打印照片攻击和视频重放攻击.

CASIA-FASD 数据集包含 50 个受试者, 3 个场景, 其中包括 150 个真实视频和 450 个攻击视频. 所有人脸都是正面的. 该数据集被分为训练集和测试集, 比例分别为 40% (20 个受试者) 和 60% (30 个受试者). 没有验证集.

Replay-Attack 数据集包含 50 个受试者的 1 200 个视频. 该数据集包括两种不同的照明环境 (控制和逆向) 和两种不同的摄像条件 (手持和固定). 数据集分为训练集 (360 个视频)、验证集 (306 个视频) 和测试集 (480 个视频).

OULU-NPU 数据集包含 4 950 个真实访问和攻击视频, 这些视频是使用 6 种移动设备 (三星 Galaxy S6 edge、HTC Desire EYE、MEIZU X5、ASUS Zenfone Selfie、Sony XPERIA C5 Ultra Dual 和 OPPO N3) 的前置摄像头在 3 个不同的场景中录制的, 场景包括不同的照明条件和背景. OULU-NPU 数据库考虑了打印和视频重放攻击这两种呈现攻击类型.

3.2 实验细节

本文使用 PyTorch 实现代码. 使用 MTCNN[32]进行人脸检测. 人脸图像和深度图都被调整为 256×256 像素大小. 与大多数人脸活体检测算法一致, 在数据增强方面采用了常见的技术, 如颜色抖动、随机旋转和归一化.

本文提出的模型分为 3 个模块. 模块 1 对应生成对抗网络生成深度图, 模块 2 对应多视图特征训练和融合, 模块 3 为分类. 在 3 个模块中, 模块 1 的输入为人脸 RGB 图像和 Groud Truth 深度图, 输出为生成的深度图. 模块 2 的输入为频率图像和模块 1 生成的深度图, 输出为融合的多视图特征. 模块 3 的输入为融合的多视图特征, 输出为分类结果 0 或 1.

模型由于分为 3 个模块, 且生成对抗网络较为难训练, 在训练 3 个数据集, 测试一个数据集所需时间较长. 在推理阶段, 由于本文使用的各模块的网络结构较为简单, 单张图像的处理时间极短, 可以满足实时性的响应. 此外, 本文还引入了在对比学习中表现良好的数据增强方法, 如随机裁剪和 CutOut, 用于 RGB 图像的增强. 网络训练的 batch size 大小为 25、学习率为 0.0003 进行训练, 共进行了 60 个 epoch. 实施了动态学习率调整策略, 如果每 20 个 epoch 验证损失没有变化, 则将学习率减小 0.1 倍. 采用 Adam 优化器来优化模型的参数. 实验在 NVIDIA RTX 3 090 显卡上进行.

3.3 消融实验

为了评估每个视图的贡献, 消融实验是必要的. 消融实验的协议在 CASIA-FASD 数据集上进行训练, 在 Replay-Attack 数据集上进行测试[33], 简称为 C→I. 表 1 评估了使用单一特征 (如 RGB 图像、低频图像和高频图像) 的影响. 结果表明, 仅使用单独的频率图像会导致较差的性能, HTER 值分别为 25.73%、33.97% 和 25.72%. 通过使用注意力机制融合高频和低频特征进行研究, 得到改进的 HTER 值为 21.5%. 最后, 在频率图像融合过程中加入 RGB 图像略微改进了性能, 为 21.2%.

表 1 消融实验

视图组合	HTER (C→I) (%)
RGB图	25.73
低频图	33.97
高频图	25.72
低频图+高频图+特征融合阶段1	21.5
RGB图+低频图+高频图+特征融合阶段1	21.2
深度图	16.69
RGB图+深度图+特征融合阶段2	16.56
低频图+深度图+特征融合阶段2	15.98
高频图+深度图+特征融合阶段2	14.25
频率图+深度图+两阶段特征融合	11.22
频率图+深度图+两阶段特征融合+有监督多视图对比学习	**9.97**

其次, 我们检验了添加深度图特征的有效性, 得到了 16.69% 的 HTER. 它优于单独使用高频和低频特征

的效果. 接着, 我们分别探索了深度图特征与 RGB 图像特征、低频特征和高频特征的组合. 这些组合得到的 HTER 分别为 16.56%、15.98% 和 14.25%, 超过了仅使用深度图特征的性能. 最后, 通过使用两阶段双线性特征融合所有特征, 获得了 11.22% 的更好实验结果. 通过添加有监督多视图对比学习获得了最佳结果, 达到了 9.97%. 这些发现表明, 频率特征和深度图特征确实提高了人脸活体检测的精确度.

3.4 不同特征融合方法对比实验

为了提高性能, 我们对比了几种特征融合方法. 对比实验的训练集是 CASIA-FASD, 测试集是 Replay-Attack, 简称为 C→I. 第 1 种方法是引入了 APM (attention perceptive module)[34]. APM 利用通道注意力和空间注意力机制进行特征融合. 第 2 种方法与第 1 种方法类似, 但是频率特征与独立权重相加, 与第 1 种方法不同. 为频率特征和深度图特征的独立加权求和分配了两个独立的权重. 第 3 种方法中[24], 首先, 多分支网络通过 squeeze-and-excitation (SE) 模块使用通道注意力. 然后, 多个分支的输出通过连接操作进行组合. 第 4 种方法是直接对图像特征进行求和. 第 5 种方法是直接对频率特征进行求和. 由于频率特征主要用作辅助分类, 它们乘以 0.5 的因子然后加到深度图特征上. 第 6 种方法是使用两阶段双线性特征融合方法, 将近似双线性池化得到的融合特征作为最终的多视角特征. 第 7 种方法仍然使用两阶段双线性特征融合方法, 但是通过近似双线性池化得到的融合多视角特征被视为残差. 这些残差与最初融合的频率特征和深度图特征使用独立可学习的权重相加. 比较实验的结果如表 2 所示, 最佳结果是通过最终的两阶段双线性方法实现的.

表 2 特征融合方法对比实验

特征融合方法	HTER(C→I) (%)
APM[34]	34.92
APM[34]+独立权重	32
拼接[24]	42
相加	16.86
带有可学习权重相加	13.25
无跳连接的两阶段特征融合	12.21
带有跳连接的两阶段特征融合	**9.97**

3.5 单一数据集实验

为了验证模型, 我们使用 OULU-NPU 数据集进行了内部实验, 涵盖了 4 个协议. 协议 I 评估了模型在不同场景下的性能. 协议 II 评估了模型对不同攻击方法的鲁棒性. 协议 III 考察了模型在面对不同成像设备 (特别是不同手机) 时的准确性. 协议 IV 综合了前面提到的各个类别. 它全面验证了模型在不同场景、攻击方法和成像设备下的性能. 表 3 呈现了对 4 个协议进行评估的结果. 在协议 I、III 和 IV 中, 融合频率图像和深度图像的本文模型在分类准确性的有效性上得到了确认. 协议 IV 涵盖了 OULU-NPU 中的所有不同因素. 在协议 IV 中取得的出色性能验证了该模型在各种环境中的综合适用性和准确性, 例如有效处理不同攻击、不同设备捕获的人脸和多样化的场景.

表 3 在 OULU-NPU 数据集上的单一数据集实验 (%)

Protocol	Method	APCER	BPCER	ACER
I	STASN[35]	1.2	2.5	1.9
	De-spoofing[36]	1.2	1.7	1.5
	STDN[37]	0.8	1.3	1.1
	CDCN[38]	0.4	1.7	1
	Auxiliary[6]	1.6	1.6	1.6
	DFA[7]	0.8	1.1	1
	Our model	0.5	0.2	**0.4**
II	STASN[35]	4.2	0.3	2.2
	De-spoofing[36]	4.2	4.4	4.3
	STDN[37]	2.3	1.6	1.9
	CDCN[38]	0.4	1.7	1.5
	Auxiliary[6]	2.7	2.7	2.7
	DFA[7]	3.8	2.1	2.9
	Our model	2.9	1.5	2.2
III	STASN[35]	4.7±3.9	0.9±1.2	2.8±1.6
	De-spoofing[36]	4.0±1.8	3.8±1.2	3.6±1.6
	STDN[37]	1.6±1.6	4.0±5.4	2.8±3.3
	CDCN[38]	2.4±1.3	3.1±1.7	2.9±1.5
	Auxiliary[6]	2.7±1.3	3.1±1.7	2.9±1.5
	DFA[7]	1.9±1.6	3.8±6.4	2.8±2.7
	Our model	1.7±1.9	3.3±3.5	**2.5±2.8**
IV	STASN[35]	6.7±10.6	8.3±8.4	7.5±4.7
	De-spoofing[36]	5.1±6.3	6.1±5.1	5.6±5.7
	STDN[37]	2.3±3.6	4.2±5.4	3.6±4.2
	CDCN[38]	4.6±4.6	9.2±8.0	6.9±2.9
	Auxiliary[6]	9.3±5.6	10.4±6.0	9.5±6.0
	DFA[7]	6.7±7.5	3.3±4.1	5.0±2.2
	Our model	4.8±4.0	1.9±1.4	**3.4±2.7**

3.6 跨数据集实验

不同数据集之间的攻击类型、上下文和成像设备变化等客观环境是不同的. 为了评估所提出模型的泛化能力, 需要进行跨数据集的实验. 因此, 本研究采用了目前最主流的跨数据集实验协议[18]进行 FAS 研究.

在跨数据集中, 在 4 个数据集中, 选择其中 3 个数据集进行训练, 剩余一个数据集进行测试. 这 4 个实验分别为 IOC→M、IOM→C、COM→I 和 ICM→O. 我们使用半错误率 HTER 和曲线下面积作为这些实验的评估指标. 表 4 的结果所示, 本文提出的模型在 4 个实验中始终优于其他多模块或者多视图方法. 结果验证了我们的模型在使用更大的数据集进行训练时的良好的泛化能力.

表 4　在 4 个公开数据集上的跨数据集实验 (%)

Method	IOC→M		IOM→C		COM→I		ICM→O	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
Binary CNN[39]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
LBP TOP[40]	36.9	70.8	42.6	61.05	49.45	49.54	53.15	44.09
Depth + rPPG[6]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
MADDG[41]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
DR-UDA[42]	16.1	—	22.2	—	22.7	—	24.7	—
DFA[7]	19.4	86.87	22.03	87.71	21.43	**88.81**	18.26	89.4
Ours	**10.21**	**96.18**	**18.45**	**88.33**	**15.92**	79.14	**15.22**	**92.02**

3.7　可视化与分析

为了使得实验结果更加直观, 本文使用 2D t-SNE[43] 将跨数据集实验的 4 个结果进行可视化, 图 4 显示了跨数据集实验中的每个实验协议下的正负样本的特征分布. 图 5 展示了在 IOC→M 实验协议下的模型推理阶段, 随机真实与欺骗样本的 Ground Truth 深度图和模型生成的深度图. 其中绿色框架中的图像为真实样本, 红色框架中的图像为欺骗样本.

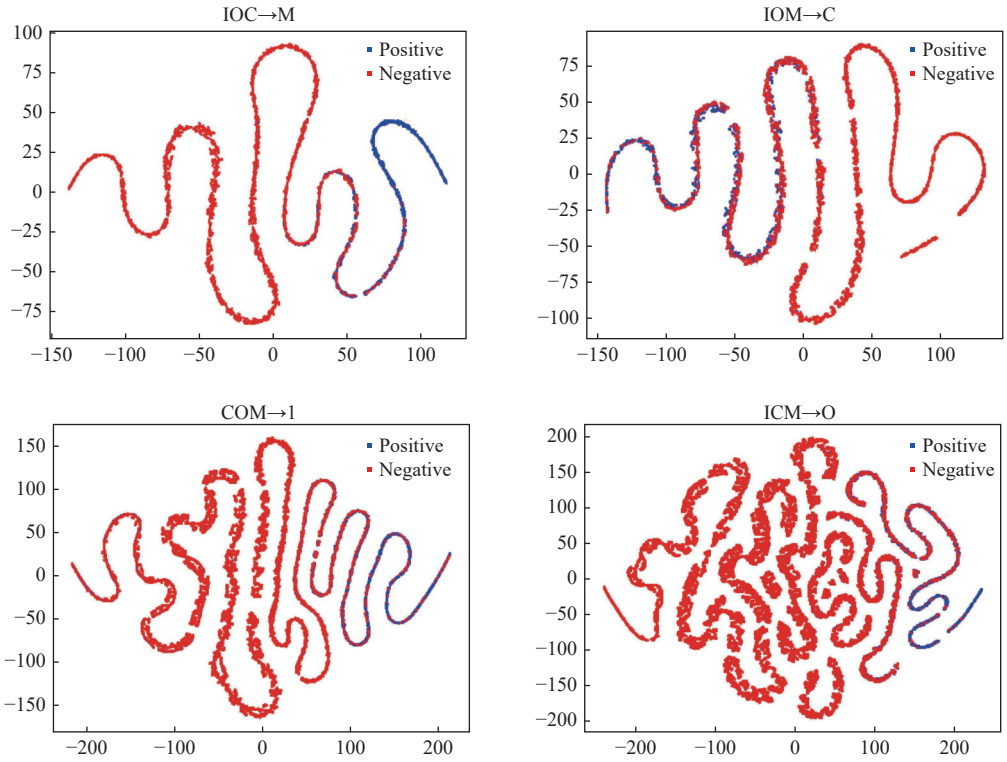

图 4　可视化正负样本特征分布

图 4 中, 正负样本的特征边界是分离的, 然而, 在 IOM→C 实验中, 由于在训练期间没有包含 CASIA 数据集, 尤其是 CASIA 数据集中的高分辨率图像, 分布边界显得模糊不清. 这表明本文所提出的模型在一定程度上依赖于数据集图像的质量. 尽管在特征融合中的权重更新中, 由于深度图的泛化能力, 提高了模型在

面对未知攻击时的泛化性, 但它并没有完全解决人脸活体检测对数据集质量的依赖性. 因此, 解决数据集质量依赖问题仍然是我们未来研究的重点.

图 5 随机样本的 Ground Truth 深度图以及生成的深度图

4 总结

本文提出了一种多分支网络, 旨在有效融合频率特征和深度图特征, 从而提高 FAS 的准确性和泛化能力. 为了增强特征分布的紧凑性, 采用了有监督对比学习方法来学习多视图特征. 为了解决不同输入导致的特征不兼容问题, 提出了一种两阶段双线性特征融合方法. 为了评估所提出的模型, 本文在 4 个公开数据集上进行了一些消融实验, 对比实验, 单一数据集和跨数据集实验. 实验结果验证了该模型在不同场景下的准确性和泛化能力. 然而, 我们的模型仍然存在一些缺点. 例如, 图像质量对我们的分类结果有影响. 为了缓解这个问题, 未来我们计划将域泛化与本文模型相结合, 尝试训练更具有泛化能力的特征.

参考文献

1 Guo JZ, Zhu XY, Zhao CX, *et al*. Learning meta face recognition in unseen domains. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6162–6171.

2 Yu ZT, Qin YX, Li XB, *et al*. Deep learning for face anti-spoofing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5609–5631.

3 Song X, Zhao X, Fang LJ, *et al*. Discriminative representation combinations for accurate face spoofing detection. Pattern Recognition, 2019, 85: 220–231. [doi: 10.1016/j.patcog.2018.08.019]

4 Asim M, Ming Z, Javed MY. CNN based spatio-temporal feature extraction for face anti-spoofing. Proceedings of the 2nd International Conference on Image, Vision and Computing. Chengdu: IEEE, 2017. 234–238.

5 Chen BL, Yang WH, Wang SQ. Face anti-spoofing by fusing high and low frequency features for advanced generalization capability. Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval. Shenzhen: IEEE, 2020. 199–204.

6 Liu YJ, Jourabloo A, Liu XM. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 389–398.

7 Wang YH, Song XN, Xu TY, *et al*. From RGB to depth: Domain transfer network for face anti-spoofing. IEEE Transactions on Information Forensics and Security, 2021, 16: 4280–4290. [doi: 10.1109/TIFS.2021.3102448]

8 Liu YJ, Liu XM. Spoof trace disentanglement for generic face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3813–3830.

9 Tian YL, Krishnan D, Isola P. Contrastive multiview coding. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 776–794.

10 Khosla P, Teterwak P, Wang C, *et al*. Supervised contrastive learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1567.

11 Jia YP, Zhang J, Shan SG, *et al*. Single-side domain generalization for face anti-spoofing. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8481–8490.

12 Dai YM, Gieseke F, Oehmcke S, *et al*. Attentional feature fusion. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3559–3568.

13 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

14 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

15 Kim JH, On KW, Lim W, *et al*. Hadamard product for low-rank bilinear pooling. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

16 Hu C, Feng ZH, Wu XJ, *et al*. Dual encoder-decoder based

generative adversarial networks for disentangled facial representation learning. IEEE Access, 2020, 8: 130159–130171. [doi: 10.1109/ACCESS.2020.3009512]

17 Feng J, Dong ZY, Shi YC, *et al*. Domain adaptation based on ResADDA model for face anti-spoofing detection. Proceedings of the 2021 International Conference on Computer Engineering and Artificial Intelligence. Shanghai: IEEE, 2021. 295–299.

18 Wang Z, Wang ZH, Yu ZT, *et al*. Domain generalization via shuffled style assembly for face anti-spoofing. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4113–4123.

19 Wang ZZ, Yu ZT, Zhao CX, *et al*. Deep spatial gradient and temporal depth learning for face anti-spoofing. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5041–5050.

20 Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 1180–1189.

21 Jaiswal A, Babu AR, Zadeh MZ, *et al*. A survey on contrastive self-supervised learning. Technologies, 2021, 9(1): 1–22.

22 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. Vienna: PMLR.org, 2020. 1597–1607.

23 Sun YY, Liu YJ, Liu XM, *et al*. Rethinking domain generalization for face anti-spoofing: Separability and alignment. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 24563–24574.

24 He D, He XP, Yuan R, *et al*. Lightweight network-based multi-modal feature fusion for face anti-spoofing. The Visual Computer, 2023, 39(4): 1423–1435.

25 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

26 Guo JZ, Zhu XY, Yang Y, *et al*. Towards fast, accurate and stable 3D dense face alignment. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 152–168.

27 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: 10.1145/3422622]

28 Zhang ZW, Yan JJ, Liu SF, *et al*. A face antispoofing database with diverse attacks. Proceedings of the 5th IAPR International Conference on Biometrics. New Delhi: IEEE, 2012. 26–31.

29 Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. Proceedings of the 2012 International Conference of Biometrics Special Interest Group. Darmstadt: IEEE, 2012. 1–7.

30 Wen D, Han H, Jain AK. Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security, 2015, 10(4): 746–761. [doi: 10.1109/TIFS.2015.2400395]

31 Boulkenafet Z, Komulainen J, Li L, *et al*. OULU-NPU: A mobile face presentation attack database with real-world variations. Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017. 612–618.

32 Zhang KP, Zhang ZP, Li ZF, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503. [doi: 10.1109/LSP.2016.2603342]

33 Boulkenafet Z, Komulainen J, Akhtar Z, *et al*. A competition on generalized software-based face presentation attack detection in mobile scenarios. Proceedings of the 2017 IEEE International Joint Conference on Biometrics. Denver: IEEE, 2017. 688–696.

34 Liu J, Zhang FY, Zhou ZY, *et al*. BFMNet: Bilateral feature fusion network with multi-scale context aggregation for real-time semantic segmentation. Neurocomputing, 2023, 521: 27–40. [doi: 10.1016/j.neucom.2022.11.084]

35 Yang X, Luo WH, Bao LC, *et al*. Face anti-spoofing: Model matters, so does data. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3502–3511.

36 Jourabloo A, Liu YJ, Liu XM. Face de-spoofing: Anti-spoofing via noise modeling. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 297–315.

37 Liu YJ, Stehouwer J, Liu XM. On disentangling spoof trace for generic face anti-spoofing. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 406–422.

38 Yu ZT, Zhao CX, Wang ZZ, *et al*. Searching central difference convolutional networks for face anti-spoofing. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020.

5294–5304.

39 Yang JW, Lei Z, Li SZ. Learn convolutional neural network for face anti-spoofing. arXiv:1408.5601, 2014.

40 de Freitas Pereira T, Komulainen J, Anjos A, *et al*. Face liveness detection using dynamic texture. EURASIP Journal on Image and Video Processing, 2014, 2014: 2. [doi: 10.1186/1687-5281-2014-2]

41 Shao R, Lan XY, Li JW, *et al*. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10015–10023.

42 Wang GQ, Han H, Shan SG, *et al*. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. IEEE Transactions on Information Forensics and Security, 2021, 16: 56–69. [doi: 10.1109/TIFS.2020.3002390]

43 Cieslak MC, Castelfranco AM, Roncalli V, *et al*. t-distributed stochastic neighbor embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. Marine Genomics, 2020, 51: 100723. [doi: 10.1016/j.margen.2019.100723]

(校对责编: 张重毅)