基于隐空间约束生成对抗网络的活体检测

陈成伟¹, 阮旺¹, 陈攀², 丁守鸿³, 谢源^{1*}, 宋海川^{1*}, 马利庄^{1*}

1. 华东师范大学计算机科学与技术学院, 上海 200062

2. 上海交通大学电子信息与电气工程学院, 上海 200240

3. 腾讯优图, 上海 200233

* 通信作者. E-mail: xieyuan8589@foxmail.com, hcsong@cs.ecnu.edu.cn, lzma@cs.ecnu.edu.cn

收稿日期: 2020-08-04; 修回日期: 2020-09-14; 接受日期: 2020-09-21; 网络出版日期: 2021-02-25

国家自然科学基金 (批准号: 61772524, 61902129, 61972157)、上海市浦江人才计划 (批准号: 19PJ1403100)、浦东新区科技发展基金 (批准号: PKJ2018-Y46)、北京市自然科学基金面上项目 (批准号: 4182067) 和上海交通大学转化医学交叉基金 (批准号: ZH2018ZDA25) 资助项目

摘要 近年来, 人脸识别技术飞速发展, 其主要应用于门禁系统和公共安防系统. 然而现有的人脸识别系统容易受到仿冒攻击 (也称为呈现攻击), 例如尝试使用用户的脸部照片、视频或者伪造的 3D 人脸去攻击人脸识别系统. 这些攻击手段给人脸识别系统带来了极具挑战性的安全问题. 因此活体检测技术的研究十分重要, 其可以使人脸识别系统免于攻击假脸的安全威胁. 目前, 大部分活体检测的方法将活体检测任务视作有监督的二分类问题, 进而努力充分提取真实人脸和攻击人脸的特征, 在单个数据集内部训练和测试可以达到很高的准确率, 但是在交叉数据集之间训练和测试往往效果不佳. 本文将活体检测任务定义为异常检测任务, 并基于此来解决之前活体检测方法存在的泛化能力差的问题. 因此本文提出了一种新颖的基于隐空间约束的深度对抗网络, 它通过半监督学习的方式进行对抗训练, 在此过程中模型不仅仅可以获得正常样本在隐空间中的分布, 还可以通过一种惩罚的方式对隐空间中正常样本的特征进行约束, 这将带来更加有效和鲁棒的活体检测效果. 测试过程中, 攻击人脸样本将被视作离群的样本, 它们相对于正常样例在隐空间中的表达具有更高的重构差. 实验表明提出的模型相较于前沿的半监督异常检测方法具备明显的优势, 并且在活体检测跨数据集和单数据集内达到了可比的效果或者目前最好的效果.

关键词 对抗网络, 活体检测, 弱监督学习, 异常检测, 人脸仿冒攻击

1 引言

自从日常生活中人们的脸部信息、指纹、虹膜等生物信息被用来解锁设备或完成授权以来, 生物特征识别在验证和安全应用中便扮演着至关重要的角色. 特别是人脸识别这种普遍使用的生物特征应

引用格式: 陈成伟, 阮旺, 陈攀, 等. 基于隐空间约束生成对抗网络的活体检测. 中国科学: 信息科学, 2021, 51: 367–382, doi: 10.1360/SSI-2020-0236
Chen C W, Yuan W, Chen P, et al. Latent regularized generative adversarial network for face spoofing detection (in Chinese). Sci Sin Inform, 2021, 51: 367–382, doi: 10.1360/SSI-2020-0236

用, 已经取得了举世瞩目的成功^[1]. 遗憾的是, 伪造的人脸攻击对人脸识别系统^[2] 带来了巨大的威胁. 呈现攻击 (presentation attacks, PA) 是最流行的人脸仿冒攻击之一, 其形式包括打印的人脸、重复播放视频的人脸以及带上面具的人脸. 因为传统的人脸识别系统对于 PA 往往表现很脆弱^[3], 多种活体检测技术已经被提出用以分辨真实的人脸和攻击的假脸. 这些技术对人脸进行活体检测, 检测验证通过之后再识别人脸. 活体检测的主要挑战在于面对各种类型的 PA 如何达到鲁棒性和泛化性.

目前大部分呈现攻击检测 (presentation attack detection) 的方法将活体检测的任务视作有监督学习下的二分类问题, 数据被分为真实人脸和攻击人脸两类. 虽然很多方法在单个数据集内部 (Intra-dataset, 即训练数据和测试数据来自同一个数据集) 的场景中获得了一些成功^[4]. 然而在真实应用中, 攻击人脸具有多样性甚至存在数据集以外未知的攻击手段, 导致大部分攻击手段并不包含于训练数据中. 因此在二分类训练中学习到的决策边界往往不能适用于真实场景下的人脸攻击. 着眼于以上目前活体检测技术的缺陷, 本文将活体检测任务定义为一个异常检测任务, 原因如下: (1) 泛化问题可以采用异常检测的方法解决, 将攻击人脸视作离群 (out-of-distribution) 样本, 这些样本比真实的样本具有更高的重构差. (2) 将所有类型的攻击人脸手段充分考虑是不实际的, 可能还会有未知类型的攻击, 而且收集攻击人脸数据用于有监督学习训练的经济和时间成本非常高. 因此本文设计了一个生成对抗网络 (generative adversarial network, GAN) 类型的异常检测框架, 其可以仅用正常类 (真实人脸) 训练的情况下, 学习目标类的隐特征空间并检测出超出空间分布范围的异常类 (攻击人脸).

本文提出一种在新型的半监督学习框架下用作活体检测任务的生成对抗网络. 与传统的生成对抗网络结构相比, 本文设计了两个关键组件, 一个是针对隐特征空间优化的约束组件和一个辅助的编码器. 除了重构差区分正负样本之外, 通过对隐特征空间进一步约束更有助于区分正常样本和离群样本. 辅助的编码器将重构后的图片再次投影到隐特征空间中的特征编码. 通过模型训练获得原始输入图像在隐特征空间中的编码和生成图像在隐特征空间中的编码, 并且以两编码之间的距离最小化为目标, 进一步优化隐空间的表达. 通过辅助的编码器再次获得投影的特征编码, 在模型优化过程中可以起到避免隐特征空间在约束条件下产生重构偏差的作用, 得到一个对于目标类更加集中的空间分布. 测试过程中, 本文在视频级别上针对活体检测设计了一个新的度量方法, 其计算视频中每一帧的异常值. 根据测试视频中的所有视频帧生成的异常值生成对应的分布. 接着用最大平均差异 (maximum mean discrepancy) 作为训练时正常人脸视频生成的参考分布与测试视频生成的分布之间差距的度量. 如果两者分布差距相近, 即认为测试视频为真实人脸视频. 预设阈值 T 根据分布之间差距来判断测试样本是否为异常视频.

本论文的主要贡献如下:

- 本文首次提出在半监督学习框架下设计对抗训练的深度网络用于活体目标检测.
- 本文不仅在图片重构空间中获取训练数据的原始属性和特征, 还用一种惩罚的方式将其限制在隐特征空间的表达, 得到更鲁棒和有效的活体检测效果.
- 本文设计了一种对于视频层面的活体检测的度量方法, 将所有帧的异常得分映射到一个分布中.
- 在若干具有挑战性的数据集上对提出的方法进行广泛的评估, 实验结果表明提出的方法的表现超过了已知目前最好的方法.

2 相关工作

首先按照传统方法到深度学习方法的顺序, 分别介绍先前的人脸活体检测的方法和异常检测方法.

2.1 人脸活体检测

纹理上的不同是一个分辨真实人脸和攻击人脸的主要线索. 大量先前的活体检测的工作基于手工纹理特征进行二分类建模 (即, SVM), 包括 LBP [5], HOG [6], DOG [7], SIFT [2] 和 SURF [8]. 尽管如此, 这些特征对于多变的环境数据并不总是有效的, 比如亮度、位置、画像以及其他因素的影响. 为了解决这些挑战, 有些工作着眼于在不同域中的活体检测问题, 例如 YCbCr 颜色空间、时间域和傅里叶频谱中. 放眼于深度学习领域, 大部分研究者试图设计深度学习模型来追求更好的效果. 卷积神经网络 (convolutional neural network, CNN) 很擅长提取和学习深度特征, Yang 等 [9] 将卷积神经网络用作活体检测的分类器, 并用不同的空间尺度的真实人脸和攻击人脸来训练模型. Xu 等 [10] 提出一种 LSTM-CNN 结构来预测视频的未来帧. 与传统方法相比, 深度学习方法在单个数据集场景下极大提升了检测准确率. 但是往往由于过拟合问题, 这些监督学习的方法在跨数据集的场景中并不具备出色的泛化表现.

2.2 异常检测

异常检测是一个计算机视觉领域中的经典问题. 当数据样本远离由训练数据定义的数据分布时, 这些数据样本被分类为异常类. 过去由于稀疏表达和字典学习方法的蓬勃发展 [11], 研究者利用稀疏表达来学习正常类的字典. 在测试过程中, 这类方法通过学习到的字典重构出原始样本, 其中重构差较大的样本视作异常类. 不过训练所需的数据标签造成的经济成本和时间成本相当高, 尤其是对于视频数据. 近年来, 基于自编码器 [12~14] 的深度学习网络被用于学习正常数据样本的模式, 并通过重构差来检测异常. 例如, 具有生成对抗网络 (GAN) 的变分自编码器 (variational autoencoders, VAE) 将学习到的低维度特征应用于异常检测任务中, 在此过程中对真实数据样本分布进行建模. 解码器将隐空间特征向量解码为重构图片. 大量基于图片重构差异进行异常检测的工作 [15, 16] 被提出. 但是计算图片重构差不是异常检测任务的初衷, 其根本目的是在异常类和正常类间找到更有区分度的隐空间表达特征. 而且在实践过程中, 由于实际正常样本分布存在多模式, 卷积自编码器倾向于提供更模糊的重构. 这些模糊将会影响基于重构差来判断的异常检测方法.

3 提出的方法

本节阐述如何使用提出的对抗训练的框架来学习目标类的原始属性和特征. 先介绍总体网络结构, 然后描述损失函数 (loss function) 中的各个部分, 最后给出具体实现方法的描述.

3.1 网络结构

本文提出的方法由数据预处理和一个对抗网络结构组成. 预处理步骤将连续的视频帧转化为光流图. 受异常检测方法 [16~18] 的启发, 图 1 中显示的 GAN 风格的结构由 4 个部分组成: 编码器-解码器部分 (记作生成子网络 $G(*)$)、鉴别器、辅助编码器和隐特征空间约束组件.

生成子网络包含一个编码器 $G_e(*)$ 和一个解码器 $G_d(*)$, 它的目的是重构输入图片来欺骗鉴别器, 同时鉴别器子网络区分真实原始图片和通过重构获得图片. 在训练过程中, 上述两个子网络相互竞争以达到高质量的重构图片使得鉴别器也无法辨认. 模型通过最小化原始图片 x 和重构图片 x' 的距离, 获取了目标类在隐特征空间中的编码, 见图 1 中的向量 z . 基于以上假设, 在测试中与正常类相比, 重构异常类会变得困难. 因为生成子网络和鉴别器在训练过程中只获得了正常样本作为输入. 为了进一步优化样本在隐特征空间 z 中的表达, 本文通过提出一个隐特征空间约束项, 通过惩罚边界以外的

图 1 (网络版彩图) 我们的框架包括一个生成器、一个鉴别器、一个隐空间约束组件和一个辅助编码器。通过相互竞争训练生成器和鉴别器, 同时进行协作以理解目标类中的基本属性和特征。隐空间约束组件进一步用于在隐特征空间中区分正常样本和离群值并给予惩罚。采用辅助编码器可以使原始输入图像的隐空间特征与生成图像的编码后的隐空间特征之间的距离最小

Figure 1 (Color online) Our framework consists of a generator, a discriminator, a latent regularizer and an auxiliary encoder. The generator and discriminator were trained by competing with each other while collaborating to understand the underlying concept in the normal class. The latent regularizer is further utilized to distinguish between normal samples and outliers in a discriminant way in latent feature space. The auxiliary encoder is adopted to minimize the distance between the bottleneck features of original input image and encoded latent feature of the generated image

样本, 最终获得更紧致的隐特征空间表达。单分类的支持向量机 (one-class SVM) 就是这样一种广泛应用于离群检测的判别模型。其主要目的是创建一个决策边界 (boundary) (超平面) 以最大化地将属于已知类和其他类的样本分离。离群样本落在超平面的另一侧。基于上述方法, 本文提出了一个全新的损失函数, 通过惩罚离群样本的方式使得目标类样本的隐空间特征更加集中。

辅助编码器 G_e' 被用作最小化原始图像的隐空间特征 z 和生成图片的隐空间特征 z' 的距离。辅助编码器与生成器中的编码器具有相同结构, 但是使用不同参数。加入辅助编码器的原因如下: (1) 仅对于正常类样本, 隐特征空间编码 z' 才能很好地从生成图片 x' 的特征向量中重构出来。(2) 由于约束项可能对隐特征空间中的分布造成偏差, 因此重构后的隐特征空间编码 z' 可以用作防止隐特征空间编码 z 产生偏差。生成器获取输入图像 x , 该编码器 G_e 通过卷积层以及紧随其后的归一化层和非饱和和激活层将其压缩为隐空间表达 z , 请参见图 1 的右侧。另一方面, 解码器 G_d 包括卷积转置层、relu 激活层、归一化层以及 tanh 激活层, 用于将隐特征向量 z 重建回图像 x 。鉴别器的结构由一系列的卷积层组成, 在图 1 的右侧也进行了说明。为了防止被生成器欺骗, 鉴别器在对抗训练过程中, 不断学习原始数据中的核心特征。鉴别器还帮助生成器在训练过程中获得更鲁棒和稳定的参数。这部分参数会进一步扩大正常样本与异常样本之间的距离。

3.2 损失函数

为了更好地训练模型, 在本文中定义一个损失函数。该损失函数包括 4 个部分, 即图像重建损失、对抗性损失、隐特征空间表达损失和离群损失:

$$\mathcal{L} = w_i \mathcal{L}_{\text{irec}} + w_a \mathcal{L}_{\text{adv}} + w_z \mathcal{L}_{\text{zrec}} + w_o \mathcal{L}_{\text{ou}}, \quad (1)$$

其中 w_i , w_a , w_z 和 w_o 是权重参数, 用于平衡每个单项损失对整体函数的影响. 对于第 1 项, 假设当生成器获取异常样本时, 它将获得较大的重构误差. 因为在训练过程中, 只有正常样本在网络中被训练, 其参数不适合重构异常样本. 对抗损失 $\mathcal{L}_{adv}$ 被用来衡量生成的图像是否成功地欺骗了鉴别器. 通过相互竞争训练生成器和鉴别器, 与此同时两者也通过相互协作学习正常样本的基本分布. 此外, 对于异常样本, 生成器的输出已经在图像空间中存在一定程度上的重构偏差, 此时使用辅助编码器将 $\mathbf{x}'$ 再次映射到隐特征向量 $\mathbf{z}'$ 时会导致其在隐特征空间上编码的重构偏差加大. 换句话说, 压缩的隐特征 $\mathbf{z}$ 与来自辅助编码器的隐特征 $\mathbf{z}'$ 之间的距离 ($\mathcal{L}_{zrec}$) 也在扩大. 在测试过程中, $\mathcal{L}_{zrec}$ 也将被视为区分是否异常的一种度量值. 此外, 最后一项为新颖的损失函数 $\mathcal{L}_{ou}$, 目的是通过惩罚超平面外的点使目标类别样本在隐特征空间中的编码变得更加集中, 导致在测试中异常和正常在隐特征空间中更可分离. 通过对图像重建空间和隐特征空间的联合约束性学习, 本文提出的模型对于活体检测任务将更加有效.

对抗损失. 对抗损失用于训练两个网络, 即生成器和鉴别器. 文献 [19] 提出了一个 ‘两人最小 – 最大’ 的博弈: 鉴别器 D 将生成的图像与真实图像区分开, 而生成器 G 通过不断的训练尝试通过重构出清晰的图像欺骗鉴别器 D . 生成器和鉴别器之间的这种对抗博弈可以表示为

$$\mathcal{L}_{adv} = \min_G \max_D (\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} [\log(1 - D(G(\mathbf{x})))]), \quad (2)$$

生成器 G 尝试学习正常类别中的基本属性和特征, 并经过训练, 最大程度地使鉴别器混淆, 使其相信其生成的样本来自真实样本的数据分布而非重构后的数据分布, 鉴别器 D 尝试区分原始图像与 G 生成的重构图像. 通过利用 $D(\mathbf{x})$ 的梯度进行训练 $\mathbf{x}$, 并使用它来优化其参数.

图像重建损失. 鉴别器试图区分真实样本和生成的样本, 而生成器试图欺骗鉴别器. 然而生成器在学习原始数据分布的过程中, 并没有通过对抗性损失朝着此目标进行优化. 一些先前的工作 [20] 提出应考虑输入图像和生成图像之间的距离. 文献 [20] 显示, 与 L2 相比, 使用 L1 产生的模糊效果更少. 因此, 本文使用 L1 损失函数通过最小化原始输入 $\mathbf{x}$ 与生成的图像 $G(\mathbf{x})$ 之间的距离优化生成器, 如下所示:

$$\mathcal{L}_{irec} = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \|\mathbf{x} - G(\mathbf{x})\|_1. \quad (3)$$

隐空间表示损失. 上述两个损失函数可以帮助生成器学习目标类中的真实属性和特征, 但是他们仅考虑优化原始输入 $\mathbf{x}$ 和生成的图像之间的距离. 通过额外使用编码器从图像空间到隐特征空间的映射, GAN 网络能够学习这种逆映射. 仅对于目标类别样本, 辅助编码器可以根据生成的图像 $\mathbf{x}'$ 很好地重建隐空间特征编码 $\mathbf{z}$. 此外, 隐空间的正则化可能会在隐特征空间中引起分布失真, 因此可以将特征表示 $\mathbf{z}'$ 视为锚, 以防止 $\mathbf{z}$ 偏移. 此外, 已经提出了很多基于图像重建误差 [17, 18, 21] 的工作来进行异常检测. 使用重构错误来检测异常不是异常检测任务的初衷. 因此, 本文考虑以最小化来自生成器的输入图像 $G_e(\mathbf{x})$ 的隐空间特征与来自辅助编码器 $G_{e'}(\mathbf{x}')$ 的重构隐空间特征之间的距离为目标, 进一步优化隐特征空间的表达. 公式如下:

$$\mathcal{L}_{zrec} = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \|G_e(\mathbf{x}) - G_{e'}(\mathbf{x}')\|_2. \quad (4)$$

离群损失. 单分类支持向量机广泛用于先前工作中的异常检测. 它的主要思想是在潜在空间中创建决策边界 (超平面), 从而使来自不同类别的样本更加可分离. 在此基础上, 本文提出了一种新颖的损失函数, 通过惩罚超平面外的点使目标类别样本的隐空间特征更加集中. 给定一组训练样本

$\mathcal{X} = \{\mathbf{x}_i^k\}_{i=1}^{N^k}$, 单分类 SVM 的基本问题可以表述为以下公式:

$$\begin{aligned} \min_{\mathbf{w}, \rho} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{vN^k} \sum_{i=1}^{N^k} \xi_i - \rho \\ \text{s.t.} \quad & \mathbf{w}^T \Phi(\mathbf{x}_i^k) \geq \rho - \xi_i, \quad \xi_i \geq 0, \end{aligned} \quad (5)$$

其中 $\mathbf{w}$ 是学习的权重向量. ρ 是偏移量. 参数 $v \in (0, 1]$ 是受控松弛变量的权重, 它调整可接受的异常值的数量. $\Phi(\cdot)$ 定义用于解决非线性分类问题, 原始输入数据 X 到特征空间 H 的非空集合, 内核函数定义为 $k(\mathbf{x}_i^k, \mathbf{x}_j^k) = \Phi(\mathbf{x}_i^k)^T \Phi(\mathbf{x}_j^k)$ 和等式 (5) 可以通过对偶形式来求解. 在实验中, 本文考虑了 rbf 内核,

$$k(\mathbf{x}_i^k, \mathbf{x}_j^k) = e^{-\frac{\|\mathbf{x}_i^k - \mathbf{x}_j^k\|^2}{2\sigma^2}}. \quad (6)$$

通过求解方程式 (5) 获得最优的 $\mathbf{w}$ 和 ρ . 离群损失可以通过计算以下公式估计:

$$\mathcal{L}_{\text{outlier}} = \rho - \mathbf{w}^T \Phi(\mathbf{x}_t^k). \quad (7)$$

3.3 实现细节

首先, 在数据预处理阶段, 以每秒 30 帧的速度从每个视频中提取帧, 并使用 FlowNet2.0 [22] 估算光流. 输出的光流图都是 RGB 图像, 如图 2 所示, 其颜色表示光流信号的强度. 所有这些图像的大小调整为 32×32 . 在训练过程中, 在训练生成器和辅助编码器的同时, 优化图像重建损失、隐空间表示损失和离群损失. 鉴别器在对抗训练的框架下更新参数. 所有组件都是随机初始化的. 本文使用自适应矩估计 (Adam) 作为优化器, 并将初始学习率设置为 0.002. 对于所有实验, 在标准训练集上进行训练, 并在测试集上进行测试. 此外, 将数据增强 (随机裁剪和水平翻转) 和归一化 (按训练图像的均值和标准差顺序相减和相除) 应用于所有训练图像. 在测试过程中, 隐空间表示损失将用于每一帧图像的异常评分. 为了获得活体检测任务的视频级别的结果, 本文考虑使用视频中每一帧的异常得分来形成一个分布, 即每个视频对应一个分布. 然后, 将训练数据中所有正常人脸视频的异常分数视为参考分布. 给出测试视频后, 通过计算获得其分布, 并计算参考分布与获得的视频之间的距离. 如果此距离大于 T (预定义的阈值), 则该测试样本将被视为异常视频.

最大平均差异 (maximum mean discrepancy, MMD) [23] 是用于比较两个分布之间距离的有效非参数度量. 给定两个分布 s 和 t , 通过使用函数 $\phi(\cdot)$ 将数据映射到再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS), 将 s 和 t 之间的 MMD 定义为

$$\text{MMD}(s, t) = \sup_{\|\phi\|_{\mathcal{H}} \leq 1} \|\mathbb{E}_{\mathbf{x}^s \sim s}[\phi(\mathbf{x}^s)] - \mathbb{E}_{\mathbf{x}^t \sim t}[\phi(\mathbf{x}^t)]\|_{\mathcal{H}}, \quad (8)$$

其中 $\mathbb{E}_{\mathbf{x}^s \sim s}[\cdot]$ 表示对分布 s 的期望, 而 $\|\phi\|_{\mathcal{H}} \leq 1$ 定义 RKHS 单位球中的一组方法. 根据 MMD 定义的统计检验, $\text{MMD}(s, t) = 0 \iff s = t$. 用 $\mathcal{D}_s = \{\mathbf{x}_i^s\}_{i=1}^M$ 和 $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^N$ 的两组样本分别从分布 s 和 t 中采样获得, MMD 的经验估计可以通过以下公式得出:

$$\text{MMD}(\mathcal{D}_s, \mathcal{D}_t) = \left\| \frac{1}{M} \sum_{i=1}^M \phi(\mathbf{x}_i^s) - \frac{1}{N} \sum_{j=1}^N \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}, \quad (9)$$

其中 $\phi(\cdot)$ 表示与内核映射相关的特征映射 $k(\mathbf{x}^s, \mathbf{x}^t) = \langle \phi(\mathbf{x}^s), \phi(\mathbf{x}^t) \rangle$, 通常定义为几个基本内核的凸组合. 在 MMD 的帮助下, 检验方法如下: 基于两种分布的样本, 一种分布是通过训练正常人脸样本而

图 2 (网络版彩图) 前两行显示正常人脸的具有的光流数据可视化的示例. 第 3 行和 4 行显示了通过支架固定的攻击假面的光流图示例. 最后两行显示使用手持假脸的照片或者视频的光流图示例. 对于每一行, 第 1 张图像是原始视频的第一帧图像, 其后是从之后的帧生成的 7 个光流图

Figure 2 (Color online) The first two rows display the live face examples with optical flow data visualization. The third and fourth rows display the fixed support attack face examples by using a stand to hold the client biometry with optical flow data visualization. The last two rows display the hand attack face examples by the attacker holding the device with optical flow data visualization. For each row, the first image is one of the frames from the video, followed by seven optical flow maps which were generated from these frames

形成的参考分布, 而另一种分布是从测试样本以相同的方式获得的. 通过在样本空间中找到连续函数 ϕ , 可以获得函数 ϕ 上不同分布的样本平均值. 将两个平均值相除得出两个分布之间的平均差. 最后, 以 MMD 作为衡量标准来确定测试视频的分类. 如果 MMD 的值小于预定阈值 T , 则认为测试样本分布接近于正常人脸参考分布; 否则, 它们是人脸攻击视频.

4 实验

本节首先介绍实验设置, 包括数据集、评估方法和更多实验细节. 之后进行验证实验和消融实验, 详细分析该方法的可行性和有效性. 最后, 通过在单个数据库训练和测试, 以及跨数据库之间的训练和测试证明提出的方法的鲁棒性.

表 1 活体检测使用的 3 个数据集的介绍
Table 1 A summary of three spoof face datasets

Database	Subject	Video	Camera	Spoof attack	Modal type
NUAA	15	24 genuine 33 spoof	Web-cam (640 × 480)	Printed photo	RGB
REPLAY ATTACK	50	200 genuine 1000 spoof	MacBook 13" camera (320 × 240)	Printed photo Display photo (mobile/HD) Replayed video (mobile/HD)	RGB
CASIA MFSD	50	150 genuine 450 spoof	Low-quality camera (640 × 480) Normal-quality camera (480 × 640) Sony NEX-5 camera (1280 × 720)	Printed photo Cut photo Replayed video (HD)	RGB

4.1 实验设置

数据集. 通常模型训练和测试来自同一数据集的训练和测试样本可以获得很高的准确度, 但是不能保证跨数据集之间的准确度也依然很高. 模型在一个数据集中进行训练, 在另一个数据集中直接进行测试, 这种实验场景更符合现实环境中的需求. 因此, 如表 1, 本文在 3 个公开的活体检测数据库上进行了数据集内部和跨数据集之间的实验: (i) NUAA 数据库^[24] 包含 15 个在不同位置和光照条件下捕获的不同目标. (ii) REPLAY ATTACK^[25] 数据集考虑了攻击中使用的不同光照条件. 通过使用攻击者的手或者固定支架握住存在假脸的屏幕或者照片. (iii) CASIA MFSD^[26] 在活体检测方面, 在图像质量、分辨率和视频长度上带来了更多挑战. 根据文献 [18] 的工作, CIFAR10 和 MNIST 数据集用于构建验证性实验, 以说明提出的方法和其他先进的异常检测方法相比的优越性. 其中一个类将被视为正常类, 而其余的则属于异常类. 总共, 分别获得了十组 MNIST 数据集和 CIFAR10 数据集, 然后仅通过在正常类别上训练模型检测测试样本是否异常.

评估指标. 本文采用的性能指标是曲线下面积 (area under the curve, AUC) 和半总错误率 (half total error rate, HTER)^[27]. HTER 是错误拒绝率 (false rejection rate, FRR) 和错误接受率 (false acceptance rate, FAR) 之和的一半. 面对某些未知攻击, 在计算指标时没有设置验证集来调整模型和阈值. 因此, 根据训练集确定阈值, 并针对所有测试样本使用固定的测试方法与在文献 [28] 中所做的实验设置保持一致.

实施细节. 在 PyTorch 中通过优化加权值 $\mathcal{L}$ (在等式 (1) 中已经定义) 实现本文提出的方法 $w_i = 1$, $w_a = 5$, $w_z = 1$ 和 $w_o = 0.05$, 根据网络搜索的方法选择这些参数, 最终产生最佳结果. 实验是在配备 NVIDIA-1080 GPU 和多核 2.1 GHz CPU 的标准 PC 上进行的.

验证性实验. 本文提出的方法比最先进的半监督异常检测器具有明显的优势, 如图 3(b) 所示. 对于 MNIST 数据集, 每次选择一个类别作为正常, 而将其余类别保留为异常类别, 从而获得十组异常检测实验. 与其他工作相比, 在 AUC 值方面, 本文提出的方法与包括 EGBAN^[16], GANomaly^[18], AnoGAN^[29] 和可变自动编码器管道 VAE^[21] 的其他前沿的方法相比有明显的效果提升. 可以通过使用图 3 中的红色曲线说明. 图 3(a) 介绍了在 CIFAR10 数据集上的比较. 图 4 描述了超参数的选择最终如何影响模型的整体性能. 从图中可以看到对于 CIFAR10 和 MNIST 数据集, 当潜在向量 z 的大小为 100 时, 可以实现最佳性能. 因此, 对于所有实验此隐特征向量的大小均保持固定.

图 3 (网络版彩图) 在 CIFAR10 和 MNIST 上进行验证性实验, 其中一个类为正常类, 其他类为异常类
Figure 3 (Color online) The experiments are conducted on (a) CIFAR10 and (b) MNIST, one class is considered as normal sample and the others are considered as abnormal sample

图 4 (网络版彩图) 在 CIFAR10 和 MNIST 数据集的实验中, 通过改变隐空间特征向量的大小, 获得最佳效果的超参数
Figure 4 (Color online) Overall performance of the model based on varying size of the latent vector in (a) CIFAR10 and (b) MNIST dataset

表 2 在单个数据集内进行训练和测试的消融实验结果

Table 2 The effect of different loss compositions is evaluated in intra dataset experiment

Loss composition	HTER (%) in intra-dataset experiment	
	NUAA	REPLAY ATTACK
$\mathcal{L}_{\text{irec}} + \mathcal{L}_{\text{adv}}$	27.7	38.1
$\mathcal{L}_{\text{irec}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{zrec}}$	19.9	10.5
$\mathcal{L}_{\text{irec}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{ou}}$	0	10.2
$\mathcal{L}_{\text{irec}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{zrec}} + \mathcal{L}_{\text{ou}}$	0	10.1

4.2 消融实验

由于所提出的架构是建立在对抗训练框架之上的, 在本小节之前已经介绍了各个损失函数. 为了

表 3 在交叉数据集之间进行训练和测试的消融实验结果

Table 3 For cross domain experiment, four loss compositions are optimized in CASIA MFSD and evaluated in REPLAY ATTACK, and vice versa

Loss composition	HTER (%)			
	Train		Test	
	REPLAY ATTACK	CASIA MFSD	CASIA MFSD	REPLAY ATTACK
$\mathcal{L}_{irec} + \mathcal{L}_{adv}$	44.8		31.5	
$\mathcal{L}_{irec} + \mathcal{L}_{adv} + \mathcal{L}_{zrec}$	39.8		25.0	
$\mathcal{L}_{irec} + \mathcal{L}_{adv} + \mathcal{L}_{ou}$	37.7		20.1	
$\mathcal{L}_{irec} + \mathcal{L}_{adv} + \mathcal{L}_{zrec} + \mathcal{L}_{ou}$	35.3		13.5

图 5 (网络版彩图) 本文提出的方法在交叉数据集之间的实验中, 获得了在三维空间中的特征表达. (a) 表示将模型在 CASIA MFSD 中训练, 在 REPLAY ATTACK 中进行测试的实验中不上离群点约束项, (b) 表示将模型在 CASIA MFSD 中训练, 在 REPLAY ATTACK 中进行测试的实验中加上离群点约束项

Figure 5 (Color online) Proposed method learned features in 3-dimensional space in cross domain experiment between CASIA MFSD and REPLAY ATTACK datasets. (a) Model is trained in CASIA MFSD dataset and tested in REPLAY ATTACK dataset without outliers constraint. (b) Model is trained in CASIA MFSD dataset and tested in REPLAY ATTACK dataset with outliers constraint

验证各个组件的有效性, 有必要进行消融研究. 将生成器、鉴别器和辅助编码器的损失函数进行不同组合在单个数据集和跨数据集之间进行实验. 如表 2 所示, 对于单个数据集内部实验, 它表明将离群损失 $\mathcal{L}_{ou}$ 和隐空间特征重建损失 $\mathcal{L}_{zrec}$ 组合在一起比单独使用 $\mathcal{L}_{ou}$ 或 $\mathcal{L}_{zrec}$ 能更好地提升效果. 在表 3 中, 即数据集间实验中, 可以看到更显著的提升. 特别地, 与 $\mathcal{L}_{zrec}$ 相比, 离群损失会进一步约束隐空间的表示, 从而获得更精准和鲁棒的活体检测器. 总的来说, 本文提出的方法设计的任何损失都可以提高检测的准确性.

为了进一步了解离群约束项的重要性, 本文将学习到的隐特征表示进行了可视化, 以进行直观的解释, 如图 5 所示, 在 CASIA MFSD 和 REPLAY ATTACK 数据集中, 3D 散点图用于可视化测试样本的潜在表示 z , 其坐标是通过 PCA 对特征向量进行降维得到的. 显然, 与不使用离群约束项相比, 通过利用离群约束项可以得到正常样本 (黄色点) 更紧密的隐空间特征. 在图 5(b) 中, 不仅可以很容易地观察到异常样本可以与正常样本分离, 而且可以看见异常样本在隐空间被映射到正常数据边界的外围.

图 6 (网络版彩图) 本文提出的方法在交叉数据集之间的实验中, 获得了在三维空间中的特征表达. (a) 表示将模型在 REPLAY ATTACK 中训练, 在 CASIA MFSD 中进行测试的实验中不加上离群点约束项, (b) 表示将模型在 REPLAY ATTACK 中训练, 在 CASIA MFSD 中进行测试的实验中加上离群点约束项

Figure 6 (Color online) Proposed method learned features in 3-dimensional space in cross domain experiment between CASIA MFSD and REPLAY ATTACK datasets. (a) Model is trained in REPLAY ATTACK dataset and tested in CASIA MFSD dataset without outliers constraint. (b) Model is trained in REPLAY ATTACK dataset and tested in CASIA MFSD dataset with outliers constraint

图 7 (网络版彩图) 显示了在具有离群约束项 (a) 和没有离群约束项 (b) 的训练模型中, 正常样本 (活脸) 和异常样本 (手持的攻击或固定支持攻击) 的异常得分分布. 横轴表示每个视频样本的异常分数. 纵轴表示每个异常得分的频率

Figure 7 (Color online) The distribution of the abnormal scores for both normal samples (live face) and abnormal samples (hand attack or fixed support attack) in the trained model with outlier constraint (a) and without outlier constraint (b). The horizontal axis denotes the abnormal scores of each video sample. The vertical axis denotes the frequency of each abnormal score

图 6 显示了在交叉数据库之间的实验 (在 REPLAY ATTACK 中进行训练以及在 CASIA MFSD 中进行测试) 中存在类似结果.

通过生成频率直方图表示图中所有测试样本的异常分数分布展示离群约束项的有效性. 如图 7

表 4 在 NUAA 数据集上通过 HTER, 将本文方法与其他方法进行的比较 (%)

Table 4 Performance comparison using accuracy measure on the NUAA dataset with other methods (%)

Method	Accuracy
Ours	99.3
ADKMM (2017) [30]	99.3
ND-CNN (2016) [4]	99.3
DS-LSP (2015) [31]	98.5
CDD (2013) [7]	97.7
DoG-LRBLR (2010) [24]	87.5
DoG-F (2004) [32]	84.5
DoG-M (2012) [26]	81.8

表 5 在 REPLAY ATTACK 数据集上通过 HTER, 将本文方法与其他方法进行的比较 (%)

Table 5 Performance comparison using HTER measure on the REPLAY ATTACK dataset with other methods (%)

Method	Development	Test
$LBP_{3 \times 3}^{u2} + x^2$ (2012) [8]	31.24	34.01
$LBP_{3 \times 3}^{u2} + LDA$ (2012) [8]	19.60	17.17
$LBP_{3 \times 3}^{u2}$ (2012) [8]	14.84	15.16
LBP+SVM (2011) [33]	13.90	13.87
DS-LBP (2015) [31]	13.73	12.50
ND-CNN (2016) [4]	—	10
ADKMM (2017) [30]	5.16	4.30
Ours	11.5	10.1

所示, 比较结果基于交叉数据集之间的实验, 在 CASIA MFSD 中进行的训练和在 REPLAY ATTACK 中进行的测试, 包括真实面孔 (绿色), 使用支架固定的存在攻击人脸的屏幕或者照片 (蓝色), 使用双手握住存在攻击人脸的屏幕或者照片 (红色). 与没有离群约束项的网络结构相比, 本文提出的模型结果可以表明以下几点: (1) 异常样本的分布 (固定支撑的攻击人脸/手持的攻击人脸) 可以与正常样本 (真实人脸) 的分布更分离. (2) 具有离群约束的数据分布, 对于正常或异常测试样本更加集中, 请参见黑色虚线将绿色区域与其他区域分开. (3) 由于离群约束项, 对应于不同异常类型的区域也更加独立, 因为蓝色区域和红色区域之间的重叠相对较小 (请参见蓝色虚线).

4.3 最新方法的比较

数据集内实验. 在 NUAA 数据集上比较了本文的方法和最新方法. 表 4 [4, 7, 24, 26, 30~32] 显示所有方法的准确度. 本文提出的半监督方法达到了与有监督学习的 ADKMM [30] 和 ND-CNN [4] 同等的最佳性能水平. 还将本文的方法与表 5 [4, 8, 30, 31, 33] 中 REPLAY ATTACK 数据集上的最新的技术进行了比较. 实现了与其他有监督的方法可比的 HTER. 该结果归因于两个因素. 首先, 需有标签的数据的二分类方法通常在单个数据库内进行训练和测试具有出色的性能. 一些比较的方法甚至使用深度信息或其他额外信息来进行活体检测. 其次, 在提出的方法中, 许多固定支撑攻击人脸被错误地识别为正常脸. 由于固定支撑的攻击人脸上没有光流信息, 因此隐空间中的特征重建损失较低.

数据集间实验. 本文按照之前的相关文献 [34~36] 在交叉数据集 (包含 CASIA MFSD 和 RE-

表 6 使用 CASIA MFSD 数据集对模型进行训练,并在 REPLAY ATTACK 数据集上进行测试,和使用 REPLAY ATTACK 数据集对模型进行训练,并在 CASIA MFSD 数据集上进行测试 (%)

Table 6 Classification performance of the proposed approach in terms of HTER (%). The algorithm is trained using the CASIA MFSD dataset and tested on the REPLAY ATTACK dataset, and vice versa

Method	Train	Test	Train	Test	Average
	CASIA MFSD	REPLAY ATTACK	REPLAY ATTACK	CASIA MFSD	
LBP (2013) [5]		47.0	39.6		43.3
LBP-TOP (2013) [5]		49.7	60.6		55.2
Motion (2013) [5]		50.2	47.9		49.1
CNN (2014) [9]		48.5	45.5		47.0
Color LBP (2018) [35]		37.9	35.4		36.7
Color Tex (2018) [35]		30.3	37.7		34.0
Color SURF (2018) [35]		26.9	23.2		25.1
Auxiliary (2018) [36]		27.6	28.4		28.0
De-Spoof (2018) [34]		28.5	41.1		34.8
GFA-CNN (2019) [37]		21.4	34.3		28.0
Proposed method		13.5	35.3		24.4

PLAY ATTACK 数据集) 上进行实验评估,其中包括对 CASIA MFSD 数据库的训练集进行训练,并对 REPLAY ATTACK 数据库的测试集进行测试. 表 6 [5, 9, 34~37] 证实了所提出的方法在包含不同类型的欺骗攻击的 REPLAY ATTACK 测试集中达到了最佳性能 (HTER = 0.135). 然后,在进行相反的实验中,在 REPLAY ATTACK 数据集的训练集上进行训练,并在 CASIA MFSD 数据集的测试集上进行测试. 本文方法在 CASIA MFSD 数据集的测试集上进行交叉测试,具有可比的效果 (HTER = 0.353). 从表 6 中,可以看到提出的方法的 HTER 优于大多数二分类有监督方法 (CNN), De-Spoof, Texture (LBP, Color Tex, Color LBP, LBP-TOP). 这表明所提出的方法可以更好地识别真假人脸之间的差异. 基于辅助深度 rPPG 的方法比提出的方法更好,这可能是由于深度信息和 rPPG 信号的帮助. 与所有先前的研究一样,本文观察到在 CASIA MFSD 上训练的模型比在重放攻击数据库上训练的模型能够更好地泛化. 推测如下: (1) 可能是因为 CASIA MFSD 数据集中的分辨率明显高于 REPLAY ATTACK 数据集中的分辨率. 高分辨率训练的模型可以比低分辨率训练的模型更好地泛化. (2) 与 REPLAY ATTACK 相比, CASIA MFSD 在收集的数据库中包含更多变化,例如,成像质量、相机与面部之间的距离、背景和攻击类型. 因此,为呈现攻击数据库优化的模型在新的采集条件下面临更多挑战. 这是本文方法和以前工作的局限性,值得进一步研究.

5 结论

为了提高人脸反欺骗的泛化能力,本文利用异常检测技术,仅在正常类上训练模型,即可学习广义特征空间. 具体而言,通过在半监督学习框架下进行对抗训练,提出了一种新颖的深度网络. 训练数据的原始属性和特征不仅可以在图像重建空间中捕获,而且还可以通过判别的方式进一步约束在隐空间中,从而导致更强大的活体检测器. 在一些数据集上进行了广泛的实验,显示了在半监督学习框架上训练的模型的高泛化能力以及使用隐空间约束项的好处.

参考文献

- 1 Galbally J, Marcel S, Fierrez J. Biometric antispoofing methods: a survey in face recognition. *IEEE Access*, 2014, 2: 1530–1552
- 2 Pan G, Sun L, Wu Z, et al. Eyeblink-based anti-spoofing in face recognition from a generic webcam. In: *Proceedings of IEEE International Conference on Computer Vision*, 2007. 1–8
- 3 Frischholz R W, Werner A. Avoiding replay-attacks in a face recognition system using head-pose estimation. In: *Proceedings of IEEE International SOI Conference*, 2003. 234–235
- 4 Alotaibi A, Mahmood A. Deep face liveness detection based on nonlinear diffusion using convolution neural network. *Signal Image Video Process*, 2017, 11: 713–720
- 5 de Freitas Pereira T, Anjos A, de Martino J M, et al. Can face anti-spoofing countermeasures work in a real world scenario? In: *Proceedings of International Conference on Biometrics*, 2013. 1–8
- 6 Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. *IEEE Trans Inform Forensics Secur*, 2016, 11: 1818–1830
- 7 Yang J, Lei Z, Liao S, et al. Face liveness detection with component dependent descriptor. In: *Proceedings of International Conference on Biometrics*, 2013. 1–6
- 8 Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: *Proceedings of the International Conference Of Biometrics Special Interest Group*, 2012. 1–7
- 9 Yang J W, Lei Z, Li S Z. Learn convolutional neural network for face anti-spoofing. *Comput Sci*, 2014, 9218: 373–384
- 10 Xu Z, Li S, Deng W. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In: *Proceedings of IAPR Asian Conference on Pattern Recognition*, 2015. 141–145
- 11 Lu C, Shi J, Jia J. Abnormal event detection at 150 FPS in MATLAB. In: *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013. 2720–2727
- 12 Chong Y S, Tay Y H. Abnormal event detection in videos using spatiotemporal autoencoder. In: *Proceedings of International Symposium on Neural Networks*, 2017. 189–196.
- 13 Wang L, Zhou F, Li Z, et al. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In: *Proceedings of IEEE International Conference on Image Processing*, 2018. 2276–2280
- 14 Baur C, Wiestler B, Albarqouni S, et al. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: *Proceedings of International MICCAI Brainlesion Workshop*, 2018. 161–169
- 15 Zong B, Song Q, Min M R, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: *Proceedings of International Conference on Learning Representations*, 2018. 1–19
- 16 Zenati H, Foo C S, Lecouat B, et al. Efficient GAN-based anomaly detection. 2018. ArXiv:1802.06222
- 17 Liu R, Fusi N, Mackey L. Teacher-student compression with generative adversarial networks. 2018. ArXiv:1812.02271
- 18 Akcay S, Atapour-Abarghouei A, Breckon T P. GANomaly: semi-supervised anomaly detection via adversarial training. In: *Proceedings of Asian Conference on Computer Vision*, 2018. 622–637
- 19 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014. 2672–2680
- 20 Isola P, Zhu J-Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1125–1134
- 21 An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. In: *Special Lecture on IE*, 2015. 2: 1–18
- 22 Ilg E, Mayer N, Saikia T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2462–2470
- 23 Yan H L, Ding Y K, Li P H, et al. Mind the class weight bias: weighted maximum mean discrepancy for unsupervised domain adaptation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2272–2281
- 24 Tan X, Li Y, Liu J, et al. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: *Proceedings of European Conference on Computer Vision*, 2010. 504–517
- 25 Chingovska I, Anjos A, Marcel S, et al. On the effectiveness of local binary patterns in face anti-spoofing. In: *Proceedings of the International Conference of Biometrics Special Interest Group*, 2012. 1–7
- 26 Zhang Z W, Yan J J, Liu S F, et al. A face antispoofing database with diverse attacks. In: *Proceedings of IAPR*

- International Conference on Biometrics, 2012. 26–31
- 27 Bengio S, Mariethoz J. A statistical significance test for person authentication. In: Proceedings of the Speaker and Language Recognition Workshop, 2004
 - 28 Liu Y J, Stehouwer J, Jourabloo A, et al. Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 4680–4689
 - 29 Schlegl T, Seebock P, Waldstein S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Proceedings of International Conference on Information Processing in Medical Imaging, 2017. 146–157
 - 30 Yu C, Jia Y. Anisotropic diffusion-based kernel matrix model for face liveness detection. 2017. ArXiv:1707.02692
 - 31 Kim W, Suh S, Han J J. Face liveness detection from a single image via diffusion speed model. IEEE Trans Image Process, 2015, 24: 2456–2465
 - 32 Li J W, Wang Y H, Tan T N, et al. Live face detection based on the analysis of Fourier spectra. In: Proceedings of SPIE, 2004. 5404: 296–303
 - 33 Maatta J, Hadid A, Pietikainen M, et al. Face spoofing detection from single images using micro-texture analysis. In: Proceedings of International Joint Conference on Biometrics, 2011. 1–7
 - 34 Jourabloo A, Liu Y, Liu X. Face de-spoofing: anti-spoofing via noise modeling. In: Proceedings of European Conference on Computer Vision, 2018. 290–306
 - 35 Boulkenafet Z, Komulainen J, Hadid A. Face anti-spoofing based on color texture analysis. In: Proceedings of International Conference on Image Processing, 2015. 2636–2640
 - 36 Liu Y J, Jourabloo A, Liu X M. Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 389–398
 - 37 Tu X G, Zhao J, Xie M, et al. Learning generalizable and identity-discriminative representations for face anti-spoofing. ACM Trans Intell Systems and Technol, 2020, 11: 60

Latent regularized generative adversarial network for face spoofing detection

Chengwei CHEN¹, Wang YUAN¹, Pan CHEN², Shouhong DING³, Yuan XIE^{1*},
Haichuan SONG^{1*} & Lizhuang MA^{1*}

1. School of Computer Science and Technology, East China Normal University, Shanghai 200241, China;

2. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;

3. Tencent Youtu Lab, Shanghai 200233, China

* Corresponding author. E-mail: xieyuan8589@foxmail.com, hcsong@cs.ecnu.edu.cn, lzma@cs.ecnu.edu.cn

Abstract Face anti-spoofing technology is critical to prevent face recognition systems from experiencing a security breach. Most of presentation attack detection (PAD) methods consider the task as a supervised binary classification problem. Many of these methods struggle to grasp adequate spoofing cues and generalize poorly. In this paper, we formulate the face anti-spoofing detection as an anomaly detection task to tackle the generalization issue. A novel deep network is proposed by using adversarial training under semi-supervised learning framework. The underlying structure of training data is captured in the image reconstruction space and can be further restricted in the space of latent representation in a discriminant manner, leading to a more robust spoof detector. In the test, the attacks are regarded as out-of-distributions samples that naturally exhibit a higher feature reconstruction error in the latent space than real samples in the dataset. Experiments show that our model is clearly superior over cutting-edge semi-supervised abnormal detectors and achieves state-of-the-art results on both intra- and inter-database testing.

Keywords adversarial networks, live face detection, semi-supervised learning, anomaly detection, face spoofing attacks

Chengwei CHEN was born in 1991. He received his master's degree in Computer Science from Liverpool University, UK, in 2017. Currently, he is pursuing his Ph.D. degree in the Department of Computer Science and Technology, East China Normal University. His research interest includes anomaly detection.

Yuan XIE was born in 1985. He is currently a professor at the School of Computer Science and Technology of East China Normal University. In 2010 and 2013, he obtained master's and doctoral degrees in Engineering from the Department of Computer, Xiamen University and the Institute of Automation, Chinese Academy of Sciences, respectively. His research interests include artificial intelligence, computer vision, and machine learning.

Haichuan SONG was born in 1986. In 2015, he graduated from the Department of Computer Science and Technology of Tsinghua University. He has engaged in research on computer graphics and computer-aided design. Since 2018, he has worked as an associate researcher at the School of Computer Science and Software Engineering of East China Normal University. His research interests include artificial intelligence, computer vision, machine learning, computer graphics, and computer animation.

Lizhuang MA was born in 1963. He received his B.S. and Ph.D. degrees from Zhejiang University, China, in 1985 and 1991, respectively. He is now a distinguished professor at the Department of Computer Science and Engineering, East China Normal University, China. His research interests include computer vision, computer-aided geometric design, computer graphics, computer animation, and theory and applications of computer graphics.