计算机科学
COMPUTER SCIENCE

具有对抗鲁棒性的人脸活体检测方法

王春东, 李泉, 付浩然, 浩庆波

引用本文

王春东, 李泉, 付浩然, 浩庆波. 具有对抗鲁棒性的人脸活体检测方法[J]. 计算机科学, 2024, 51(6A): 230400022-7.

WANG Chundong, LI Quan, FU Haoran, HAO Qingbo. Face Anti-spoofing Method with Adversarial Robustness [J]. Computer Science, 2024, 51(6A): 230400022-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400022

具有对抗鲁棒性的人脸活体检测方法

王春东　李　泉　付浩然　浩庆波
天津理工大学计算机科学与工程学院　天津 300384
天津理工大学"智能计算及软件新技术"天津市重点实验室　天津 300384
(michael3769@163.com)

摘　要　现有人脸活体检测方法在深度神经网络的支持下已获得优秀的检测能力,但面临对抗样本攻击时仍呈现脆弱性。针对此问题,引入胶囊网络(Capsule Network,CapsNet)提出一种具有对抗鲁棒性的人脸活体检测方法 FAS-CapsNet:通过 CapsNet 及其图像重建机制保留特征间关联,过滤样本中的对抗扰动;根据皮肤与平面介质的反射性质差异,以 Retinex 算法增强图像光照特征,增大活体与非活体人脸类间距离的同时破坏对抗扰动模式,进而提升模型准确性与鲁棒性。在 CASIA-SURF 数据集上进行实验可知:FAS-CapsNet 对正负样本的检测准确率为 87.344%,对比模型中最高准确率为 78.917%,说明 FAS-CapsNet 具备充分的常规活体检测能力。为进一步验证模型鲁棒性,基于 CASIA-SURF 测试集生成两种对抗样本数据集并进行实验:FAS-CapsNet 在两数据集上的检测准确率分别为 84.552%和 79.042%,较常规检测准确率下降 3.197%和 9.505%;对比模型在两数据集上的最高准确率分别为 74.938%和 41.667%,较常规检测下降 5.042%和 47.201%。可见 FAS-CapsNet 受对抗扰动影响更小,具有显著的对抗鲁棒性优势。

关键词:人脸活体检测;对抗鲁棒性;胶囊网络;Retinex;对抗样本
中图分类号　TP391

Face Anti-spoofing Method with Adversarial Robustness

WANG Chundong,LI Quan,FU Haoran and HAO Qingbo
School of Computer Science and Engineering,Tianjin University of Technology,Tianjin 300384,China
Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology,Tianjin University of Technology,Tianjin 300384,China

Abstract　The existing face anti-spoofing methods based on deep neural networks perform excellently now,but they are absolute weak when facing adversarial examples. To solve the problem,capsule network(CapsNet) is introduced to propose an adversarial robust method called FAS-CapsNet. The capsule structure and reconstruction mechanism of CapsNet are utilized to retain the correlation between features and filter the adversarial perturbations in images. The Retinex algorithm is utilized to enhance illumination features which show the difference of reflection properties between skin and planar medium,increasing the between-class distance of living and spoof faces and destroying the very adversarial perturbation modes in images,improving the accuracy and robustness of FAS-CapsNet. Experiments on CASIA-SURF show that the spoofing detection accuracy of FAS-CapsNet is 87.344%,and the highest accuracy of comparison models is 78.917%,which demonstrates that FAS-CapsNet is capable to solve general face anti-spoofing problems. This paper further generates two adversarial datasets from CASIA-SURF validation set to verify the robustness of each model. The accuracy of FAS-CapsNet on the two datasets is 84.552% and 79.042% respectively,which decreases by 3.197% and 9.505% compared to the previous results. The highest accuracy of comparison models on adversarial datasets is 74.938% and 41.667% respectively,which is 5.042% and 47.201% lower than that of the conventional detection. It proves that FAS-CapsNet is significantly robust in adversarial attacks.

Keywords　Face anti-spoofing,Adversarial robustness,CapsNet,Retinex,Adversarial examples

1　引言

目前,人脸识别技术已广泛应用于移动支付、门禁、安检等场景,在给大众生活带来便利的同时也引发了新的安全问题。由于人脸特征具有唯一性和不易更改性,一旦被盗用就将导致与其关联的众多信息系统失守,严重威胁用户的隐私信息和财产账户安全,且用户在未来可能难以注册和使用其他人脸认证系统。诸多案例[1-2]表明:利用他人的人脸特征欺骗认证系统并非难事。只有从技术角度开展研究,探索可靠的特征保护和系统防御方法,才能有效提升人脸认证系统的安全性与可信度。

呈现攻击(Presentation Attack)是威胁人脸认证系统的攻击方式之一,攻击者常将盗取、伪造的合法用户人脸特征展示在系统摄像头前来欺骗认证系统。这类攻击的主要形式有照片演示攻击、视频重放攻击和 3D 面具攻击[3-4],其都是通过非活体媒介(纸张、屏幕、硅胶面具等)携带目标用户特征实

基金项目:科技助力经济 2020 重点专项(SQ2020YFF0413781)
This work was supported by the Science and Technology for Economy 2020 Key Project of China(SQ2020YFF0413781).
通信作者:李泉(shadow.lee20@qq.com)

施的。因此，可以利用人脸活体检测（Face Anti-Spoofing，FAS）方法检查被测人脸的活体特征，从而有效防御此类攻击。

随着深度学习的快速发展，基于深度神经网络（Deep Neural Network，DNN）的人脸活体检测方法大量涌现并表现出卓越的检测性能。然而，DNN 本身存在易受对抗样本干扰的弱点，这使对抗样本攻击成为现有活体检测方法面临的一大安全威胁[5-7]。为此，本文重点针对此问题展开研究，考察现有方法在对抗攻击中的表现，并提出一种具有显著对抗鲁棒性的人脸活体检测方法。本文主要工作有：

（1）引入并优化 CapsNet[8]，提出具有显著对抗鲁棒性的静默式人脸活体检测方法 FAS-CapsNet：在单帧非活体人脸图像中含有对抗扰动的条件下，仍可维持较高、较稳定的活体检测准确性。

（2）采用 Retinex 滤波算法增强（1）中工作：一方面破坏对抗扰动的特定噪声模式，保护图像原始信息，进一步提升模型的对抗鲁棒性；另一方面通过提取图像的光照特征反映活体人脸与非活体人脸间的光反射性质差异，增大样本类间距离，进一步提升检测准确性。

2 相关工作

现有活体检测方法可根据用户参与形式分为交互式与静默式。交互式检测方法[9-11]通过向用户施放“张嘴”“眨眼”等指令，捕捉动作并比对其与指令的一致性来判断待测人脸是否为活体。但对用户而言，若在响应指令时出现失误（如未在时限内完成所有指令），将导致认证过程中断并重启，操作体验不佳，因而逐渐被静默式方法替代。

静默式活体检测无需用户参与动作交互，而是直接根据实时捕捉视频中的某帧或某几帧图像来判定待测人脸活性。由于照片、视频在翻拍后通常会改变或丢失细节信息（如边缘模糊、亮度变化），可通过衡量图像的纹理[12-20]、色彩[13-20]、质量[4,20-22]等增强活体与非活体人脸图像间的差异，再利用统计分析方法进行分类判别。这些方法多基于人工设定的特征模式，性能更取决于设计者的知识经验。

而在深度学习的助力下，通过 DNN 提取的图像深层特征为静默活体检测带来重大的性能突破。Zhang 等[23]发布了大规模多模态人脸活体检测数据集 CASIA-SURF，并基于深度残差网络（Deep Residual Network，ResNet）提出一种多模态融合的活体检测方法，在该数据集上获得 2.4% 的平均误判率。Zhang 等[24]提出一种极轻量的多模态级联融合模型 FeatherNets，将 CASIA-SURF 上的平均误判率降至 0.13%。Shen 等[25]提出一种多流卷积神经网络架构，运用各模态图像的补丁级数据集训练对应子模型，并加入模态特征擦除机制缓解过拟合，在 CASIA-SURF 上的平均误判率仅为 0.0985%。大量研究[23-29]证明了 DNN 在活体检测领域的应用价值，也使基于 DNN 的静默式检测方法成为主流研究方向。

然而 Akhtar 等[30]指出，DNN 极易受到对抗样本的威胁，图像中细微的对抗扰动足以导致模型预测结果的完全错误。对人脸活体检测这种二分类任务而言，分类错误将使其从一种安防机制转变为重大系统漏洞。Ma 等[5]提出一种针对活体检测的对抗样本生成算法，基于人眼视觉特性加入扰动间距约束，只对图像的极少维度进行扰动便生成不易察觉的对抗样本，使人脸照片被错判为活体。Zhang 等[6]更提出面向物理域的对抗样本生成算法，成功欺骗以 DNN 为核心的活体检测模块。可见，基于 DNN 的活体检测方法在对抗样本攻击中呈现明显的鲁棒性不足。

针对上述问题，本文结合 CapsNet 在对抗样本防御中的潜能[8,31-33]提出一种具有显著对抗鲁棒性的人脸活体检测方法 FAS-CapsNet。通过胶囊结构保留特征间的关联，可降低局部扰动对模型在全局特征识别和决策过程中的影响，且在训练中采用的图像重建机制也能有效弱化对抗扰动。此外，本文以 Retinex 图像增强算法提高活体与非活体人脸图像间的区分度，在提升活体检测准确率的同时过滤检测图像中的部分对抗扰动，降低对抗攻击的成功率，进而帮助提升模型的对抗鲁棒性。

3 FAS-CapsNet 人脸活体检测方法

FAS-CapsNet 方法流程如图 1 所示，分为预处理、光照特征提取、深度特征提取和识别分类 4 个步骤。为适应静默检测的实时性要求，减少计算开销，本文采用单模态 RGB 图像训练活体检测模型，并将数据集样本预处理为 32×32 灰度图，以提升模型训练效率。处理后的图像经 Retinex 算法提取光照特征，再由 CapsNet 模块结合光照特征进一步构造深度特征胶囊。CapsNet 模块包含两个胶囊层：初级胶囊层（Primary Capsule Layer）和 R-S 胶囊层（Real-Spoof Capsule Layer）。低层胶囊完成向高层的动态路由匹配后，使用 Softmax 分类器对 R-S 胶囊层输出向量进行概率预测。

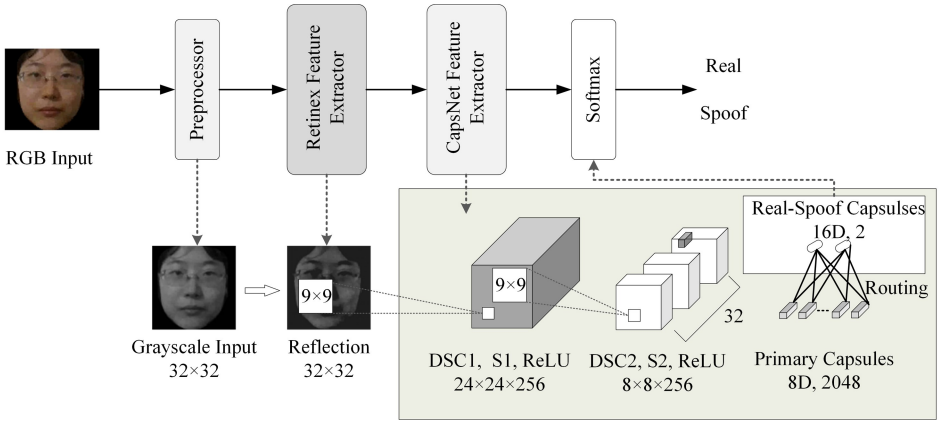


图 1 FAS-CapsNet 人脸活体检测方法流程

Fig. 1 Flowchart of FAS-CapsNet face anti-spoofing method

3.1 光照特征提取

由于真实人脸具有立体结构及独特的皮肤质地，与纸张、屏幕等平面媒介相比具有截然不同的光反射性质，故本文根据 Land[34] 提出的 Retinex 理论提取图像中的反射光特征，进一步提升分类器对样本的识别能力。

Retinex 理论认为：人眼、相机接收到的图像实际是由照射物体的光线与物体表面反射的光线共同作用形成。因此，可将一幅图像 $f(x,y)$ 表示为：

$$f(x,y)=L(x,y)\cdot R(x,y) \tag{1}$$

其中，$L(x,y)$ 为图像 f 的入射光分量(Lumination Part)，$R(x,y)$ 为反射光分量(Reflection Part)。由于入射光分量的形成受多种不稳定因素影响，没有精确的计算方式，多使用滤波算法提取图像的低频部分作为近似量

$$\widetilde{L}(x,y)\approx H(x,y)*f(x,y) \tag{2}$$

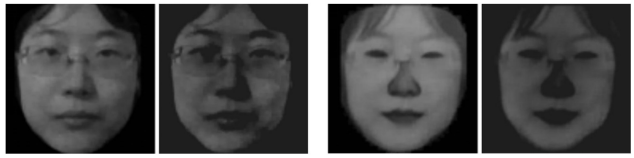
其中，$*$ 表示卷积运算，$H(x,y)$ 为滤波函数，本文中为高斯滤波器：

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

反射光分量 R 也以近似值表示为：

$$\begin{aligned}\widetilde{R}(x,y)&=\lg R(x,y)\\&=\lg f(x,y)-\lg\widetilde{L}(x,y)\\&=\lg f(x,y)-\lg(G(x,y,\sigma)*f(x,y))\end{aligned} \tag{4}$$

经 Retinex 分解出的分量图像如图 2 所示。各子图中左图为原始图像，右图为反射光分量图像。对比真实人脸和伪造人脸的反射光分量图像可见：前者各区域间存在轻微亮度变化且过渡平滑；而由于照片本身是平面结构，后者中大面积区域呈现几乎一致的亮度水平，且照片被剪除区域内亮度整体偏低，脸颊、下颌附近未能像真实人脸一样产生自然的亮度变化。



(a)真实人脸　　(b)伪造人脸

图 2　真伪人脸的反射光图像示例

Fig. 2　Reflection image of real and spoof faces

3.2 深度特征胶囊生成与动态路由匹配

本文通过 CapsNet 提取图像深层特征，以胶囊形式将特征成组表示和传递，保留内在关联信息，从而降低模型受特征级对抗扰动的影响程度。同时，为提高模型检测效率，本文采用深度可分离卷积(Depthwise Separable Convolution，DSC)对 CapsNet 进行轻量化优化，使模型卷积核参数量压缩为原先的 1.63%，大大降低了计算开销。

应用 DSC 的 CapsNet 结构如图 3 所示。DW Conv1 层对反射光分量 $\widetilde{R}$ 进行步长为 1 的 9×9 逐通道卷积，生成尺寸为 24×24×1 的中间特征图 F_1'，再由 PW Conv1 层进行逐点卷积，将其扩展为 256 通道，获得在更高维空间的特征映射。升维后特征图尺寸为 24×24×256，至此完成第一次 DSC 过程，其卷积核参数量共计 9×9×1+1×1×256=337，仅是常规卷积(9×9×1×256=20 736)的 1.63%。之后进行第二次 DSC，DW 卷积的滑动步长增加为 2，得到尺寸为 8×8×256 的特征图 F_2，卷积核参数量也是常规卷积参数量的 1.63%。可见，优化后的 CapsNet 模型体量明显减轻，能有效减少计算开销。

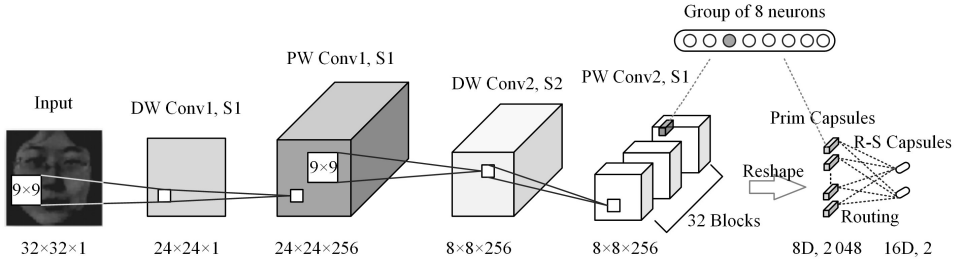


图 3　应用深度可分离卷积的 CapsNet 结构

Fig. 3　Architecture of CapsNet with depthwise separable convolution

对 F_2 进行张量整形生成深度特征胶囊，以 8 通道为单位将所有特征图通道均等拆分为 32 组，各组内按像素位拆分为 8×8=64 个 8 维向量胶囊，则初级胶囊层共有 64×32=2048 个单元，记为：

$$U^{(\mathrm{Prim})}=\{u_0,\cdots,u_i,\cdots,u_{2047}\},u_i\in R^8 \tag{5}$$

同样，R-S 胶囊层可表示为：

$$U^{(\mathrm{R\text{-}S})}=\{u_0,u_1\},u_j\in R^{16} \tag{6}$$

其中，$u_0^{(\mathrm{R\text{-}S})}$ 表示伪造人脸对应的 Spoof 胶囊，$u_1^{(\mathrm{R\text{-}S})}$ 表示真实人脸对应的 Real 胶囊。低层胶囊 $u^{(\mathrm{Prim})}$ 与高层胶囊 $u^{(\mathrm{R\text{-}S})}$ 间按照算法 1 所示动态路由算法进行特征匹配传递，$u^{(\mathrm{R\text{-}S})}$ 的输出向量则通过 Softmax 分类器进行类别概率预测。

算法 1　动态路由算法

1. procedure ROUTING($\hat{u}_{j|i}$, r, l)
2. for all capsule u_i in layer l and capsule u_j in layer(l+1): $b_{ij}\leftarrow 0$
3. for r iterations do:
4. for all u_i in layer l: $c_i\leftarrow$ Softmax(b_i)
5. for all u_j in layer(l+1): $s_j\leftarrow\sum_i c_{ij}\hat{u}_{j|i}$
6. for all u_j in layer(l+1): $v_j\leftarrow$ squash(s_j)
7. for all u_i in layer l and u_j in layer(l+1): $b_{ij}\leftarrow b_{ij}+\hat{u}_{j|i}\cdot v_j$
9. return v_j

3.3 分类器与损失函数

本文采用 Softmax 分类器进行类别预测。对于 R-S 胶囊层的输出向量 v_j，可将图像 f 属于真实人脸或伪造人脸的概率表示为：

$$p(v_j)=\mathrm{Softmax}(v_j)=\frac{\exp(v_j)}{\sum_{k=0}^{1}\exp(v_k)} \tag{7}$$

模型训练所用损失函数为：

$$loss=\sum L_j+\theta\cdot MSE \tag{8}$$

其中，$\theta=0.0005$，L_j 为边际损失(Margin Loss)：

$$L_j = T_j \max(0, m^+ - \| v_j \|)^2 + \lambda(1 - T_j)\max(0, \| v_j \| - m^-)^2 \tag{9}$$

当且仅当由 $p(v_j)$ 所得预测类别标签 $\hat{y}$ 与真实标签 y 一致时，$T_j = 1$。边际设置为 $m^+ = 0.9$，$m^- = 0.1$。MSE 是原始图像 f 与重建图像 $\tilde{f}$ 间的 L2 损失，$\tilde{f}$ 由图 4 所示解码器对 $v_j^{(R\text{-}S)}$ 重建获得。

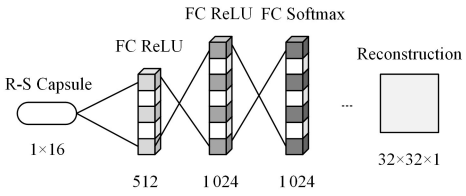


图 4 FAS-CapsNet 解码器结构

Fig. 4 Architecture of decoder of FAS-CapsNet

4 实验与分析

本文实验基于 NVIDIA GeForce RTX 2060 显卡 GPU(6 GB)，算法部分使用深度学习框架 TensorFlow 1.11、科学计算库 NumPy 1.19.2 和计算机视觉库 OpenCV-Python 4.5.5.62 实现。使用式(8)所示损失函数及 Adam 优化器训练 FAS-CapsNet 模型，初始学习率设置为 0.0001，batch size 为 64，Retinex 滤波参数 σ 根据表 1 实验情况设置为 15，共在训练集上迭代 40 轮。

表 1 采用不同滤波参数 σ 时的模型性能

Table 1 Model performance with different σ

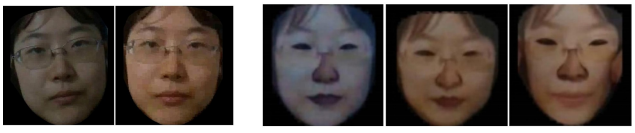
(%)

	σ						
	3	6	9	12	15	18	21
ACC	65.760	74.875	78.271	80.354	**82.313**	79.261	75.920
HTER	25.639	19.489	17.500	15.940	**14.226**	16.528	18.141

对比模型采用 VGGNet、MobileNetV1[35]、经典 CapsNet、LBP-CapsNet 和 FeatherNets(A& B)。其中，CNN 类架构包括 VGGNet 及两种轻量化架构 MobileNetV1 和 FeatherNets；CapsNet 类架构包括经典 CapsNet 和 LBP-CapsNet。在 LBP-CapsNet 中，图像被送入 CapsNet 前先应用了局部二值模式(Local Binary Pattern，LBP)算子提取纹理特征。

4.1 数据集与评价指标

实验采用人脸活体检测公开数据集 CASIA-SURF 和 CASIA-SURF-CeFA。其中，CASIA-SURF 数据集用于训练 FAS-CapsNet 模型，样本覆盖如图 5 所示的多种照片假体呈现攻击形式(手持、绑定佩戴、照片局部区域剪除)以及不同的环境光照情况。



(a)真实人脸 (b)伪造人脸

图 5 CASIA-SURF 数据集样本示例

Fig. 5 Real and spoof face images of CASIA-SURF

如表 2 所列，实验以静默活体检测中最常用的 RGB 图像作为训练数据，覆盖 300 个不同的受试对象，共包含 8492 张真实人脸图像和 20324 张欺骗攻击图像；验证集包含 100 个不同对象的 2994 张真实人脸图像与 6614 张欺骗攻击图像。

表 2 CASIA-SURF 数据集详情(实验选用部分)

Table 2 Details of CASIA-SURF(experimental part)

数据集	真实人脸	伪造人脸	图像数量	采集对象
训练集(RGB)	8942	20324	29266	300
验证集(RGB)	2994	6614	9608	100

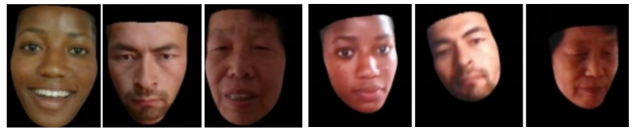
此外，本文采用 BIM[36] 和 One-Pixel[37] 两种对抗样本生成算法基于 CASIA-SURF 验证集生成两个对抗样本集(见图 6)，以评估模型的对抗鲁棒性(训练模型时不使用对抗样本)。

(a)原始图像 (b)BIM (c)One-Pixel

图 6 对抗样本图像示例

Fig. 6 Images of adversarial examples

CASIA-SURF-CeFA 是一个大型跨人种数据集，包含对 1607 个来自非洲、东亚和中亚地区的不同人种对象采集的图像和视频数据。该数据集用于测试各模型的跨数据域泛化能力。样本示例图如图 7 所示。



(a)真实人脸 (b)伪造人脸

图 7 CASIA-SURF-CeFA 数据集样本示例

Fig. 7 Examples of CASIA-SURF-CeFA

实验涉及的评估指标包括：检测准确率(Accuracy，ACC)、错误接受率(False Acceptance Rate，FAR)、错误拒绝率(False Rejection Rate，FRR)以及半错误率(Half Total Error Rate，HTER)。其中，HTER 的计算方式为

$$HTER = \frac{FAR + FRR}{2} \tag{10}$$

4.2 对比实验

实验采用 CASIA-SURF 验证集对各模型进行照片攻击检测性能评估。

如表 3 所列，FAS-CapsNet 对常规样本的 *ACC* 为 87.344%，较次优模型 FeatherNetB 超出 10.678%；*HTER* 为 11.268%，较次优模型 LBP-CapsNet 降低 37.302%；*FAR* 也较次优模型 FeatherNetA 降低了 39.911%；*FRR* 指标则未达最佳。该模型综合表现优于其他模型。

表 3 各模型在 CASIA-SURF 数据集上的性能对比

Table 3 Performance omparison of each model on CASIA-SURF

(%)

模型	*ACC*	*FAR*	*FRR*	*HTER*
VGGNet	77.042	30.856	5.486	18.171
MobileNetV1	70.406	36.167	15.055	25.611
CapsNet	72.635	38.496	**2.743**	20.620
LBP-CapsNet	76.719	32.053	3.881	17.972
FeatherNetA	78.875	24.868	12.847	18.857
FeatherNetB	78.917	25.866	10.505	18.186
FAS-CapsNet(Ours)	**87.344**	**14.943**	7.594	**11.268**

同时，FAS-CapsNet 和 LBP-CapsNet 在 *ACC* 上都较

CapsNet 有明显提升，综合指标均有优化，说明预先增强图像特征有利于 CapsNet 后续的拟合过程。而 Retinex 算法提取的光照特征在实验中比 LBP 纹理特征更能反映真伪人脸间的差异，使得 FAS-CapsNet 较 LBP-CapsNet 的 *ACC* 提升 13.849%，HTER 降低 37.302%，指标优化效果明显。

4.3 对抗鲁棒性实验

本文基于 VGG16 模型生成 BIM 和 One-Pixel 两种迁移性较高的对抗样本图像。如表 4 所列，FAS-CapsNet 在两个对抗样本数据集上的重要指标 ACC 和 HTER 都优于其他对比模型。

表 4 各模型在对抗样本数据集上的性能对比

Table 4 Performance comparison of each model on adversarial examples dataset

(%)

模型	*ACC*		*HTER*	
	@BIM	@One-Pixel	@BIM	@One-Pixel
VGGNet	50.698	66.000	37.364	26.070
MobileNetV1	76.697	37.563	32.691	47.194
CapsNet	78.833	46.708	18.869	38.922
LBP-CapsNet	61.542	69.302	29.691	23.104
FeatherNetA	75.468	37.521	20.659	45.840
FeatherNetB	74.938	41.667	21.157	42.858
FAS-CapsNet(Ours)	**84.552**	**79.042**	**17.127**	**19.808**

结合表 3 可知，VGGNet 受 BIM 样本攻击前后 *ACC* 下降幅度和 *HTER* 升高幅度最大，其他 CNN 类模型的指标变化情况也反映了这些对抗样本对 CNN 具有普遍的干扰作用。而在受到 One-Pixel 样本攻击后，所有模型性能均出现下降，且 CNN 类模型下降幅度最大。这说明基于单像素点的黑盒攻击在具有隐蔽性的同时也有着更强的扰动能力，会对人脸认证系统造成严重威胁。相比之下，FAS-CapsNet 在 BIM 数据集上的 *ACC* 高于全部对比模型，且较 4.2 节下降幅度最低，仅降低 3.197%；其 HTER 指标虽较原先有所升高，但仍低于其他对比模型。

受 One-Pixel 样本扰动程度较低的模型有 FAS-CapsNet，LBP-CapsNet，VGGNet 和 CapsNet。CNN 类中受扰动最小的 VGGNet 使用深层次常规卷积，较另外 3 个轻量化架构感知到的图像范围更大，特征更复杂，因而不易被单像素点噪声大幅扰动。CapsNet 类架构则因胶囊结构保留了特征的位置、朝向等辅助信息，在网络内部形成稳定的特征关联模式，使其在受单像素点扰动时可通过胶囊中的其他元素进行校正，以维持自身鲁棒性。结合其在 BIM 样本攻击下的稳定表现可以说明：FAS-CapsNet 在对抗样本攻击中仍可维持稳定良好的活体检测能力，具有显著的对抗鲁棒性优势。

4.4 消融实验

为验证模型中所用各项策略的有效性和必要性，本节对采用的 CapsNet，Retinex 和 DSC 3 项策略进行了单项消融实验，结果如表 5 所列（“w/o”表示消融某项策略）。

表 5 单项策略消融实验

Table 5 Single strategy ablation experiment

	ACC/%	*HTER*/%	*ACC@BIM*/%	*ACC@One-Pixel*/%	*Time*/s
FAS-CapsNet	**84.906**	**14.844**	82.385	**78.000**	11.201
(w/o) CapsNet	73.271	19.874	72.531	62.187	**10.702**
(w/o) Retinex	76.458	19.613	79.448	39.116	15.777
(w/o)DSC	76.281	17.643	**82.958**	70.583	18.901

消融 CapsNet 时以 VGGNet 作为替代分类器，模型在正常样本和对抗样本验证集上的 *ACC* 均大幅下降，尤其是在 One-Pixel 样本攻击下，下降幅度达到 20.273%，证明 CapsNet 对提升模型对抗鲁棒性具有实质性作用。而这一环节中检测用时的减少，则是由于 VGG 的池化操作减少了特征表达及传递过程中的计算开销。

消融 Retinex 后，模型在 One-Pixel 数据集上的 *ACC* 下降最明显，证明 Retinex 在增强特征的同时对细微对抗扰动有明显过滤作用，进一步提升了模型的对抗鲁棒性。

消融 DSC 后可见，模型对等规模数据的检测用时明显多于原模型，反映了 DSC 对轻减模型体量有重要作用，可以有效降低计算开销，更符合静默式活体检测的实时性要求。

4.5 泛化性实验

在真实应用场景中，人脸特征的分布情况比训练模型使用的数据集更加复杂。为进一步检验 FAS-CapsNet 对更复杂数据分布的泛化效果，本节采用样本量更大、包含更多人种特征的 CASIA-SURF-CeFA 数据集进行跨数据域泛化性实验。如表 6 所列，VGGNet，CapsNet 和 LBP-CapsNet 3 个模型的 ACC 和 HTER 指标较优，均应用常规卷积，其他模型则采用 DSC。可见，轻量化会伴随着一定程度的特征表达能力下降，影响模型的跨数据域泛化能力。

表 6 各模型跨数据集泛化性能对比

Table 6 Comparison of cross-dataset generalization

(%)

模型	*ACC*	*FAR*	*FRR*	*HTER*
VGGNet	**61.345**	37.509	39.701	**38.605**
MobileNetV1	41.898	56.605	59.466	58.035
CapsNet	58.040	46.705	**37.636**	42.171
LBP-CapsNet	51.632	41.296	54.813	46.554
FeatherNetA	30.417	52.602	86.063	69.332
FeatherNetB	42.604	25.698	88.157	56.927
FAS-CapsNet(Ours)	41.952	**13.671**	98.488	56.079

而 VGGNet 中的池化操作去除了图像中的冗余特征，使模型更倾向于表示人脸的关键特征；且 VGGNet 通过渐进式卷积升维获得了更高维度的观察视角，可表示更深层的空间特征，从而令 VGGNet 获得优于 CapsNet 的泛化性。如何平衡模型的鲁棒性和泛化性，则是未来工作的重要研究部分。

结束语 本文针对现有人脸活体检测方法在对抗样本防御方面的不足，引入并改进 CapsNet 提出一种具有显著对抗鲁棒性的静默式活体检测方法 FAS-CapsNet；并通过实验证明：FAS-CapsNet 在对抗样本攻击中仍可保持与常规攻击条件下相当的活体检测准确性，即本文采用各项策略对提升模型对抗鲁棒性具备有效性。

实验也反映出该模型对未知特征分布的泛化检测能力较差，这可能是因为 CapsNet 在以更大的感受野收集特征时，对图像各邻域采用了同等权重。实际上，图像的语义特征更多集中于图像中部，边缘的语义特征则很少，卷积所得特征图中心点的信息强度较边缘更高。当图像边缘和中心被 CapsNet 同等接受时，冗余的背景信息会削弱其对重要特征的表达能力，造成模型欠拟合。而感受野中的权重参数呈高斯分布，通常在从中心向外延伸时迅速衰减，使得感受野中心对输出特征点的影响远大于边缘，有效感受野区域仅占整体的一小部分[38]。感受野的这一特性加剧了 CapsNet 模型对关键特征

表达不足的问题，下一步工作可从特征表达方式入手，如结合中心差分卷积(Central Difference Convolution)[39]来描述感受野边缘与中心的联系，以增强模型整体对细粒度特征的表示和泛化能力。

此外，近年来智能换脸算法、视频生成算法快速发展，攻击者正利用这些技术实施更隐蔽的欺骗攻击，给人脸认证带来新的威胁。我们需要尽快对这些攻击形式展开研究，找到更强健的活体检测方法，从而使人脸认证系统更加安全可信。

参 考 文 献

[1] YAO Y J, MA X C, GONG W, et al. In the era of face authentication, is your "face" still safe[J]. Policy Research & Exploration, 2017(23): 28-29.

[2] LIU S Y. Criminal regulation of cracking face recognition and authentication from the perspective of cooperative crime[J]. Journal of Henan University of Economics and Law, 2021, 36(4): 41-51.

[3] JIANG F L, LIU P C, ZHOU X D. A review on face anti-spoofing[J]. Acta Automatica Sinica, 2021, 47(8): 1799-1821.

[4] GALBALLY J, MARCEL S. Face anti-spoofing based on general image quality assessment[C]// 2014 22nd International Conference on Pattern Recognition. IEEE, 2014: 1173-1178.

[5] MA Y K, WU L F, JIAN M, et al. Algorithm to generateadversarial examples for face-spoofing detection[J]. Journal of Software, 2019, 30(2): 469-480.

[6] ZHANG B, TONDI B, BARNI M. Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability[J]. Computer Vision and Image Understanding, 2020, 197-198: 102988.

[7] MIAO J. Risk challenge and regulatory research of face recognition facing "easy to crack"[J]. Journal of Information Security Research, 2021, 7(10): 984.

[8] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]// Proc. of 31st Conference on Neural Information Processing Systems (NIPS). MIT Press, 2017: 3859-3869.

[9] SINGH A K, JOSHI P, NANDI G C. Face recognition with liveness detection using eye and mouth movement[C]// 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014). IEEE, 2014: 592-597.

[10] MA Y X, TAN L, DONG X, et al. Interactive livenessdetection algorithm for VTM[J]. Computer Engineering, 2019, 45(3): 256-261.

[11] ZHANG J, ZHANG N N. Research on interactive face detection based on optimized feature extraction[J]. Computer Engineering and Applications, 2019, 55(13): 193-200.

[12] MÄÄTTÄ J, HADID A, PIETIKÄINEN M. Face spoofing detection from single images using micro-texture analysis[C]// 2011 International Joint Conference on Biometrics (IJCB). IEEE, 2011: 1-7.

[13] BOULKENAFET Z, KOMULAINEN J, HADID A. Face anti-spoofing based on color texture analysis[C]// 2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015: 2636-2640.

[14] FENG J, DONG Z Y, LIU T T, et al. Face anti-spoofingliveness detection combining DQ_ CoALBP with LPQ descriptors[J]. Computer Engineering and Applications, 2022, 58(14): 134-143.

[15] ZHOU J, SHU K, ZHAO D, et al. Domain adaptation based person-specific face anti-spoofing using color texture features[C]// Proceedings of the 2020 5th International Conference on Machine Learning Technologies. 2020: 79-85.

[16] SHU X, TANG H, HUANG S. Face spoofing detection based on chromatic ED-LBP texture feature[J]. Multimedia Systems, 2021, 27: 161-176.

[17] ZHOU J, SHU K, LIU P, et al. Face anti-spoofing based on dynamic color texture analysis using local directional number pattern[C]// 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 4221-4228.

[18] ZHANG L B, PENG F, QIN L, et al. Face spoofing detection based on color texture Markov feature and support vector machine recursive feature elimination[J]. Journal of Visual Communication and Image Representation, 2018, 51: 56-69.

[19] SHU X, TANG H, YANG X B, et al. Research on face anti-spoofing algorithm based on DQ_LBP[J]. Journal of Computer Research and Development, 2020, 057(): 1508-1521.

[20] DANIEL N, ANITHA A. Texture and quality analysis for face spoofing detection[J]. Computers & Electrical Engineering, 2021, 94(1): 107293.

[21] CHANG H H, YEH C H. Face anti-spoofing detection based on multi-scale image quality assessment[J]. Image and Vision Computing, 2022, 121: 104428.

[22] WEN D, HAN H, JAIN A K. Face spoof detection with image distortion analysis[J]. IEEE Transactions on Information Forensics & Security, 2015, 10(4): 746-761.

[23] ZHANG S, WANG X, LIU A, et al. A dataset and benchmark for large-scale multi-modal face anti-spoofing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 919-928.

[24] ZHANG P, ZOU F, WU Z, et al. FeatherNets: Convolutional neural networks as light as feather for face anti-spoofing[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019.

[25] SHEN T, HUANG Y. FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing[C]// The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2019). IEEE, 2019.

[26] PADNEVYCH R, SEMEDO D, CARMO D, et al. Improving face liveness detection robustness with deep convolutional generative adversarial networks[C]// 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022.

[27] LI X, WU W, LI T, et al. Face face liveness detection based on parallel CNN[J]. Journal of Physics: Conference Series, 2020, 1549(4): 042069.

[28] SHEKHAR S, PATEL A, HALOI M, et al. An ensemble model for face liveness detection[EB/OL]. (2022-01-19) [2023-02-17]. https://arxiv.org/abs/2201.08901.

[29] YU Z, WAN J, QIN Y, et al. NAS-FAS: Static-dynamic central difference network search for face anti-spoofing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(9): 3005-3023.

[30] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: A survey[D]. Ithaca: Cornell Uni-

versity,2018.

[31] FROSST N,SABOUR S,HINTON G. DARCCC:Detecting adversaries by reconstruction from class conditional capsules[EB/OL]. (2018-11-16) [2023-02-17]. https://arxiv.org/abs/1811.06969.

[32] QIN Y,FROSST N,SABOUR S,et al. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions [C]// International Conference on Learning Representations. 2020.

[33] QIN Y,FROSST N,RAFFEL C,et al. Deflecting adversarial attacks[EB/OL]. (2020-02-18) [2023-02-17]. https://arxiv.org/abs/2002.07405.

[34] LAND E. Lightness and retinex theory[J]. Journal of the Optical Society of America,1971,61(1):1-11.

[35] HOWARD A G,ZHU M,CHEN B,et al. MobileNets:Efficient convolutional neural net-works for mobile vision applications [EB/OL]. (2017-04-17) [2023-02-17]. https://arxiv.org/abs/1704.04861.

[36] KURAKIN A,GOODFELLOW I,BENGIO S,et al. Adversarial machine learning at scale[EB/OL]. (2017-02-11) [2023-02-17]. https:// arxiv.org/abs/1611.01236.

[37] SU J,VARGAS D V,KOUICHI S. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation,2019,23(5):828-841.

[38] LUO W,LI Y,URTASUN R,et al. Under-standing the effective receptive field in deep convo-lutional neural networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016:4905-4913.

[39] YU Z,ZHAO C,WANG Z,et al. Searching central difference convolutional networks for face anti-spoofing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:5295-5305.

WANG Chundong, born in 1969, Ph.D, professor, is a member of CCF (No. 16230M). His main research interests include big data and smart computing security, network security situation awareness, etc.

LI Quan, born in 1997, postgraduate. Her main research interests include face anti-spoofing, capsule networks, etc.