

SAPD: 面向遮挡人脸识别的对抗补丁 分割防御方法

王浩轩, 黄瑞阳, 刘宏基, 汪浣沙

(信息工程大学, 河南 郑州 450001)

摘要: 针对现有对抗补丁人脸识别防御方法的泛化能力差、难以抵御强对抗扰动的问题, 提出一种基于分割的对抗补丁攻击的防御方法(SAPD)。该方法构建融合注意力机制的 U-Net 分割网络, 结合多尺度混合损失函数实现像素级补丁定位; 提出形态学掩码平滑(MMS)策略, 有效消除扰动的同时不破坏面部结构; 设计遮挡区域正则化(OAR)损失函数, 在零对抗样本训练条件下增强模型对遮挡的鲁棒性, 保持对干净数据的识别性能。在添加对抗补丁的 LFW 数据集上实验表明, 该方法对眼镜状补丁、口罩状补丁 1 及口罩状补丁 2 的识别准确率分别达到 92.50%、96.16% 及 95.67%。此外, 该方法在 MobileFaceNet、FaceNet 和 IResNet50 等主流人脸识别模型上均有性能提升, 证明其具有较强的兼容性、良好的泛化能力与实用价值。

关键词: 对抗防御; 对抗补丁; 分割; 修复; 人脸识别; 深度学习; 遮挡

中图分类号: TP391.1

文献标识码: A

文章编号: 1671-0673(2026)02-0199-09

SAPD: Segmentation-Based Adversarial Patch Defense Framework for Occluded Face Recognition

WANG Haoxuan, HUANG Ruiyang, LIU Hongji, WANG Huansha

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the issues of poor generalization and weak robustness against strong adversarial perturbations in existing face recognition defense methods against adversarial patches, a segmentation-based adversarial patch defense (SAPD) framework is proposed. In this method, a U-Net segmentation network fused with an attention mechanism is constructed, and a multi-scale hybrid loss function is employed to realize pixel-level patch localization. A morphological mask smoothing (MMS) strategy is introduced to effectively remove perturbations while preserving facial structure. An occluded-area regularization (OAR) loss function is further designed to improve occlusion robustness without adversarial training, maintaining clean-data accuracy. The experiments conducted on the LFW dataset with the added adversarial patches showed that the recognition accuracies of the method for the eyeglass-shaped patches, mask-shaped patch 1, and mask-shaped patch 2 reach 92.50%, 96.16%, and 95.67%, respectively. Furthermore, consistent performance gains are observed across mainstream face recognition models such as MobileFaceNet, FaceNet, and IResNet50, verifying the strong compatibility, generalization capability, and practical value of the proposed approach.

Key words: adversarial defense; adversarial patch; segmentation; inpainting; face recognition; deep learning; occlude

深度学习驱动的人脸识别系统凭借强大的特征表达能力,已成为智能门禁、移动支付和公共安全等关键场景的核心技术。以 ArcFace^[1]为例,其在 LFW(Labeled Faces in the Wild)干净数据集上的识别准确率达 99.83%,显著超越人类水平。然而,其安全性正受到现实世界中局部对抗补丁攻击的严重威胁:攻击者可将打印的对抗补丁附着于眼镜^[2]、帽子^[3]或口罩^[4-5]等日常物品上,诱使系统误判身份。在此类攻击下,识别准确率常从 99% 以上骤降至 10% 以下。此类攻击因成本低、隐蔽性强且破坏性高,极易绕过安防机制,造成非授权访问与身份伪造,对金融与公共安全构成重大风险。

当前针对物理局部对抗补丁的防御策略主要分为 3 类。1) 通过对抗训练引入对抗样本提升鲁棒性,但依赖大量攻击数据,泛化能力差,且损害干净样本性能;2) 特征蒸馏法,通过教师模型迁移清洁特征,但难以有效应对遮挡或扰动区域带来的干扰;3) 检测—分割方法,定位补丁再修复图像,虽具实用性,但存在两大缺陷:分割精度不足以及修复后边界残留扰动,导致防御失效。

综上,现有防御方法未能解决对抗训练泛化能力差、高精度分割缺失、对抗扰动去除不彻底等 3 大核心挑战,难以满足实际部署需求。为此,提出一种基于分割的对抗补丁攻击的防御方法(Segmentation-Based Adversarial Patch Defense, SAPD),构建了一种无需对抗样本训练的鲁棒人脸识别框架。其核心包括:1) 构建基于 U-Net^[6]的高精度分割模块,结合注意力门(Attention Gates, AG)^[7]实现对抗补丁的像素级定位;2) 提出形态学掩模平滑(Morphological Mask Smoothing, MMS)算法,优化掩码边界,确保对抗扰动被完全消除;3) 设计遮挡鲁棒识别机制,通过遮挡区域正则化(Occluded-Area Regularization, OAR)损失函数引导模型聚焦未遮挡区域,在保留干净性能的同时提升遮挡下的识别准确率。

1 相关工作

1.1 遮挡环境下的人脸识别

利用可见面部区域进行识别,可在一定程度上缓解遮挡带来的负面影响。此类方法被学界称为遮挡感知人脸识别(Occlusion-Aware Face Recognition, OAFR)^[8]。当前主流 OAFR 方法普遍采用“先检测、后识别”范式,即首先定位遮挡区域,再从未被遮挡部分提取判别性特征用于身份匹配。

早期工作中^[9-11],将人脸划分为若干非重叠块,训练支持向量机(Support Vector Machine, SVM)分类器判断各块是否被遮挡,并在匹配时融合未遮挡区域特征。在此基础上,文献[12]提出对抗遮挡感知人脸检测(Adversarial Occlusion-Aware Face Detection, AOFD),进一步引入对抗掩码机制,通过随机屏蔽局部区域迫使模型关注核心判别结构,并辅以分割分支处理不完整人脸。文献[13]基于区域重要性评估,降低模型对易遮挡区域的依赖。文献[14]提出一种联合遮挡检测与分割模块,用于识别潜在被遮挡的人脸组件并生成对应的遮挡掩码。

1.2 对抗补丁攻击的防御

现有对抗攻击防御方法可分为 3 类。早期方案如 JPEG^[15]、数字水印(Digital Watermark, DW)^[16]和局部梯度平滑(Local Gradient Smoothing, LGS)^[17]均因依赖简化的信号假设而难以应对真实物理攻击。JPEG 假设扰动集中于高频,但物理对抗补丁常含低频结构,且压缩图像后会破坏部分人脸纹理;DW 则依赖对抗扰动引发异常激活,而现代攻击可生成视觉自然的扰动,其响应与正常遮挡难以区分;LGS 在训练中平滑高梯度区域,但模糊了人脸本身的关键特征,降低了干净样本的识别率。这些方法本质上将对抗扰动视为“非自然噪声”,忽视了其在物理世界中的语义伪装能力,因而无法抵御高强度、高逼真的对抗补丁攻击。

为提升鲁棒性,后续方法转向认证防御与对抗训练,但仍然存在明显局限性。文献[18]提出区间边界传播,通过在网络中传播输入边界实现认证鲁棒性。类似的,针对遮挡攻击的防御方法^[19]是基于对抗训练的方法,在面对投影梯度下降(Projected Gradient Descent, PGD)^[20]等强迭代攻击时表现出更强的鲁棒性。文献[21]则提出一种基于图像变分辨率敏感不一致性的对抗性人脸检测方法,通过超分辨率变换过程中敏感度变化,检测人脸攻击。文献[22]提出二阶对抗样本生成机制,通过在原始对抗样本上叠加二次扰动构造更强训练样本,以提升模型对已知攻击的抵抗力。然而,文献[18]在复杂场景下的人脸识别任务中准确率大幅下降;文献[19]对未见攻击类型的泛化能力仍存疑;文献[21]主要聚焦于检测数字攻击,无法防御物理攻击;文献[22]仅对引入训练集的攻击方式具有鲁棒性,对于未知攻击几乎无效。

近年来,基于检测—中和和分割—修复的方法成为主流,但是这种思路模型的识别准确率、泛

化能力面临挑战。文献[23]提出 PatchGuard,利用小感受野和异常值掩蔽机制定位并中和对抗补丁,但需要对主干分类器架构进行大量修改。其改版 PatchCleanser^[24],采用两阶段判断机制,兼容任意分类器。然而,二者均在面对大面积或复杂形状补丁时性能显著下降。文献[25]提出分割与补全方法(Segment and Complete, SAC),采用基于图像重建的防御范式,首先检测并分割出对抗补丁区域,然后通过图像修复去除扰动。但该方法分割精度较低,且对多样化补丁形状和尺寸的泛化能力较差。文献[26]提出 PatchTracker,通过追踪对抗补丁在深层特征图中的高响应区域并动态掩蔽,实现对已知补丁位置的抑制。而 PatchTracker 依赖补丁在特征空间产生显著激活这一假设,在面对低强度或伪装性强的物理补丁时,其响应与正常人脸区域难以区分,导致漏检或误抑,防御性能不稳定。文献[27]提出一种面向多样对抗补丁的鲁棒自适应防御方法(Robust and Adaptive Defense Against Diverse Adversarial Patches, RADAP),进一步改进这个思路,结合 FCutout 和 F-patch 策略,并引入傅里叶空间采样掩码的分割器,增强了人脸识别中的遮挡鲁棒性;同时,引入边缘感知二值交叉熵(Edge-Aware Binary Cross-Entropy, EBCE)

损失函数提高补丁检测精度。尽管 F-patch 实现了更高的分割精度,但其分割与填充(Split and Fill, SAF)策略会改变原有的对抗补丁形状,覆盖由原本未被干扰的有效人脸特征区域,削弱了模型的整体表征能力。综上,以上方法虽摆脱了对抗训练依赖,却陷入高精度分割难实现、低质量修复必伤身份的困境,一旦分割不准或修复策略僵化,防御过程本身即成为性能瓶颈。

2 SAPD 框架

为解决对抗补丁防御技术的泛化能力差、识别准确率较低等缺点,提出一种面向对抗补丁攻击的人脸识别防御方法,如图1所示。其中: X 表示输入的对抗补丁图像; M 表示定位的补丁二值掩码; I 表示经过 MMS 策略修复后的对抗补丁图像;通过 OAR 混合损失函数(其值为 L_{OAR}),完成遮挡条件下的人脸识别。SAPD 包含3个关键阶段:1)高精度分割,基于 U-Net 的分割网络,结合注意力机制,实现对抗补丁的像素级定位;2)形态学平滑,利用形态学操作消除掩码边界的孤立噪声,彻底清除残余扰动;3)遮挡鲁棒识别,采用零对抗样本训练策略,引入 OAR 损失函数,像素级引导模型聚焦非遮挡区域,增强遮挡下的判别能力。

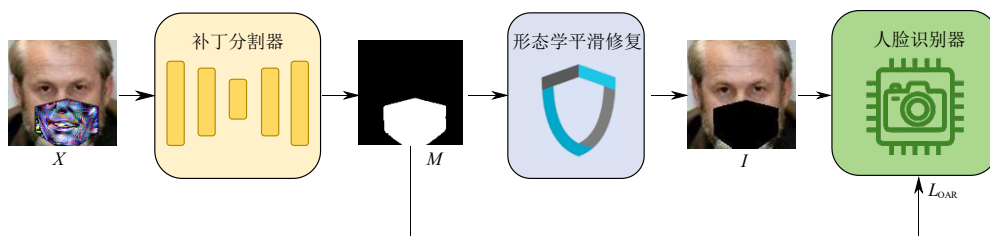


图1 SAPD 整体防御流程框架

2.1 分割网络

受到面向图像分割的注意力边界感知网络(Attention Boundary-Aware Network for Image Segmentation, ABANet)^[28]的启发,以 U-Net 构建分割网络,通过引入 BCE 损失函数提升边界分割精度,增强对不同形状补丁的泛化能力。

2.1.1 网络架构

采用“分割子网+精分子网”两阶段架构,如图2所示。

1)分割子网。前4个阶段以 ResNet-34^[29]为编码器主干,第5、6个阶段各含512通道,由3个残差块组成。输入经64个3×3卷积核初步提取特征。桥接层由3层3×3膨胀卷积构成,在无下采样条件下扩大感受野。解码器对称设计,共6个阶段:前

两个阶段含3×3卷积层+双线性上采样;后4个阶段在跳跃连接中嵌入AG,动态增强高分辨率细节响应,提升边界感知能力。

2)精分子网。以轻量化4层 U-Net 为架构,以原始 RGB 图像与粗掩码的通道拼接结果为输入,逐级细化边界,有效抑制锯齿伪影。

3)多分支输出设计。受 ABANet 多尺度策略启发,设计8个输出分支: d_b (桥接层)、 $d_2 \sim d_7$ (解码阶段2~7)、 d_{out} (精炼输出)。所有分支经 Sigmoid 激活函数生成概率图。训练时,由四重混合损失联合监督,施加多尺度一致性约束,加速收敛并提升边界精度;推理时,通过加权聚合多层级预测,融合多尺度上下文,确保分割在结构完整性与局部准确性上的平衡。

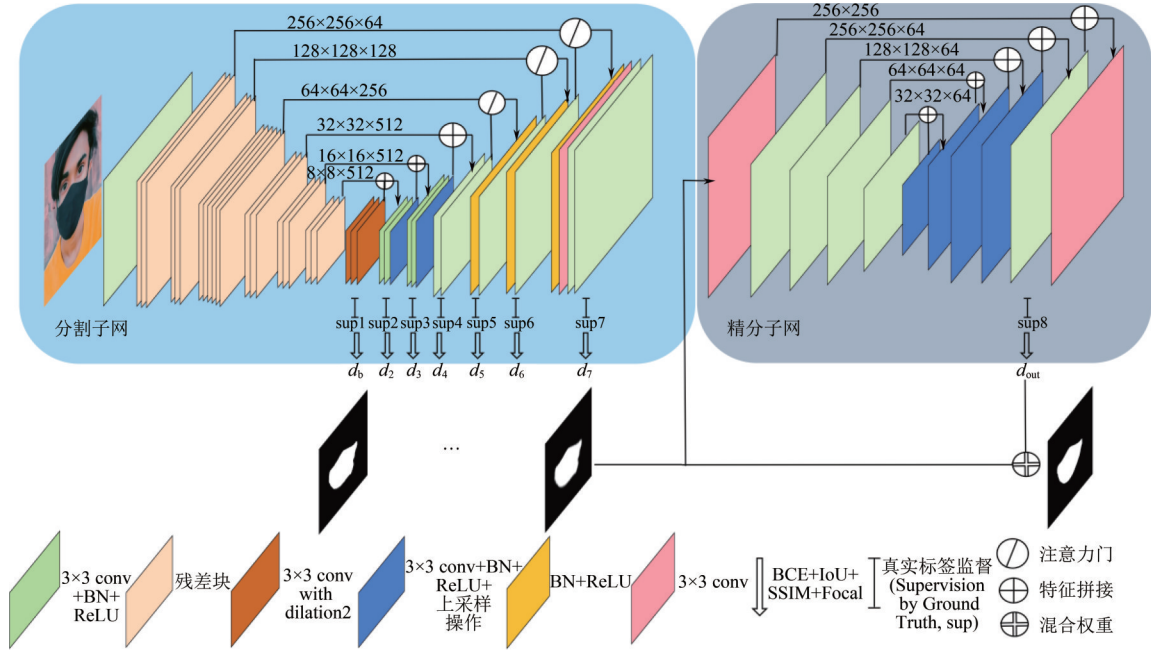


图2 分割网络架构

2.1.2 损失函数

ABANet 是由 Focal、结构相似性指数 (Structural Similarity Index Measure, SSIM) 与交并比 (Intersection over Union, IoU) 构成的复合损失函数。Focal 可缓解遮挡与非遮挡区域间的像素级类别不平衡,提升边缘定位精度;IoU 则通过直接优化预测掩码与真实掩码的空间重叠度,改善边界对齐性能。为进一步增强像素级判别能力,引入 BCE 损失,度量预测概率与二值标签间的逐像素分类误差,其定义如式(1)所示:

$$L_{BCE} = -\frac{1}{B \cdot H \cdot W} \sum_{b=1}^B \sum_{x=1}^H \sum_{y=1}^W [G(b, x, y) \cdot \ln(p(b, x, y)) + (1 - G(b, x, y)) \cdot \ln(1 - p(b, x, y))] \quad (1)$$

式中: B 表示批次大小; H 、 W 表示图像的高和宽; b 表示批次中图像的索引; $G(b, x, y)$ 表示真实掩码, 1 表示遮挡区域, 0 表示正常未遮挡区域; $p(b, x, y)$ 表示预测遮挡概率。

四重混合损失函数定义如式(2)所示:

$L_{mix} = \lambda_{BCE} L_{BCE} + \lambda_{SSIM} L_{SSIM} + \lambda_{IoU} L_{IoU} + \lambda_{Focal} L_{Focal} \quad (2)$

式中, 各项损失权重为 $\lambda_{BCE} = 0.3$ 、 $\lambda_{SSIM} = 0.2$ 、 $\lambda_{IoU} = 0.2$ 、 $\lambda_{Focal} = 0.3$ 。BCE 提供像素级监督, 但易导致边界碎片化; SSIM 从亮度、对比度和结构 3 个方面保持掩码形态连贯性。在原损失函数的基础上引入 SSIM 损失, 二者协同兼顾像素精度与结构保真度。对 BCE 和 SSIM 损失提供更高的权重比例, 增强模型对边界区域的分割精度。

基于上述的 8 个分支输出结构, 总训练损失为

各分支损失的加权和, 其定义如式(3)所示:

$$L_{final} = \sum_{k=1}^8 w_k \cdot L_{total}^{(k)} \quad (3)$$

式中: $L_{total}^{(k)}$ 表示在第 k 个分支的混合损失; w_k 对应输出分支权重。其中: 精分子网最终输出 (d_{out}) 的权重最高设为 0.2; 深解码层 ($d_5 \sim d_7$) 与桥阶层 (d_b) 的权重为 0.12; 浅层 ($d_2 \sim d_4$) 的权重依次为 0.05、0.07、0.1, 以渐进增强早期语义。

2.1.3 训练策略

采用 Adam 优化器, 初始学习率为 10^{-3} , 动量参数 $\beta_1=0.9$ 、 $\beta_2=0.999$, 权重衰减系数为 0.000 5。在前 5 个训练周期内, 采用线性预热策略; 随后, 引入余弦退火调度器, 周期性地调整学习率; 与此同时, 叠加一个指数衰减调度器, 其衰减因子设为 0.9, 以确保在训练后期学习率持续平稳下降。

2.2 形态学掩模平滑策略

为彻底消除对抗补丁并保持面部视觉一致性, 提出 MMS 策略。该策略包含以下两个阶段。

1) 边界追踪。基于分割网络输出的二值掩模 $M(M(x, y)=1$ 为遮挡区域), 对每行扫描获取左右边界列坐标。设定边界阈值 $\Delta y = 3$, 若某行边界跨度小于阈值, 即可判定该点为孤立噪点, 将掩模值设为 (0, 3), 对剩余行边界坐标进行平滑, 采用 3 点滑动平均滤波消除边界抖动, 得到初步优化的边界坐标 $\hat{y}_{left}(x)$ 和 $\hat{y}_{right}(x)$, 如式(4)和式(5)所示:

$$\hat{y}_{left}(x) = \frac{1}{4} (y_{left}(x-1) + 2y_{left}(x) + y_{left}(x+1)) \quad (4)$$

$$\hat{y}_{\text{right}}(x) = \frac{1}{4} (y_{\text{right}}(x-1) + 2y_{\text{right}}(x) + y_{\text{right}}(x+1)) \quad (5)$$

2)形态学处理。形态学处理的核心在于结构元素(Structure Element, SE)的定义。本文采用一个3×3的方形结构元素 S ,即

$$S = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

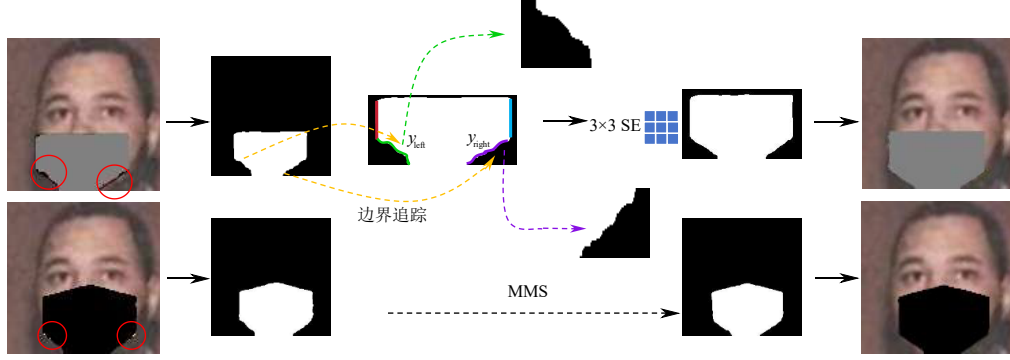


图3 MMS策略

2.3 遮挡识别鲁棒性提升

尽管扰动可被清除,处理后图像仍存在自然遮挡,因此模型需具备遮挡鲁棒性。OAR框架如图4所示,引导模型聚焦未遮挡区域,采用零对抗样本训练,避免过拟合攻击模式,保留干净数据性。

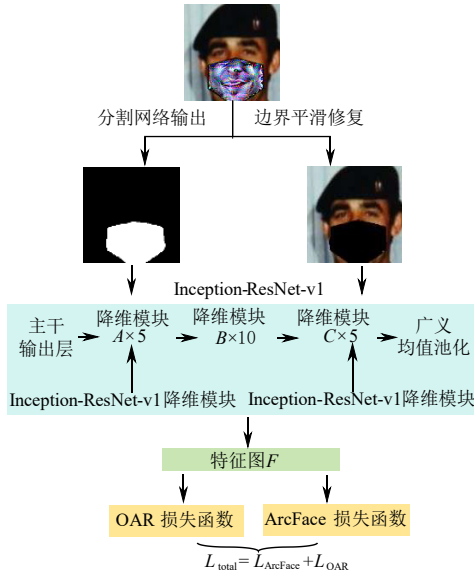


图4 OAR模型整体架构

基于FaceNet^[30]框架,保留Inception-ResNet-v1^[31]的1~16层卷积层;将全局平均池化替换为广义均值池化(Generalized Mean Pooling, GeM)^[32]以增强遮挡鲁棒性;用ArcFace损失替代不稳定的三元组损失。设主干输出特征图 $F \in \mathbf{R}^{H \times W \times C}$,对每个位置 (i, j) 计算通道 L_2 范数的平方作为局部激活强度,以度量局部激活强度,如式(6)所示:

腐蚀:仅当 S 与掩码 M 完全重叠为1时,中心像素保留为1,否则置0,消除孤立点。

膨胀:只要 S 与 M 有任意重叠为1,中心像素设为1,修复锯齿边缘。

MMS整体流程如图3所示,通过定义结构元素对掩模边界进行平滑操作,有效去除边缘孤立噪点、平滑轮廓,同时保留遮挡主体结构。

$$E(i, j) = \sum_{c=1}^C F(i, j, c)^2 \quad (6)$$

利用掩码 M ,将特征区域分别划分为含有对抗补丁的遮挡区域(Occluded)和不含对抗补丁的干净区域(Visible),如式(7)和式(8)所示:

$$E_{\text{Occluded}} = \sum_{(i, j) \in U_{\text{Occluded}}} E(i, j) \quad (7)$$

$$E_{\text{Visible}} = \sum_{(i, j) \in U_{\text{Visible}}} E(i, j) \quad (8)$$

定义OAR损失函数如式(9)所示:

$$L_{\text{OAR}} = \lambda \cdot \frac{E_{\text{Occluded}}}{E_{\text{Visible}} + \varepsilon} \quad (9)$$

式中: $\varepsilon > 0$ 用于避免除零错误; λ 控制正则化强度。通过将OAR损失函数最小化,引导模型降低遮挡区域的特征响应。

最终的优化目标由ArcFace分类损失与OAR正则项联合构成,如式(10)所示:

$$L_{\text{total}} = L_{\text{ArcFace}} + L_{\text{OAR}} \quad (10)$$

3 实验

对SAPD在人脸识别系统中的防御性能进行全面评估。评估其对不同形状与位置的对抗补丁的防御效果;与当前最先进的防御方法进行对比;测试在不同扰动强度下的鲁棒性;通过消融实验分析分割模块与OAR模型各自作用。所有实验均在NVIDIA GeForce RTX 3080 Ti GPU上完成,软件环境为Python 3.9.16、PyTorch 1.10.1和CUDA 11.7。

3.1 实验设置

在人脸识别任务中,所有图像均对齐至 112×112 分辨率。训练集与测试集示例如图5和图6所示。



图5 用于分割网络的训练集、验证集、测试集



图6 用于遮挡人脸识别模型的训练集

分割网络采用口罩人脸分割数据集(Masked Face Segmentation Dataset, MFSD)^[33]和眼镜检测(Glass Detector)^[34]数据集,其中包含多样化的遮挡人脸样本;对于遮挡人脸识别模型(OAR),利用MaskTheFace^[35]工具对CelebA-HQ^[36]数据集进行处理(CelebA-HQ-M),生成遮挡人脸图像。相关统计数据如表1所示。

表1 训练集

训练集	图片数量	包含遮挡图片数量	是否为真实遮挡物
MFSD	11 601	11 601	是
Glass Detector	391 570	31 084	否
CelebA-HQ-M	30 000	30 000	否

分割网络在MFSD与MeGlass^[37]的测试子集上进行评估;针对对抗补丁分割任务,使用LFW-A数据集,该数据集通过Face-Sec^[38]工具包在LFW基础上注入对抗补丁构建而成;在遮挡鲁棒性评估方面,采用LFW-M数据集,该数据集通过将MaskTheFace应用于LFW生成,涵盖多种口罩类型。相关统计数据如表2所示。

表2 测试集

测试集	图片数量	包含遮挡图片数量	是否为真实遮挡物
MFSD	3 761	3 761	是
MeGlass	6 840	6 840	是
LFW-A	13 233	13 233	否
LFW-M	13 233	13 233	否

为评估SAPD对不同形状对抗补丁的泛化能力,采用3种补丁形状。图7展示了对抗补丁攻击所使用的二值补丁掩码。1)口罩状补丁1,覆盖口部及鼻下区域,遮挡约50%的面部特征区域;2)口罩状补丁2,遮挡鼻梁以下所有区域,覆盖约70%的面部特征;3)眼镜状补丁,模拟真实眼镜框架。本文全面评估了SAPD对形状与位置多样化的对抗补丁的鲁棒性。

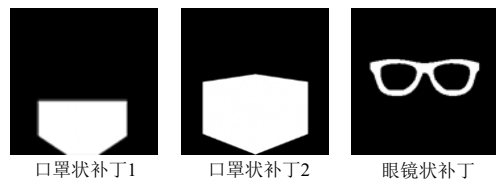


图7 3种对抗补丁

采用PGD攻击生成对抗补丁。为全面评估SAPD在不同威胁等级下的鲁棒性,测试3种扰动强度:低、中、高。默认攻击配置:迭代次数为100;步长 $\alpha=0.007$;扰动强度 $\varepsilon=0.3$;动量因子为0.9。

为评估跨模型泛化能力,在3种主流人脸识别模型上分别开展实验:MobileFaceNet^[39]、FaceNet^[30]和IRResNet50^[29]。将SAPD与4种代表性基线防御方法进行对比,分别为JPEG、LGS、SAC和RADAP。在人脸识别任务中,均采用错误率为0.1%时的正确率(TAR@0.001FAR)作为评估指标。

3.2 SAPD对比分析

不同人脸识别模型的实验结果如表3所示。在中等强度的对抗扰动下,在MobileFaceNet上,SAPD对眼镜状补丁、口罩状补丁1和口罩状补丁2的识别准确率达到84.41%、93.76%和92.81%,显著优于SAC,并大幅超越JPEG与LGS,也优于最新提出的RADAP,尤其在口罩状补丁1和口罩状补丁2场景下的识别准确率提升更为明显。此外,SAPD在不同模型上均保持稳定的性能,展现出跨模型迁移能力。不同于JPEG和LGS因严重性能退化而难以应对强扰动,SAPD采用基于MMS的图像修复策略,能够更自然地恢复遮挡区域,同时有效保留面部结构与特征连续性。与RADAP所采用的SAF策略相比,SAPD的MMS图像修复方法在不改变遮挡区域底层面部结构的前提下,实现了更自然的视觉重建,并更好地保持了特征连续性。

实际场景中,攻击者常采用不同强度的扰动来攻击模型。选用低、中、高3级扰动生成的LFW数据集进行测试,所选人脸识别模型均为经过训练具有遮挡鲁棒性的OAR模型,结果如表4所示。

表3 不同防御方法在主流人脸识别模型上的表现

识别模型	防御方法	识别准确率/%		
		眼镜状补丁	口罩状补丁1	口罩状补丁2
MobileFaceNet ^[39]	JPEG	24.47	22.43	25.36
	LGS	32.83	50.40	52.70
	SAC	14.08	81.80	83.66
	RADAP	82.85	92.40	91.67
	SAPD	84.41	93.76	92.81
FaceNet ^[30]	JPEG	24.47	24.80	18.23
	LGS	59.82	58.23	53.57
	SAC	11.61	86.86	87.53
	RADAP	85.63	89.10	89.43
	SAPD	86.70	90.80	90.90
IResNet50 ^[29]	JPEG	22.06	4.26	5.16
	LGS	55.83	18.40	19.37
	SAC	24.46	83.83	83.04
	RADAP	83.93	89.95	88.51
	SAPD	83.96	90.76	89.10

表4 不同攻击强度下SAPD的防御性能

识别模型	攻击方式	参数设置	识别准确率/%		
			眼镜状补丁	口罩状补丁1	口罩状补丁2
OAR	PGD	$\varepsilon=0.03, \alpha=0.003$	92.23	95.83	94.18
		$\varepsilon=0.05, \alpha=0.005$	92.50	96.16	95.67
		$\varepsilon=0.1, \alpha=0.01$	92.17	95.16	93.73

经SAPD防御后,对口罩状补丁1的识别准确率,在弱攻击下,达到95.83%;在中等强度攻击下,仍保持96.16%;即使在强攻击下,准确率仍稳定在95.16%。对口罩状补丁2的识别准确率,在弱攻击下,达到94.18%;在中等强度攻击下,达到95.67%;即使在强攻击下,准确率仍达到93.73%。这一结果表明,SAPD具备强大的抗干扰力。

综上所述,SAPD不仅在静态遮挡场景下优于现有防御方法,而且在动态扰动条件下展现出卓越的稳定性。该结果验证了SAPD作为一种通用、高效且实用的防御框架,能够有效应对人脸识别系统所面临的物理对抗补丁攻击。

3.3 SAPD消融实验

为验证所提出SAPD框架中各组件的有效性,围绕3个核心模块开展消融实验:分割网络、MMS图像修复策略和OAR损失函数。所有模型均在相同参数下使用LFW-Masked-Att数据集进行训练,以确保公平比较。主要评估指标为错误率为0.1%时的正确率(TAR@0.001FAR)。

1)分割网络消融实验。评估高精度分割网络

对SAPD的作用,结果如表5所示。采用SAC的粗粒度分割(仅粗略定位遮挡区域)时,性能受限:在口罩状补丁1和口罩状补丁2上的准确率分别为89.86%和89.50%;将其替换为ABANet分割网络后,准确率分别提升至92.17%和90.53%,表明精确的边界检测对于有效去除扰动至关重要;本文MMS策略,通过形态学平滑消除锯齿状边缘,准确率进一步提升至96.16%和95.67%。因此,像素级分割精度是彻底消除对抗噪声的基础前提。

表5 分割网络消融实验

识别模型	分割网络	修复方式	识别准确率/%		
			眼镜状补丁	口罩状补丁1	口罩状补丁2
OAR	ABANet	SAC	6.94	89.86	89.50
		MMS	46.73	92.17	90.53
		本文方法	92.50	96.16	95.67

2)MMS图像修复策略消融实验。在保持相同分割主干网络的前提下,评估不同图像修复策略的影响,结果如表6所示。SAC在口罩状补丁1上仅取得83.63%的准确率,主要由于边界残留噪声导致扰动去除不彻底;SAF通过考虑语义一致性将准确率提升至93.90%,但其修复过程会扭曲遮挡区域的主体结构,无意中掩盖了有效的面部特征;相比之下,MMS结合形态学去噪,保留遮挡区域的原始形状,还确保了面部纹理一致性,在口罩状补丁1和口罩状补丁2上的准确率分别达到96.16%和95.67%。上述结果表明,消除边界孤立噪点和分割残留对于正确识别人脸至关重要。

表6 修复策略消融实验

识别模型	分割网络	修复方式	识别准确率/%		
			眼镜状补丁	口罩状补丁1	口罩状补丁2
OAR	本文方法	SAC	16.67	83.63	82.86
		SAF	83.50	93.90	93.64
		MMS	92.50	96.16	95.67

3)OAR损失函数消融实验。最后,验证OAR损失在提升遮挡鲁棒性方面的作用,结果如表7所示。采用标准ArcFace损失训练的基线人脸识别模型在口罩状补丁1上的准确率为93.49%;引入OAR损失后,准确率提升至96.16%,提高了2.67个百分点。该提升源于两个机制:OAR损失引导网络聚焦于未遮挡的判别性区域(如眼睛、眉毛),降低对受损区域的依赖;在梯度层面引入正则化,防

止模型在训练过程中过拟合特定遮挡模式。因此,OAR损失不仅提升了识别准确率,还增强了模型在部分遮挡条件下的可解释性与稳定性。

表7 OAR损失函数消融实验

识别模型	分割网络	修复方式	神经网络	损失函数	识别准确率/%	
					口罩状补丁1	口罩状补丁2
FR	本文方法	MMS	Inception-ResNet-v1	ArcFace	93.49	92.33
OAR				ArcFace+OAR	96.16	95.67

3.4 实验结论

实验结果验证了SAPD的有效性与鲁棒性。在LFW-Masked-Att基准上,SAPD在中等强度扰动下,对眼镜状补丁、口罩状补丁1和口罩状补丁2的识别准确率分别达到92.50%、96.16%和95.67%,显著优于JPEG、LGS、SAC和RADAP等方法;在低、中、高3种攻击强度下,SAPD均保持稳定性能,准确率波动不超过2个百分点,展现出对自适应攻击者的强抗扰能力;此外,SAPD在MobileFaceNet、FaceNet和IResNet50等多种网络架构上均能一致提升识别准确率,且无需针对特定模型进行调优。

4 结束语

本文提出的SAPD防御框架,通过分割—修复—识别的协同机制,有效解决了对抗补丁与自然遮挡的防御难题,在无需对抗训练的前提下实现了鲁棒性与强泛化能力,并且保证了干净样本的识别准确率。这一设计验证了像素级精准定位与结构保持式修复在物理攻击防御中的关键作用。然而,当前方法仍依赖高质量分割掩码,在极端遮挡、多人物场景、复杂背景干扰下性能可能下降;同时,MMS策略对非规则形状补丁的适应性有待提升;且在无对抗遮挡条件下的人脸识别准确率仍无法达到普通样本的水准。未来工作将聚焦于:1)引入自监督或弱监督机制降低对真值掩码的依赖;2)探索动态形态学操作以适配任意形状扰动;3)应提升模型在遮挡条件下的人脸识别能力,达到普通样本下人脸识别的准确率。

参考文献:

[1] DENG J K, GUO J, XUE N N, et al. ArcFace: additive angular margin loss for deep face recognition [C] // Proceedings of the 2019 IEEE/CVF Conference on Computer

Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020:4685-4694.

- [2] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition [C] // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM, 2016:1528-1540.
- [3] KOMKOV S, PETIUSHKO A. AdvHat: real-world adversarial attack on ArcFace face ID system [C] // Proceedings of the 2020 25th International Conference on Pattern Recognition. Piscataway, USA: IEEE, 2021:819-826.
- [4] LIU X L, SHEN F R, ZHAO J, et al. RSTAM: an effective black-box impersonation attack on face recognition using a mobile and compact printer [DB/OL]. (2022-06-25) [2026-01-02]. <https://arxiv.org/abs/2206.12590>.
- [5] ZOLFI A, AVIDAN S, ELOVICI Y, et al. Adversarial mask: real-world universal adversarial attack on face recognition models [M] // AMINI M R, CANU S, FISCHER A, et al. Machine learning and knowledge discovery in databases. Cham, Switzerland: Springer, 2023:304-320.
- [6] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [M] // NAVAB N, HORNEGGER J, WELLS W, et al. Medical image computing and computer-assisted intervention. Cham, Switzerland: Springer, 2015:234-241.
- [7] OKTAY O, SCHLEMPER J, LE FOLGOC L, et al. Attention U-Net: learning where to look for the pancreas [DB/OL]. (2018-04-11) [2026-01-02]. <https://arxiv.org/abs/1804.03999>.
- [8] ZENG D, VELDHUIS R, SPREEUWERS L. A survey of face recognition techniques under occlusion [J]. IET Biometrics, 2021,10(6):581-606.
- [9] CHEN Z H, XU T R, HAN Z Y. Occluded face recognition based on the improved SVM and block weighted LBP [C] // Proceedings of the 2011 International Conference on Image Analysis and Signal Processing. Piscataway, USA: IEEE, 2011:118-122.
- [10] MIN R, HADID A, DUGELAY J L. Improving the recognition of faces occluded by facial accessories [C] // Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway, USA: IEEE, 2011:442-447.
- [11] 张思思,左信,刘建伟.深度学习中的对抗样本问题[J].计算机学报,2019,42(8):1886-1904.
- [12] CHEN Y J, SONG L X, HE R. Masquer hunter: adversarial occlusion-aware face detection [DB/OL]. (2017-09-15) [2026-01-02]. <https://arxiv.org/abs/1709.05188>.
- [13] SÁEZ TRIGUEROS D, MENG L, HARTNETT M. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss [J]. Image and Vision Computing, 2018,79:99-108.

- [14] MIN R, HADID A, DUGELAY J L. Efficient detection of occlusion prior to robust face recognition[J]. The Scientific World Journal, 2014, 2014: No.519158.
- [15] DZIUGAITE G K, GHAHRAMANI Z, ROY D M. A study of the effect of JPG compression on adversarial images[DB/OL]. (2016-08-02)[2026-01-02]. <https://arxiv.org/abs/1608.00853>.
- [16] HAYES J. On visible adversarial perturbations & digital watermarking[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, USA: IEEE, 2018: 1597-1604.
- [17] NASEER M, KHAN S, PORIKLI F. Local gradients smoothing: defense against localized adversarial attacks[C]//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision. Piscataway, USA: IEEE, 2019: 1300-1307.
- [18] CHIANG P Y, NI R K, ABDELKADER A, et al. Certified defenses for adversarial patches[DB/OL]. (2020-03-14)[2026-01-02]. <https://arxiv.org/abs/2003.06693>.
- [19] WU T, TONG L, VOROBAYCHIK Y. Defending against physically realizable attacks on image classification[DB/OL]. (2019-09-20)[2026-01-02]. <https://arxiv.org/abs/1909.09552>.
- [20] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[DB/OL]. (2017-06-19)[2026-01-02]. <https://arxiv.org/abs/1706.06083>.
- [21] 刘宏基, 黄瑞阳, 汪浣沙. 一种利用变分辨率敏感度不一致性的对抗性人脸检测方法[J]. 信息工程大学学报, 2025, 26(2): 175-181.
- [22] 钱亚冠, 张锡敏, 王滨, 等. 基于二阶对抗样本的对抗训练防御[J]. 电子与信息学报, 2021, 43(11): 3367-3373.
- [23] XIANG C, BHAGOJI A N, SEHWAG V, et al. PatchGuard: a provably robust defense against adversarial patches via small receptive fields and masking[DB/OL]. (2020-05-17)[2026-01-02]. <https://arxiv.org/abs/2005.10884>.
- [24] XIANG C, MAHLOUJIFAR S, MITTAL P. PatchCleanser: certifiably robust defense against adversarial patches for any image classifier[C]//Proceedings of the 31st USENIX Security Symposium. Berkeley, USA: USENIX Association, 2022: 2065-2082.
- [25] LIU J, LEVINE A, LAU C P, et al. Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 14953-14962.
- [26] 肖镇杰, 黄诗瑀, 叶锋, 等. 基于 PatchTracker 的对抗补丁防御算法[J]. 网络与信息安全学报, 2024, 10(1): 169-180.
- [27] LIU X L, SHEN F R, ZHAO J, et al. RADAP: a robust and adaptive defense against diverse adversarial patches on face recognition[J]. Pattern Recognition, 2025, 157: No.110915.
- [28] REZVANI S, FATEH M, KHOSRAVI H. ABANet: attention boundary-aware network for image segmentation[J]. Expert Systems, 2024, 41(9): No.e13625.
- [29] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 770-778.
- [30] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2015: 815-823.
- [31] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the 31th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2017: 4278-4284.
- [32] RADENOVIC F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with No human annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655-1668.
- [33] REZVANI S, YASIN R. MFSD: masked face segmentation dataset for face-related tasks[EB/OL]. (2024-03-17)[2026-01-02]. <https://github.com/sadjadrz/MFSD>.
- [34] MANTAS B, HERMANN T. Glasses Detector[EB/OL]. (2023-07-25)[2026-01-02]. <https://github.com/mantasu/glasses-detector>.
- [35] ANWAR A, RAYCHOWDHURY A. Masked face recognition for secure authentication[DB/OL]. (2020-08-25)[2026-01-02]. <https://arxiv.org/abs/2008.11104>.
- [36] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[DB/OL]. (2017-10-27)[2026-01-02]. <https://arxiv.org/abs/1710.10196>.
- [37] GUO J Z, ZHU X Y, LEI Z, et al. Face synthesis for eye-glass-robust face recognition[M]//ZHOU J, WANG Y H, SUN Z N, et al. Biometric recognition. Cham, Switzerland: Springer, 2018: 275-284.
- [38] TONG L, CHEN Z Z, NI J C, et al. FACESEC: a fine-grained robustness evaluation framework for face recognition systems[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 13249-13258.
- [39] CHEN S, LIU Y, GAO X, et al. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices[M]//ZHOU J, WANG Y H, SUN Z N, et al. Biometric recognition. Cham, Switzerland: Springer, 2018: 428-438.

(编辑: 李志豪)