面向指纹室内定位的高鲁棒性集成对抗训练方法

张学军[1]，李梅[1,2]，陈惠[1]，王国华[1]
（1.兰州交通大学电子与信息工程学院，甘肃 兰州 730070；2.兰州信息科技学院计算机与人工智能学院，甘肃 兰州 730300）

摘　要：针对指纹室内定位模型容易遭受对抗样本攻击以及传统对抗训练资源开销大、泛化能力弱等问题，提出了一种基于数据增强与蒸馏技术的集成对抗防御方法EDEAD。该方法利用数据蒸馏技术改善增广数据的质量，融合提前停止算法节省训练成本，并引入相干性梯度对齐损失项增强子模型对抗响应一致性的同时保持模型间的多样性，以降低对抗样本在定位模型间的可转移性和提升整个室内定位系统的鲁棒性及泛化能力。实验结果表明，在抵御强大黑盒攻击时，EDEAD方法相比于传统高鲁棒性的集成策略GAL和DVERGE，分别节省了30.6%和26.1%的时间开销，同时提升了70.6%和28.3%的定位精度。这验证了所提EDEAD方法在保证高鲁棒性的同时实现了效率优化。

关键词：室内定位；集成对抗训练；黑盒攻击；鲁棒性
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025138

High-robustness integrated adversarial training method for fingerprint-based indoor localization systems

ZHANG Xuejun[1], LI Mei[1,2], CHEN Hui[1], WANG Guohua[1]
1. School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
2. School of Computer and Artificial Intelligence, Lanzhou University of Information Science and Technology, Lanzhou 730300, China

Abstract: In response to the vulnerability of fingerprint-based indoor positioning models to adversarial sample attacks, as well as the high resource overhead and limited generalization ability of traditional adversarial training, an ensemble adversarial defense method based on data augmentation and distillation, named EDEAD, was proposed. In EDEAD, the data distillation technique was employed to improve the quality of the augmented data and the early stopping algorithm was used to save training costs. Additionally, a coherence gradient alignment loss term was introduced to enhance adversarial response consistency among sub-models while maintaining inter-model diversity. This effectively reduced the transferability of adversarial samples among different positioning models and enhanced the robustness and generalization of the entire indoor positioning system. Experimental results show that under strong black-box attacks, comparing to the traditional high-robustness ensemble strategies GAL and DVERGE, EDEAD reduces the time overhead by 30.6% and 26.1%, respectively, while improving positioning accuracy by 70.6% and 28.3%. These findings verify that EDEAD optimizes computational efficiency while maintaining high robustness.

Keywords: indoor localization, ensemble adversarial training, black-box attack, robustness

收稿日期：2025–05–26；**修回日期：**2025–08–03
基金项目：国家自然科学基金资助项目（No.61762058）；甘肃省重点研发计划基金资助项目（No.25YFFA089）；甘肃省教育厅产业支撑基金资助项目（No.2022CYZC-38）；甘肃省自然科学基金资助项目（No.25JRRA190）
Foundation Items：The National Natural Science Foundation of China (No.61762058), The Key Research and Development Project of Gansu Province (No.25YFFA089), Industrial Support Project of the Education Department of Gansu Provincial (No.2022CYZC-38), The Natural Science Foundation of Gansu Province (No.25JRRA190)

0 引言

随着无线通信和物联网技术的快速发展，室内定位技术伴随着位置服务（LBS, location-based service）技术的兴起，极大地增强了物联网提供服务的能力，改善了物联网的整体服务质量和用户体验，已在智慧城市、智慧园区、智慧家庭、智慧养老、智慧司法等诸多领域得到了广泛应用[1]。为了提供高质量的室内定位服务，学术界和工业界对室内定位技术进行了深入研究，提出了基于 Wi-Fi、射频识别、超宽带、蓝牙及数据图像等技术的室内定位系统[2]。在众多定位系统中，基于 Wi-Fi 接收信号强度（RSS, received signal strength）指纹的室内定位系统[3]，因其成本低、硬件要求低、部署方便、覆盖范围广等优点，已成为最具吸引力的室内定位解决方案之一，并已培育了许多商业应用。

Bahl 等[4]提出了一种基于射频识别的系统 radar，通过测量无线 RSS 指纹来定位和跟踪用户在室内的位置，是早期开创性作品。Mosleh 等[5]提出了一种基于粒子群优化的 K 最近邻算法，在去除信号不稳定性的同时，快速实现指纹室内定位。随着智能手机及物联网设备的快速普及，指纹室内定位技术通过结合无线信号特征以及传感器数据，实现更可靠的室内定位服务[6-7]。然而，RSS 信号受多径效应、衍射等影响，存在多种无线电波的干扰，使其在传播过程中出现不稳定的抖动，导致指纹室内定位的精度不高[8]。为克服复杂室内环境带来的不利影响，深度学习（DL, deep learning）技术被引入指纹室内定位系统[9]。尽管融合 DL 技术可以提取 RSS 指纹信号数据的复杂特征，从而提升定位性能，但针对定位系统的安全问题仍缺乏系统性探讨，这主要根源于 Wi-Fi 媒体的开放性和分类器的固有缺陷（如易遭受对抗样本攻击）。Patil 等[10]研究发现指纹室内定位系统易受对抗攻击，并使定位误差从 7.62 m 急剧增加到 170.41 m，致使其服务质量下降，若应用于工业领域，此类偏差将引发严重的安全事故。由此可见，室内定位技术虽然提高了人类在室内环境中与智能设备进行位置交互的精确度，但对抗样本的存在严重阻碍了指纹室内定位系统在真实物理世界中的应用[11]。

针对以上问题，本文从数据增强和模型优化双维度出发，提出了一种基于数据增强与蒸馏技术的集成对抗防御（EDEAD, enhancement and distillation ensemble adversarial defense）方法，以提升指纹室内定位系统在对抗场景下的稳定性和鲁棒性。本文主要贡献如下。

1) 利用蒸馏机制优化增强样本的质量，提升模型的定位精度和泛化能力。同时，结合扰动空间维度的量化分析，有效抑制对抗样本在不同定位模型之间的转移性，从而在提升对抗鲁棒性的同时，维持较高的原始定位精度。

2) 构建了一个面向 RSS 指纹定位系统的对抗防御框架，融合数据增强、数据蒸馏与对抗训练策略，在多种白盒和黑盒攻击下均表现出良好的适应性和鲁棒性。

3) 从扰动幅度、攻击类型等多个角度进行评估，在公开数据集 UJIIndoorLoc 上的实验结果表明，本文 EDEAD 方法在应对强攻击时，仍保持较高的定位精度，且相比传统的集成策略 ADP、GAL、DVERGE 与 FASTEN 分别提升了 23.1%、40.9%、12.8% 和 6.9%，验证了该方法在复杂室内环境下的抗干扰能力与实际应用潜力。

1 相关工作

1.1 对抗攻击

人工智能的安全问题日益受到关注，尤其是深度神经网络（DNN, deep neural network）存在的固有缺陷，使其容易受到对抗性攻击的威胁[12]。对抗性攻击主要指攻击者在原始数据 x 上添加人眼难以察觉的微小扰动 α，生成对抗样本 $x' = x + \alpha$，从而导致 DNN 模型 $f(\cdot)$ 预测出错，即 $f(x') \neq y$，其中 y 为样本的真实标签。这一现象揭示了 DNN 对输入扰动的高度敏感性，暴露了其在实际应用中的潜在安全隐患。

近年来，随着 DNN 在室内定位系统中的广泛应用，研究者发现这些定位模型在面对对抗样本攻击时普遍存在鲁棒性不足的问题。Szegedy 等[13]首次提出对抗样本的概念，并指出深度神经网络对于输入中微小扰动具有高度敏感性。Goodfellow 等[14]提出了快速梯度符号法（FGSM, fast gradient sign method），通过一次梯度计算即可生成具有攻击性的对抗样本。Madry 等[15]提出的投影梯度下降（PGD, projected gradient descent）法被认为是最强的一阶白盒攻击方法之一。针对黑盒场景，Papernot 等[16]引入了替代模型策略，提升了对抗样本的

可转移性。Dong 等[17]通过引入动量迭代快速梯度符号法（MI-FGSM, momentum iterative fast gradient sign method），增强了攻击路径的稳定性。Wu 等[18]进一步提出跳过梯度法（SGM, skip gradient method），重点研究了 ResNet 中残差连接对对抗样本迁移性的影响，证明了 SGM 在黑盒场景下的攻击性能更强。上述攻击方法在不同的网络结构和训练数据集上都表现出极高的错分率，给室内定位系统的安全性带来了极大的威胁。

1.2　对抗防御

1.2.1　对抗训练

对抗训练（AT, adversarial training）是目前最常用的防御策略之一，它通过将对抗样本融入训练过程以提升 DNN 模型的鲁棒性。文献[15]提出用一阶攻击方法 PGD 生成的对抗样本进行对抗训练，并将其形式化为 min-max 优化问题，缓解了在高维空间中获取最优扰动时梯度计算困难的问题。在对抗训练的研究基础上，防御蒸馏作为一种扩展策略被提出，用于进一步提升模型鲁棒性。文献[16]利用 DNN 模型知识来改变生成对抗样本时的梯度，使得对抗样本更难以生成，并在保证模型性能的前提下降低了计算资源需求。尽管上述方法在增强模型鲁棒性方面取得了一定的成效，但它们主要集中于提高输入层的抗干扰能力，普遍存在计算成本高、生成干净样本的准确性低、训练过程中对数据和模型复杂性要求高等问题，难以适用于复杂室内定位场景。为此，张等[8]从 RSS 指纹信号特征出发，设计了一种可抵御对抗攻击的室内定位模型，提高模型鲁棒性的同时保证了用户的隐私。Yan 等[19-20]提出了一种采用深度卷积生成对抗网络的防御策略[19]，并开发了一个抵御对抗攻击的安全框架 adv-LG[20]，其主要由基于 Transformer 的生成对抗网络和清洁模块组成，从模型优化角度消除对抗扰动，以增强基于通道状态信息的定位方法的安全性。然而，这些防御策略大多都侧重于增强单个模型，忽略了多个模型间潜在的信息交互，难以适用于实际场景。

1.2.2　集成对抗训练

为了提高对抗训练在实际场景中的防御效果，集成防御策略被提出，它通过整合多个子模型的预测结果，提高模型多样性并降低对抗样本的可转移性。集成对抗防御策略大致可分为数据层面的优化和模型层面的优化，具体如下。

1) 基于优化的多模型集成

集成方法 ADP[21]和 GAL[22]从模型优化的角度使每个子模型多样化。其中，ADP 利用正则化器最大化子模型之间非最大输出散度来鼓励模型多样性，使对抗样本难以在各个子模型之间转移；GAL 通过扩大输入梯度的余弦距离，降低共享对抗子空间的维数，从而提高集成模型的对抗鲁棒性。为了实现最优集成，这两项研究结合了多样性度量和数据验证技术，通过最小化式(1)，在提高干净样本精度和对抗鲁棒性的同时，增强了模型的泛化能力。

$$\min_{\theta} L_{\theta}(x,y) - \beta \cdot R(F,x) \rightarrow (x,y) \sim D \tag{1}$$

其中，L 表示交叉熵损失，θ 表示集成模型 F 训练时的学习参数，(x,y) 表示从 RSS 指纹数据 D 中采样的标签对，R 是一个正则化项，用来测量集成模型的多样性，β 表示平衡因子，用来衡量干净样本和对抗样本的比例。

2) 基于数据增强的多模型集成

目前的大多数研究更倾向于利用数据增强的思想实现更高的分类性能，DVERGE[23]、TRS[24]及 FASTEN[25]等集成防御方法是在优化过程中生成增强数据，以提高模型多样性，如式(2)所示。

$$\min_{\theta} L_{\theta}(x,y) - \beta \cdot R(F,x) \rightarrow (x,y) \sim A(F,D) \tag{2}$$

其中，A 是依赖于 (F,D) 的数据增广方法。DVERGE 方法采用特征蒸馏生成高质量的增广数据，并以交叉模型的方式优化训练参数。TRS 采用 PGD 攻击方法生成对抗样本从而作为增广数据，并使用梯度信息平滑模型的决策边界。FASTEN 方法利用循环增广技术 A 来降低模型训练的复杂度，并提出一种新的正则化器 R 来增强模型多样性。

基于模型优化的集成方法通过引入正则化项、优化模型参数或权重分布等手段增强模型鲁棒性，并提升了定位模型的抗干扰能力;但其未充分考虑在线阶段中对抗扰动的动态特性，且难以兼顾实时性与攻击强度的变化，导致在实际部署中存在适应性不足的问题。基于数据增强的集成方法通过简单的仿射变换、噪声注入及图像处理等技术扩充数据，虽然解决了数据量不足的问题，提高了模型在离线训练阶段的泛化能力，但这类方法通常缺乏对环境敏感特性的建模能力，难以动态响应 RSS 指纹

数据对多样性和真实性的需求。针对以上挑战，本文提出了双维度优化策略，从模型层面提升对抗扰动的动态响应能力，从数据层面增强RSS指纹数据的多样性与真实性，实现离线训练和在线定位两阶段的联合防御，有效提高模型的泛化能力与鲁棒性，特别是在面对多类型对抗攻击时展现出更强的稳定性与适应性。

2　方法设计

2.1　系统架构

在线定位服务过程中，用户提交RSS指纹数据及请求服务时，可能遭受对抗样本攻击，干扰室内定位模型的正常训练与预测，导致楼层分类大幅度出错[26]。为此，本文提出了一种适用于指纹室内定位的集成对抗防御方法，采用“离线训练+在线推理”的双阶段架构，兼顾鲁棒性与实时性，系统架构如图1所示。

离线训练阶段，终端设备采集RSS指纹数据（含Wi-Fi、蓝牙等信号类型），数据包含电磁噪声、多径、遮挡等干扰，并模拟伪造信号/恶意注入攻击。为适配CNN等室内定位的模型结构及提升计算效率，通过哈达玛积（Hadamard product）将预处理后的RSS指纹转化为统一尺寸的灰度图像输入。该方式也简化了后续计算流程，在一定程度上降低了资源和时间开销。基于该数据训练初始模型，进一步结合数据增强与典型对抗攻击方法（如FGSM、PGD、MIM等）生成混合训练集，并利用数据蒸馏筛选高质量指纹样本，通过提前停止算法选取“友好对抗样本”参与训练，提升模型的鲁棒性与泛化能力。在线推理阶段，服务器部署经过对抗训练的模型，对用户上传的RSS指纹灰度图像（含潜在对抗样本）进行一次前向推理，不需要动态扰动生成或进行复杂反向传播，计算量小、响应时延低，满足实际LBS系统对实时性的需求。

2.2　数据蒸馏与多模型协同优化的防御方法设计

2.2.1　数据增强与蒸馏优化

本文选择以指纹图像标签y为条件信息的条件生成对抗网络（CGAN, conditional generative adversarial network），生成更贴近真实RSS数据分布特征的新数据[26]。考虑到原始RSS指纹样本中可能存在信号强度极弱或不相关的AP数据，在预处理阶段首先剔除与定位位置弱相关的RSS值，从而提升指纹图像的质量与稳定性。随后，利用哈达玛积对RSS指纹向量进行图像化转化，将其重塑为统一大小的二维灰度图输入，以适配室内定位模型结构并增强建模能力。每张灰度图均结合建筑和楼层标签作为条件输入，用于指导CGAN生成器合成具有可控结构特征的RSS指纹图像，最终构建更真实的增强数据集。

在具体实现中，CGAN模块参考文献[26]进行

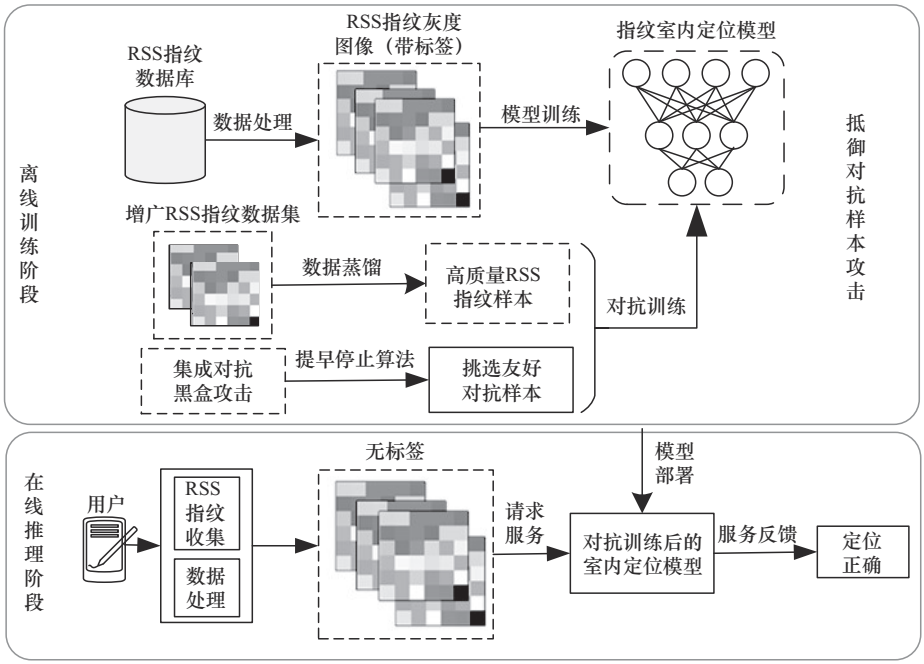


图1　系统架构

设计与构建，其生成器以随机噪声向量与标签信息为输入，输出模拟RSS特征的增强样本；判别器以原始与生成样本为输入，并联合标签信息完成真伪判别。通过对抗训练优化生成器与判别器的损失函数，使生成样本逐步逼近真实指纹数据分布。考虑到该模块并非本文研究重点，相关网络结构与超参数配置在此不再赘述，具体可参考文献[26]。

然而，文献[25]指出提升模型鲁棒性关键在于增广数据的质量，而非仅靠数量；过多且低质量的增强样本甚至可能削弱模型性能，并带来过高的训练成本。为此，本文进一步引入了蒸馏机制，对增广数据进行质量优化：通过不断从原始、初始化增强及其历史增强数据中提炼知识，逐步改善增广数据的质量。将训练历史数据中回收的先验知识视为教师，提取与原始输入相对应的非鲁棒特征，并将这些特征作为指导新生学习增强数据的知识。相较于单纯依赖大规模数据增强方法，蒸馏机制具有三方面突出优势：一是能更好地适应室内复杂场景中的多径效应与噪声干扰，提炼关键特征；二是降低对海量冗余样本的依赖，从而节约资源与计算开销；三是提升模型在跨场景、异构信号环境中的泛化与稳定性。数据蒸馏过程如图2所示。

因此，对于每个训练步骤t，从大量增广后的真实训练数据x_t^A中获取一小批次的高质量数据$x_t^M, M \ll A$，具体过程如式(3)所示。

$$x_t^M = x + \alpha g(x_t^A) + (1-\alpha) g(x_{t-1}^A) \tag{3}$$

在此框架下，引入平衡参数α调节原始数据与增强数据在特征提取过程中的权重比例。本文设定$\alpha = \dfrac{2}{255}$，旨在利用原始样本稳定性的同时，充分发挥增强数据的潜在贡献。蒸馏阶段采用软目标策略，并设置蒸馏温度$T = 4.0$（无量纲），该设定参考相关文献并经小规模验证确定，以增强教师模型Logits中的结构信息保留能力。直观地说，$g(x_t^A)$表示从当前初始化中得出的非鲁棒特征知识，$g(x_{t-1}^A)$表示从训练历史中积累的增强知识。离线训练的目标是寻找模型的最优权重参数，并使蒸馏后的数据集x_t^M及其定位标签y上学习到的每个定位模型f_i的交叉熵损失最小，具体如式(4)所示。

$$\arg\min_{\theta} L(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_\theta(x_t^M, y) \tag{4}$$

其中，θ为初始化的模型参数，n为聚合后的RSS指纹数据特征集大小，$\arg\min_{\theta} L(\theta)$表示为模型的优化目标。对于每一步骤$t$，用蒸馏后的数据$x_t^M$及每个采样的初始权重$\theta$，使用梯度下降更新当前参数$\theta^*$，用于在线定位阶段的楼层分类和位置预测，具体如式(5)所示。

$$\theta^* = \theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} L(x_t^M, y) \tag{5}$$

其中，η表示学习率，并利用更新后的权重参数θ^*重新训练模型，得到使式(4)最小化的高质量RSS指纹数据。

该集成训练方法在提取易受攻击的非鲁棒RSS指纹数据方面表现出优异的泛化性能，不仅能有效提高集成模型的鲁棒性，而且随着训练的进行，可以持续优化增强后指纹数据的自然性和分布一致性，避免定位模型学习到不真实的指纹特征。然而，这种集成训练方式也可能会陷入一个困境，即会导致所有定位模型都倾向于学习相似的、易受攻击的非鲁棒特征，这会显著降低模型的多样性和鲁棒性。为此，本文考虑使用对抗性示例来进一步增强鲁棒性，结合蒸馏后的高质量RSS指纹数据x_t^M和对抗样本x_t'更新集成模型，并融入对抗训练的思

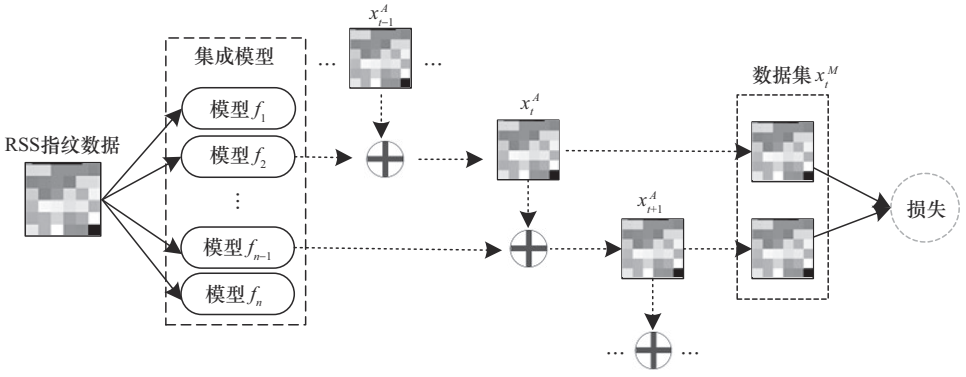


图2　数据蒸馏过程

想，即模型优化目标可重新表示为

$$\arg\min_{\theta^*} L(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}[\,L_{\theta^*}(x_t^M,y) + L_{\theta^*}(x_t',y)\,] \quad (6)$$

然而，在集成训练过程中，无限制引入对抗样本可能导致训练不稳定，甚至引发对抗扰动与干净样本的特征空间混叠，影响模型的泛化能力与收敛效果。为此，在训练阶段引入一种基于提前停止算法[27]的对抗样本选择机制。该机制通过实时评估对抗样本的扰动幅度与梯度变化率，筛选出已充分逼近决策边界的样本作为有效训练对抗样本，避免在扰动空间中过度偏移引起样本分布混乱。

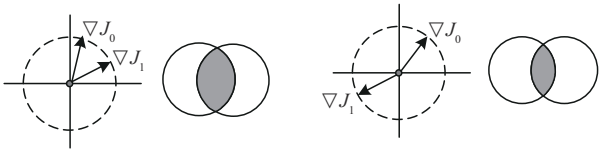
具体而言，设定扰动幅度阈值$\varepsilon_{\max}$，当生成的对抗样本满足$\|x'-x\|_2 \geqslant \varepsilon_{\max}$时，即认为该样本已逼近分类边界，满足对抗训练的最优要求，随后停止该样本的扰动更新过程，作为最终对抗样本纳入损失优化中。这种提前停止策略一方面控制了训练开销，另一方面有效减少了干净样本与对抗样本间的不合理交叉重叠，进一步提升了对抗训练的收敛效率与模型整体鲁棒性。

2.2.2 对抗空间维度度量与优化

集成学习的目标是在保持模型内部相似性的同时，实现模型间的多样性，以确保室内定位系统的稳定性和鲁棒性。通常情况下，对抗示例存在于RSS指纹数据的有限扰动空间中，为此，量化扰动空间的维度成了目前有效的解决办法。然而目前的梯度对齐对抗子空间[28]和局部内在维数[29]等方法，无法测量RSS指纹数据扰动空间的维度，原因在于RSS数据扰动空间不连续且含多个不可微分处理模块，导致计算开销过高。因此，本文设计了一种适用于复杂室内定位场景且计算成本低廉的扰动空间维度测量方法，并在训练过程中引入正则化项，即相干性梯度对齐损失（CGAL, coherence gradient alignment loss）。特别地，CGAL的设计不仅依赖于含有多径干扰的训练数据，还结合了室内定位中常见的多径效应与局部信号遮挡等复杂环境因素，通过约束模型间扰动空间的重叠度，提高系统在多径和信号波动环境中的鲁棒性。

在室内定位场景中，RSS指纹数据通常来自多个信号源（如Wi-Fi或蓝牙），并受周围环境因素的影响较大（墙壁、家具和人员的遮挡等）。为了生成能同时欺骗多个模型的对抗样本，这些样本需位于各模型共享的扰动空间中，而集成模型的扰动空间维度与单模型扰动空间重叠区域成正比。下面以室内定位模型f_0和f_1为例，探讨其扰动空间的重叠情况，它们相对于RSS指纹数据x的损失梯度表示为$\nabla_x J_0(\theta,x,y)$和$\nabla_x J_1(\theta,x,y)$，其描述了x被扰动的方向，以最大限度地增加损失函数。当f_0和f_1的梯度相互对齐时，对扰动的响应呈正相关，即损失J_0的增加可能导致J_1增加，这表明f_0和f_1在面对对抗扰动时具有相似的脆弱性，并享有较大的扰动空间，梯度对齐的关系示意如图3(a)所示。相反，当梯度未对齐时，两模型不太可能被相同的扰动所欺骗，导致共享扰动空间的维数较低，梯度未对齐的关系示意如图3(b)所示。通过计算2个模型梯度的余弦相似度，可量化扰动空间的重叠程度，从而评估定位模型的鲁棒性，如式(7)所示。

$$\mathrm{Sim}(\nabla_x J_0,\nabla_x J_1) = \frac{\nabla_x J_0 \cdot \nabla_x J_1}{\|\nabla_x J_0\| \cdot \|\nabla_x J_1\|} \quad (7)$$



(a) 梯度对齐 (b) 梯度未对齐

图3 梯度对齐和重叠的关系示意

余弦相似度的取值范围为[–1,1]。理想情况下，希望2个室内定位模型的余弦相似度为–1，即梯度方向完全相反，使扰动空间无重叠。该思想可以推广到N个模型的集合，并期望它们的梯度向量$\{\nabla_x J_i\}_{i=1}^N$尽可能地错位。本文通过评估这些向量集合的相干性来衡量最大余弦相似度，如式(8)所示。

$$\mathrm{Coh}(\{\nabla_x J_i\}_{i=1}^N) = \max_{\substack{a,b\in\{1,\cdots,N\}\\ a\neq b}} \mathrm{Sim}(\nabla_x J_a,\nabla_x J_b) \quad (8)$$

由于相干性计算涉及非平滑函数，使用梯度下降优化算法会很慢。因此，本文利用平滑近似提高收敛速度。具体来说，使用LogSumExp代替式(8)中的最大值运算，得到式(9)所示的相干性梯度对齐损失CGAL，从而测量扰动空间的重叠程度。

$$\mathrm{CGAL} = \lg\left(\sum_{1\leqslant a<b\leqslant N} \exp(\mathrm{Sim}(\nabla_x J_a,\nabla_x J_b))\right) \quad (9)$$

当集成模型中输入x的CGAL值较低时，表明很难生成能够同时欺骗多个模型的对抗性示例x'。

为了使集成模型具有较低的CGAL值，本文尝试在训练期间将其作为正则化项加入损失函数中，得到如式(10)所示的损失函数。

$$\arg\min_{\theta^*} L(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}\left[L_{\theta^*}(x_t^M, y) + L_{\theta^*}(x_t', y)\right] + \lambda \mathrm{CGAL} \quad (10)$$

式(10)中第一部分是集合中每个模型对干净样本的平均交叉熵损失和对抗训练过程中生成的对抗样本的损失，第二部分表示CGAL损失。其中，λ是一个超参数，用于平衡干净样本损失与梯度对齐项的重要性。本文在小规模验证集上进行了启发式调参，最终选取$\lambda = 0.7$，兼顾了干净样本的准确率与对抗鲁棒性的提升效果。该方法可有效降低对抗样本在模型间的可转移性，从而提升整个室内定位系统的对抗鲁棒性。

本文EDEAD方法的具体流程如算法1所示。

算法1　EDEAD方法框架

输入　子模型数量n，RSS指纹数据集D，批次大小b，最大训练轮次e

输出　集成定位模型F

1) for i = 1 to n do
2)　初始化集成模型F的每个子模型f_i
3) end for
4) # 数据增强与蒸馏优化
5) for k = 1 to e do
6)　for j = 1 to b do
7)　从数据集D中随机选择指纹标签对(x,y)
8)　使用条件生成对抗网络*CGAN*生成增强数据x_t^A（式(3)）
9)　for i = 1 to n do;
10)　使用梯度下降优化模型权重参数θ^*（式(5)）
11)　计算交叉熵损失$L(\theta^*)$
12)　计算相干性梯度对齐损失CGAL（式(9)）
13)　更新并最小化总损失$L(\theta^*)$（式(10)）
14)　end for
15)　end for
16) end for

3　实验结果与分析

3.1　环境设置及数据集

1) 实验使用PyTorch库来构建抵御对抗样本攻击的室内定位模型。操作系统为Windows 10，RAM大小为64 GB，GPU为GeForce RTX 3060。所有集成模型的训练轮次epoch都为200，对于SGD优化器，动量参数设置为0.9，初始学习率为0.1，在120~170轮次动态调整学习率。对于Adam优化器，初始学习率设为0.1，权重衰减设置为0.000 1。

2) 为了验证所提方法的有效性，RSS指纹数据选520维的UJIIndoorLoc数据集[30]和35维的Mall_Wi-Fi数据集[31]，并将完整数据集划分为70%的训练集和30%的测试集。

3.2　不同集成防御方法的时间开销分析

本节分析了不同防御方法在UJIIndoorLoc数据集上的训练时间和测试时间（白盒攻击和黑盒攻击），测试时间以PGD-10攻击为例进行分析，如表1所示。

表1　不同防御方法的训练开销分析

方法	训练时间/s	白盒测试时间/s	黑盒测试时间/s
Baseline	2 907	2 216	2 100
ADP	3 180	2 846	2 940
GAL	4 336	9 640	4 420
DVERGE	9 766	2 229	4 140
TRS	10 095	3 900	5 820
FASTEN	4 920	4 222	2 340
本文方法	5 460	4 920	3 060

从表1可见，Baseline方法因未引入正则化与数据增强，训练与测试时间均最短。基于优化的集成方法（ADP和GAL）训练耗时略高于Baseline方法，但GAL因RSS指纹扰动空间缺乏连续性，测试开销显著增大；基于数据增强的方法（DVERGE和TRS）因多步生成高质量数据，训练成本急剧增加，其中TRS还因梯度平滑性与相似性权重失衡导致扩展性受限。相比之下，FASTEN方法通过循环增强策略加快数据生成，节省了约一半的训练成本。本文方法在训练阶段虽需计算CGAL正则项并

引入余弦相似度，但通过提前停止策略减少冗余迭代，较DVERGE节省了44.1%训练耗时；在线推理阶段仅需单次前向传播，无对抗扰动生成或复杂后续处理，黑盒测试时间较GAL降低30.6%，显著优于多数对比方法。此外，基于实验室自研的Android RSS指纹采集APP，目前已实现真实环境下的RSS指纹数据采集与本地预处理，为后续线上推理实验奠定了基础。虽然当前阶段尚未完成端到端推理流程的真实验证与功耗测试，但结合计算流程的简洁性与表1的实验结果，可初步证明本文方法满足室内定位场景对实时性与资源消耗的需求。综合来看，本文方法在保持高鲁棒性的同时，实现了训练–推理成本的良好平衡。

3.3　白盒攻击鲁棒性

对于白盒攻击，选用基于10步的投影梯度下降迭代法（PGD-10）和经典的快速梯度符号法（FGSM）评估定位模型的鲁棒性，步长设为$\frac{\varepsilon}{5}$，随机开始设置为5，并在UJIIndoorLoc和Mall_Wi-Fi数据集上评估不同集成方法的对抗鲁棒性。

1) UJIIndoorLoc数据集上的对抗鲁棒性评估

实验比较了在UJIIndoorLoc数据集上Baseline、ADP[21]、GAL[22]、DVERGE[23]、FASTEN[25]以及本文方法在没有对抗训练和引入对抗训练下的白盒鲁棒性，此处以3个子模型为例进行评估，无对抗训练PGD-10攻击下不同方法的白盒鲁棒性如图4所示，对抗训练PGD-10攻击下不同方法的白盒鲁棒性如图5所示。

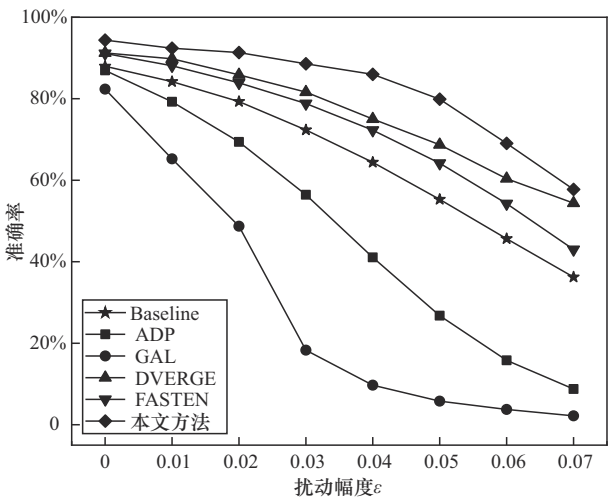


图4　无对抗训练PGD-10攻击下不同方法的白盒鲁棒性（UJIIndoorLoc数据集）

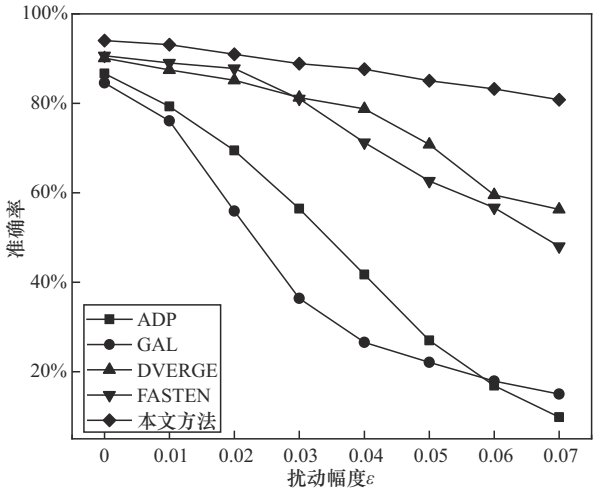


图5　对抗训练PGD-10攻击下不同方法的白盒鲁棒性（UJIIndoorLoc数据集）

由图4和图5可以看出，在UJIIndoorLoc数据集抵御PGD-10强攻击时，基于优化集成方法（ADP和GAL）的白盒鲁棒性显著弱于数据增强的集成方法（DVERGE和FASTEN）。当扰动幅度ε增大到0.07时，定位精度均低于20%，表明优化集成方法难以有效抵御强攻击。Baseline方法通过移除正则化与数据增强，实现最短训练与测试时间，但未引入对抗训练导致在PGD-10攻击下精度骤降至36.2%。DVERGE能捕获子模型的非鲁棒特征，并引导其学习不同特征子空间，有效阻止对抗样本的迁移，防御效果更强，但训练成本较高（如表1所示）。FASTEN虽降低了数据增强成本，但可能无法全面覆盖复杂RSS指纹数据分布，抵御攻击时稳定性略逊于DVERGE。由于TRS方法在复杂室内定位数据集上难以实现有效集成训练，本文选择不在实验中展示该方法的对比性能。相比而言，本文方法结合友好对抗训练与数据蒸馏，在扰动幅度ε增大到0.07时，仅牺牲0.4%干净样本精度，仍能保持80.8%的定位精度。尽管高维扰动空间测量过程增加了部分计算开销，但在实际复杂场景下依然展现出更优的稳定性，因此，综合性能优于现有方法。

2) Mall_Wi-Fi数据集上的对抗鲁棒性评估

本节比较了在Mall_Wi-Fi数据集上各种集成方法在没有对抗训练和引入对抗训练下的白盒鲁棒性，以3个子模型为例进行评估，无对抗训练PGD-10攻击下不同方法的白盒鲁棒性如图6所示，对抗训练PGD-10攻击下不同方法的白盒鲁棒性如图7所示。

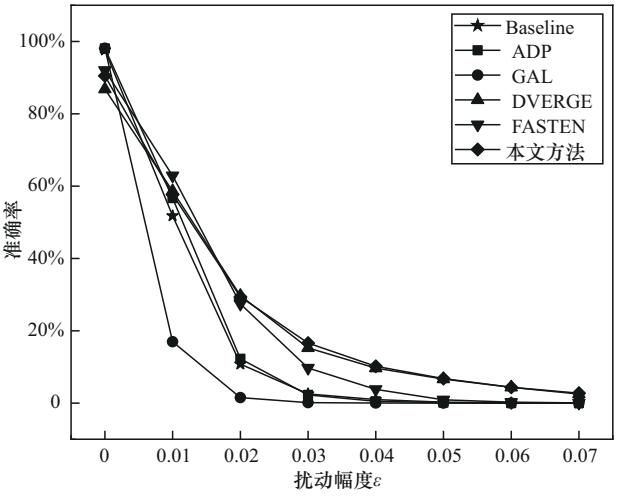


图6　无对抗训练PGD-10攻击下不同方法的白盒鲁棒性（Mall_Wi-Fi数据集）

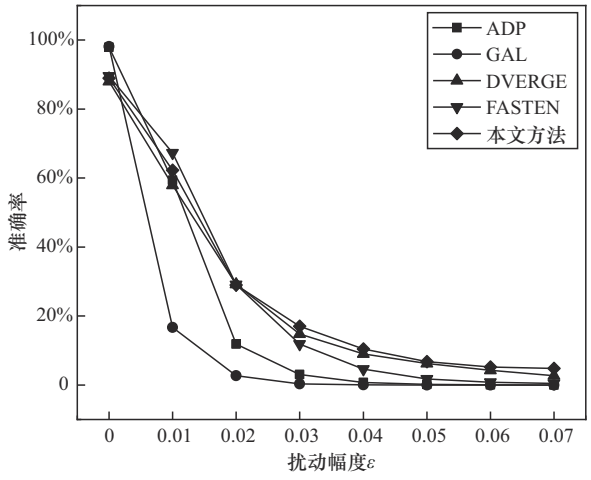


图7　对抗训练PGD-10攻击下不同方法的白盒鲁棒性（Mall_Wi-Fi数据集）

在Mall_Wi-Fi数据集上的实验结果表明，不同集成方法的鲁棒性在面对强大的PGD-10攻击时，定位精度都随着扰动幅度ε的增大而急剧下降，但对抗训练的引入仍然提升了定位模型的鲁棒性。其中，Baseline方法和基于优化的方法（ADP和GAL）在干净样本上的定位精度分别高达97.6%、98.1%和98.1%，但在加入小扰动后定位精度急剧下降，当扰动幅度为0.07时，精度低至0，定位完全失效。基于数据增强的方法（DVERGE和FASTEN）相比于基于优化的方法，在损失了一定干净样本精度的同时，在不同扰动幅度下显著提升了白盒鲁棒性，尤其在同等扰动幅度ε= 0.02时，FASTEN相比GAL方法定位精度提升了45.9%，加入对抗训练后提升了50.7%；DVERGE相比于其他方法相对稳定，即使攻击强度很大，定位依然有效。而本文EDEAD方法相比DVERGE方法，在加入对抗训练后，不论是在干净样本上的准确性，还是对抗样本上的鲁棒性，虽然仅提升了0.96%和0.29%，但值得注意的是，在白盒测试环境下，本文方法显著降低了54.7%的测试时间开销。

3) FGSM攻击下不同方法的白盒鲁棒性对比

本节对比了不同方法在FGSM攻击下的白盒鲁棒性，并在UJIIndoorLoc和Mall_Wi-Fi这2种数据集上进行了实验分析，如表2所示。

从表2可以看出，本文方法在面对快速梯度符号法FGSM攻击时，展现出了显著的鲁棒性优势。在复杂的高维数据集UJIIndoorLoc上，加入对抗训练的本文方法在攻击强度为0.07时仍能保持90.1%的定位准确率，较FASTEN和DVERGE方法分别提升了19.5%和23.9%，在无对抗训练的情况下，其定位准确率仍高于其他方法，证明本文方法在复杂环境下具有更强的抗干扰能力。在Mall_Wi-Fi数据集上，本文方法加入对抗训练后的高强度攻击下定位准确率仅达25.4%，但是相比ADP、GAL、DVERGE、FASTEN方法分别提升了22.8%、2.4%、15.5%、21.9%。这表明在较简单的数据集上，尽管本文方法相比其他集成防御方法提供了一定的防御优势，但由于数据集本身简单和对抗攻击的高效性，集成防御方法的优势未能充分体现，反而因过度的防御策略导致定位精度的损失。值得注意的是，GAL方法在加入对抗训练后表现极其不稳定，添加扰动之后模型精度快速下降，这可能是因为模型在训练过程中存在梯度对齐失效或过拟合的问题，导致防御策略未能有效适应对抗攻击。而本文方法通过集成对抗训练与数据蒸馏技术，有效平衡了模型多样性和鲁棒性，在强攻击强度下仍能保持高定位精度，但计算成本较高，尤其在扰动空间维度测量过程中，可能会限制大模型部署效率。因此，尽管本文方法在高维数据集和复杂环境下展现了优异的防御性能，但在简单数据集或对计算效率有更高要求的应用场景中，仍具有很好的优化潜力。

3.4　黑盒攻击鲁棒性评估

通过5种黑盒攻击评估定位模型的鲁棒性，包括FGSM、MIM、SGM、投影梯度下降迭代法PGD-10及PGD-100，并分别在UJIIndoorLoc和Mall_Wi-Fi数据集上进行实验分析。

表2 FGSM攻击下不同方法的白盒鲁棒精度

数据集	方法	不同ε的对抗训练下的白盒鲁棒精度					不同ε的无对抗训练下的白盒鲁棒精度				
		干净样本精度	0.01	0.03	0.05	0.07	干净样本精度	0.01	0.03	0.05	0.07
UJIIndoorLoc	Baseline	—	—	—	—	—	88.0%	85.7%	79.8%	72.2%	63.1%
	ADP	86.1%	82.2%	72.8%	61.4%	49.4%	86.7%	82.5%	72.9%	62.0%	49.1%
	GAL	90.8%	55.6%	20.7%	14.9%	14.0%	92.1%	80.6%	51.6%	33.4%	22.8%
	DVERGE	84.5%	82.5%	77.5%	72.2%	66.2%	89.8%	88.0%	83.4%	78.1%	71.5%
	FASTEN	90.1%	88.5%	84.2%	78.6%	71.6%	91.1%	89.1%	83.8%	77.4%	70.2%
	本文方法	94.9%	93.7%	92.7%	91.1%	90.1%	97.0%	92.3%	90.6%	85.6%	78.4%
Mall_Wi-Fi	Baseline	—	—	—	—	—	87.3%	26.8%	9.4%	3.7%	1.4%
	ADP	97.5%	75.7%	17.1%	5.4%	2.6%	97.6%	75.5%	15.6%	5.11%	2.11%
	GAL	98.1%	77.5%	41.7%	28.7%	23.0%	98.0%	75.9%	38.0%	22.7%	15.6%
	DVERGE	86.8%	63.0%	28.1%	14.9%	9.9%	88.8%	64.3%	27.8%	13.8%	9.2%
	FASTEN	95.0%	75.3%	22.9%	8.2%	3.5%	95.1%	75.7%	23.0%	8.8%	4.1%
	本文方法	98.1%	84.6%	47.6%	31.5%	25.4%	98.2%	83.4%	46.5%	30.4%	23.1%

1) UJIIndoorLoc数据集

本节对比了不同攻击方法在UJIIndoorLoc数据集上的黑盒鲁棒精度，如表3所示，评估各个集成方法面对不同类型攻击时的防御效果。

针对5种攻击方法的综合评估表明，对抗训练的引入显著提升了模型的防御能力，但不同防御方法的性能差异显著。基于优化的集成方法ADP和GAL的鲁棒性较弱，尤其在没有引入对抗训练时，在

表3 UJIIndoorLoc数据集下不同攻击方法的黑盒鲁棒精度

方法	攻击方法	不同ε的对抗训练下的黑盒鲁棒精度				不同ε的无对抗训练下的黑盒鲁棒精度			
		0.01	0.03	0.05	0.07	0.01	0.03	0.05	0.07
ADP	FGSM	82.2%	72.8%	61.4%	49.4%	82.5%	72.9%	62.0%	49.1%
	PGD-10	79.5%	59.9%	33.6%	16.7%	79.4%	57.6%	28.7%	10.1%
	MIM	79.2%	59.4%	34.0%	17.8%	79.1%	56.8%	28.9%	12.1%
	PGD-100	78.8%	56.4%	27.8%	11.0%	78.8%	53.3%	22.2%	6.4%
	SGM	81.5%	69.1%	51.2%	34.8%	79.9%	63.0%	43.7%	29.9%
GAL	FGSM	78.7%	53.1%	37.8%	26.7%	55.6%	20.8%	14.9%	14.1%
	PGD-10	72.1%	38.4%	19.9%	11.2%	28.6%	7.9%	6.9%	8.2%
	MIM	75.7%	34.8%	12.6%	7.4%	28.2%	9.6%	10.4%	11.2%
	PGD-100	70.2%	30.4%	15.6%	9.1%	22.6%	4.5%	2.3%	2.1%
	SGM	65.9%	46.8%	27.4%	18.4%	46.2%	15.0%	13.7%	13.6%
DVERGE	FGSM	88.0%	83.4%	78.1%	71.5%	87.9%	83.7%	78.3%	71.1%
	PGD-10	87.3%	80.2%	69.2%	53.5%	87.3%	80.5%	69.0%	52.9%
	MIM	87.2%	80.1%	69.4%	54.9%	87.2%	80.4%	69.2%	54.5%
	PGD-100	87.1%	79.4%	66.4%	48.5%	87.1%	79.6%	66.3%	48.2%
	SGM	87.8%	82.5%	74.5%	62.2%	87.7%	82.7%	74.2%	61.9%
FASTEN	FGSM	88.5%	84.2%	78.6%	71.6%	89.1%	83.8%	77.4%	70.2%
	PGD-10	87.8%	80.9%	68.8%	51.2%	88.1%	79.1%	65.2%	45.0%
	MIM	87.8%	80.7%	69.0%	52.8%	88.0%	78.9%	65.3%	47.3%
	PGD-100	87.7%	80.4%	67.3%	45.4%	87.9%	77.8%	61.4%	37.4%
	SGM	88.3%	83.4%	75.5%	63.6%	88.8%	82.3%	72.8%	59.0%
本文方法	FGSM	92.0%	90.0%	87.4%	84.1%	88.0%	83.7%	78.3%	71.1%
	PGD-10	91.9%	89.7%	86.4%	81.1%	87.3%	80.5%	69.0%	52.9%
	MIM	91.9%	89.7%	86.5%	81.5%	87.2%	80.4%	69.2%	54.5%
	PGD-100	92.0%	89.7%	86.2%	80.3%	87.1%	79.6%	66.3%	48.2%
	SGM	92.0%	89.8%	86.5%	81.8%	87.7%	82.7%	74.2%	61.9%

PGD-10攻击下，ADP方法随扰动幅度增至0.07时定位精度骤降至10.1%，GAL方法降至8.2%，即使引入对抗训练技术，二者定位精度也仅达16.7%和11.2%；尤其是GAL方法在加入对抗训练技术后随扰动幅度增大定位精度极其不稳定，可能因为在高扰动下对抗样本的生成失效，导致子模型集成时响应不稳定。相比之下，基于数据增强的DVERGE方法在对抗训练后面对强大的PGD-100攻击仍能保持48.5%的定位精度，分别高于ADP、GAL、FASTEN方法37.5%、39.4%、3.1%，其优势源于差异化特征学习阻断了对抗样本的迁移，但训练成本较高。

从攻击类型的影响来看，多步迭代攻击对模型的破坏性显著高于单步攻击，例如，在无对抗训练时，ADP在面对PGD-100攻击时定位精度为11.0%，而在单步攻击FGSM下其定位精度可达49.4%。DVERGE则表现出更强的适应性，即使没有引入对抗训练技术，在面对SGM攻击时其定位精度仍达61.9%。值得注意的是，本文方法通过融合友好对抗训练及数据蒸馏技术，在保证高鲁棒性的同时实现了效率优化，面对多类型的攻击且扰动幅度为0.07时，定位精度都高达80%以上。虽然数据蒸馏技术在高维扰动空间的计算成本仍待优化，但其通过提升样本质量，在MIM、PGD-100等强攻击场景中展现出了稳定的防御效能，无对抗训练不同攻击下的黑盒鲁棒性如图8所示，对抗训练下不同攻击下的黑盒鲁棒性如图9所示。综上，本文方法从扰动幅度、攻击方法多角度来看，都能有效实现最佳性能，但计算效率与鲁棒性的协同优化仍是未来需研究的核心挑战。

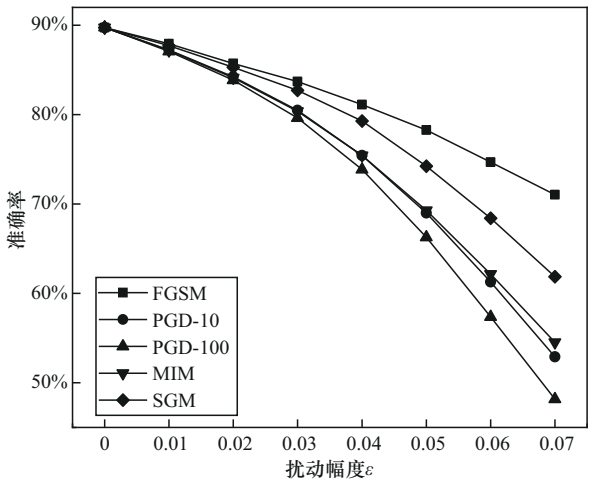


图8　无对抗训练不同攻击下的黑盒鲁棒性

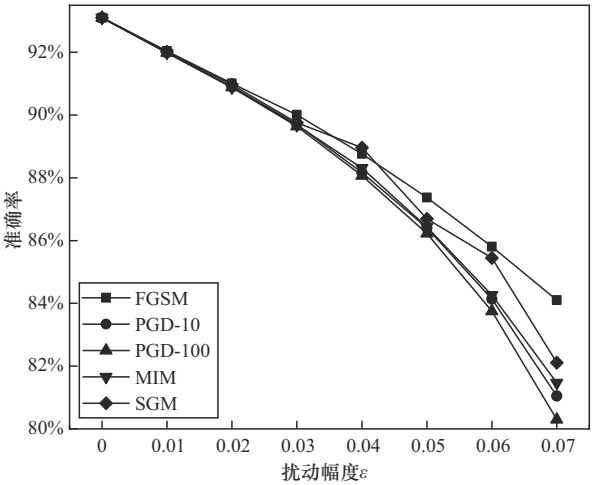


图9　对抗训练不同攻击下的黑盒鲁棒性

2) Mall_Wi-Fi数据集

本节对比了不同攻击方法在Mall_Wi-Fi数据集上的黑盒鲁棒精度，如表4所示，评估各个集成方法面对不同类型攻击时的防御效果。

从前面的分析可观察到，相比于UJIIndoorLoc数据集上不同攻击方法的黑盒鲁棒性，Mall_Wi-Fi数据集的鲁棒性显著下降，可能归因于集成策略对于简单小型数据集并非必要，从而影响了模型的泛化能力，尽管如此，本文方法相较于其他方法，仍然有明显的性能提升。尤其在加入对抗训练技术后，面对强大的PGD-100攻击时，即使扰动幅度较弱（ε=0.01），本文方法相比于ADP、GAL、DVERGE、FASTEN等方法分别提升了23.1%、40.9%、12.8%、6.9%。在面对FGSM攻击时，即使扰动幅度很强（ε=0.07），本文方法仍能达到19.3%的定位精度，而其他方法ADP、DVERGE、FASTEN的定位精度都接近0，使定位系统基本失效，无法完成高质量的定位服务；GAL方法也可达到10.1%的定位精度；这是因为本文方法的梯度对齐损失能有效测量扰动子空间，且在面对弱攻击时，并没有出现梯度对齐失效的问题。由此可见，对抗训练对于高强度攻击的防御更有效，本文方法通过数据增强及模型优化显著提升了黑盒场景下的鲁棒性。

3.5　消融实验

本节以Baseline方法为起点，以本文方法的性能为参照。通过3组消融实验验证各组件对模型性能的影响：1) 移除数据蒸馏模块（w/o distill）；2) 禁用早停机制（w/o earlystop）；3) 取消梯度对齐损失（w/o CGAL）。消融实验结果对比如表5所示。

表4 Mall_Wi-Fi数据集下不同攻击方法的黑盒鲁棒精度

方法	攻击方法	不同ε的对抗训练下的黑盒鲁棒精度				不同ε的无对抗训练下的黑盒鲁棒精度			
		0.01	0.03	0.05	0.07	0.01	0.03	0.05	0.07
ADP	FGSM	45.4%	10.0%	3.1%	1.0%	44.1%	9.4%	3.7%	1.4%
	PGD-10	38.6%	3.2%	0.3%	0	37.8%	2.9%	0.2%	0
	MIM	39.5%	3.8%	0.6%	0.1%	38.9%	3.7%	0.1%	0
	PGD-100	36.3%	1.8%	0.1%	0	35.8%	1.5%	0	0
	SGM	36.8%	2.7%	0.4%	0.1%	36.8%	2.5%	0	0
GAL	FGSM	61.3%	37.6%	26.4%	20.6%	56.5%	32.6%	15.6%	10.1%
	PGD-10	27.2%	2.7%	0.7%	0.2%	25.2%	2.6%	0.6%	0.4%
	MIM	32.4%	4.6%	1.0%	0.4%	28.9%	4.2%	0.9%	0.7%
	PGD-100	18.5%	0.5%	0	0	17.1%	0.6%	0.1%	0
	SGM	42.3%	10.1%	5.3%	3.2%	37.0%	9.9%	5.1%	3.0%
DVERGE	FGSM	50.2%	19.7%	9.4%	4.8%	53.8%	8.7%	1.9%	0.7%
	PGD-10	47.4%	12.5%	4.0%	0.9%	47.2%	3.4%	0.5%	0
	MIM	47.6%	12.6%	4.1%	1.3%	48.8%	3.8%	0.6%	0
	PGD-100	46.6%	10.5%	2.3%	0.1%	44.8%	1.9%	0.2%	0
	SGM	47.4%	12.5%	3.8%	0.7%	44.8%	2.3%	0.2%	0
FASTEN	FGSM	54.6%	15.3%	4.8%	1.3%	54.8%	15.4%	4.8%	1.1%
	PGD-10	53.1%	11.1%	2.1%	0.4%	53.0%	11.6%	2.7%	0.3%
	MIM	53.6%	11.8%	2.5%	0.5%	53.5%	11.9%	3.1%	0.4%
	PGD-100	52.5%	10.3%	1.5%	0.1%	52.5%	10.3%	2.1%	0.1%
	SGM	52.6%	10.2%	1.3%	0.1%	52.5%	10.5%	2.1%	0.1%
本文方法	FGSM	65.3%	30.2%	16.3%	11.4%	68.0%	29.0%	22.9%	19.3%
	PGD-10	60.9%	18.4%	8.6%	4.4%	64.0%	18.7%	8.2%	4.1%
	MIM	61.0%	18.7%	8.9%	4.3%	64.1%	19.5%	8.0%	4.1%
	PGD-100	59.4%	15.8%	5.8%	1.7%	62.9%	16.0%	6.2%	2.0%
	SGM	61.0%	18.1%	7.0%	3.0%	63.7%	17.7%	7.1%	3.6%

表5 消融实验结果对比

方法	干净样本精度	FGSM	PGD	MIM	SGM
Baseline	88.0%	63.1%	38.2%	39.6%	52.0%
w/o distill	89.9%	81.0%	76.8%	77.3%	80.2%
w/o earlystop	92.9%	83.5%	79.5%	80.0%	81.3%
w/o CGAL	92.2%	78.6%	73.2%	74.0%	77.5%
本文方法	93.1%	84.1%	81.0%	81.5%	81.8%

实验结果表明，数据蒸馏模块对模型泛化能力具有关键作用，移除后导致干净样本准确率下降3.2%，FGSM防御性能下降3.1%；梯度对齐损失CGAL对抵御多步迭代攻击尤为重要，取消该组件使本文方法对PGD攻击的鲁棒性显著下降7.8%，这验证了梯度相干性约束对增强决策边界鲁棒性有效；而早停机制虽然对干净样本的精度影响较小，但能有效节省大量训练时间并维持模型稳定，不同攻击场景下的鲁棒性波动控制在1.3%以内。完整方法对于MIM攻击的鲁棒性较Baseline提升了41.9%，其中CGAL贡献最大，数据蒸馏和早停机制次之，这说明所提各个模块在集成框架下具备分工明确、协同增强的防御效能。

4 结束语

本文基于数据增强与蒸馏技术的集成对抗防御EDEAD方法，有效解决了复杂室内定位系统中对抗样本攻击带来的威胁问题。实验结果表明，本文方法在高维且特征丰富的UJIIndoorLoc数据集上，能够有效降低对抗样本的可转移性，在面对多类型攻击时定位精度均高于80%，并且在节省训练与测试开销的同时，相比传统集成防御方法展现出更优的鲁棒性和效率平衡。

本文方法采用离线训练与在线推理相结合的双阶段架构，兼容Wi-Fi、蓝牙等异构信号。在线阶段仅需单次前向传播即可处理真实环境扰动，已在Wi-Fi场景验证了有效性。实时性优势主要源于架构的简洁性，后续将基于Android APP开展多信源的实测验证。本文EDEAD方法在不同数据集上的性能存在差异，例如在采样点较少的Mall_Wi-Fi数据集上防御效果略低于UJIIndoorLoc，这可能与数据密度和信号稳定性有关，也表明方法的跨场景泛化仍有提升空间。未来将结合多样化数据源，优化

模型轻量化和快速部署能力，并通过 Android 采集 APP 开展更多真实环境测试，以进一步保障室内定位系统的安全与可靠性。

参考文献：

[1] 张学军, 何福存, 盖继扬, 等. 边缘计算下指纹室内定位差分私有联邦学习模型[J]. 计算机研究与发展, 2022, 59(12): 2667-2688.
ZHANG X J, HE F C, GAI J Y, et al. A differentially private federated learning model for fingerprinting indoor localization in edge computing[J]. Journal of Computer Research and Development, 2022, 59(12): 2667-2688.

[2] ZHU X Q, QU W Y, QIU T, et al. Indoor intelligent fingerprint-based localization: principles, approaches and challenges[J]. IEEE Communications Surveys & Tutorials, 2020, 22(4): 2634-2657.

[3] HSIEH C H, CHEN J Y, NIEN B H. Deep learning-based indoor localization using received signal strength and channel state information[J]. IEEE Access, 2019, 7: 33256-33267.

[4] BAHL P, PADMANABHAN V N. RADAR: an in-building RF-based user location and tracking system[C]//Proceedings of the IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2000: 775-784.

[5] MOSLEH M F, ABD-ALHAMEED R A, QASIM O A. Indoor positioning using adaptive KNN algorithm based fingerprint technique[J]. Social Informatics and Telecommunications Engineering, 2018, 263: 13-21.

[6] LIU J H, ZENG B S, LI S N, et al. MLA-MFL: a smartphone indoor localization method for fusing multisource sensors under multiple scene conditions[J]. IEEE Sensors Journal, 2024, 24(16): 26320-26333.

[7] KARABEY AKSAKALLI I, BAYINDIR L. Enhancing indoor localization with room-to-room transition time: a multi-dataset study[J]. Applied Sciences, 2025, 15(4): 1985.

[8] 张学军, 鲍俊达, 何福存, 等. 抵御对抗样本攻击的指纹室内定位方法[J]. 北京航空航天大学学报, 2022, 48(11): 2087-2101.
ZHANG X J, BAO J D, HE F C, et al. A fingerprint indoor localization method against adversarial sample attacks[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(11): 2087-2101.

[9] XUE J Q, ZHANG J, GAO Z Y, et al. Enhanced Wi-Fi CSI fingerprints for device-free localization with deep learning representations[J]. IEEE Sensors Journal, 2023, 23(3): 2750-2759.

[10] PATIL M, WANG X Y, WANG X Y, et al. Adversarial attacks on deep learning-based floor classification and indoor localization[C]//Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2021: 7-12.

[11] WANG Z Z, SHU X, WANG Y, et al. A feature space-restricted attention attack on medical deep learning systems[J]. IEEE Transactions on Cybernetics, 2023, 53(8): 5323-5335.

[12] 张剑, 周侠, 张一然, 等. 基于雅可比显著图的电磁信号快速对抗攻击方法[J]. 通信学报, 2024, 45(1): 180-193.
ZHANG J, ZHOU X, ZHANG Y R, et al. Electromagnetic signal fast adversarial attack method based on Jacobian saliency map[J]. Journal on Communications, 2024, 45(1): 180-193.

[13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[14] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv Preprint, arXiv: 1706.06083, 2017.

[16] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2016: 582-597.

[17] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.

[18] WU D X, WANG Y S, XIA S T, et al. Skip connections matter: on the transferability of adversarial examples generated with ResNets[J]. arXiv Preprint, arXiv: 2002.05990, 2020.

[19] YAN Q L, XIONG W, WANG H M. Secure indoor localization against adversarial attacks using DCGAN[J]. IEEE Communications Letters, 2025, 29(1): 130-134.

[20] YAN Q L, XIONG W, WANG H M. TransGAN-based secure indoor localization against adversarial attacks[J]. IEEE Internet of Things Journal, 2025, 12(5): 5918-5930.

[21] PANG T Y, XU K, DU C, et al. Improving adversarial robustness via promoting ensemble diversity[C]//Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2019: 4970-4979.

[22] KARIYAPPA S, MOINUDDIN K Q. Improving adversarial robustness of ensembles with diversity training[J]. arXiv Preprint, arXiv: 1901.09981, 2019.

[23] YANG H R, ZHANG J Y, DONG H L, et al. DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020). Massachusetts: MIT Press, 2020: 5505-5515.

[24] YANG Z L, LI L Y, XU X J, et al. TRS: transferability reduced ensemble via encouraging gradient diversity and model smoothness[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021). Massachusetts: MIT Press, 2021:17642-17655.

[25] HUANG L F, HUANG Q, QIU P C, et al. FASTEN: fast ensemble learning for improved adversarial robustness[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2565-2580.

[26] 张学军, 席阿友, 加小红, 等. 基于深度学习的指纹室内定位对抗样本攻击研究[J]. 计算机工程, 2024, 50(10): 228-239.
ZHANG X J, XI A Y, JIA X H, et al. Study on adversarial sample attacks on deep learning based fingerprinting indoor localization[J]. Computer Engineering, 2024, 50(10): 228-239.

[27] ZHANG J F, XU X L, HAN B, et al. Attacks which do not kill training make adversarial learning stronger[C]//Proceedings of the International Conference on Machine Learning(ICML). New York: ACM Press, 2020: 11278-11287.

[28] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[C]//Proceedings of the 6th Interna-

tional Conference on Learning Representations (ICLR 2018). Vancouver: ICLR, 2018: 1894-1913.

[29] MA X J, LI B, WANG Y S, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[C]//Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver: ICLR, 2018: 402-416.

[30] TORRES-SOSPEDRA J, MONTOLIU R, MARTÍNEZ-USÓ A, et al. UJIIndoorLoc: a new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems[C]//Proceedings of the 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN). Piscataway: IEEE Press, 2014: 261-270.

[31] JIANG X L, CHEN Y Q, LIU J F, et al. FSELM: fusion semi-supervised extreme learning machine for indoor localization with Wi-Fi and Bluetooth fingerprints[J]. Soft Computing, 2018, 22(11): 3621-3635.

[作者简介]

张学军（1977– ），男，宁夏中宁人，博士，兰州交通大学教授、博士生导师，主要研究方向为网络安全、数据隐私与机器学习等。

李梅（2000– ），女，甘肃武威人，兰州交通大学硕士生，主要研究方向为室内定位、对抗样本攻击及防御等。

陈惠（2002– ），女，甘肃天水人，兰州交通大学硕士生，主要研究方向为室内定位、对抗攻防、隐私保护等。

王国华（1980– ），男，内蒙古呼和浩特人，博士，兰州交通大学副教授、硕士生导师，主要研究方向为网络入侵检测与防御、信息物理系统安全理论等。