面向自动调制分类的域自适应对抗防御方法

杨研蝶，林云，徐路平，张思成，李奎贤，韩宇
（哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001）

摘　要：基于深度学习的自动调制分类模型易受对抗样本攻击，在信道环境动态变化、信号标签难以获取的实际场景中面临更严峻的对抗安全威胁。针对这一问题，提出一种基于多域分布对齐的域自适应对抗防御方法。首先，通过相位旋转数据增强策略，丰富模型可学习的判别性特征与域不变特征。其次，构建双判别器结构，减少目标域原始信号和对抗信号与源域之间的特征分布差异。然后，结合高置信度伪标签引入对比学习约束，利用源域类别锚点增强目标域的类内紧凑性和类间分离性。最后，采用一致性约束减少目标域原始信号与对抗信号的输出差异。在公开和仿真数据集上的实验结果表明，与现有方法相比，所提方法在多种对抗攻击下均展现出优异的域适应性与对抗鲁棒性，可有效提升复杂电磁环境中自动调制分类系统的可靠性与安全性。

关键词：频谱监测；自动调制分类；无监督域自适应；对抗鲁棒性
中图分类号：TN971
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026030

Domain adaptive adversarial defense method for automatic modulation classification

Yang Yandie, Lin Yun, Xu Luping, Zhang Sicheng, Li Kuixian, Han Yu
College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China

Abstract: Deep learning-based automatic modulation classification (AMC) models are vulnerable to adversarial example attacks, posing severe adversarial security threats in practical scenarios characterized by dynamic channel conditions and limited availability of signal labels. To address this issue, a multi-domain distribution alignment based on domain-adaptive adversarial defense method was proposed. Firstly, a phase rotation data augmentation strategy was used to enrich the discriminative and domain-invariant features learned by the model. Secondly, a dual-discriminator architecture was constructed to reduce the feature distribution discrepancy between the original and adversarial signals in the target domain and those in the source domain. Thirdly, contrastive learning constraints were introduced in conjunction with high-confidence pseudo-labels, leveraging source domain class anchor to enhance intra-class compactness and inter-class separability in the target domain. Finally, a consistency constraint was employed to reduce the output discrepancy between original and adversarial signals in the target domain. Experimental results on both public and simulated datasets demonstrate that, compared with existing methods, the proposed method exhibits superior domain adaptability and adversarial robustness under various adversarial attacks, effectively enhancing the reliability and security of AMC systems in complex electromagnetic environments.

Keywords: spectrum monitoring, automatic modulation classification, unsupervised domain adaptation, adversarial robustness

收稿日期：2025–11–30；**修回日期：**2026–01–28
通信作者：张思成，2015080325@hrbeu.edu.cn
基金项目：中央高校基本科研业务费专项资金资助项目（No.3072025YY0801）；黑龙江省博士后基金资助项目（No.3236340036）；国家自然科学基金资助项目（No.62201172）
Foundation Items: The Fundamental Research Funds for the Central Universities (No.3072025YY0801), Heilongjiang Postdoctoral Fund (No.3236340036), The National Natural Science Foundation of China (No.62201172)

0 引言

随着无线通信技术的快速发展，用频设备的广泛部署以及通信用户数量的持续增长，使电磁空间日益拥挤，导致频谱资源紧缺[1]。频谱监测与管理系统能够实现对开放电磁环境中用频活动的实时感知与识别，在优化频谱资源分配、缓解信号干扰和维护频谱秩序方面发挥重要作用。自动调制分类（automatic modulation classification, AMC）作为频谱监测系统的关键技术之一，能够识别特定频段信号的调制类型[2-3]。由于无线通信系统通常采用特定的调制方式进行信号传输，通过调制类型识别可为信号来源与分析提供依据，对监测频段内无线设备用频行为的合法性以及识别潜在的异常或干扰信号具有重要意义。近年来，深度学习（deep learning, DL）[4]凭借其在特征自动提取与复杂模式表征方面的显著优势，被广泛应用于AMC任务，有效提升了复杂电磁环境下调制信号识别的准确性和可靠性[5]。

尽管DL技术显著提升了AMC的识别性能，但Sadeghi等[6]发现，在原始信号中添加精心设计且难以察觉的微小扰动，即可误导AMC模型输出错误的分类结果，并验证了开放电磁空间背景下对抗攻击的可行性。随后，针对AMC模型的各种对抗攻击方法相继被提出[7-9]，对电磁频谱安全构成了严重威胁[10]。攻击者可利用微小扰动影响频谱监测模型的识别结果，导致将合法信号误判为非法信号而引发误报，或将非法信号伪装成合法信号以规避监测并恶意占用频谱资源。Lin等[9]研究了快速梯度符号法（fast gradient sign method, FGSM）[11]、投影梯度下降（projected gradient descent, PGD）法[12]和基本迭代法（basic iterative method, BIM）[13]等基于梯度的对抗攻击方法对卷积神经网络（convolutional neural network, CNN）调制识别模型的攻击效果，揭示了AMC模型对对抗攻击的脆弱性，并指出迭代攻击通常优于单步攻击。Kokalj-Filipovic等[14-15]采用Carlini-Wagner（CW）优化算法开展目标对抗攻击，通过将无线信号伪装成特定调制类型，展示了对抗扰动对信号识别准确性的影响。为应对上述对抗安全威胁，研究者提出了特征去噪、对抗训练、随机平滑等多种对抗防御方法[16-18]。对抗训练（adversarial training, AT）作为应用最广泛的防御方法，通过在训练过程中引入对抗样本增强模型鲁棒性，可实现对多种对抗攻击的有效防御。Madry等[12]将对抗训练建模为最小–最大优化问题，通过使用PGD攻击迭代生成对抗样本进行训练，有效改善模型的对抗鲁棒性。McClintick等[19]首次在真实信号传输环境中开展信号分类的对抗训练研究，验证了该方法在实际环境中的可行性及其在增强无线系统安全方面的潜力。

然而，现有的对抗防御方法普遍依赖训练集与测试集满足独立同分布假设，并要求模型在充分标注的数据上进行监督训练。当测试数据分布偏离训练数据且缺乏标签时，传统全监督的AMC对抗防御方法的性能将显著下降。在实际电磁环境中，信道受多径衰落、阴影效应等因素影响呈高度动态性，导致不同信道条件下的接收信号分布存在明显差异。此外，调制信号虽易于采集，但其标注成本较高，获取大规模标签数据仍十分困难。因此，基于DL的AMC模型在实际部署中面临更严峻的对抗安全挑战。为应对分布偏移与标签缺失导致的性能退化，无监督域适应（unsupervised domain adaptation, UDA）被广泛研究[20-21]。通常将具备标注的数据域定义为源域，将待识别且缺乏标签的数据域定义为目标域。UDA方法通过最小化源域与目标域之间的特征分布差异，实现边缘分布对齐，从而促进模型知识从源域向目标域的有效迁移[22]。早期的域对齐方法通常通过特定的距离度量缩小域间分布差异。Long等[23]提出深度自适应网络（deep adaptation network, DAN）结构，通过将任务特定层的隐藏表示嵌入再生核希尔伯特空间，显式匹配不同领域的均值嵌入学习可迁移特征。受对抗学习启发，基于域判别器的方法通过对抗训练使特征提取器学习域不变特征。Ganin等[24]提出了域对抗神经网络（domain-adversarial neural network, DANN）方法，该方法作为典型的端到端框架，利用梯度反转层（gradient reversal layer, GRL）实现域判别器与特征提取器的对抗训练，使特征提取器学习域不变特征，已成为当前域适应研究的基准方法。

传统UDA方法主要关注提升目标域的分类性能，往往忽视模型在目标域的对抗鲁棒性。为弥补这一不足，部分研究开始探索鲁棒UDA方法，旨在域迁移过程中增强模型的对抗鲁棒性。现有

鲁棒 UDA 方法主要包括鲁棒特征蒸馏[25]与对抗训练[26-28]两类。鲁棒特征蒸馏方法利用外部预训练的鲁棒模型在 UDA 过程中提取鲁棒特征，以提升目标域模型的鲁棒性。Awais 等[25]通过引入鲁棒特征适应模块，将预训练鲁棒模型的中间特征与 UDA 模型的特征进行对齐，以提升模型在目标域的对抗鲁棒性。然而，该方法的蒸馏效果依赖预训练模型的架构与性能，且训练鲁棒教师模型通常需要较高的计算成本，限制了该方法在实际场景中的适用性。对抗训练方法先在源域和目标域上进行域对齐训练，再利用预训练模型生成伪标签，在目标域上开展对抗训练以获得鲁棒的目标域模型。Lo 等[26]提出了无监督域适应的对抗鲁棒训练（adversarially robust training for unsupervised domain adaptation, ARTUDA）方法，通过自监督对抗训练，利用 KL 散度在目标域数据上生成对抗样本并进行训练。但该方法生成的对抗样本不能保证在对抗训练中的内部最大化，可能导致鲁棒性优化不足。Zhu 等[27]提出了基于元自训练的鲁棒无监督域适应（meta self-training for robust unsupervised domain adaptation, SRoUDA）方法，该方法采用随机掩码策略对目标域无标注数据进行增强，并基于源域预训练模型生成目标域伪标签，进而以伪标签为监督信号开展目标域对抗训练，同时引入元学习机制迭代优化源域模型，从而提升伪标签质量。Wang 等[28]提出了散度感知对抗训练（divergence-aware adversarial training, DART）框架，通过引入面向目标域对抗损失的泛化界限，构建统一的跨域防御框架。基于对抗训练的方法普遍依赖伪标签进行监督，容易受到伪标签的噪声干扰，难以充分发挥对抗训练的潜力，从而限制其鲁棒性提升效果。

针对信道环境变化下 AMC 模型易受对抗样本攻击的问题，本文提出一种基于多域分布对齐的域自适应对抗防御（multi-domain distribution alignment based domain-adaptive adversarial defense, MDDA）方法。首先，通过相位旋转的数据增强策略扩展信号的分布覆盖范围，提高模型的跨域泛化能力。其次，通过双判别器结构将目标域原始特征与对抗特征分别对齐至源域特征空间，实现多域特征对齐。在此基础上，引入对比学习与输出一致性约束，进一步优化目标域决策边界并增强无标签条件下的对抗鲁棒性。实验结果表明，MDDA 方法在多种迁移场景和对抗攻击下均能有效提升目标域的调制分类性能，并进一步验证无监督域适应模型对对抗攻击的脆弱性以及开展鲁棒域自适应研究的必要性。本文的主要贡献如下。

1) 引入基于相位旋转的数据增强策略。通过在训练阶段对源域与目标域信号施加随机相位旋转，扩展训练样本分布范围，丰富特征表示的多样性，使模型学习到更加稳健的判别性特征。

2) 设计多域分布对齐的域自适应机制。构建双判别器的域自适应模型，分别对齐目标域原始信号与对抗信号的特征分布，有效减小跨域特征分布差异。同时，结合高置信度伪标签引入对比学习约束，将源域类别特征作为锚点，增强特征空间的类内紧凑性与类间分离性，优化目标域的决策边界。此外，通过对目标域原始信号和对抗信号的输出施加一致性约束，进一步提高无标签条件下的对抗鲁棒性。

3) 在公开数据集及多种典型信道数据集上进行实验，并在不同攻击场景、攻击类型和扰动幅度下验证所提方法的有效性。实验结果表明，所提方法在不同信道环境下均能够显著提升模型对目标域的分类准确性和对抗鲁棒性。

1　问题表述

在开放动态的电磁空间中，存在对抗攻击威胁的频谱监测系统，如图 1 所示。在该系统中，合法设备在目标频段内采用多种调制方式进行无线通信，频谱监测设备对环境中的无线信号进行接收，并将其输入基于 DL 的 AMC 模型，以识别信号的调制类型，从而实现对用频行为的感知与监管。在开放电磁环境中，攻击者可通过构造精心设计的对抗扰动并将其叠加到发射信号上，误导频谱监测系统中的 AMC 模型，使其将合法信号误判为非法信号，或将非法信号伪装成合法信号，从而达到规避监管或恶意占用频谱资源的目的。此外，受信道条件变化和环境时变性的影响，频谱监测系统在不同场景下接收到的信号分布往往存在显著差异，从而引入典型的域偏移问题。本文正是在上述系统场景下，研究在对抗攻击和信道变化共同存在的条件下如何提升 AMC 模型在目标域环境中的鲁棒性与泛化能力。

图1　频谱监测系统

为了形式化描述调制信号分类任务，本文首先对相关符号进行定义。令 $X \subseteq R^{\mathrm{D}}$ 为调制信号的数据空间，Y 为对应的调制类型的标签空间 $P_{XY}(x,y)$，为 $X \times Y$ 的联合概率分布。将由大量带标签调制信号样本构成的分布定义为源域，表示为 $P^{\mathrm{s}}_{XY}(x,y)$；将仅包含未标注调制信号样本的分布定义为目标域，表示为 $P^{\mathrm{t}}_{XY}(x,y)$。源域和目标域均由相同调制类型的无线通信信号组成，在语义标签空间中具有相关性，但由于无线信道环境不同，源域与目标域的分布不同，即 $P^{\mathrm{s}}_{XY}(x,y) \neq P^{\mathrm{t}}_{XY}(x,y)$。

基于上述定义，带标签的源域数据集表示为 $\mathcal{S}=\{(x^{\mathrm{s}}_i,y^{\mathrm{s}}_i)\}^{N_{\mathrm{s}}}_{i=1}\sim P^{\mathrm{s}}_{XY}(x,y)$，其中，$N_{\mathrm{s}}$ 为源域带标签信号样本的数量；无标签的目标域数据集表示为 $\mathcal{T}=\{(x^{\mathrm{t}}_i)\}^{N_{\mathrm{t}}}_{i=1}\sim P^{\mathrm{t}}_{XY}(x,y)$，其中，$N_{\mathrm{t}}$ 为目标域无标签信号样本的数量。在目标域中，对抗信号 $x^{\mathrm{t}}_{\mathrm{a}}$ 由对抗攻击者在原始信号 x^{t} 上添加微小扰动 δ 生成，相应的目标域对抗数据集表示为 $\mathcal{T}_{\mathrm{a}}=\{(x^{\mathrm{t}}_{\mathrm{a},i})\}^{N_{\mathrm{t}}}_{i=1}$。鲁棒的无监督域适应调制分类任务的目标是学习一个分类器 f，使其在缺乏目标域标签的条件下，能够对目标域中的原始信号 x^{t} 及其对应的对抗信号 $x^{\mathrm{t}}_{\mathrm{a}}$ 实现正确的调制类型识别。因此，本文研究的鲁棒无监督域适应问题可形式化表示为

$$\max_{f} \mathbb{E}_{(x)\in\{\mathcal{S},\mathcal{T},\mathcal{T}_{\mathrm{a}}\}} f(x)=y \tag{1}$$

2　基于多域分布对齐的域自适应对抗防御方法

本节首先概述 MDDA 方法框架，随后分别介绍源域 AMC 模型预训练、基于相位旋转的数据增强以及多域分布对齐的域自适应机制，最后给出推理阶段的输出。

2.1　MDDA 方法框架

本文提出的 MDDA 方法框架如图 2 所示，旨在提升 AMC 模型在信道环境变化条件下的对抗鲁棒性。整体而言，MDDA 由源域 AMC 模型预训练、基于相位旋转的数据增强以及多域分布对齐的域自适应 3 个部分组成，其训练过程包括预训练阶段与域自适应阶段。在预训练阶段，利用源域信号样本和标签训练源域 AMC 模型。该模型由源域特征提取器 E_{s} 和源域分类器 C_{s} 组成，其中，E_{s} 用于学习具有类别可分性的源域特征，C_{s} 基于所提取的特征实现对源域调制信号的准确分类。在域自适应阶段，首先对源域和目标域信号进行随机相位旋转变换，以扩展训练样本的分布范围，为跨域特征对齐提供更丰富的特征表示，并增强模型对相位变化的鲁棒性。多域分布对齐的域自适应机制包括特征提取器 E、分类器 C、原始信号域判别器 D_{c} 以及对抗信号域判别器 D_{a}，其中，E 用于同时学习源域与目标域的判别性特征与域不变特征，C 对源域和目标域信号进行调制分类，D_{c} 专注于原始信号的域对齐，D_{a} 用于对抗信号的域对齐，将原始信号和对抗信号的对齐过程解耦，能够避免梯度干扰，更加有效地减少域间特征差异。

需要说明的是，预训练阶段与域自适应阶段中的特征提取器和分类器采用相同的网络结构。多域分布对齐的域适应机制基于 DANN[24]构建，在两个域判别器前均引入 GRL，并采用相同的 GRL 参数以保证原始信号与对抗信号的域对齐过程能够协同优化。在反向传播过程中，域判别器的梯度经 GRL 反转后传递至特征提取器，从而促使模型学习生成对域不可分的原始信号特征和对抗信号特征，实现双判别器驱动的鲁棒域自适应对齐。

2.2　源域 AMC 模型预训练

通过在带标签的源域信号样本上对源域 AMC 模型进行监督训练，E_{s} 学习类内聚合、类间分离的判别性源域特征，为后续跨域特征对齐提供类别锚点。将源域信号 x^{s} 依次输入 E_{s} 和 C_{s}，可以获得

$$f^{\mathrm{s}}=E_{\mathrm{s}}(x^{\mathrm{s}}),p^{\mathrm{s}}=C_{\mathrm{s}}(f^{\mathrm{s}}) \tag{2}$$

其中，f^{s} 为源域信号的深度特征表示，p^{s} 为对应的类别预测概率。随后，使用交叉熵损失函数 $\mathrm{CE}(\cdot)$ 计算 x^{s} 的分类损失。

$$\ell^{\mathrm{s}}_{\mathrm{cls}}=\mathrm{CE}(p^{\mathrm{s}},y^{\mathrm{s}}) \tag{3}$$

其中，y^{s} 为 x^{s} 的调制类别标签。

为了进一步增强特征空间的判别性，本文在源域 AMC 预训练阶段引入对比损失，通过拉近同类

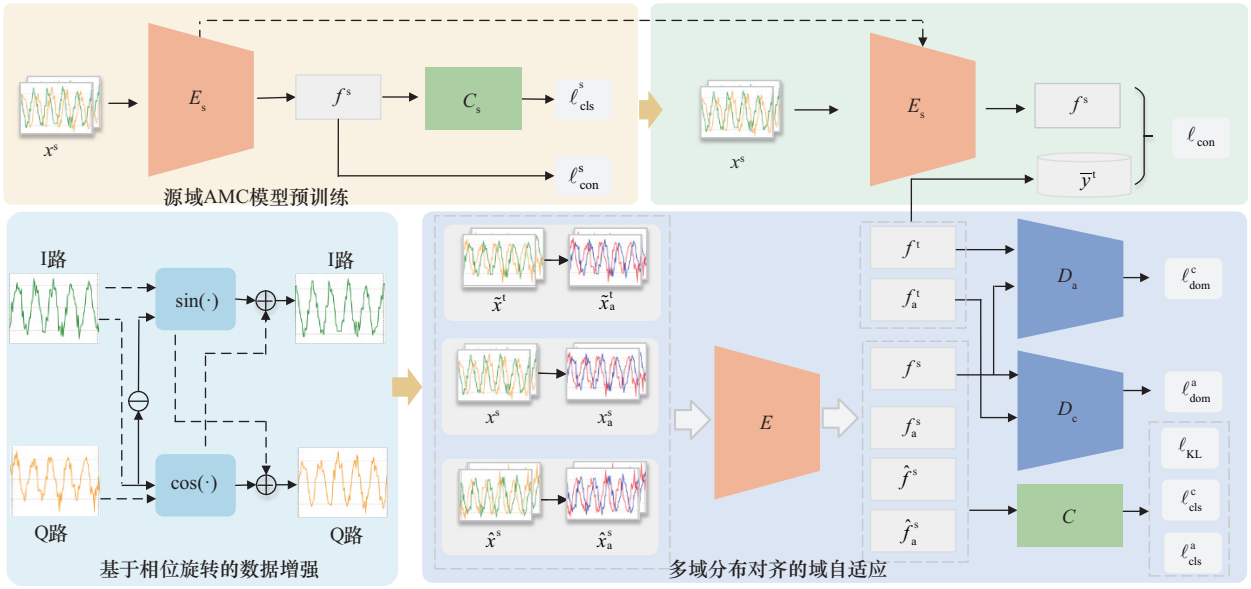

图2　MDDA 方法框架

样本特征、推远异类样本特征，优化源域决策边界。对于输入的源域信号 x_i^{s}、x_j^{s} 以及对应的标签 y_i^{s}、y_j^{s}，对应的对比损失表示为

$$\ell_{\mathrm{con}}^{\mathrm{s}} = \mathop{\mathbb{E}}_{1\leqslant i,j\leqslant N_{\mathrm{s}}} [\mathbb{I}_{[y_i^{\mathrm{s}}=y_j^{\mathrm{s}}]}\cdot \mathcal{D}_{ij}^{\mathrm{s}\,2} + \mathbb{I}_{[y_i^{\mathrm{s}}\neq y_j^{\mathrm{s}}]}\cdot \max(0, m-\mathcal{D}_{ij}^{\mathrm{s}})^2] \tag{4}$$

其中，$\mathbb{I}_{[\cdot]}$为指示函数，当[·]内条件成立时取值为1，否则为0；$\mathcal{D}_{ij}^{\mathrm{s}} = \left\| f_i^{\mathrm{s}} - f_j^{\mathrm{s}} \right\|_2$为两个特征之间的欧氏距离；$m>0$为预设的阈值，用于约束不同类别样本间的特征距离。因此，源域AMC模型的优化目标可以表示为

$$\min_{E_{\mathrm{s}},C_{\mathrm{s}}} \ell_{\mathrm{cls}}^{\mathrm{s}} + \lambda_{\mathrm{s}}\ell_{\mathrm{con}}^{\mathrm{s}} \tag{5}$$

其中，λ_{s}为权重系数，用于平衡分类损失和对比损失。源域AMC模型预训练完成后，保留E_{s}提取后续对比学习的类别锚点，同时将预训练参数用于E和C的初始化。

2.3　基于相位旋转的数据增强

为了增强模型在不同信道条件下的泛化能力，并为域自适应阶段提供更丰富的特征分布，本文引入基于相位旋转的数据增强策略。相位旋转能够模拟无线信道中常见的相位偏移效应，从而扩展训练样本所覆盖的信号分布范围，并引导模型学习对相位变化具有不变性的判别特征。

对于复数调制信号，通常表示为$x = \boldsymbol{I} + \mathrm{j}\boldsymbol{Q}$，其中，$\boldsymbol{I}$和$\boldsymbol{Q}$分别表示信号的同相（in-phase, I）分量和正交（quadrature, Q）分量。相位旋转相当于将复数信号乘以单位复数$\mathrm{e}^{\mathrm{j}\theta}$，其中，$\theta$为旋转角度，即在复平面上逆时针旋转原始信号的角度。根据欧拉公式$\mathrm{e}^{\mathrm{j}\theta} = \cos(\theta) + \mathrm{j}\sin(\theta)$，旋转后的$\boldsymbol{I}$和$\boldsymbol{Q}$分量分别表示为

$$\hat{\boldsymbol{I}} = \cos(\theta)\cdot\boldsymbol{I} - \sin(\theta)\cdot\boldsymbol{Q} \tag{6}$$

$$\hat{\boldsymbol{Q}} = \sin(\theta)\cdot\boldsymbol{I} + \cos(\theta)\cdot\boldsymbol{Q} \tag{7}$$

在训练阶段，本文使用随机相位旋转的数据增强方法，旋转角度θ从$(0, 2\pi)$上的均匀分布中采样，以模拟无线信道中可能出现的任意相位偏移。通过引入随机相位扰动，模型能够在训练过程中学习到更为多样的信号分布，引导模型学习更加稳健和泛化的决策边界，以准确分类目标域的调制信号。源域和目标域的原始信号分别表示为$x^{\mathrm{s}} = (\boldsymbol{I}^{\mathrm{s}},\boldsymbol{Q}^{\mathrm{s}})$和$x^{\mathrm{t}} = (\boldsymbol{I}^{\mathrm{t}},\boldsymbol{Q}^{\mathrm{t}})$，经相位旋转增强后的信号分别表示为$\hat{x}^{\mathrm{s}} = (\hat{\boldsymbol{I}}^{\mathrm{s}},\hat{\boldsymbol{Q}}^{\mathrm{s}})$和$\hat{x}^{\mathrm{t}} = (\hat{\boldsymbol{I}}^{\mathrm{t}},\hat{\boldsymbol{Q}}^{\mathrm{t}})$，其中，$x^{\mathrm{s}}$和$\hat{x}^{\mathrm{s}}$具有相同的类别标签$y^{\mathrm{s}}$。$x^{\mathrm{s}}$和$\hat{x}^{\mathrm{s}}$的特征空间存在细微差异，对源域信号进行相位旋转增强，可以在标签监督下扩展训练样本分布，促使模型学习到对相位变化具有鲁棒性的分类特征。对目标域信号进行相位旋转增强，则可丰富目标域的特征表示，有助于减小源域和目标域在特征空间中的分布差异。

2.4　多域分布对齐的域自适应机制

多域分布对齐的域自适应机制旨在学习具备判别性和域不变性的鲁棒特征表示，通过减小源域与目标域在特征空间中的分布差异，实现对目标域原始信号及其对抗信号的准确分类。该阶段的训练由

多对抗域自适应、对比学习和一致性约束 3 个部分组成。下面分别给出其原理和训练过程。

1) 多对抗域自适应

本节将目标域原始信号及其相位增强信号统一记为目标域信号 $\tilde{x}^{\mathrm{t}}=\{x^{\mathrm{t}},\hat{x}^{\mathrm{t}}\}$。需要注意的是，源域增强信号仅用于分类监督以增强源域判别性，在域对齐过程中，源域分布由源域原始信号特征表示。

首先，将 x^{s}、$\hat{x}^{\mathrm{s}}$ 以及 $\tilde{x}^{\mathrm{t}}$ 同时输入 E，得到对应的特征表示。

$$f^{\mathrm{s}}=E\left(x^{\mathrm{s}}\right),\hat{f}^{\mathrm{s}}=E\left(\hat{x}^{\mathrm{s}}\right),\tilde{f}^{\mathrm{t}}=E\left(\tilde{x}^{\mathrm{t}}\right) \tag{8}$$

其中，f^{s}、$\hat{f}^{\mathrm{s}}$ 和 $\tilde{f}^{\mathrm{t}}$ 分别为源域原始信号、源域增强信号和目标域信号的深度特征表示。将上述特征送入 C。

$$p^{\mathrm{s}}=C\left(f^{\mathrm{s}}\right),\hat{p}^{\mathrm{s}}=C\left(\hat{f}^{\mathrm{s}}\right),\tilde{p}^{\mathrm{t}}=C\left(\tilde{f}^{\mathrm{t}}\right) \tag{9}$$

其中，p^{s}、$\hat{p}^{\mathrm{s}}$ 和 $\tilde{p}^{\mathrm{t}}$ 分别为源域原始信号、源域增强信号和目标域信号的类别概率向量。因此，原始信号的分类损失可以表示为

$$\ell_{\mathrm{cls}}^{\mathrm{c}}=\mathrm{CE}\left(p^{\mathrm{s}},y^{\mathrm{s}}\right)+\mathrm{CE}\left(\hat{p}^{\mathrm{s}},y^{\mathrm{s}}\right) \tag{10}$$

为实现原始信号的领域对齐，仅将 f^{s} 和 $\tilde{f}^{\mathrm{t}}$ 输入 D_{c}。

$$p_{\mathrm{dc}}^{\mathrm{s}}=D_{\mathrm{c}}\left(f^{\mathrm{s}}\right),\tilde{p}_{\mathrm{dc}}^{\mathrm{t}}=D_{\mathrm{c}}\left(\tilde{f}^{\mathrm{t}}\right) \tag{11}$$

其中，$p_{\mathrm{dc}}^{\mathrm{s}}$ 和 $\tilde{p}_{\mathrm{dc}}^{\mathrm{t}}$ 分别是源域信号和目标域信号的原始域判别概率。基于此，原始信号的域判别损失表示为

$$\ell_{\mathrm{dom}}^{\mathrm{c}}=\frac{1}{N_{\mathrm{s}}}\sum_{i=1}^{N_{\mathrm{s}}}\ln\left(p_{\mathrm{dc},i}^{\mathrm{s}}\right)+\frac{1}{2N_{\mathrm{t}}}\sum_{i=1}^{2N_{\mathrm{t}}}\ln\left(1-\tilde{p}_{\mathrm{dc},i}^{\mathrm{t}}\right) \tag{12}$$

为了提升模型的鲁棒性并实现鲁棒性的跨域迁移，本文对源域和目标域分别生成对抗样本。源域样本具有真实标签，采用传统的 PGD 方法生成对抗样本；目标域缺少标签，采用自监督方式[29]来生成对抗样本。具体而言，源域对抗信号 $x_{\mathrm{a}}^{\mathrm{s}}$ 由 PGD 迭代生成。

$$x_{\mathrm{a},0}^{\mathrm{s}}=x^{\mathrm{s}}+\xi \tag{13}$$

$$x_{\mathrm{a},k+1}^{\mathrm{s}}=\mathrm{Clip}_{\eta,x^{\mathrm{s}}}\left(x_{\mathrm{a},k}^{\mathrm{s}}+\alpha\cdot\mathrm{Sign}\left(\nabla_{x_{\mathrm{a},k}^{\mathrm{s}}}\mathrm{CE}\left(C\left(E\left(x_{\mathrm{a},k}^{\mathrm{s}}\right)\right),y^{\mathrm{s}}\right)\right)\right) \tag{14}$$

其中，ξ 是初始化的随机噪声，k 为迭代次数，α 为扰动步长，$\mathrm{Clip}_{\eta,x^{\mathrm{s}}}(\cdot)$ 为投影函数，用于约束扰动幅度满足 $\left\|x_{\mathrm{a},k+1}^{\mathrm{s}}-x^{\mathrm{s}}\right\|_p\leqslant\eta$，$\|\cdot\|_p$ 为 l_p 范数，本文取 $p=\infty$，η 表示对抗扰动大小。经过适当次数的迭代后，即可获得 $x_{\mathrm{a}}^{\mathrm{s}}$。源域增强信号的对抗信号 $\hat{x}_{\mathrm{a}}^{\mathrm{s}}$ 采用相同的生成过程，在这里不作赘述。

目标域对抗信号 $\tilde{x}_{\mathrm{a}}^{\mathrm{t}}$ 使用分类器对目标域信号的输出作为监督，利用 KL 散度 $\mathrm{KL}(\cdot)$ 迭代生成。

$$\tilde{x}_{\mathrm{a},0}^{\mathrm{t}}=\tilde{x}^{\mathrm{t}}+\xi \tag{15}$$

$$\tilde{x}_{\mathrm{a},k+1}^{\mathrm{t}}=\mathrm{Clip}_{\eta,x^{\mathrm{t}}}\left(\tilde{x}_{\mathrm{a},k}^{\mathrm{t}}+\alpha\cdot\mathrm{Sign}\left(\nabla_{\tilde{x}_{\mathrm{a},k}^{\mathrm{t}}}\mathrm{KL}\left(C\left(E\left(\tilde{x}_{\mathrm{a},k}^{\mathrm{t}}\right)\right),C\left(E\left(\tilde{x}^{\mathrm{t}}\right)\right)\right)\right)\right) \tag{16}$$

随后，将 $x_{\mathrm{a}}^{\mathrm{s}}$、$\hat{x}_{\mathrm{a}}^{\mathrm{s}}$ 以及 $\tilde{x}_{\mathrm{a}}^{\mathrm{t}}$ 同时输入 E，得到对应的对抗特征。

$$f_{\mathrm{a}}^{\mathrm{s}}=E\left(x_{\mathrm{a}}^{\mathrm{s}}\right),\hat{f}_{\mathrm{a}}^{\mathrm{s}}=E\left(\hat{x}_{\mathrm{a}}^{\mathrm{s}}\right),\tilde{f}_{\mathrm{a}}^{\mathrm{t}}=E\left(\tilde{x}_{\mathrm{a}}^{\mathrm{t}}\right) \tag{17}$$

其中，$f_{\mathrm{a}}^{\mathrm{s}}$、$\hat{f}_{\mathrm{a}}^{\mathrm{s}}$ 和 $\tilde{f}_{\mathrm{a}}^{\mathrm{t}}$ 分别表示源域对抗信号、源域增强信号的对抗信号和目标域对抗信号的深度特征表示。将 $f_{\mathrm{a}}^{\mathrm{s}}$、$\hat{f}_{\mathrm{a}}^{\mathrm{s}}$、$\tilde{f}_{\mathrm{a}}^{\mathrm{t}}$ 送入 C 可以得到

$$p_{\mathrm{a}}^{\mathrm{s}}=C\left(f_{\mathrm{a}}^{\mathrm{s}}\right),\hat{p}_{\mathrm{a}}^{\mathrm{s}}=C\left(\hat{f}_{\mathrm{a}}^{\mathrm{s}}\right),\tilde{p}_{\mathrm{a}}^{\mathrm{t}}=C\left(\tilde{f}_{\mathrm{a}}^{\mathrm{t}}\right) \tag{18}$$

其中，$p_{\mathrm{a}}^{\mathrm{s}}$、$\hat{p}_{\mathrm{a}}^{\mathrm{s}}$ 和 $\tilde{p}_{\mathrm{a}}^{\mathrm{t}}$ 分别是源域对抗信号、源域增强信号的对抗信号和目标域对抗信号的类别概率，因此，源域对抗训练的分类损失可以表示为

$$\ell_{\mathrm{cls}}^{\mathrm{a}}=\mathrm{CE}\left(p_{\mathrm{a}}^{\mathrm{s}},y^{s}\right)+\mathrm{CE}\left(\hat{p}_{\mathrm{a}}^{\mathrm{s}},y^{s}\right) \tag{19}$$

现有研究表明[30]，鲁棒模型的鲁棒性具有可迁移性。基于此，本文对源域样本及其增强样本同时进行对抗训练，促使特征提取器学习具有泛化能力的鲁棒性特征表示，从而提升鲁棒性在跨域场景中的迁移效果。同样地，为对齐对抗信号的跨域特征分布，仅 f^{s} 和 $\tilde{f}_{\mathrm{a}}^{\mathrm{t}}$ 送入 D_{a}。

$$p_{\mathrm{da}}^{\mathrm{s}}=D_{\mathrm{a}}\left(f^{\mathrm{s}}\right),\tilde{p}_{\mathrm{da}}^{\mathrm{t}}=D_{\mathrm{a}}\left(\tilde{f}_{\mathrm{a}}^{\mathrm{t}}\right) \tag{20}$$

其中，$p_{\mathrm{da}}^{\mathrm{s}}$ 和 $\tilde{p}_{\mathrm{da}}^{\mathrm{t}}$ 分别是源域原始信号和目标域对抗信号的对抗域判别概率。基于此，可以计算对抗信号的域判别损失。

$$\ell_{\mathrm{dom}}^{\mathrm{a}}=\frac{1}{N_{\mathrm{s}}}\sum_{i=1}^{N_{\mathrm{s}}}\ln\left(p_{\mathrm{da},i}^{\mathrm{s}}\right)+\frac{1}{2N_{\mathrm{t}}}\sum_{i=1}^{2N_{\mathrm{t}}}\ln\left(1-\tilde{p}_{\mathrm{da},i}^{\mathrm{t}}\right) \tag{21}$$

双判别器结构将原始信号和对抗信号进行对齐解耦，避免混合对齐导致的梯度干扰，从而提升域自适应的有效性。通过对抗信号域判别器将目标域对抗信号对齐至源域特征空间，有效减小两域的特征分布差异，实现对目标域对抗样本的准确分类。

2) 对比学习

域判别器实现源域与目标域的整体分布对齐，但忽略了不同类别之间的分布对齐。为弥补这一不

足，本文引入对比学习以实现类别级的特征对齐。通过预训练的源域特征提取器 E_{s} 提取源域类别特征 f^{s} 作为锚点，并且采用分类器对 $\tilde{x}^{\mathrm{t}}$ 的预测作为伪标签 $\tilde{y}^{\mathrm{t}}=\operatorname{argmax}(\tilde{p}^{\mathrm{t}})$。然后按置信度降序排列，保留置信度较高的样本以过滤低置信度噪声。筛选后的目标域信号数据集表示为 $\bar{\mathcal{T}}^{\mathrm{t}}=\{(\bar{x}_i^{\mathrm{t}},\bar{y}^{\mathrm{t}})\}_{i=1}^{\bar{N}_{\mathrm{t}}}$，其中，$\bar{N}_{\mathrm{t}}$ 为筛选后的目标域信号样本的数量。因此目标域的对比损失可以表示为

$$\ell_{\mathrm{con}}^{\mathrm{c}}=\mathop{\mathbb{E}}_{1\leqslant i,j\leqslant \bar{N}_{\mathrm{t}}}\left[\mathbb{I}_{\left[\bar{y}_i^{\mathrm{t}}=y_j^{\mathrm{s}}\right]}\cdot \mathcal{D}_{ij}^{\mathrm{t}\,2}+\mathbb{I}_{\left[\bar{y}_i^{\mathrm{t}}\neq y_j^{\mathrm{s}}\right]}\cdot \max(0,m-\mathcal{D}_{ij}^{\mathrm{t}})^2\right] \tag{22}$$

$$\ell_{\mathrm{con}}^{\mathrm{a}}=\mathop{\mathbb{E}}_{1\leqslant i,j\leqslant \bar{N}_{\mathrm{t}}}\left[\mathbb{I}_{\left[\bar{y}_i^{\mathrm{t}}=y_j^{\mathrm{s}}\right]}\cdot \mathcal{D}_{\mathrm{a},ij}^{\mathrm{t}\,2}+\mathbb{I}_{\left[\bar{y}_i^{\mathrm{t}}\neq y_j^{\mathrm{s}}\right]}\cdot \max(0,m-\mathcal{D}_{\mathrm{a},ij}^{\mathrm{t}})^2\right] \tag{23}$$

$$\ell_{\mathrm{con}}=\ell_{\mathrm{con}}^{\mathrm{c}}+\ell_{\mathrm{con}}^{\mathrm{a}} \tag{24}$$

其中，$\ell_{\mathrm{con}}^{\mathrm{c}}$ 和 $\ell_{\mathrm{con}}^{\mathrm{a}}$ 分别表示目标域信号和目标域对抗信号的对比损失，ℓ_{con} 表示目标域的总对比损失，$\mathcal{D}_{ij}^{\mathrm{t}}=\left\| f_i^{\mathrm{t}}-f_j^{\mathrm{s}}\right\|_2$ 和 $\mathcal{D}_{\mathrm{a},ij}^{\mathrm{t}}=\left\| f_{\mathrm{a},i}^{\mathrm{t}}-f_j^{\mathrm{s}}\right\|_2$ 分别为目标域原始信号和对抗信号的特征与源域类别锚点之间的欧氏距离。

经过监督预训练的 E_{s} 能够提取具有类别判别性的深度特征，为目标域样本提供可靠的类别锚点。在此基础上，对比损失依据高置信度伪标签的指引，将目标域样本特征与同类源域样本特征拉近、与不同类别的源域样本特征推远，增加了目标域特征的类内紧凑性与类间分离性。此外，该损失函数同时约束目标域原始信号与对抗信号，有效减小原始信号与对抗信号之间的特征差异。基于源域特征的引导，可防止对抗信号将原始信号牵引至错误类别，从而优化决策边界，提升模型对目标域对抗样本的鲁棒性。

3) 一致性约束

为进一步提升目标域无标签条件下的对抗鲁棒性，本文借鉴对抗训练的核心思想，引入自监督的一致性约束，使目标域对抗信号输出与其对应的目标域信号输出保持一致。

$$\ell_{\mathrm{KL}}=\mathrm{KL}(\tilde{p}_{\mathrm{a}}^{\mathrm{t}},\tilde{p}^{\mathrm{t}}) \tag{25}$$

考虑到分类任务和域对齐任务的优化方向存在差异，本文采用交替对抗优化策略。首先分别更新域判别器 D_{c}、D_{a}，提升其域判别能力，对齐目标域和源域之间的特征分布。

$$\min_{D_{\mathrm{c}}}\ell_{\mathrm{dom}}^{\mathrm{c}},\ \min_{D_{\mathrm{a}}}\ell_{\mathrm{dom}}^{\mathrm{a}} \tag{26}$$

然后，固定域判别器的参数，更新特征提取器 E 和分类器 C，在分类损失、域对齐损失、对比损失和一致性损失的约束下，学习域不变的鲁棒判别特征。

$$\min_{E,C}\lambda_{\mathrm{cls}}\cdot(\ell_{\mathrm{cls}}^{\mathrm{c}}+\ell_{\mathrm{cls}}^{\mathrm{a}})+\lambda_{\mathrm{dom}}\cdot(\ell_{\mathrm{dom}}^{\mathrm{c}}+\ell_{\mathrm{dom}}^{\mathrm{a}})+\lambda_{\mathrm{KL}}\cdot \ell_{\mathrm{KL}}+\lambda_{\mathrm{con}}\cdot\ell_{\mathrm{con}} \tag{27}$$

其中，λ_{cls}、λ_{dom}、λ_{KL}、λ_{con} 为各损失项的权重系数，交替优化策略可以提高训练的稳定性。

综上所述，MDDA 方法的训练流程如算法 1 所示。

算法 1　MDDA 方法

输入　源域信号 x^{s} 和标签 y^{s}，目标域信号 $\tilde{x}^{\mathrm{t}}$，权重系数 λ_{s}、λ_{cls}、λ_{dom}、λ_{KL}、λ_{con}，对抗参数 α、η、K

输出　源域 AMC 模型组件 E_{s}、C_{s}，鲁棒域适应模型组件 E、C、D_{c}、D_{a}

1) 当算法未收敛时，重复执行
2)　提取源域特征 f^{s} 并获得类别概率 p^{s}
3)　通过式(3)和式(4)分别计算源域分类损失 $\ell_{\mathrm{cls}}^{\mathrm{s}}$ 和对比损失 $\ell_{\mathrm{con}}^{\mathrm{s}}$
4)　通过式(5)更新 E_{s} 和 C_{s}
5) 重复直至结束
6) 初始化 $E\leftarrow E_{\mathrm{s}}$、$C\leftarrow C_{\mathrm{s}}$
7) 当算法未收敛时，重复执行
8)　采样旋转角度 $\theta\sim U(0,2\pi)$，通过式(6)和式(7)生成增强信号 $\hat{x}^{\mathrm{s}}$ 和 $\hat{x}^{\mathrm{t}}$
9)　构建目标域信号 $\tilde{x}^{\mathrm{t}}=\{x^{\mathrm{t}},\hat{x}^{\mathrm{t}}\}$
10)　提取特征 f^{s}、$\hat{f}^{\mathrm{s}}$、$\tilde{f}^{\mathrm{t}}$，获得类别概率 p^{s}、$\hat{p}^{\mathrm{s}}$、$\tilde{p}^{\mathrm{t}}$ 和判别概率 $p_{\mathrm{dc}}^{\mathrm{s}}$、$\tilde{p}_{\mathrm{dc}}^{\mathrm{t}}$
11)　通过式(10)和式(12)分别计算原始信号的分类损失 $\ell_{\mathrm{cls}}^{\mathrm{c}}$ 和域判别损失 $\ell_{\mathrm{dom}}^{\mathrm{c}}$
12)　通过式(13)和式(14)生成源域对抗信号 $x_{\mathrm{a}}^{\mathrm{s}}$ 和 $\hat{x}_{\mathrm{a}}^{\mathrm{s}}$
13)　通过式(15)和式(16)生成目标域对抗信号 $\tilde{x}_{\mathrm{a}}^{\mathrm{t}}$
14)　提取特征 $f_{\mathrm{a}}^{\mathrm{s}}$、$\hat{f}_{\mathrm{a}}^{\mathrm{s}}$、$\tilde{f}_{\mathrm{a}}^{\mathrm{t}}$，获得类别概率 $p_{\mathrm{a}}^{\mathrm{s}}$、$\hat{p}_{\mathrm{a}}^{\mathrm{s}}$、$\tilde{p}_{\mathrm{a}}^{\mathrm{t}}$ 和判别概率 $p_{\mathrm{da}}^{\mathrm{s}}$、$\tilde{p}_{\mathrm{da}}^{\mathrm{t}}$
15)　通过式(19)和式(21)计算对抗信号的分类损失 $\ell_{\mathrm{cls}}^{\mathrm{a}}$ 和域判别损失 $\ell_{\mathrm{dom}}^{\mathrm{a}}$
16)　通过式(26)更新 D_{c} 和 D_{a}
17)　通过式(24)和式(25)分别计算对比损失

ℓ_{con}和KL损失ℓ_{KL}

18)　通过式(27)更新E和C

19) 重复直至结束

20) 返回E、C、D_{c}、D_{a}

2.5　推理阶段输出

在推理阶段，通过E和C对接收的目标域信号进行分类，实现对原始信号和对抗信号的准确分类。

$$p = C\big(E(x)\big), y = \operatorname{argmax} p \tag{28}$$

3　实验结果及分析

本节对所提方法进行实验评估。首先介绍实验采用的数据集、对比方法和实验设置。然后在各种对抗攻击场景下验证所提方法的有效性并与现有方法进行对比，进一步通过消融实验分析各模块对模型性能的影响。最后，分析模型在更低信噪比（signal-to-noise ratio, SNR）条件下的性能表现以及所提方法的计算复杂度。

3.1　实验设置

1) 数据集

本文使用公开数据集[31]和仿真数据集[20]验证所提方法的有效性。公开数据集包括RML2016.10a和RML2016.04c，这两个数据集均包含20种不同的信噪比和11种调制类型，包括二进制相移键控（binary phase shift keying, BPSK）、正交相移键控（quadrature phase shift keying, QPSK）、八进制相移键控（8-level phase shift keying, 8PSK）、连续相位频移键控（continuous phase frequency shift keying, CPFSK）、高斯频移键控（Gaussian frequency shift keying, GFSK）、四阶脉冲幅度调制（4-level pulse amplitude modulation, 4PAM）、16阶正交幅度调制（16-level quadrature amplitude modulation, 16QAM）、64阶正交幅度调制（64-level quadrature amplitude modulation, 64QAM）、调幅双边带（amplitude modulation-double sideband, AM-DSB）、调幅单边带（amplitude modulation-single sideband, AM-SSB）、宽带调频（wideband frequency modulation, WBFM）。每个样本由IQ两路信号组成，每条信号包含128个采样点。RML2016.10a每种调制类型在每个SNR下包含1 000个样本，共220 000个样本；RML2016.04c的样本数在不同调制类型下不完全一致，总样本数为162 060个。实验选取SNR为18 dB的样本，使用RML2016.04c作为源域、RML2016.10a作为目标域。

对于仿真数据集，采用的典型信道类型包括加性白高斯噪声信道（additive white Gaussian noise channel, AWGN）、莱斯衰落信道（Rician fading channel, Rician）、瑞利衰落信道（Rayleigh fading channel, Rayleigh）。本文对公开数据集生成代码[31-32]中的信道仿真模块进行修改，生成上述3种信道条件下的调制信号样本，主要仿真参数如表1所示。仿真数据集的信号类型、数量和信噪比与RML2016.10a数据集保持一致。实验中将Rician信道数据集作为源域，AWGN信道数据集和Rayleigh信道数据集分别作为目标域进行测试。每个数据集中80%的样本用于训练，20%的样本用于测试。

表1　　3种信道条件的主要仿真参数

信道类型	最大多普勒频移/Hz	时延/ms	增益	瑞利信道因子
AWGN	0	[0, 0, 0]	[1, 1, 1]	—
Rician	0.6	[0, 0.8, 1.6]	[1,0.9,0.4]	5
Rayleigh	1	[0, 0.8, 1.6]	[1,0.9,0.4]	—

2) 对比方法

为全面评估所提方法的有效性，本文选取3种基线方法进行对比。

① DANN：典型的无监督域自适应方法，未考虑对抗鲁棒性。

② AT+UDA：生成源域对抗信号，与目标域信号进行无监督域自适应训练。

③ UDA+AT：通过无监督域适应训练伪标签生成模型，并结合目标域信号和伪标签进行对抗训练。

在此基础上，本文进一步与现有鲁棒UDA方法进行对比，包括ARTUDA、SRoUDA、DART。其中，SRoUDA、DART中使用随机掩码的数据增强，为适应信号数据集，将其替换为相位旋转的数据增强方法。

3) 实验设置

实验在配备Intel Core i9-12900K CPU（3.20 GHz）和NVIDIA GeForce RTX 3080 GPU的设备上进行，对抗样本生成及模型训练过程均在单块GPU上完成。模型基于PyTorch框架实现，训练批次大小为128。优化器采用Adam，初始学习率设定为0.001，

动量参数选择0.5、0.9。GRL参数$\lambda_{p'}$设置为动态变化的，其变化公式为$\lambda_{p'}=\frac{2}{1+\exp(-\gamma\cdot p')}-1$，其中，$p'$表示迭代进程相对值，即当前迭代次数与总迭代次数的比率，γ为常数10。学习率调度使用余弦退火策略，最小学习率为1×10^{-6}。损失函数的权重参数λ_{cls}、λ_{dom}、λ_{KL}、λ_{con}分别设置为0.25、0.5、1、0.005。对抗训练阶段，扰动幅度设置为0.05，迭代步长为0.02，迭代次数为10。

3.2　基于梯度的对抗攻击测试

1) 白盒攻击

本节评估MDDA方法在白盒攻击下的防御性能。在白盒攻击中，攻击者已知目标模型的网络结构、参数权重及输入梯度，并可利用目标域真实标签生成对抗样本。实验采用FGSM、PGD和BIM这3种典型的基于梯度的攻击方法，扰动幅度分别设置为0.025、0.050、0.075和0.100，其他攻击参数与对抗训练阶段保持一致。图3展示了以RML2016.04c为源域、RML2016.10a为目标域时各方法的分类准确率随扰动幅度的变化情况。若无特殊说明，以下实验均在RML2016.04c→RML2016.10a的迁移场景进行。

由图3可知，未使用数据增强的4种方法在无攻击条件下目标域分类准确率较低；PGD和BIM的攻击效果优于FGSM，这是因为迭代攻击可通过多次梯度更新在扰动约束内寻找更有效的对抗扰动。在白盒攻击下，标准DANN模型的性能随扰动幅度增加而急剧下降，表明无监督域适应模型对对抗攻击具有明显脆弱性。AT+UDA通过源域对抗训练可以将鲁棒性迁移至目标域，但迁移效果相对有限。UDA+AT、SRoUDA和DART均先训练伪标签生成模型，再利用伪标签监督目标域的对抗训练，然而此类方法的防御性能依赖伪标签的生成质量。ARTUDA则在域对齐过程中加入自监督对抗训练以增强鲁棒性。上述方法均具有一定防御效果，验证了目标域对抗训练策略的有效性。本文方法在3种对抗攻击下均优于对比方法，数据增强策略有效提升了目标域原始信号的分类准确率，双判别器结构将目标域原始信号和对抗信号同时对齐至源域特征空间，并引入对比学习优化目标域决策边界，提升了模型的分类性能与对抗鲁棒性。

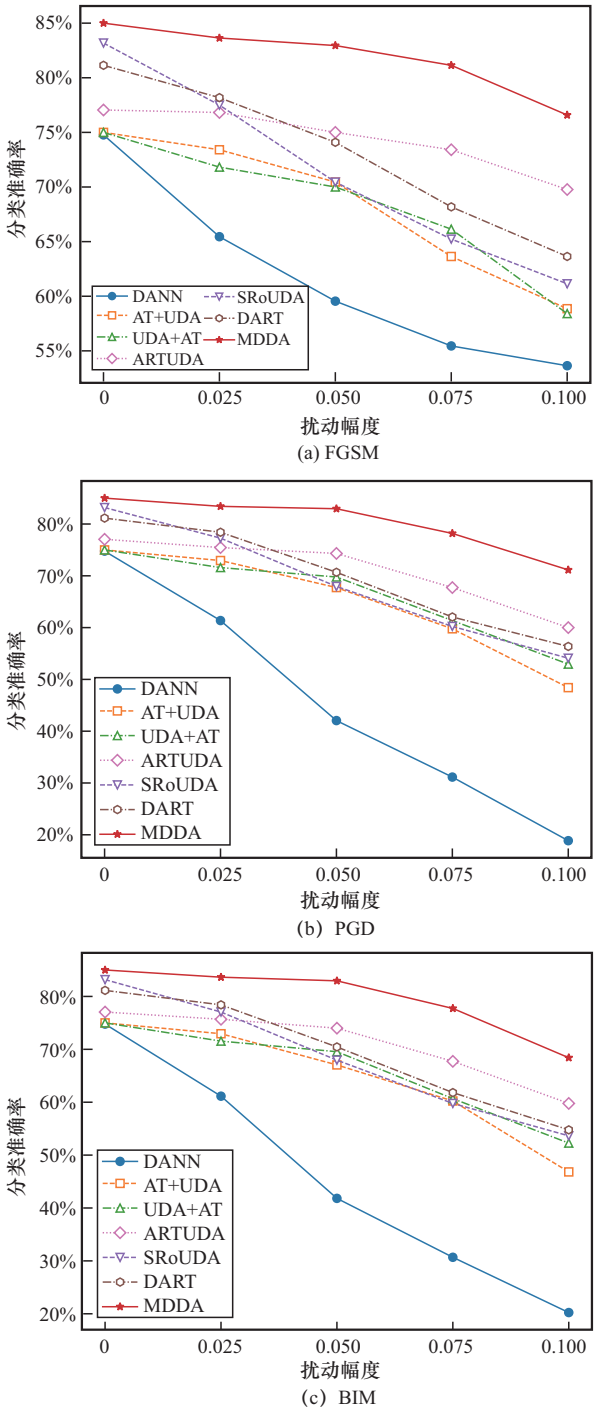

图3　RML2016.04c→RML2016.10a迁移场景下白盒攻击的分类准确率

2) 黑盒攻击

黑盒攻击相较于白盒攻击更符合实际应用场景。黑盒攻击通常假设攻击者无法获取目标模型的结构与参数信息，仅能基于替代模型生成对抗样本并迁移攻击目标模型。本文采用CNN作为替代模型进行DANN域对齐训练，以生成可迁移的对抗扰动。图4展示了黑盒攻击场景下各方法在FGSM、

BIM 和 PGD 攻击下的分类准确率随扰动幅度的变化情况。

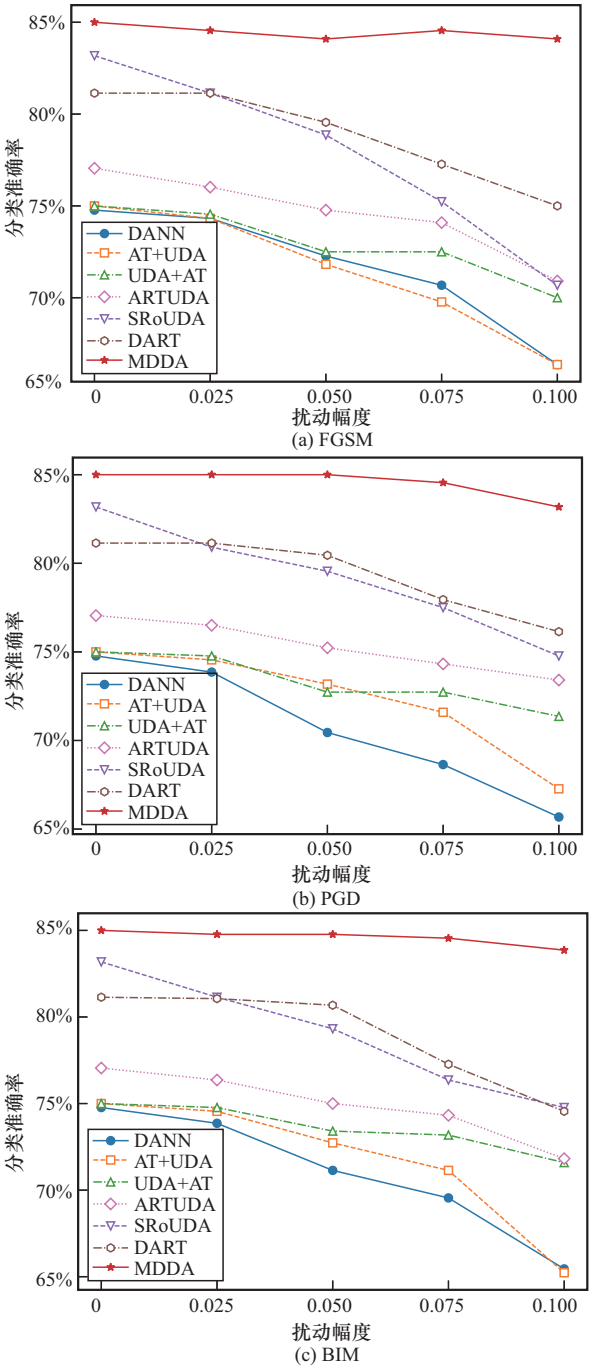

图 4　RML2016.04c→RML2016.10a 迁移场景下黑盒攻击的分类准确率

从图 4 可以看出，由于缺乏目标模型的先验信息，黑盒攻击的整体攻击效果弱于白盒攻击。在黑盒攻击场景下，FGSM 在不同扰动幅度下的攻击效果普遍略强于 PGD 与 BIM。这可能是因为 PGD 和 BIM 在替代模型上进行迭代优化时容易过拟合其决策边界，从而导致对抗样本的跨模型迁移性不足。现有的鲁棒 UDA 方法在黑盒攻击下均优于基准方法。值得注意的是，在 FGSM 黑盒攻击中，AT+UDA 的鲁棒性甚至低于未考虑对抗鲁棒性的标准 DANN，反映了仅依赖源域对抗训练的防御方法在黑盒场景下的局限性。相比之下，本文方法在黑盒攻击下表现出显著优势。当扰动幅度为 0.100 时，FGSM、PGD 与 BIM 3 种攻击下的分类准确率仅分别下降 0.91%、1.82% 和 1.14%，充分验证了本文方法在黑盒场景下的对抗鲁棒性。

3.3　基于优化的对抗攻击测试

本节进一步评估各方法对基于优化的 CW 攻击的防御性能。与基于梯度的攻击方法不同，CW 攻击通过优化过程逐步逼近模型的决策边界，在保证攻击成功的前提下寻找最小扰动。实验中，CW 攻击的迭代步长设置为 0.01，最大迭代次数分别设置为 2、4、6 和 8，目标置信度设置为 0。图 5(a) 和图 5(b) 分别展示了白盒和黑盒攻击场景下各方法分类准确率随迭代次数的变化趋势。

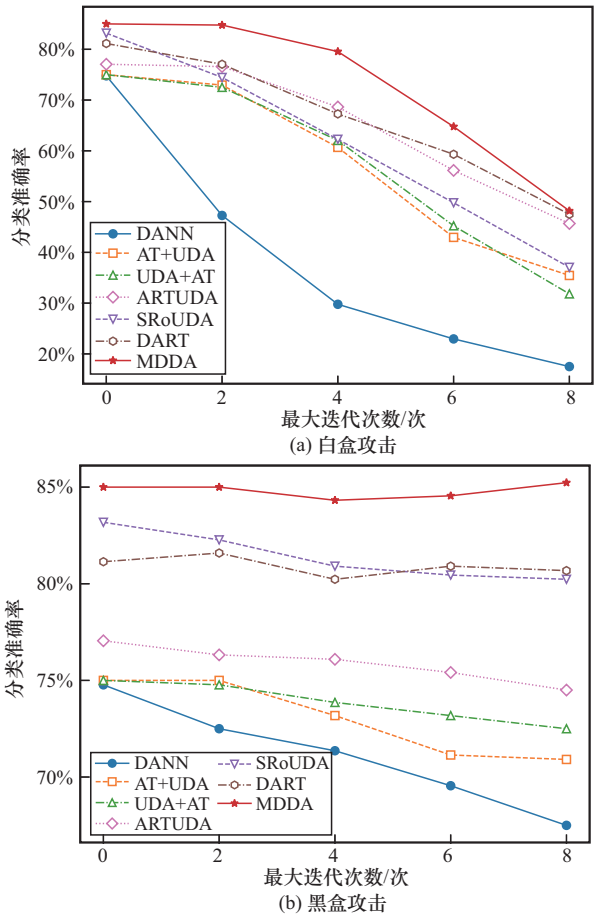

图 5　RML2016.04c→RML2016.10a 迁移场景下 CW 攻击的分类准确率

在白盒攻击场景下，CW 攻击的效果明显优于基于梯度的攻击方法。由于 CW 攻击通过优化机制直接逼近模型的决策边界，即使迭代次数较少，也能显著降低标准 DANN 模型的分类准确性。在各防御方法中，DART 与 ARTUDA 表现相近，仅次于本文方法。在黑盒攻击场景下，CW 攻击的效果明显弱于基于梯度的攻击方法。虽然 CW 攻击对标准 DANN 仍具有一定攻击能力，但对考虑对抗鲁棒性的防御方法影响十分有限，攻击前后的分类性能几乎无明显变化。这主要是因为 CW 攻击高度依赖目标模型的结构与参数信息，在缺乏先验知识的黑盒场景下难以有效逼近目标模型的决策边界。上述实验结果表明，MDDA 方法无论是在基于梯度的攻击还是在基于优化的攻击下，均展现出优异的防御性能。

3.4　不同信道场景下的对抗攻击测试

本节评估 MDDA 方法在仿真数据集上的对抗防御性能，实验构建了 Rician→AWGN 和 Rician→Rayleigh 两种典型信道迁移场景。表 2 和表 3 分别为不同信道场景下白盒攻击和黑盒攻击的分类准确率比较。白盒攻击中，FGSM、PGD 和 BIM 的扰动幅度设置为 0.050，CW 攻击的最大迭代次数设置为 4；黑盒攻击中，为更充分地评估各方法的鲁棒性差异，将 FGSM、PGD 和 BIM 的扰动幅度增大至 0.100，CW 攻击的最大迭代次数增加至 10。其余实验设置与 RML2016.04c→RML2016.10a 迁移实验保持一致。

从表 2 与表 3 可以看出，MDDA 方法在两种信道迁移场景下的防御效果明显优于其他对比方法。在白盒攻击下，当面对 PGD 攻击时，标准 DANN 在两种信道迁移场景下的分类准确率分别下降 57.95% 和 48.45%，MDDA 方法仅下降 4.09% 和 3.59%，表明其对目标域对抗扰动的稳健性。UDA+AT 在两种信道迁移场景下均具有较好的鲁棒性，进一步验证了基于伪标签的对抗训练能够提升模型防御能力，但其效果受伪标签质量的制约。ARTUDA 通过自监督对抗训练提升了对抗鲁棒性，但仅依赖该机制会导致原始信号的分类性能下降。

表 2　不同信道场景下白盒攻击的分类准确率比较

方法	Rician→AWGN					Rician→Rayleigh				
	无攻击	FGSM	PGD	BIM	CW	无攻击	FGSM	PGD	BIM	CW
DANN	83.27%	51.55%	25.32%	24.50%	15.77%	81.59%	57.95%	33.14%	33.45%	23.00%
AT+UDA	82.41%	74.64%	71.23%	71.23%	51.09%	77.41%	73.05%	72.50%	72.77%	68.59%
UDA+AT	83.23%	81.59%	81.00%	80.95%	78.91%	81.18%	76.27%	74.27%	74.09%	68.18%
ARTUDA	81.82%	77.95%	73.68%	74.09%	71.23%	75.00%	66.77%	64.09%	63.86%	59.05%
SRoUDA	85.14%	78.00%	75.14%	75.23%	62.14%	80.82%	73.27%	69.23%	68.95%	56.00%
DART	85.05%	82.86%	80.59%	80.41%	77.95%	82.18%	74.95%	72.55%	72.41%	68.00%
MDDA	**86.95%**	**83.59%**	**82.68%**	**82.86%**	**81.77%**	**87.36%**	**85.68%**	**83.64%**	**83.36%**	**82.18%**

表 3　不同信道场景下黑盒攻击的分类准确率比较

方法	Rician→AWGN					Rician→Rayleigh				
	无攻击	FGSM	PGD	BIM	CW	无攻击	FGSM	PGD	BIM	CW
DANN	83.27%	63.45%	68.36%	69.05%	80.50%	81.59%	61.41%	65.82%	67.36%	79.77%
AT+UDA	82.41%	73.14%	68.36%	73.45%	82.09%	77.41%	73.73%	74.27%	73.91%	76.68%
UDA+AT	83.23%	81.95%	68.36%	82.82%	83.02%	81.18%	77.50%	78.86%	78.50%	80.86%
ARTUDA	81.82%	78.50%	68.36%	78.86%	81.36%	75.00%	69.86%	69.68%	69.73%	72.95%
SRoUDA	85.14%	82.23%	68.36%	82.23%	84.55%	80.82%	77.41%	77.68%	77.14%	80.09%
DART	85.05%	**83.91%**	68.36%	84.66%	85.00%	82.18%	78.05%	79.45%	78.73%	81.64%
MDDA	**86.95%**	83.18%	**85.00%**	**84.95%**	**86.55%**	**87.36%**	**84.18%**	**85.95%**	**85.82%**	**87.23%**

相比之下，MDDA方法在引入自监督约束的同时，通过源域特征引导目标域原始信号与对抗信号的对齐，有效缓解了自监督一致性约束对原始信号分类性能的负面影响。在黑盒攻击场景下，以FGSM攻击为例，标准DANN在两种信道迁移场景下的分类准确率分别下降19.82%和20.18%，AT+UDA在两种场景下的性能分别下降9.25%和3.68%，其他方法均表现出较强的鲁棒性。CW黑盒攻击由于缺乏目标模型的先验信息，对各方法的影响更有限。MDDA方法在两种信道迁移场景下均保持显著优势，充分验证了该方法在不同信道场景下具有良好的域适应性和对抗鲁棒性。

3.5　消融实验

本节通过消融实验验证数据增强、双判别器结构、对比学习约束及一致性约束对MDDA方法的作用，结果如表4所示，其中，w/o D_c表示仅保留一个域判别器进行域对齐，w/o D_c& ℓ_{con}表示在使用单域判别器的同时移除对比损失。实验中采用PGD和CW两种对抗攻击方法，PGD攻击的扰动幅度设置为0.050，CW攻击的最大迭代次数设置为4。

表4　MDDA在不同消融条件下的分类准确率比较

方法	无攻击	PGD	CW
MDDA	85.00%	**82.95%**	**79.55%**
w/o $\hat{x}^s$	82.05%	78.18%	74.32%
w/o $\hat{x}^t$	**85.23%**	80.23%	77.27%
w/o ℓ_{KL}	84.32%	81.82%	75.23%
w/o ℓ_{con}	83.18%	78.64%	77.05%
w/o D_c	84.55%	82.05%	77.50%
w/o D_c&ℓ_{con}	79.55%	72.73%	71.36%

由表4可知，各组件对模型性能均有积极作用。在数据增强方面，移除源域数据增强后，目标域的泛化性能和对抗鲁棒性均明显下降，表明源域数据增强有助于特征提取器学习泛化特征，从而更有效地将源域的鲁棒性迁移至目标域；移除目标域数据增强后，原始信号的分类准确率虽略有提升，但对抗鲁棒性明显下降，说明目标域数据增强通过丰富样本多样性，可优化对抗域对齐效果。在自监督机制方面，移除ℓ_{KL}后，模型在对抗攻击下的性能明显下降，进一步验证了自监督对抗训练机制的有效性。在对比学习方面，移除ℓ_{con}后，模型在原始和对抗信号上的性能均大幅下降，表明ℓ_{con}可以通过源域特征，有效减小目标域原始信号与对抗信号在同类特征之间的距离，同时增大不同类特征之间的距离。在判别器结构方面，虽然仅使用单一判别器的分类准确率下降不明显，但是当同时移除对比学习时，性能下降十分明显，这说明单一判别器的域对齐效果不佳，进一步表明双判别器与对比学习具有协同增效作用，前者实现域间对齐，后者优化类内紧凑性与类间分离性，两者结合可显著提升模型的分类性能与对抗鲁棒性。

3.6　不同SNR下的性能验证

为进一步验证不同域适应方法在更低信噪比条件下的鲁棒性表现，本文在SNR=10 dB条件下对3种典型迁移场景中的白盒对抗攻击分类性能进行了对比分析，实验结果如表5所示。对抗样本由PGD方法生成，扰动幅度设置为0.050。

表5　10 dB下不同信道场景的白盒攻击的分类准确率比较

方法	RML2016.04c→RML2016.10a		Rician→AWGN		Rician→Rayleigh	
	无攻击	PGD	无攻击	PGD	无攻击	PGD
DANN	74.55%	26.82%	78.82%	28.91%	81.05%	22.95%
AT+UDA	73.27%	65.23%	80.45%	61.23%	79.73%	71.86%
UDA+AT	75.68%	65.00%	77.86%	75.68%	81.09%	73.82%
ARTUDA	78.18%	**75.68%**	78.82%	71.59%	81.00%	76.09%
SRoUDA	78.86%	65.95%	84.14%	73.91%	81.64%	69.59%
DART	78.91%	66.91%	84.64%	77.27%	82.95%	74.82%
MDDA	**81.82%**	72.73%	**87.00%**	**86.32%**	**84.00%**	**76.45%**

从无攻击结果可以看出，在10 dB条件下，各种方法在3种迁移场景中仍能保持相对稳定的分类性能，不同方法之间存在性能差异，相比之下，MDDA方法在3种迁移场景下均取得了较高的无攻击分类准确率，表明其在更低信噪比环境中具备较好的目标域调制信号分类能力。在PGD白盒攻击下，不同方法的鲁棒性差异明显，DANN方法在对抗攻击下性能显著下降，引入鲁棒增强机制的方法在一定程度上缓解了对抗扰动导致的性能下降。MDDA方法在3种迁移场景下均取得了较为优异的对抗鲁棒性，特别地，在Rician→AWGN迁移场景中，MDDA方法在PGD攻击下的分类准确率仅下

降0.68%，显著优于其他对比方法；在RML2016.04c→RML2016.10a和Rician→Rayleigh场景中，其性能下降幅度分别为9.09%和7.55%，同样保持在较低水平。上述结果表明，MDDA方法能够在不同信道迁移场景下有效防御对抗扰动，具有良好的鲁棒性和泛化能力。

3.7　复杂度分析

为验证本文方法的实际部署可行性，本节从计算复杂度和实时性两个方面进行分析。计算复杂度指标包括每秒浮点运算次数（floating point operation per second, FLOPS）和模型参数量，实时性指标采用单样本推理时间。需要说明的是，双判别器仅在训练阶段使用，推理阶段仅保留特征提取器和分类器。另外，在计算FLOPS时，采用的输入批量大小设定为128，与模型训练阶段的输入设置保持一致。

由于所有对比方法均采用相同的基线网络结构，因此各方法的FLOPS和参数量保持一致，分别为3.72×10^6和287.83×10^3。表6给出了各方法在3种迁移场景下的推理时间对比。在RML2016.04c→RML2016.10a迁移场景中，MDDA方法的单样本推理时间为0.024 6 ms，较DANN缩短了0.005 8 ms，表现出最优的推理效率。在Rician→AWGN和Rician→Rayleigh两种信道迁移场景中，MDDA方法的推理时间分别为0.032 1 ms和0.033 7 ms，虽略高于DANN和ARTUDA，但相较于两种鲁棒域自适应方法DART和SRoUDA仍具有明显优势。综合来看，MDDA方法在显著提升对抗鲁棒性的同时，保持了较低的计算开销，各场景下的推理时间均在0.034 ms以内，能够满足部署的实时性需求。

表6　不同模型的推理时间

方法	RML2016.04c→RML2016.10a/ms	Rician→AWGN/ms	Rician→Rayleigh/ms
DANN	0.030 4	0.032 8	**0.030 2**
AT+UDA	0.028 0	0.033 0	0.034 9
UDA+AT	0.026 3	0.034 2	0.034 0
ARTUDA	0.028 7	**0.031 5**	0.032 0
SRoUDA	0.029 1	0.035 7	0.037 0
DART	0.031 6	0.036 9	0.036 9
MDDA	**0.024 6**	0.032 1	0.033 7

4　结束语

本文提出了MDDA方法，以提升信道动态变化下AMC模型的鲁棒性与适应性。通过相位旋转的数据增强策略，丰富模型可学习的判别性特征与域不变特征，从而有效增强其跨域泛化能力。在此基础上，设计了多域分布对齐的域自适应机制，构建了双判别器结构，通过对源域与目标域的原始信号和对抗信号分别进行特征对齐，显著减小了跨域特征分布差异。同时，引入对比学习，利用源域类别特征作为锚点，优化目标域的类间决策边界。此外，通过自监督一致性约束确保目标域原始信号与对抗信号在输出空间保持一致性，进一步提升模型的对抗鲁棒性。在公开数据集RML2016.10c、RML2016.10a以及3种典型信道模型数据集上的实验结果表明，MDDA方法在多种攻击场景与不同扰动强度下均能有效抵御对抗攻击，显著提升无监督条件下AMC模型的对抗鲁棒性。然而，本文基于闭集假设，在实际开放的无线通信环境中，接收信号的调制类型往往无法被训练集完全覆盖，测试阶段不可避免地会出现未知类别。后续工作将面向开放集场景，进一步探索同时具备开放集识别能力与对抗防御能力的AMC方法。

参考文献：

[1] 梁应敞, 谭俊杰, Dusit Niyato. 智能无线通信技术研究概况[J]. 通信学报, 2020, 41(7): 1-17.
Liang Y C, Tan J J, Niyato D. Overview on intelligent wireless communication technology[J]. Journal on Communications, 2020, 41(7): 1-17.

[2] Tu Y, Lin Y, Hou C B, et al. Complex-valued networks for automatic modulation classification[J]. IEEE Transactions on Vehicular Technology, 2020, 69(9): 10085-10089.

[3] 张思成, 林云, 涂涯, 等. 基于轻量级深度神经网络的电磁信号调制识别技术[J]. 通信学报, 2020, 41(11): 12-21.
Zhang S C, Lin Y, Tu Y, et al. Electromagnetic signal modulation recognition technology based on lightweight deep neural network[J]. Journal on Communications, 2020, 41(11): 12-21.

[4] Peng S L, Sun S J, Yao Y D. A survey of modulation classification using deep learning: signal representation and data preprocessing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7020-7038.

[5] Lin Y, Zha H R, Tu Y, et al. GLR-SEI: green and low resource specific emitter identification based on complex networks and fisher pruning[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(5): 3239-3250.

[6] Sadeghi M, Larsson E G. Physical adversarial attacks against end-to-end autoencoder communication systems[J]. IEEE Communications Letters, 2019, 23(5): 847-850.

[7] Bao Z D, Lin Y, Zhang S C, et al. Threat of adversarial attacks on DL-based IoT device identification[J]. IEEE Internet of Things Journal, 2022, 9(11): 9012-9024.

[8] 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.

[9] Lin Y, Zhao H J, Ma X F, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2021, 70(1): 389-401.

[10] 张思成, 张建廷, 杨研蝶, 等. 电磁频谱人工智能模型的对抗安全威胁综述[J]. 无线电通信技术, 2024, 50(1): 1-13.
Zhang S C, Zhang J T, Yang Y D, et al. Review of adversarial security threats to electromagnetic spectrum artificial intelligence models[J]. Radio Communications Technology, 2024, 50(1): 1-13.

[11] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: OpenReview. net, 2015: 1035-1045.

[12] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the 6th International Conference on Learning Representations. San Diego: OpenReview.net, 2018: 1-23.

[13] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[C]//Proceedings of the 5th International Conference on Learning Representations- Workshop Track Proceedings. San Diego: OpenReview.net, 2017: 1-14.

[14] Kokalj-Filipovic S, Miller R, Morman J. Targeted adversarial examples against RF deep classifiers[C]//Proceedings of the ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2019: 6-11.

[15] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[16] Bao Z D, Zhang S C, Yang S L, et al. PFRTF: a robust training framework to counter adversarial attacks in signal classification for next-G consumer electronics[J]. IEEE Transactions on Consumer Electronics, 2025, 71(1): 1235-1248.

[17] Zhang S C, Yang Y D, Zhou Z Y, et al. DIBAD: a disentangled information bottleneck adversarial defense method using Hilbert-Schmidt independence criterion for spectrum security[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 3879-3891.

[18] Zhang S C, Lin Y, Yu J R, et al. HFAD: homomorphic filtering adversarial defense against adversarial attacks in automatic modulation classification[J]. IEEE Transactions on Cognitive Communications and Networking, 2024, 10(3): 880-892.

[19] McClintick K W, Harer J, Flowers B, et al. Countering physical eavesdropper evasion with adversarial training[J]. IEEE Open Journal of the Communications Society, 2022, 3: 1820-1833.

[20] Wang S, Xing H T, Wang C X, et al. SigDA: a superimposed domain adaptation framework for automatic modulation classification[J]. IEEE Transactions on Wireless Communications, 2024, 23(10): 13159-13172.

[21] 刘文学, 苗昫宸, 杨超三, 等. 基于双置信度融合机制的半监督信号调制识别方法[J]. 通信学报, 2025, 46(9): 229-240.
Liu W X, Miao X C, Yang C S, et al. Semi-supervised signal modulation recognition method based on dual-confidence fusion mechanism[J]. Journal on Communications, 2025, 46(9): 229-240.

[22] 张晓林, 李阳, 孙溶辰. 脉冲噪声下基于域自适应的调制识别[J]. 哈尔滨工程大学学报, 2024, 45(9): 1840-1847.
Zhang X L, Li Y, Sun R C. Modulation recognition based on domain adaptation under impulsive noises[J]. Journal of Harbin Engineering University, 2024, 45(9): 1840-1847.

[23] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the 32nd International Conference on Machine Learning. New York: ACM Press, 2015: 97-105.

[24] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. Journal of machine learning research, 2016, 17(59): 1-35.

[25] Awais M, Zhou F W, Xu H, et al. Adversarial robustness for unsupervised domain adaptation[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 8548-8557.

[26] Lo S Y, Patel V M. Exploring adversarially robust training forUnsupervised domain adaptation[C]//16th Asian Conference on Computer Vision. Berlin: Springer, 2023: 561-577.

[27] Zhu W Q, Yin J L, Chen B H, et al. SRoUDA: meta self-training for robust unsupervised domain adaptation[J]. Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM Press, 2023: 3852-3860.

[28] Wang Y J, Hazimeh H, Ponomareva N, et al. DART: a principled approach to adversarially robust unsupervised domain adaptation[C]//Proceedings of the 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Piscataway: IEEE Press, 2025: 773-796.

[29] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy[C]//International conference on machine learning. New York: ACM Press, 2019: 7472-7482.

[30] Shafahi A, Saadatpanah P, Zhu C, et al. Adversarially robust transfer learning[C]//Proceedings of the 8th International Conference on Learning Representations. San Diego: OpenReview.net, 2020: 6278-6291.

[31] O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognition networks[C]//International conference on Engineering Applications of Neural Networks. Berlin: Springer, 2016: 213-226.

[32] O'shea T J, Corgan J, Clancy T C. Open radioML synthetic benchmark dataset[R]. 2026.

[作者简介]

杨研蝶（2000– ），女，河南南阳人，哈尔滨工程大学博士生，主要研究方向为电磁信号识别、对抗样本防御。

张思成（1996– ），男，山东临沂人，哈尔滨工程大学师资博士后，主要研究方向为电磁频谱智能感知模型对抗攻防与安全评测。

林云（1980– ），男，黑龙江哈尔滨人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为人工智能、深度学习、信号识别和智能评测等。

李奎贤（1998– ），男，河南安阳人，哈尔滨工程大学博士生，主要研究方向为频谱资源分配、智能化电磁管理。

徐路平（2002– ），男，河南南阳人，哈尔滨工程大学硕士生，主要研究方向为信号检测与处理、信号识别模型对抗攻防等。

韩宇（1986– ），男，黑龙江哈尔滨人，博士，哈尔滨工程大学副教授，主要研究方向为电磁环境数据挖掘与智能模型建模、电磁频谱地图生成等。