基于层间激活演化一致性的对抗样本检测方法

林云[1]，王佳丽[1]，张思成[1]，魏晓磊[2]
（1.哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001；2.中国运载火箭技术研究院，北京 丰台 100071）

摘要：自动调制分类作为电磁频谱感知的核心技术，是频谱监测等应用中感知环节的关键任务，但其对抗脆弱性对系统安全构成严峻威胁。针对现有对抗样本检测方法仅分析静态特征、忽视对抗扰动在网络前向传播过程中动态累积效应的问题，本文提出一种基于层间激活演化一致性的对抗样本检测方法。通过提取目标网络多个关键层的激活表征，训练自编码器学习各层激活流形，在重构空间中构建样本与类平均激活的距离矩阵，依据该矩阵的层间演化一致性实现对抗样本检测。实验结果表明，所提方法在两个数据集、五种典型攻击及多个信噪比条件下平均AUC超过90%，有效提升了基于深度学习的频谱感知在对抗样本攻击下的安全性。

关键词：电磁频谱感知；智能感知安全；对抗样本检测；层间激活演化；类平均激活
中图分类号：TN918.91
文献标志码：A
doi：10.11959/j.issn.1000

Adversarial Example Detection Method Based on Inter-layer Activation Evolution Consistency

Lin Yun[1], Wang Jiali[1], Zhang Sicheng[1], Wei Xiaolei[2]
1. School of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
2. Command Automation, The Second Artillery Engineering College, Xi'an, Shaanxi, 710025

Abstract: Automatic modulation classification, a core technique for electromagnetic spectrum sensing, serves as a key sensing task in applications such as spectrum monitoring, yet its adversarial vulnerability poses a severe security threat. To address the limitations that existing adversarial example detection methods only analyze static features while ignoring the dynamic cumulative effects of adversarial perturbations during forward propagation, an adversarial example detection method based on inter-layer activation evolution consistency was proposed. Activation representations from multiple key network layers were extracted, and autoencoders were trained to learn their activation manifolds. Distance matrices between samples and class-average activations were constructed in the reconstruction space, and adversarial examples were detected based on their inter-layer evolution consistency. Experimental results show that an average AUC exceeding 90% was achieved across two datasets, five typical attacks, and multiple signal-to-noise ratios, effectively enhancing the security of deep-learning-based spectrum sensing against adversarial attacks.

Key words: electromagnetic spectrum sensing, intelligent sensing security, adversarial example detection, inter-layer activation evolution, class average activation

收稿日期：2026–XX–XX；修回日期：XXXX–XX–XX
通信作者：张思成，2015080325@hrbeu.edu.cn
基金项目：黑龙江省博士后面上基金（No.3236340036）；国家自然科学基金资助项目（No.U23A20271）
Foundation Items：The Heilongjiang Postdoctoral Fund (No.3236340036), The National Natural Science Foundation of China (No. U23A20271)

0 引言

电磁频谱作为信息时代的核心战略资源，已成为支撑国家安全与社会发展的重要基础[1-3]。随着6G通信、低空经济等新兴应用的快速发展，电磁环境日益复杂，频谱监测与安全监管面临更高要求，亟需更加智能化、自适应的频谱认知技术[4]。面向电磁频谱的感知任务涵盖调制识别、波形分类、辐射源个体识别等，其中自动调制分类（Automatic Modulation Classification，AMC）是判别通信体制、解析信号调制类型的前置环节，处于频谱监管与识别处理链路的关键位置[5]。近年来，深度神经网络凭借端到端的特征学习能力，显著提升了AMC的识别准确率与处理效率，已成为该任务的主流技术[6]。

然而，深度神经网络固有的脆弱性对其分类能力构成严峻威胁。攻击者在电磁信号中注入微小扰动即可形成对抗样本，诱使模型产生高置信度的误判[7]。在频谱认知以及具身智能等应用中，感知环节处于处理链路的最前端，一旦其输出因对抗样本出现偏差，该偏差将沿后续决策链路逐级传递并放大，显著增加系统的安全风险[8]。因此，构建面向AMC系统的高可靠对抗样本防护机制，是保障基于深度学习的频谱感知可信运行的关键任务。

针对对抗样本攻击威胁，现有防御策略主要分为对抗防御与对抗样本检测两类[9]。其中，对抗防御旨在通过加固模型提升其鲁棒性，但难以对各类攻击实现完全免疫，仍可能存在对抗样本绕过防御机制并干扰模型判决；对抗样本检测则在分类前过滤可疑样本，无需修改原有模型架构，且可与其他防御手段协同构建多层次防御体系，已成为AMC系统对抗攻击防护的重要技术路线之一。

现有对抗样本检测方法大致可分为三类。基于统计分布的方法通过分析输入或特征空间的分布特性区分正常与对抗样本，主要包括核密度估计[10]、局部内在维度分析[11]、马氏距离度量[12]及主成分统计量[13]等；基于网络不变性的方法利用模型对正常样本的预测一致性或表征稳定性进行检测[14-17]。这两类方法均聚焦于特定层级或特征空间的静态差异，未能捕捉扰动在前向传播中的动态变化。基于层间演化特性的方法则通过分析样本在各层的传播路径与激活模式检测异常[18-20]，但现有工作多以误差特征作为单一判别特征，对弱扰动敏感度受限，且常需逐层训练独立模块或依赖启发式搜索确定层组合，计算开销大、部署困难。

为捕捉对抗扰动在网络前向传播中的动态累积效应，本文提出一种基于层间激活演化一致性的对抗样本检测方法（Adversarial Example Detection Method Based on Inter-layer Activation Evolution Consistency，IAEC-AD）。该方法基于样本激活表征在目标网络各关键层间的关联模式进行设计。具体地，在各关键层训练Wasserstein自编码器（Wasserstein AutoEncoder，WAE）建模激活流形，计算重构样本与各类平均激活的距离，构建以层数和类别数为维度的跨层类平均激活距离矩阵。最后，训练支持向量机（Support Vector Machine，SVM）学习该矩阵的层间演化模式，实现对抗样本检测，本文的主要贡献如下。

1）提出一种基于层间激活演化一致性的对抗样本检测方法，将检测依据由单层静态特征扩展至跨层动态表征变化，有效刻画了对抗扰动在网络前向传播过程中的逐层累积效应。

2）提出一种基于WAE激活重构与类平均激活距离矩阵的检测特征构造方法，通过建模样本在各层相对于各类别中心的连续距离变化，保留了完整的类别关联信息，显著增强了对弱扰动对抗样本的检测敏感度。

3）在两个公开数据集RML2016.10a和RML22上，针对五种典型对抗攻击及多个信噪比条件开展实验，所提方法在两个数据集上的平均AUC均超过90%，在高信噪比条件下最高可达98.69%，充分验证了所提方法的有效性。

1 相关工作

1.1 基于统计分布的对抗样本检测

基于统计分布的检测方法假设对抗样本与正常样本在输入或特征空间服从不同分布，通过建模正常样本分布并识别偏离实现检测。Feinman等人[10]结合贝叶斯不确定性估计与深层特征核密度估计表征分布偏移；Ma等人[11]利用局部内在维度刻画样本邻域的几何特性进行判别；Lee等人[12]对深层特征建立类条件高斯模型，以马氏距离度量样本到类中心的偏离；Li等人[13]提取卷积层主成分统计量构建级联分类器。然而，上述方法主要聚焦于特定层级特征空间的静态分布差异，由于不同攻击算法

产生的扰动在统计特性上差异显著，针对特定攻击训练的检测器泛化能力有限。

1.2　基于网络不变性的对抗样本检测

基于网络不变性的检测方法利用深度神经网络对正常样本的预测一致性或表征稳定性，通过观测模型在输入变换下的行为差异识别对抗样本。Zhang等人[14]对输入施加像素遮蔽等变换，利用中间层标签序列的波动检测异常；Gao等人[15]采用双自编码器结合互信息估计比较潜在空间与输出空间的分布一致性；徐东伟等人[16]通过VMD去噪前后模型输出的相似性与置信度差值进行融合检测；Wang等人[17]基于特征归因差异并结合三重对比策略缓解信号噪声干扰。然而，这类方法对网络架构依赖较强，且主要依赖输出层或特定层级的行为，未考虑扰动在网络中的动态传播。

1.3　基于层间演化特性的对抗样本检测

基于层间演化特性的检测方法通过分析样本在神经网络各层激活空间中的行为变化识别对抗样本，这类方法的理论基础是深度神经网络对扰动的逐层累积放大效应[21]。深度网络由多层非线性变换级联构成，输入端的微小扰动经前向传播后会逐层累积，在深层产生的表征偏离远大于浅层。受此影响，对抗样本在浅层贴近原始类别，随层级加深逐渐向错误类别偏移，而正常样本在各层激活空间中类别保持稳定[23]。Wójcik等人[18]提出基于层级自编码器重构误差的检测方法（Layer-wise Autoencoder Detection，AE-Layers），在目标网络多个中间层分别训练自编码器学习正常样本激活的流形表示，通过各层重构误差判别对抗样本。Katzir等人[19]提出基于激活空间行为的检测方法（Spatial Behavior in Activation Spaces，SPBAS），在各层激活空间中构建k近邻分类器，为输入样本生成跨层标签序列，通过序列似然度检测异常。Mumcu等人[20]提出基于层间非均匀影响的检测方法（Non-uniform Impact on Layer Analysis，NILA），训练轻量级回归模型从浅层特征预测深层特征，利用预测误差反映对抗样本攻击对不同层的影响。

上述方法存在共同局限，AE-Layers和NILA以重构误差作为检测特征，误差作为单一标量值，丢失了类别维度信息，当扰动较弱时检测敏感度下降；SPBAS虽引入类别标签，但离散标签序列损失了样本与各类别的连续距离结构，且需为每层训练独立的k近邻分类器，计算复杂度较高。

针对上述局限，本文提出基于层间激活演化一致性的对抗样本检测方法，以类平均激活距离矩阵作为检测特征。相比误差类特征，距离矩阵保留了样本与各类别中心的完整距离关系，对弱扰动具有更高的检测敏感度；相比离散标签序列，连续距离度量避免了量化粗化导致的信息损失；基于残差块输出的层选择策略无需为每层单独训练检测模块，有效降低了计算开销。

2　基于层间激活演化一致性的对抗样本检测方法

本节提出基于层间激活演化一致性的对抗样本检测方法（IAEC-AD），其核心在于将样本在网络各层相对各类别中心的连续距离变化，作为对抗扰动跨层累积效应的统一表征。为获得稳定且可比较的跨层距离结构，先通过WAE对各层激活进行

流形约束与去噪，再构建类平均激活距离矩阵，最终由检测分类器学习该矩阵的层间演化模式以识别对抗样本。基于上述设计，本节构建了如图1所示的IAEC-AD检测框架，主要包括多层激活特征提取、基于WAE的激活重构、类平均激活距离矩阵构建以及检测判别四个主要部分。

2.1　多层激活特征提取

深度神经网络对对抗样本的脆弱性源于其层级化非线性变换的固有特性[23]。根据Lipschitz连续性理论，若网络第 k 层变换的Lipschitz常数为 L_k，则L层深度网络复合映射的全局Lipschitz常数上界为

$$Lip(f) \leqslant \prod_{k=1}^{L} L_k \tag{1}$$

上式给出了输入扰动在第层特征空间中传播幅度的理论上界。随着网络层级加深，多层非线性映射的局部增益不断累积，使输入端微小扰动在深层产生显著特征偏移的可能性增加[22]。该上界并不表示任意扰动均会被单调放大，但对抗扰动沿分类损失增大的方向构造，更容易在传播过程中形成累积偏移。因此，单一层级难以完整描述扰动的传播过程，本文从多个关键层提取激活，以追踪其跨层动态演化。

本文选用残差网络（Residual Networks，ResNet）作为目标网络进行方法验证，并选取各残

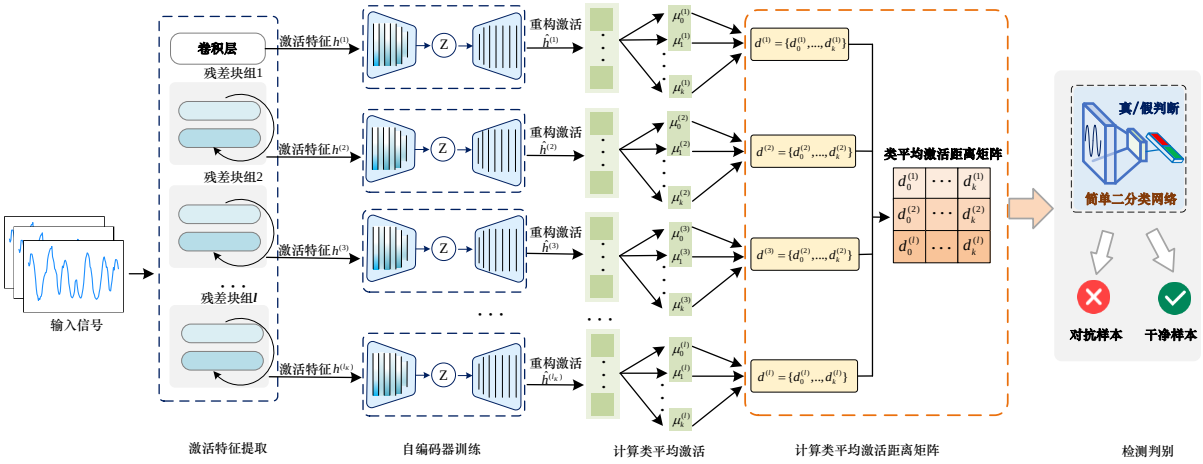


图1　基于层间激活演化一致性的对抗样本检测方法框架图

差块的输出端作为激活提取位置。相较于仅含残差支路非线性变换结果的中间激活，残差块输出端融合了非线性特征变换与恒等映射，同时包含扰动的非线性处理分量与直接传递分量，能更综合地反映扰动在该层级的累积状态，因而作为激活演化的观测点更为合理。设目标网络由 L 个功能模块组成，定义激活提取层集合为 $\Omega=\{l_1,l_2,\ldots,l_K\}$，其中 K 为提取层数量，具体的激活提取位置如图2所示。对于输入样本 $\boldsymbol{x}$，第 l_k 层的激活表征可形式化表示为

$$\boldsymbol{h}^{(l_k)}=f^{(l_k)}(\boldsymbol{x})\in R^{C_{l_k}\times H_{l_k}\times W_{l_k}} \tag{2}$$

其中，$f^{(l_k)}(\cdot)$ 表示网络从输入层到第 l_k 层的复合映射，C_{l_k}、H_{l_k}、W_{l_k} 分别为该层激活的通道数、高度和宽度。具体的提取过程采用前向钩子机制实现，在目标层注册回调函数，当网络执行前向传播时自动捕获并保存该层的输出张量，无需修改原有网络结构，从而保持目标分类器的完整性。

各层激活的空间维度存在差异，直接进行跨层比较存在困难。为此，本文对提取的激活特征进行自适应平均池化，将激活张量的空间维度压缩为1×1，随后展平为一维向量，形式化表示为

$$\tilde{\boldsymbol{h}}^{(l_k)}=\mathrm{Flatten}\left(\mathrm{AdaptiveAVgPool}\left(\boldsymbol{h}^{l_k}\right)\right)\in \mathrm{R}^{d_{l_k}} \tag{3}$$

其中，$d_{l_k}=C_{l_l}$ 为池化后的特征维度。经过上述处理，各层激活被统一表示为固定格式的一维向量，便于后续的自编码器重构与距离计算。

2.2　基于自编码器的激活重构

经过上述多层激活提取后，各层激活被表示为一维向量。然而，原始激活空间中直接计算距离面临流形结构失真及冗余噪声干扰等问题，难以准确反映样本的类别关系。为解决上述问题，本文引入自编码器对各层激活进行重构。自编码器通过编码器与解码器的瓶颈结构，迫使网络在低维潜在空间中保留输入的核心信息，同时滤除冗余成分。设第 l_k 层的自编码器由编码器 $E^{(l_k)}$ 和解码器 $D^{(l_k)}$ 组成，对于输入激活 $\tilde{\boldsymbol{h}}^{(l_k)}$，编码过程将其映射到低维潜在空间，解码过程再从潜在表示重构回原始维度

$$\boldsymbol{z}^{(l_k)}=E^{(l_k)}\left(\tilde{\boldsymbol{h}}^{(l_k)}\right),\quad \hat{\boldsymbol{h}}^{(l_k)}=D^{(l_k)}\left(\boldsymbol{z}^{(l_k)}\right) \tag{4}$$

其中，$\boldsymbol{z}^{(l_k)}$ 为潜在表示，$\hat{\boldsymbol{h}}^{(l_k)}$ 为重构后的激活。由于各层激活的维度和分布特性存在差异，本文为每个提取层单独训练一个自编码器。

在自编码器类型选择上，本文采用 WAE。普通

自编码器对潜在空间缺乏分布约束，不同样本的编码分布不规则，影响后续距离度量的稳定性。变分自编码器（Variational Autoencoder，VAE）虽引入分布约束，但其基于KL散度的逐样本强约束会迫使各类别编码向同一先验分布聚集，造成不同类别的潜在表示发生混叠，破坏类平均激活的判别力。WAE改用最大均值差异（Maximum Mean Discrepancy，MMD）作为分布度量，仅要求所有样本的聚合编码分布与先验匹配，而非约束单个样本的编码形态。这一机制在保持潜在空间整体分布规则的同时，允许各类别维持相对独立的聚类结构，有效保证了类平均激活距离度量的稳定性。WAE的训练目标函数为

$$\mathcal{L}=\mathcal{L}_{\mathrm{recon}}+\lambda\cdot\mathrm{MMD}^2(q_z,p_z) \tag{5}$$

其中，$\mathcal{L}_{\mathrm{recon}}=\|\tilde{\boldsymbol{h}}^{(l_k)}-\hat{\boldsymbol{h}}^{(l_k)}\|_2^2$ 重构损失，保证编码与

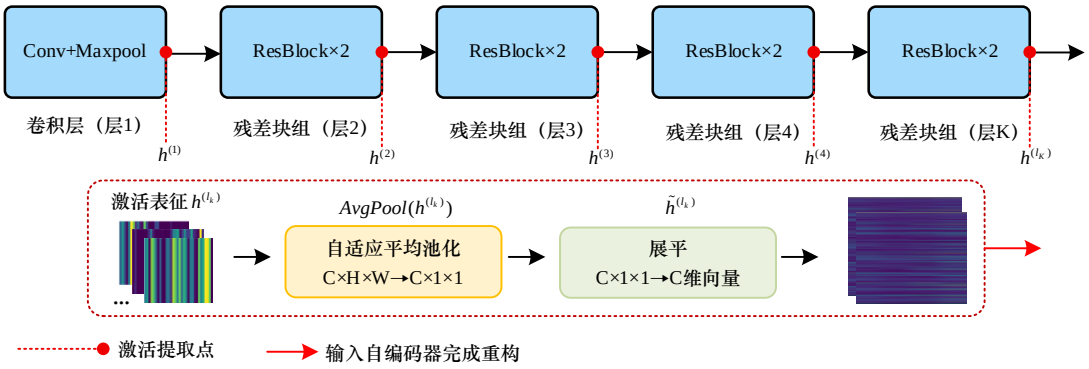


图2　激活提取过程

解码过程中信息的有效保留；q_z表示编码器产生的聚合潜在分布，p_z表示先验分布，第二项通过最大均值差异约束二者接近；λ为平衡系数。

编码与解码过程通过非线性流形投影滤除冗余噪声，使重构激活重构激活$\hat{\boldsymbol{h}}^{(l_k)}$相比原始激活$\tilde{\boldsymbol{h}}^{(l_k)}$具有更稳定的流形结构；重构空间保持与原始激活相同的维度，避免了潜在空间维度过低可能导致的信息损失。

2.3　类平均激活距离矩阵构建

深度网络中间层激活往往具有极高的维度，直接在该空间中计算距离易受维度灾难影响，导致不同类别间的欧氏距离测度失去区分性。为此，本文引入主成分分析（Principal Components Analysis，PCA）对重构激活进行特征降维。在降维后的空间中，为量化样本在网络各层的类别变化，本文引入类平均激活作为各类别在激活空间中的参考中心。正常样本经过网络处理后，其各层特征应靠近所属类别的中心。而对抗样本由于扰动的逐层累积效应，其特征会逐渐偏离原始类别中心。通过计算样本到各类平均激活的距离，可以量化样本与各类别之间的关联程度，进而追踪这种关联关系随网络层级的动态变化。

类平均激活及PCA投影矩阵均基于训练集中的正常样本计算，对每个类别、每个提取层分别进行。设训练集中所有正常样本在第l_k层的重构激活均值为$\bar{\boldsymbol{h}}^{(l_k)}$，对于任意输入样本$\boldsymbol{x}$，其在第$l_k$层的重构激活$\hat{\boldsymbol{h}}^{(l_k)}$经降维后的特征向量$\boldsymbol{p}^{(l_k)}$计算为

$$\boldsymbol{p}^{(l_k)}(\boldsymbol{x}) = \left(\hat{\boldsymbol{h}}^{(l_k)}(\boldsymbol{x}) - \bar{\boldsymbol{h}}^{(l_k)}\right)\boldsymbol{W}_{PCA}^{(l_k)} \tag{6}$$

设$\mathcal{S}_c$为属于类别c的训练样本集合，类别c在第l_k层的平均激活定义为

$$\boldsymbol{\mu}_c^{(l_k)} = \frac{1}{|\mathcal{S}_c|}\sum_{\boldsymbol{x}\subset\mathcal{S}_c}\boldsymbol{p}^{(l_k)}(\boldsymbol{x}) \tag{7}$$

样本激活与类平均激活之间的距离采用欧氏距离度量。由于各层激活已经过WAE重构与PCA降维，特征空间在数值尺度和分布特性上更加稳定，适合直接使用欧氏距离进行度量。样本在第l_k层到类别c中心的距离定义为

$$d_{k,c} = \|\boldsymbol{p}^{(l_k)} - \boldsymbol{\mu}_c^{(l_k)}\|_2 \tag{8}$$

对于输入样本$\boldsymbol{x}$，计算其在每个提取层到每个类别中心的距离，构成类平均激活距离矩阵。类平均激活仅由正常样本得到，在检测时保持固定。记从输入到第k层PCA特征的复合映射为g_k，其局部Lipschitz常数为G_k。对于满足$\|\boldsymbol{\delta}\|_2 \le \varepsilon$的扰动，定义样本在第层的类别距离间隔为

$$\begin{aligned} m_k(\boldsymbol{x}) &= \min_{c\neq y}\left[d_{k,c}(\boldsymbol{x}) - d_{k,y}(\boldsymbol{x})\right] \\ \left|d_{k,c}(\boldsymbol{x}+\boldsymbol{\delta}) - d_{k,c}(\boldsymbol{x})\right| &\le G_k\varepsilon \\ m_k(\boldsymbol{x}+\boldsymbol{\delta}) &\ge m_k(\boldsymbol{x}) - 2G_k\varepsilon \end{aligned} \tag{9}$$

当$m_k(\boldsymbol{x}) > G_k$时，该层最近类别能够保持稳定；随着网络加深，扰动传播的累计敏感性增大，类别距离间隔可能逐渐缩小并发生最近类别翻转。因此，对抗样本容易形成由浅层接近原类别、向深层逐渐偏移的距离演化模式，而正常样本的类别距离关系相对稳定。

由于不同层激活的数值尺度存在差异，本文对距离矩$\boldsymbol{D}$中的元素$d_{k,c}$按层进行独立标准化

$$\tilde{d}_{k,c} = \frac{d_{k,c} - m_k}{s_k} \tag{10}$$

其中，$\tilde{d}_{k,c}$为标准化后的距离，m_k和s_k分别为第k层所有距离值的均值和标准差。距离值的均值m_k以及标准差s_k计算方式为

$$m_k = \frac{1}{C}\sum_{c=1}^{C} d_{k,c} \tag{11}$$

$$s_k = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(d_{k,c} - m_k)^2} \tag{12}$$

其中，C为类别数。标准化后的距离矩阵便于后续检测分类器对跨层模式的学习。

2.4　对抗样本检测

对抗样本检测阶段，将标准化距离矩阵按层级顺序向量化，并作为特征输入二分类SVM。该特征保留了样本与各类别中心的距离关系及其随网络层级的变化，使检测器能够识别对抗扰动引起的跨层异常偏移。自编码器和类平均激活仅使用正常样本训练，SVM则以正常样本为正类、对抗样本为负类，学习两类样本在层间演化特征空间中的判别边界。训练用对抗样本由FGSM、BIM、PGD、DeepFool和C&W五种攻击生成，训练集与测试集按类别分层划分且互不重叠。测试时，构建输入样本的标准化距离矩阵，并由SVM输出检测结果，整个检测过程无需修改目标分类网络的结构和参数。完整流程如算法1所示。

算法1 IAEC-AD对抗样本检测算法流程

输入：目标分类网络F，正常训练样本集χ，对抗训练样本集$\mathcal{X}_{adv}$，待测试样本x^*

输出：检测结果（正常样本/对抗样本）

训练阶段：

选定激活提取层集合$\mathcal{L}=\{l_1,l_2,\cdots,l_K\}$

FOR each layer $l_K \in \mathcal{L}$ DO

　提取正常样本激活$\mathcal{H}_k \leftarrow \left\{\boldsymbol{h}^{(l_k)}(\boldsymbol{x}) \mid \boldsymbol{x}\in\chi\right\}$并训练自编码器

　基于自编码器输出重构激活$\hat{\boldsymbol{h}}^{(l_k)}$拟合该层PCA投影矩阵$\mathbf{W}_{PCA}^{(l_k)}$及均值$\bar{\boldsymbol{h}}^{(l_k)}$

END FOR

FOR each layer $l_K \in \mathcal{L}$ DO

　FOR each class $c \in \{1,...,C\}$ DO

　　将正常样本重构激活投影至PCA降维空间，并在该空间计算类平均激活中心$\boldsymbol{\mu}_c^{(l_k)}$

　END FOR

END FOR

对正常样本$\mathcal{X}$与对抗样本$\mathcal{X}_{adv}$构建标准化距离矩阵特征，以上述带标签特征训练二分类SVM，得到检测判别边界

测试阶段：

FOR each layer $l_K \in \mathcal{L}$ DO

　提取$\boldsymbol{x}^*$在l_k层的激活$\boldsymbol{h}^{(l_k)}(\boldsymbol{x}^*)$

　经自编码器重构得$\hat{\boldsymbol{h}}^{(l_k)}(\boldsymbol{x}^*)$

　计算$\hat{h}^{(l_k)}(x^*)$在该层的PCA降维特征向量$\boldsymbol{p}^{(l_k)}(\boldsymbol{x}^*)$

　FOR each class $c \in \{1,...,C\}$ DO

　　计算距离$d_{k,c}=d(\boldsymbol{p}^{(l_k)},\boldsymbol{\mu}_c^{(l_k)})$

　END FOR

END FOR

构建距离矩阵$\boldsymbol{D}_{x^*}\in\boldsymbol{R}^{K\times C}$并层内标准化得$\tilde{\boldsymbol{D}}_{x^*}$

RETURN 检测结果（正常样本/对抗样本）

3　仿真分析

本节对IAEC-AD开展系统性实验评估，首先介绍实验设置，接下来进行对比实验、可视化分析、消融实验及推理时间对比。

3.1　实验设置

本文实验基于RML2016.10a[24]和RML22[25]两个公开的调制信号数据集开展。RML2016.10a包含8种数字调制和3种模拟调制共11种类型，每种调制在-20dB至18dB范围内以2dB间隔采样，每个信噪比下包含1000个样本。RML22数据集同样包含11种调制类型，样本长度与RML2016.10a一致。本文分别选取0dB和10dB开展实验，其中0dB用于模拟低信噪比及噪声影响明显的频谱监测场景，10 dB用于模拟接收质量较好的常规监测场景，以验证不同工作状态下的检测性能。目标分类模型采用针对调制信号特性改进的ResNet34网络。

实验对提取的各层分别训练WAE学习激活流形，WAE训练采用Adam优化器，学习率设为0.001，批量大小设为32，训练轮数设为100轮，MMD损失权重设为0.001。各层激活经WAE重构后，通过PCA降维至64维以缓解高维空间中的距离度量退化问题。检测分类器采用SVM，检测分类器采用SVM，其参数通过验证集确定。

对抗样本生成采用FGSM[26]、BIM[27]、PGD[28]、DeepFool[29]和C&W[30]五种攻击方法。其中FGSM为单步梯度攻击，BIM和PGD为多步迭代攻击；为在保证扰动隐蔽性的同时使其足以误导分类模型，将FGSM的扰动幅度设为0.1，即对抗扰动的能量约为原始信号能量的1%[23]；BIM以及PGD的扰动幅度设为0.05，迭代次数为40，步长为0.01；DeepFool通过迭代寻找最小扰动边界，迭代次数设为50，过冲系数设为0.01；C&W将对抗样本生成建模为约束优化问题，置信度设为0.7，迭代次数设为50。DeepFool及C&W的优化损失在约第40次迭代后变化已小于1%，故取50次迭代即可保证收敛。为验证所生成对抗样本的有效性，本节给出目标ResNet34模型在干净样本与五种攻击下的分类性能，如表1所示。其中，原始准确率(Ori ACC)为模型对未受扰动样本的分类准确率，

攻击后准确率(Adv ACC)为模型受对抗攻击后的分类准确率。

表1　目标模型攻击前后识别准确率

数据集	信噪比/dB	Ori ACC/%	攻击方式	Adv ACC/%
RML2016.10a	0	83.5	FGSM	24.75
			BIM	12.05
			PGD	12
			DeepFool	16.15
			C&W	0.3
	10	84.6	FGSM	22.5
			BIM	21.9
			PGD	20.25
			DeepFool	16.15
			C&W	1
RML22	0	81.32	FGSM	29.55
			BIM	16.27
			PGD	16
			DeepFool	13.59
			C&W	9.95
	10	86.95	FGSM	46.86
			BIM	33.55
			PGD	33.02
			DeepFool	13.59
			C&W	20.05

为全面评估检测性能，本文采用准确率（ACC）、AUC、F1分数、精确率（Precision）和召回率（Recall）五项指标。其中，F1分数为精确率和召回率的调和平均，综合反映检测器在两者上的均衡表现；AUC为ROC曲线下面积，ROC曲线以假阳性率为横轴、真阳性率为纵轴，表示检测器在不同决策阈值下的性能表现，AUC越接近1表示检测性能越好。

3.2　对比实验

为全面评估所提方法的检测性能，本节在RML2016.10a和RML22两个数据集上，针对FGSM、BIM、PGD、DeepFool和C&W五种攻击，将IAEC-AD与四种代表性检测方法进行对比。对比方法包括基于VMD去噪特征融合的检测方法[16]（VMDFF）、基于层级自编码器重构误差的检测方法[18]（AE-Layers）、基于激活空间行为的检测0dB和10dB条件下的ROC曲线分别如图3和图4所示，在RML22数据集0dB和10dB条件下的ROC方法[19]（SPBAS）以及基于层间非均匀影响的检测曲线分别如图5和图6所示。

从图3在RML2016.10a数据集0dB条件下的ROC曲线可以看出，IAEC-AD对应的红色实线在五种攻击下均最贴近左上角，整体AUC均值超过90%，其中C&W攻击下达到98.51%。从图4在10dB条件下的结果可以看出，面对BIM攻击，对比方法曲线显著偏离最优边界，AUC仅为68.41%，而本文方法ROC曲线仍贴近左上角，AUC达到93.91%。从图5在RML22数据集0dB条件下的结果可以看出，对比方法曲线整体下降明显，而本文方法曲线仍保持上升，五种攻击平均AUC达到90.04%。从图6在RML22数据集10dB条件下的结果可以看出，本文方法在所有攻击下的AUC值最大，平均AUC达到93.76%。从机理看，FGSM为单步攻击，BIM和PGD通过多步迭代累积扰动，DeepFool和C&W则通过逼近分类边界实现误导。IAEC-AD在不同攻击下均保持较稳定的检测性能，说明跨层距离矩阵能够捕捉不同攻击方式造成的共同表征偏移，降低检测器对特定扰动特征的依赖。

3.3　层间演化的可视化分析

为直观验证对抗扰动在网络前向传播中的跨层累积效应，本节通过可视化样本至11个类别中心的距离矩阵，分析正常与对抗样本的跨层激活差异。图7展示了正常样本及PGD、C&W攻击下对抗样本的距离矩阵热力图。图中每行对应一个网络层（由浅至深），每列对应一个调制类别；绿色虚线框与红色实线框分别标记样本的真实类别与各层距离最小的类别。

观察图7可知，图7(a)中正常样本在各层的最近类别保持高度一致，绿框与红框几乎完全重合，形成垂直色带。相比之下，图7(b)和图7(c)中的对抗样本在浅层表现正常，但随层数加深，红框逐渐偏移，至层4和层5已完全转移至错误类别。这一现象直观证实了对抗扰动经非线性级联作用后在深层产生的显著类别偏移，也突显了利用距离矩阵记录完整层间演化过程的必要性。

基于上述观察，本文定义跨层类别一致性指标（Layer-wise Class Consistency，LCC）量化样本在

各层最近类别的稳定程度，设样本$\boldsymbol{x}$在第k层的最近类别为$\hat{c}_k = \arg\min_c \boldsymbol{D}_{k,c}$，真实类别为$c^*$，则LCC定义为各层最近类别与真实类别一致的比例

$$LCC(x) = \frac{1}{K}\sum_{k=1}^{K} I(\hat{c}_k = c^*) \tag{13}$$

其中，$I(\cdot)$为指示函数，K为提取层数。LCC取值范围为[0,1]，值越高表示样本在各层的类别归属越稳定，反之则说明层间类别变化越明显。

图8展示了正常样本与PGD、C&W攻击下对抗样本的LCC分布箱线图。从图8可以看出，正常样本的LCC均值约为0.8，表明其在多数网络层中保持了正确的预测。相比之下对抗样本的LCC发生显著下降，其中C&W攻击与PGD攻击下的均值分别降至0.632与0.624。正常样本与对抗样本在

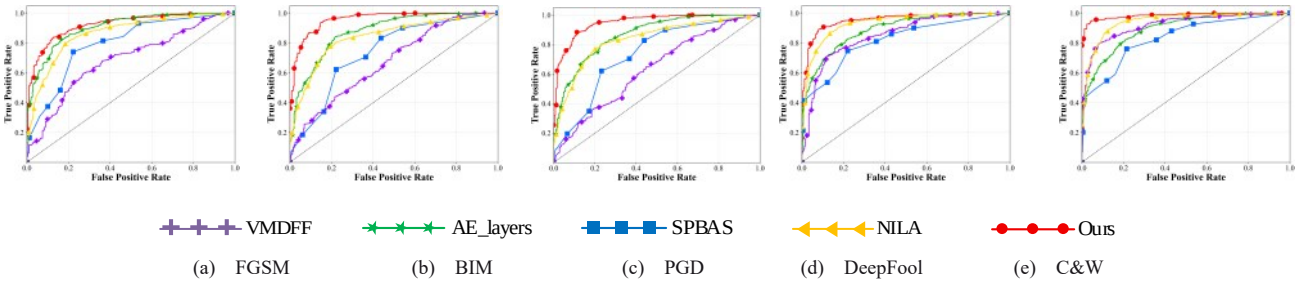


图3　RML2016.10a数据集对抗样本检测ROC曲线（SNR = 0 dB）

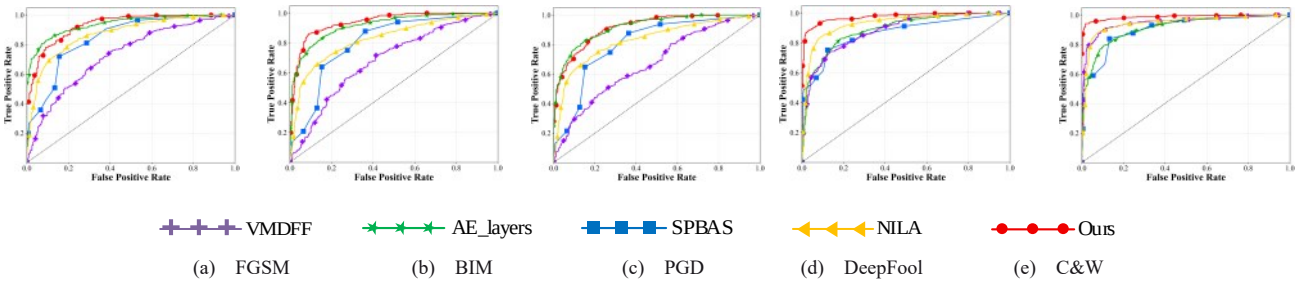
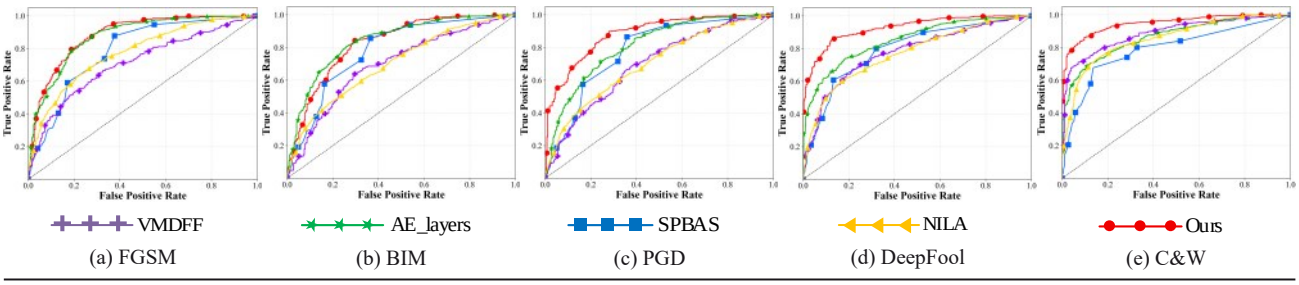


图4　RML2016.10a数据集对抗样本检测ROC曲线（SNR = 10 dB）

图5　　RML22数据集对抗样本检测ROC曲线（SNR = 0 dB）



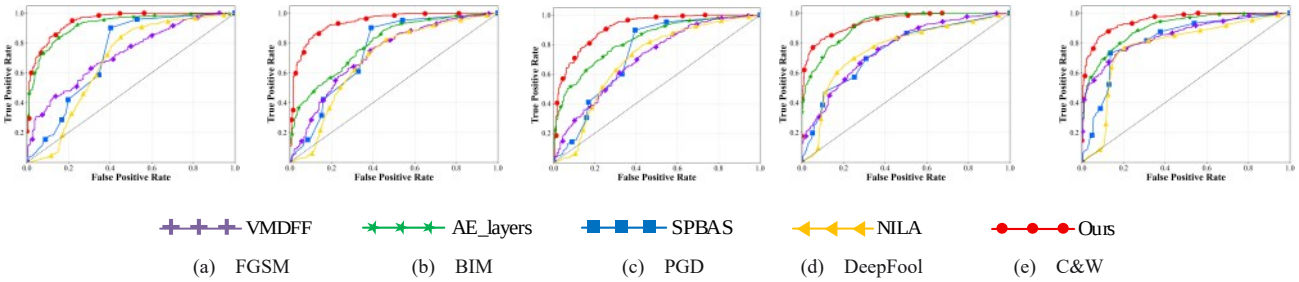


图6　RML22数据集对抗样本检测ROC曲线（SNR = 10 dB）

LCC分布上的差异表明，对抗扰动会破坏样本跨层类别关系的稳定性，与式（9）所描述的类别距离间隔逐渐缩小及最近类别翻转机理相一致。

3.4　算法综合性能评估

为全面评估所提IAEC-AD方法在不同数据集和攻击类型下的检测性能，本节给出ACC、AUC、F1分数、Precision和Recall五项指标在两个数据集上的详细结果，如表2所示。

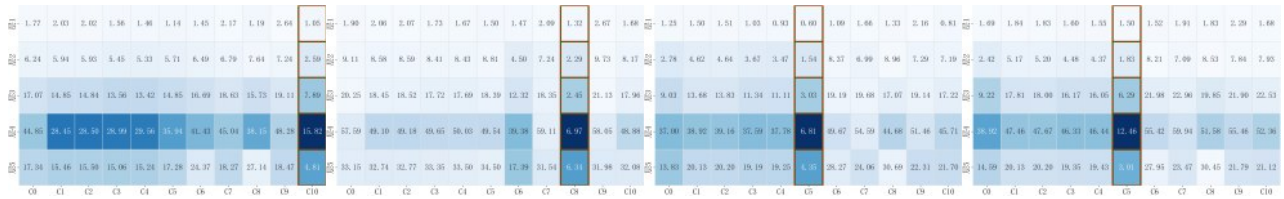
从表2可以看出，IAEC-AD方法在两个数据集的低信噪比与高信噪比条件下均展现出稳定的检测性能。在RML2016.10a数据集上，0dB条件下五种攻击的平均AUC达到95.20%，平均F1分数为92.43%；10dB条件下平均AUC达到95%，平均F1分数为91.96%；其中C&W攻击在两种信噪比下的检测性能均最优，0dB和10dB条件下AUC分别达到98.51%和98.69%，这是由于C&W攻击虽然扰动幅度小，但其优化目标导致样本在深层激活空间产生更显著的偏移，更易被跨层距离特征捕捉。RML22包含更接近实际无线传播环境的信道影响。在0 dB和10 dB条件下，IAEC-AD的平均AUC分别达到90.04%和93.76%，平均F1分数分别为87.87%和91.09%。结果表明，在信道噪声与对抗扰动共同存在的低信噪比监测场景中，所提方法仍具有较好的区分能力；随着信号质量提高，正常样本的层间类别关系更加稳定，检测性能进一步提升。

3.5　消融实验

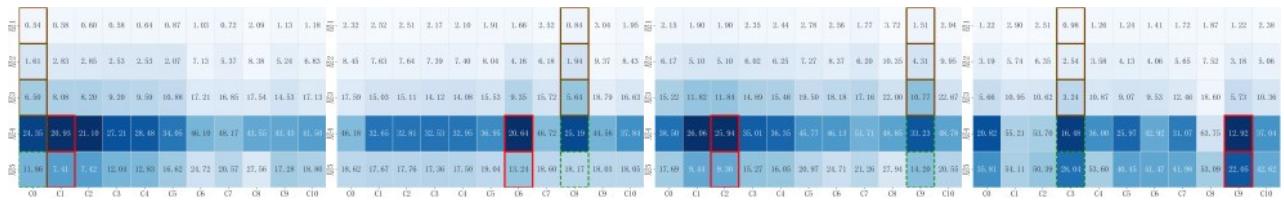
为系统验证所提方法关键设计的有效性，本节从两方面开展消融实验。首先是层选择策略，验证跨层激活演化相比单层静态特征的优势；其次是激活重构所采用的自编码器类型，验证WAE相比其他重构方式的必要性。

3.5.1　层选择策略消融

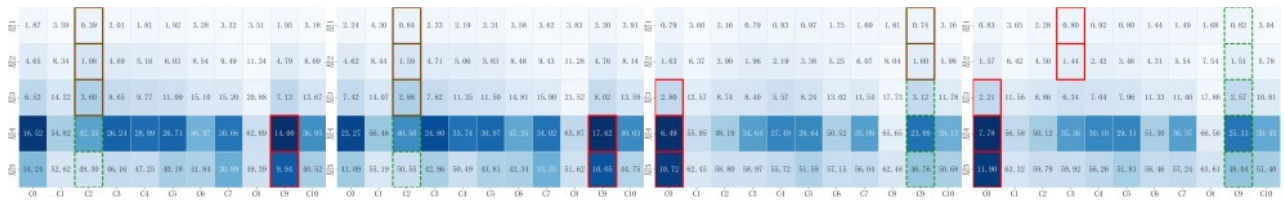
为验证层间激活演化分析的有效性，本节设计消融实验对比不同层选择策略的检测性能。如3.1节所述，本文在ResNet34的各残差块输出端提取激活，共计5个提取层。消融实验选取次深层（层4）和最深层（层5）作为单层策略的代表，层4和层5分别代表网络的次深层与最高语义层特征，有助于捕捉对抗扰动经跨层非线性放大后产生的显著结构偏移，是现有基于单层特征检测方法最常采用的观测节点。本节对比以下三种策略：（1）全层策略：使用全部5层；（2）层4策略：仅使用次深层；（3）层5策略：仅使用最深层。实验结果如表3



(a)　正常样本的层级类平均激活距离热力图



(b)　PGD对抗样本的层级类平均激活距离热力图



(c)　C&W对抗样本的层级类平均激活距离热力图

图7　正常样本与对抗样本的层级类平均激活距离热力图

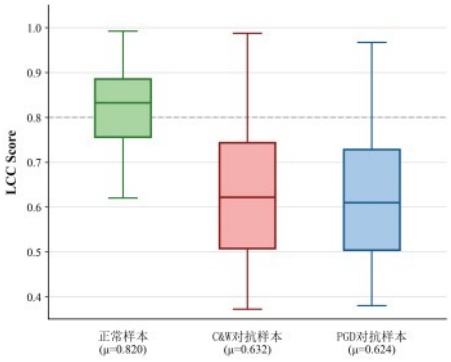


图8　正常样本与不同攻击下对抗样本的LCC分布箱线图

所示。

从表3可以看出，基于层间激活演化一致性的检测策略在两个数据集的低信噪比与高信噪比条件下均优于单层静态特征。在RML2016.10a数据集上，0dB条件下全层策略的平均AUC为95.20%，相比层4和层5分别提升4.28%和2.20%；10dB条件下全层策略的平均AUC为94.99%，相比层4和层5分别提升3.13%和3.37%。在RML22数据集上，0dB和10dB条件下全层策略的平均AUC分别为90.04%和93.76%，同样优于两种单层策略。全层策略联合利用浅层细节特征与深层类别特征，能够完整表征扰动的传播过程。DeepFool和C&W攻击下层5与全层策略差距相对较小，说明其可检测偏移主要体现在深层表征中；而FGSM和BIM攻击下全层策略优势更加明显，表明浅层与深层信息具有互补作用。

3.5.2　自编码器消融

为验证WAE对激活重构的必要性及其相比其他自编码器的优势，本节在保持层选择、PCA降维、类平均激活距离矩阵构建与SVM判别等其余环节完全一致的前提下，仅替换激活重构模块，对比五种重构方式的检测性能：（1）原始激活，不使用自编码器，直接在经同样PCA降维的原始激活上构建距离矩阵；（2）普通自编码器（AE），以重构误差为唯一目标训练的标准自编码器；（3）变分自编码器（VAE），引入KL散度约束潜在分布的自

表2　IAEC-AD方法在两个数据集上的详细检测性能

数据集	信噪比/dB	攻击方式	ACC	AUC	F1	Precision	Recall
RML2016.10a	0	FGSM	0.8607	0.9214	0.899	0.8703	0.9296
		BIM	0.874	0.9313	0.9084	0.8815	0.937
		PGD	0.8907	0.9624	0.9213	0.8856	0.96
		DeepFool	0.9018	0.9596	0.9261	0.929	0.9232
		C&W	0.956	0.9851	0.9667	0.9756	0.958
	10	FGSM	0.8636	0.9316	0.9	0.8806	0.9202
		BIM	0.8673	0.9391	0.9027	0.8827	0.9235
		PGD	0.8512	0.917	0.8896	0.8799	0.8994
		DeepFool	0.9237	0.9747	0.9434	0.9338	0.9532
		C&W	0.9506	0.9869	0.9625	0.9725	0.9528
RML22	0	FGSM	0.8319	0.8833	0.8773	0.8543	0.9017
		BIM	0.7813	0.8627	0.8422	0.8114	0.8755
		PGD	0.8329	0.8823	0.8784	0.8535	0.9049
		DeepFool	0.8467	0.9262	0.8862	0.8782	0.8942
		C&W	0.8786	0.9474	0.9094	0.9042	0.9147
	10	FGSM	0.8832	0.9365	0.9143	0.8957	0.9337
		BIM	0.8778	0.9331	0.91	0.8942	0.9263
		PGD	0.8602	0.909	0.8998	0.8616	0.9415
		DeepFool	0.8727	0.9478	0.9051	0.8994	0.9107
		C&W	0.9001	0.9614	0.9253	0.922	0.9287

编码器；（4）无 MMD 约束的 WAE，去除 MMD、仅保留重构损失；（5）完整 WAE，本文采用的以 MMD 约束聚合后验分布的自编码器。实验结果如表 4 所示。

从表 4 可以看出，完整 WAE 在两个数据集、不同信噪比及五种攻击下的平均 AUC 均为最高。以 RML2016.10a 数据集 0dB 条件为例，完整 WAE 的五种攻击平均 AUC 为 95.20%，高于原始激活、普通 AE、VAE 与无 MMD 约束的 WAE。不使用自编码器的原始激活策略在多数条件下表现最低，表明引入自编码器经非线性去噪与流形投影后，流形约束与重构过程能够提升类中心距离的稳定性和判别力，MMD 约束能够规范潜在空间分布，增强类中心结构的稳定性。完整 WAE 在不同数据集、信噪比与攻击类型下均保持最优且性能波动最小，表现出最强的稳定性，验证了本文采用 WAE 进行激活重构的合理性与有效性。

3.6　检测时间对比

为论证所提方法的部署可行性，本节先分析算法时间复杂度，再结合实测加以验证。设提取层数为 K、类别数为 C、各层激活经自适应平均池化后的特征维度为 d，自编码器潜在空间维度为 h，PCA 维度为 p、目标网络单次前向开销为 $\mathcal{O}(F)$。训练阶段需训练 K 个自编码器、逐层拟合 PCA 并计算各类平均激活，该开销在离线阶段一次性完成。推理阶段对单样本仅需一次目标网络前向 $\mathcal{O}(F)$、K 次自编码器重构 $\mathcal{O}(K \cdot dh)$ 与 PCA 投影 $\mathcal{O}(K \cdot dp)$ 及到 C 个类中心的距离计算 $\mathcal{O}(KCp)$ 及 SVM 判别 $\mathcal{O}(KC)$。

本节在统一环境下对比五种方法的单样本平均检测时间，硬件平台为配备 Intel Core i5-13490F 与 NVIDIA RTX5060 的计算节点，软件环境为基于 CUDA12.8 的 PyTorch2.7.1。检测时间为端到端单样本耗时，涵盖目标网络前向传播、WAE 重构、PCA 投影、距离矩阵计算与 SVM 判别全过程；测试 batch size 设为 1，计时前进行 50 次预热并在每次 GPU 同步后统计，于 1000 个测试样本上重复 10 次取均值，结果如表 5 所示。从表 5 可以看出，NILA 方法由于仅使用轻量级回归模型，推理速度最快，单样本平均耗时仅 2.32ms；AE-Layers 与本文 IAEC-AD 方法均涉及自编码器重构过程，时间开销相近，分别为 5.85ms 和 6.41ms，IAEC-AD 略慢，主要源于在重构空间中需额外计算各层到 11 个类别中心的距离；VMDFF 方法耗时 8.73ms，主要来源于 VMD 分解的迭代优化过程；SPBAS 方法耗时最长，达 15.64ms，其基于 KNN 分类器，需计算待测样本与训练集所有样本的距离。

综合检测性能与时间开销，本文 IAEC-AD 方法在保持所有对比方法中最优 AUC 的同时，单样本推理时间约为 6.4ms。具备接入在线频谱监测处理链路的近实时部署潜力，在检测性能与计算开销之间取得了较好平衡。

为进一步验证上述复杂度分析，本节将单样本推理耗时按处理阶段分解，各阶段独立计时，结果

表 3　　层选择策略消融实验 AUC 结果

数据集	信噪比/dB	层策略	FGSM	BIM	PGD	DeepFool	C&W
RML2016.10a	0	全层	0.9214	0.9313	0.9624	0.9596	0.9851
		层 4	0.9040	0.9014	0.9014	0.8983	0.9407
		层 5	0.8915	0.8978	0.9090	0.9654	0.9865
	10	全层	0.9316	0.9391	0.917	0.9747	0.9869
		层 4	0.9083	0.8752	0.881	0.9601	0.9682
		层 5	0.8897	0.8763	0.8645	0.9703	0.98
RML22	0	全层	0.8833	0.8627	0.8823	0.9262	0.9474
		层 4	0.844	0.8433	0.8078	0.8944	0.9254
		层 5	0.8794	0.8315	0.808	0.8802	0.9323
	10	全层	0.9365	0.9331	0.909	0.9478	0.9614
		层 4	0.9056	0.8991	0.9032	0.9354	0.9292
		层 5	0.9147	0.9025	0.9189	0.918	0.9447

表4 自编码器消融实验AUC结果

数据集	信噪比/dB	重构方式	FGSM	BIM	PGD	DeepFool	C&W
RML2016.10a	0	原始激活	0.8415	0.8421	0.8453	0.8328	0.8537
		AE	0.8607	0.9197	0.8977	0.8696	0.9051
		VAE	0.8966	0.9235	0.9256	0.9102	0.92
		WAE（无MMD）	0.9105	0.9212	0.9376	0.8909	0.9269
		WAE（本文）	0.9214	0.9313	0.9624	0.9596	0.9851
	10	原始激活	0.7768	0.8408	0.8357	0.8265	0.8112
		AE	0.9292	0.905	0.9051	0.9526	0.955
		VAE	0.9111	0.9061	0.9118	0.9187	0.9333
		WAE（无MMD）	0.9196	0.9127	0.9112	0.9175	0.9356
		WAE（本文）	0.9316	0.9391	0.917	0.9747	0.9869
RML22	0	原始激活	0.8185	0.8005	0.8787	0.8925	0.9393
		AE	0.832	0.8429	0.8617	0.9021	0.9388
		VAE	0.8139	0.8322	0.8699	0.8903	0.9385
		WAE（无MMD）	0.8371	0.856	0.8761	0.8953	0.9276
		WAE（本文）	0.8833	0.8627	0.8823	0.9262	0.9474
	10	原始激活	0.9065	0.911	0.8723	0.9382	0.9285
		AE	0.9322	0.9205	0.8843	0.9143	0.9484
		VAE	0.9323	0.9176	0.8913	0.9249	0.9559
		WAE（无MMD）	0.9121	0.918	0.8831	0.8662	0.8806
		WAE（本文）	0.9365	0.9331	0.909	0.9478	0.9614

表5 5种方法的单样本检测时间对比

方法名称	检测机制	单样本检测时间/ms
NILA	基于轻量级特征回归预测	2.32
AE-Layers	基于自编码器重构误差	5.85
IAEC-AD	基于层级类平均激活距离	6.41
VMDFF	基于信号去噪与特征融合	8.73
SPBAS	基于激活空间KNN分类	15.64

如表6所示。从表6可以看出，目标网络前向传播与WAE激活重构两项合计占比约91%，构成推理开销的主体；而PCA投影、类中心距离计算与SVM判别均为轻量操作，三者合计仅9%。这表明本文方法在原始分类网络前向开销之上引入的额外检测代价很小，且这些操作的规模仅由提取层数K、类别数C与PCA维度p决定；即便在持续运行、参考数据不断累积的频谱感知系统中，其推理开销也不会随之膨胀，因而具有良好的部署可行性。

表6 时间复杂度分解表

处理阶段	对应复杂度项	耗时/ms	占比/%
目标网络前向传播	$\mathcal{O}(F)$	2.05	32
WAE激活重构	$\mathcal{O}(K \cdot dh)$	3.8	59.3
PCA投影	$\mathcal{O}(K \cdot dp)$	0.3	4.7
类平均激活距离	$\mathcal{O}(KCp)$	0.16	2.5
标准化与SVM判别	$\mathcal{O}(KC)$	0.1	1.6

4 结束语

本文面向电磁频谱感知环节，针对基于深度学习的自动调制分类技术易受对抗样本攻击的问题，提出了一种基于层间激活演化一致性的对抗样本检测方法（IAEC-AD）。该方法通过提取目标网络多个关键层的激活表征，并利用自编码器学习激活流形，有效捕捉对抗扰动在网络前向传播中的动态累积效应；通过构建样本与各类平均激活的距离矩阵表征层间演化特性，克服了现有方法依赖单层静态特征的局限。实验结果表明，IAEC-AD在两个数

据集、五种典型攻击以及两种信噪比下的平均AUC均超过90%，单样本推理时间约为6.4ms，在检测性能与计算开销之间取得了较好平衡。未来工作将进一步探索基于阈值自适应机制的对抗样本检测方法。

参考文献：

[1] 阮天宸, 吴启晖, 赵世瑾, 等. 认知学习: 电磁频谱空间机器学习新范式[J]. 电子学报, 2023, 51(06): 1430-1442.
Ruan T C, Wu Q H, Zhao S J, et al. Cognitive learning: a new paradigm for machine learning in electromagnetic spectrum environment[J]. Acta Electronica Sinica, 2023, 51(06): 1430-1442.

[2] Lin Y, Wang M, Zhou X, et al. Dynamic spectrum interaction of UAV flight formation communication with priority: a deep reinforcement learning approach[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(3): 892-903.

[3] 杨健, 严牧, 赵研, 等. 群体频谱智能理论及架构研究[J]. 电子科技大学学报, 2025, 54(06): 850-865.
Yang J, Yan M, Zhao Y, et al. Research on theory and architectural framework of swarm spectrum intelligence[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(06): 850-865.

[4] 董超, 崔灿, 贾子晔, 等. 面向低空智联网的多维信息统一表征技术综述[J]. 电子与信息学报, 2025, 47(05): 1215-1229.
Dong C, Cui C, Jia Z Y, et al. Survey of unified representation technology of multi-dimensional information for low altitude intelligent network[J]. Journal of Electronics & Information Technology, 2025, 47 (05): 1215-1229.

[5] Lin Y, Tu Y, Dou Z, et al. Contour stella image and deep learning for signal recognition in the physical layer[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(1): 34-46.

[6] Xu Z W, Han G J, Zhu H B, et al. A lightweight specific emitter identification model for IIoT devices based on adaptive broad learning[J]. IEEE Transactions on Industrial Informatics, 2023, 19(5): 7066-7075.

[7] 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(01): 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(01): 1-13.

[8] Shafique M, Hanif M A, Guesmi A, et al. Adversarial and backdoor threats in autonomous-vehicle and embodied-AI systems[C]//Proceedings of the 2025 IEEE 38th International System-on-Chip Conference (SOCC). Dubai, United Arab Emirates: IEEE, 2025: 1-6.

[9] Zhao Y, Xu K, Wang H, et al. Adversarial attacks and defenses in deep learning: a review[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106-116.

[10] Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts[EB/OL]. (2017-03-01) [2025-12-01]. https://arxiv. org/abs/1703.00410.

[11] Ma X, Li B, Wang Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018.

[12] Lee K, Lee K, Lee H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[C]//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada: Curran Associates, Inc., 2018: 7167-7177.

[13] Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 5775-5783.

[14] Zhang S, Chen S, Hua C, et al. LSD: adversarial examples detection based on label sequences discrepancy[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5133-5147.

[15] Gao S, Wang R, Wang X, et al. Detecting adversarial examples on deep neural networks with mutual information neural estimation[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20 (6): 5168-5181.

[16] 徐东伟, 蒋斌, 陈嘉峻, 等. 基于特征融合的电磁信号对抗样本检测方法[J]. 电波科学学报, 2024, 39(05): 926-933.
Xu D W, Jiang B, Chen J J, et al. An electromagnetic signal adversarial sample detection method based on feature fusion[J]. Chinese Journal of Radio Science, 2024, 39(05): 926-933.

[17] Wang W, Zhu L, Gu Y, et al. Adversarial samples detection based on feature attribution and contrast in modulation recognition[J]. IEEE Communications Letters, 2024, 28(11): 2483-2487.

[18] Wójcik B, Morawiecki P, Śmieja M, et al. Adversarial examples detection and analysis with layer-wise autoencoders[C]//Proceedings of the IEEE International Conference on Tools with Artificial Intelligence. Washington, USA: IEEE, 2021: 1322-1326.

[19] Katzir Z, Elovici Y. Detecting adversarial perturbations through spatial behavior in activation spaces[C]//Proceedings of the International Joint Conference on Neural Networks. Budapest, Hungary: IEEE, 2019: 1-9.

[20] Mumcu F, Yilmaz Y. Universal and efficient detection of adversarial data through nonuniform impact on network layers[J/OL]. Transactions on Machine Learning Research, 2025[2025-12-01]. https://openreview.net/forum?id= 0CY5APFnFI.

[21] Zühlke M M, Kudenko D. Adversarial robustness of neural networks from the perspective of Lipschitz calculus: a survey[J]. ACM Computing Surveys, 2025, 57(6): 1-41.

[22] Fazlyab M, Robey A, Hassani H, et al. Efficient and accurate estimation of Lipschitz constants for deep neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019: 11423-11434.

[23] Lin Y, Zhao H, Ma X, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2020, 70(1): 389-401.

[24] O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognition networks[C]//Proceedings of the International Conference on Engineering Applications of Neural Networks. Aberdeen, UK: Springer, 2016: 213-226.

[25] Sathyanarayanan V, Gerstoft P, Gamal A E. RML22: realistic dataset generation for wireless modulation classification[J]. IEEE Transactions on Wireless Communications, 2023, 22(11): 7663-7675.

[26] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015.

[27] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[C]//Proceedings of the International Conference on Learning

Representations. Toulon, France, 2017.

[28] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018.

[29] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 2574-2582.

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

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。

王佳丽（2002– ），女，黑龙江哈尔滨人，哈尔滨工程大学硕士研究生，主要研究方向为电磁信号识别、对抗样本检测等。

魏晓磊（1978– ），男，江西高安人，中国运载火箭技术研究院工程师，主要研究方向为指挥自动化系统安全防护、军事计算机网络通信对抗。

张思成（1996– ），男，山东临沂人，博士，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。