# 基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法

李东阳[①②] 王林元*[①] 彭进先[②] 马德魁[①] 闫 镔[①]

[①](信息工程大学成像与智能处理实验室 郑州 450001)

[②](中国人民解放军63611部队 库尔勒 841000)

**摘 要：**随着深度神经网络在信号检测任务的应用，神经网络易受到对抗样本攻击的脆弱性也受到了广泛关注。针对无法获取模型内部信息的信号检测网络黑盒攻击场景，该文提出一种基于稀疏子空间采样的黑盒查询对抗攻击方法。该方法将信号样本检测数量消失比例作为判断攻击是否成功的约束条件，构造信号检测网络对抗样本攻击模型，参考跳步跳跃攻击(HSJA)算法设计基于决策边界的信号检测网络黑盒查询对抗攻击方法求解该模型，以生成信号对抗样本。为了进一步改善查询效率，该文根据信号对抗扰动特点构建稀疏子空间采样进行查询攻击，即在生成对抗样本时，按照一定比例选择具有较大幅度的信号分量，并仅在这些选定的分量上添加扰动。实验结果表明，在信号目标消失数量比例0.3的决策边界下，稀疏子空间采样黑盒对抗攻击方法使得信号检测网络mAP值降低了43.6%、召回率降低了41.2%。与全空间采样方法相比，稀疏子空间采样方法攻击成功率提升了2.5%，且对抗扰动平均能量比降低了3.47%。稀疏子空间采样攻击方法可以使得信号检测网络性能明显下降，相较于全空间采样具有攻击成功率更高、扰动强度更小等优势。

**关键词：**信号检测网络；信号对抗样本；黑盒查询攻击；稀疏子空间采样

**中图分类号：**TN918; TP391 **文献标识码：**A **文章编号：**1009-5896(2025)08-2808-11

**DOI**: 10.11999/JEIT241019 **CSTR**: 32379.14.JEIT241019

## 1 引言

深度神经网络(Deep Neural Networks, DNN)在多个领域飞速发展，在计算机视觉领域的目标检测任务中更是取得了巨大的成功[1–3]。通过将目标检测网络运用于信号处理领域，发展出的信号检测网络相比于传统信号检测方法具有抗噪性强、检测效率高等特点[4–8]。随着DNN在信号检测任务中广泛应用，其自身的脆弱性，即易受到对抗样本攻击的特点[9–12]，使得信号检测网络同样面临严重的安全威胁。对抗样本攻击比传统的利用随机噪声干扰的攻击效果更为显著，且所需的扰动功率更小。近年来，对抗样本攻击已成为无线信号安全领域的一个重要研究方向[13–16]。

信号检测是信号处理领域的重要研究内容，传统的方法有匹配滤波法[17,18]、循环平稳法[19]、特征值检测法[20]和能量检测法[21]等，传统方法重点关注于信号是否存在，而对信号参数的估计研究较少。基于深度学习的信号检测研究通常是将信号数据经过傅里叶变换为时频图，并利用图像目标检测网络完成检测和识别任务，得到信号目标的存在与否、出现时间、中心频率和带宽等信息[4–8]。当前针对信号检测网络的对抗样本研究较少，Li等人[16]定义了信号检测网络信号端对抗样本生成模型，提出了白盒场景下的信号对抗样本生成方法，该模型利用目标检测网络的梯度攻击方法对时频图开展攻击，通过逆傅里叶变换将时频图对抗扰动恢复到信号域生成对抗样本。上述对抗攻击方法在白盒场景下可以有效地实施攻击，但在仅可获取信号检测网络的信号输入和检测结果的黑盒场景下，由于网络的具体结构与内部参数无法获取，难以对处于中间过程的时频图求解对抗扰动，因此无法生成对抗样本。

黑盒对抗攻击方法的研究主要集中在图像领域，常见的黑盒攻击方法可以分为基于查询的攻击方法和基于迁移的攻击方法。基于迁移的攻击方法需要通过训练替代模型来实施攻击[22–24]，而基于查询的攻击方法则通过向目标模型发送特定的查询样本，观察模型的输出，并分析不同查询样本与输出之间的关联关系，从而生成有效的对抗样本。基于查询的攻击方法进一步分为基于分数的攻击方法[25–27]和基于决策边界的攻击方法[28–31]。基于分数的攻击方法以输出的类别分数作为查询对象进行攻击，而基于决策边界的攻击方法则仅需访问模型的最终输出类别来实施攻击。常见的基于决策边界的黑盒攻击方法有基于决策边界的攻击方法(Boundary Attack)[28]、符号优化攻击(Sign-opt)[29]、跳步跳跃攻击(HopSkipJumpAttack, HSJA)[30]、基于决策边界

收稿日期：2024-11-14；改回日期：2025-06-28；网络出版：2025-07-04
*通信作者： 王林元 wanglinyuanwly@163.com
基金项目：国家自然科学基金(62271504)
Foundation Item: The National Natural Science Foundation of China (62271504)

的高效查询攻击(Query-Efficient Boundary-based blackbox Attack, QEBA)[31]等。Boundary Attack是该方向的首篇研究，该方法主要针对L2约束下的目标攻击，算法主要过程为：利用二分搜索将样本投影至决策边界，利用随机扰动查询估计梯度方向，选择合适的步长至扰动成功，再通过二分搜索将扰动样本投影回边界。随后的决策边界对抗攻击方法研究主要集中在提高查询估计梯度的效率，Cheng等人[29]进行了梯度符号估计研究，Chen等人[30]应用蒙特卡罗采样方法进行梯度估计，Li等人[31]通过从图像特征的具有代表性的子空间(包括线性插值、低频空间和主成分)中采样来改进梯度估计。

受图像领域查询对抗攻击相关研究的启发，本文提出一种基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法。该方法将信号样本检测目标的消失比例作为判断攻击是否成功的约束条件，以缩小向信号样本添加的扰动能量作为优化目标，考虑信号对抗扰动倾向于添加在幅度较大信号分量上的特点，构建信号扰动的稀疏采样子空间，通过在信号扰动的稀疏采样子空间进行采样来向信号添加扰动，根据输出的信号检测结果来估计梯度方向，经多轮迭代生成对抗样本。

## 2　基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法

本节详细介绍基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法。首先结合信号检测网络对抗攻击目标提出信号检测网络对抗攻击模型，随后参考HSJA算法，设定决策边界并刻画待优化求解问题，提出基于决策边界的信号检测网络黑盒查询对抗攻击方法。更进一步，为了改善施加扰动的有效性，依据原始信号分量幅度值与对抗扰动大小的相关性构建扰动的稀疏采样子空间，提出基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法。

### 2.1　信号检测网络对抗攻击模型

本节主要介绍信号检测网络对抗攻击模型，首先介绍信号检测网络的结构和检测结果，随后结合实际场景给出信号检测网络对抗攻击模型。

#### 2.1.1　信号检测网络结构及检测结果

信号检测网络的结构如图1所示，由信号时频图转换模块和目标检测网络级联而成。时频图转换模块以信号数据为输入，将其转换为包含幅度和相位信息的复数矩阵，随后将幅度信息映射为信号时频图。图2(a)所示为信号时频图，图像的像素值对应信号在对应时间和频率处的功率谱幅度，白色长条为信号目标。

目标检测网络级联在时频图转换模块之后，其输入为信号数据转换后的信号时频图，输出为检测到的信号类别和位置信息，信号时频图检测结果如图2(b)所示。图中信号目标的标签和数字分别为信号类别和置信度，检测框位置信息的横坐标对应时频图中信号目标的时间信息，纵坐标对应信号目标的频率信息。

由此级联结构搭建的信号检测网络，可以直接读取信号数据，并输出检测到的信号类别、频率和时间信息，实现信号端到端的检测。信号检测网络的数学表达如下：设信号检测网络模型为$C(\cdot)$，信号序列$\boldsymbol{x}$为输入，检测出的信号目标的数量为$c$，即$c=C(\boldsymbol{x})$。

#### 2.1.2　信号检测网络对抗攻击模型

信号检测作为信号处理任务的重要组成部分，主要目标是通过分析信号数据判断信号目标的存在性。在传统的信号检测任务中，通常采用真正例(True Positive, TP)、假正例(False Positive, FP)以及假反例(False Negative, FN)来评价信号检测方法的效果：TP是指被正确检测到的信号目标数量；FP代表被错误检测为信号目标的数量；而FN是指实际存在但未被检测到的信号目标数量。评价指标则通常使用命中率$R_{\mathrm{H}}$ (Hit Rate)和虚警率$R_{\mathrm{F}}$ (Alarm Rate)

$$\begin{aligned} R_{\mathrm{H}} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ R_{\mathrm{F}} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \end{aligned} \tag{1}$$

而在图像目标检测任务中，通常使用精确率(Precision, $P$)和召回率(Recall, $R$)来评价目标检测网络性能

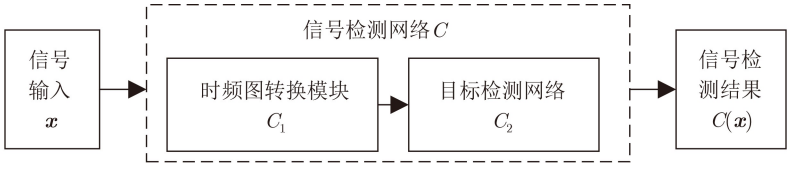



图 1 信号检测网络结构图

(a) 信号时频图 (b) 检测结果

图 2 信号时频图及检测结果

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$
$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

从检测的评价指标来看，目标检测所采用的精确率$P$和召回率$R$与信号检测任务中的命中率$R_{\mathrm{H}}$和虚警率$R_{\mathrm{F}}$具有对应关系。当使用目标检测网络处理信号检测任务时，可以考虑使用目标检测网络的评价指标来评价信号检测的检测效果。

需要注意的是，在实际的信号检测任务中，低命中率所带的危害要远远大于高虚警率。高虚警率下检测到的异常信号目标较易被察觉，通常也不易对正常信号目标产生有效干扰；低命中率的表现则是部分信号目标漏检，相对而言不易被察觉，且会影响到后续的信号处理。因此本文将信号检测网络对抗攻击的攻击效果设定为使得部分信号目标检测消失。

此外，在前期关于信号调制识别网络对抗样本的研究中，Sadeghi等人[13]证明了信号调制识别网络面对信号对抗样本的脆弱性，并提出了信号调制识别对抗样本的定义如式(3)所示

$$\arg\min_{\boldsymbol{r}_{\boldsymbol{x}}}\|\boldsymbol{r}_{\boldsymbol{x}}\|_2,\ \text{s.t.}\ \hat{l}(\boldsymbol{x},\theta)\neq l(\boldsymbol{x}+\boldsymbol{r}_{\boldsymbol{x}},\theta),\ \boldsymbol{x}+\boldsymbol{r}_{\boldsymbol{x}}\in\mathcal{X} \tag{3}$$

其中，$\boldsymbol{r}_{\boldsymbol{x}}$为对抗扰动，$\hat{l}(\boldsymbol{x},\theta)$为信号样本$\boldsymbol{x}$的识别标签，该定义中将调制识别网络识别错误作为约束，以扰动的二范数表示扰动的能量，并将其作为优化目标。

基于信号检测网络的对抗攻击效果并参考信号调制识别网络的对抗扰动大小的度量方法，本文将针对信号检测网络的攻击目标定义为：在满足特定比例信号目标消失的约束条件下，最小化信号扰动的L2范数。由此构建的信号检测网络对抗攻击模型如式(4)所示

$$\min||\boldsymbol{\delta}_r||_2,\ \text{s.t.}\ c'=C(\boldsymbol{x}')\leq C(\boldsymbol{x})\times(1-\alpha),\ \boldsymbol{\delta}_r=\boldsymbol{x}'-\boldsymbol{x} \tag{4}$$

其中，$\boldsymbol{x}$为原始信号样本，$\boldsymbol{x}'$为信号对抗样本，$\boldsymbol{\delta}_r$为向信号样本$\boldsymbol{x}$添加的扰动，$C(\cdot)$为信号检测网络，$C(\boldsymbol{x})$输出为信号样本$\boldsymbol{x}$检测的信号数量，$\alpha$为信号添加对抗扰动$\boldsymbol{\delta}_r$攻击后消失的信号检测数量$C(\boldsymbol{x})-C(\boldsymbol{x}')$与原始信号样本检测数量$C(\boldsymbol{x})$比例的阈值，此时剩余的信号目标数量比例则应不超过$1-\alpha$。该模型以使得一定比例的信号目标数量消失作为攻击的约束条件，以最小化信号扰动的能量作为优化目标，求解对抗扰动$\boldsymbol{\delta}_r$。

### 2.2 基于决策边界的信号检测网络黑盒查询对抗攻击方法：

2.1节对信号检测网络结构、检测结果和攻击目标进行了分析，并提出了信号检测网络对抗攻击模型。在黑盒攻击场景下，无法直接访问信号检测网络的内部信息，仅能根据信号检测网络的信号输入和检测结果来求解对抗样本，本文基于HSJA算法[30]设计基于决策边界的信号检测网络黑盒查询对抗攻击方法：根据模型式(4)定义中的约束条件，将信号对抗扰动攻击后的信号检测数量消失比例作为决策边界，通过多轮查询降低信号扰动强度，生成信号对抗样本。

#### 2.2.1 信号检测网络对抗攻击决策边界及优化问题：

在识别任务的对抗攻击研究中，通常以类别的概率作为决策边界的基准，而在信号检测对抗攻击中，我们将对抗攻击后的信号检测数量消失比例作为决策边界，具体如下：

信号检测网络模型为$C(\cdot)$，$\boldsymbol{x}$为输入的信号样本，设$c$为检测出的信号目标的个数 $c=C(\boldsymbol{x})$，以叠加扰动的信号检测数量相比原始信号检测数量的消失比例阈值$\alpha$作为决策边界条件，即给定输入$\boldsymbol{x}'$满足

$$c'=C(\boldsymbol{x}')\leq C(\boldsymbol{x})\times(1-\alpha) \tag{5}$$

由此定义决策函数$S_{\boldsymbol{x}}(\boldsymbol{x}')$为

$$S_{\boldsymbol{x}}(\boldsymbol{x}'):=C(\boldsymbol{x})\times(1-\alpha)-C(\boldsymbol{x}') \tag{6}$$

其中$\alpha$为叠加扰动的信号样本消失的检测目标数量与原始信号样本的检测目标数量比例所能接受的最低值，当$S_{\boldsymbol{x}}(\boldsymbol{x}')\geq 0$时，认为攻击成功。在该决策函数下，攻击成功和不成功的决策边界为

$$\mathrm{bd}(S_{\boldsymbol{x}}):=\mathrm{bd}\left\{\boldsymbol{x}'\,|S_{\boldsymbol{x}}(\boldsymbol{x}')=0\right\} \tag{7}$$

该决策边界的含义为生成的信号对抗样本的检测数量$c'$与原始干净信号样本检测数量$c$的比例等于$1-\alpha$。

在完成信号检测网络对抗攻击决策边界定义后，引入符号函数$\phi_{\boldsymbol{x}}(\cdot)$来表示攻击的成功，即为

$$\phi_{\boldsymbol{x}}(\boldsymbol{x}') := \mathrm{sign}(S_{\boldsymbol{x}}(\boldsymbol{x}')) = \begin{cases} 1, & S_{\boldsymbol{x}}(\boldsymbol{x}') \geq 0 \\ -1, & \text{其他} \end{cases} \tag{8}$$

此函数在查询过程中可以直接进行访问，只需获得叠加扰动的信号检测结果即可得到$\phi_{\boldsymbol{x}}(\boldsymbol{x}')$的结果。

在完成决策边界及攻击成功的定义后，本文设置优化目标如下：设信号对抗样本为$\boldsymbol{x}'$，当$\phi_{\boldsymbol{x}}(\boldsymbol{x}')=1$时，使得信号对抗样本$\boldsymbol{x}'$与信号干净样本$x$的距离最小，公式表达为

$$\underset{\boldsymbol{x}'}{\mathrm{argmin}}\, d(\boldsymbol{x},\boldsymbol{x}'),\ \text{s.t.}\ \phi_{\boldsymbol{x}}(\boldsymbol{x}')=1 \tag{9}$$

其中$d$为对抗样本$\boldsymbol{x}'$与干净样本$\boldsymbol{x}$距离的度量函数。

信号样本通常使用能量来衡量扰动的大小，参考信号调制识别网络的对抗扰动大小的度量方法，本文使用2范数来表示扰动的大小，上述问题可具体表述为

$$\underset{\boldsymbol{x}'}{\mathrm{argmin}} \|\boldsymbol{x}-\boldsymbol{x}'\|_2^2,\ \text{s.t.}\ \phi_{\boldsymbol{x}}(\boldsymbol{x}')=1 \tag{10}$$

#### 2.2.2 信号检测网络黑盒查询对抗攻击方法流程：

信号检测网络黑盒查询对抗攻击方法包括样本初始化、梯度方向估计、步长搜索、信号样本更新等操作，经多轮迭代生成信号对抗样本。下面为该方法的具体描述：

(1)信号样本初始化至决策边界附近且攻击成功：对信号样本$\boldsymbol{x}$叠加随机噪声$\boldsymbol{\delta}$进行初始化操作，得到决策边界附近的信号样本$\boldsymbol{x}_0$，该样本$\boldsymbol{x}_0$满足$\phi_{\boldsymbol{x}}(\boldsymbol{x}_0)=1$。具体过程如下：

引入随机噪声$\boldsymbol{\delta}$，使得其满足攻击成功的条件，即$S_{\boldsymbol{x}}(\boldsymbol{\delta})\geq 0$，随后将$\boldsymbol{\delta}$与原始信号样本$\boldsymbol{x}$进行叠加操作：$\boldsymbol{x}_0=\beta\boldsymbol{x}+(1-\beta)\boldsymbol{\delta}$。式中$\beta$为叠加比例，初始值设置为0.5，在[0,1]的区间内二分搜索，通过判断叠加后得到的$\hat{\boldsymbol{x}}_0=\hat{\beta}\boldsymbol{x}+(1-\beta)\boldsymbol{\delta}$是否攻击成功，对搜索区间进行二分操作。若满足$\phi_{\boldsymbol{x}}(\hat{\boldsymbol{x}}_0)=1$，则将$\hat{\beta}$的搜索区间取小的1/2，若$\phi_{\boldsymbol{x}}(\hat{\boldsymbol{x}}_0)=0$，则$\hat{\beta}$的搜索区间取大的1/2，直到搜索区间小于等于0.001且$\phi_{\boldsymbol{x}}(\hat{\boldsymbol{x}}_0)=1$，停止搜索，得到的$\hat{\boldsymbol{x}}_0$即为初始化结果$\boldsymbol{x}_0$。经过初始化所得到的信号样本$\boldsymbol{x}_0$在决策边界$S_{\boldsymbol{x}}$附近且在$\phi_{\boldsymbol{x}}(\boldsymbol{x}_0)=1$的一侧。

(2)估计梯度方向：设$\boldsymbol{x}_t$为第$t$轮迭代的信号样本，在求解信号检测网络对抗样本的过程中，由于无法直接访问梯度$\nabla S_{\boldsymbol{x}}(\boldsymbol{x}_t)$，只能访问符号函数$\phi_{\boldsymbol{x}}(\boldsymbol{x}_t)$，因此我们参考HSJA方法[30]，在决策边界$\mathrm{bd}(S_{\boldsymbol{x}}):=\mathrm{bd}\{\boldsymbol{x}'|S_{\boldsymbol{x}}(\boldsymbol{x}')=0\}$附近使用蒙特卡罗及方差下降的方法引入替代梯度

$$\hat{\nabla} S_{\boldsymbol{x}}(\boldsymbol{x}_t,\varepsilon) := \frac{1}{B-1}\sum_{b=1}^{B}(\phi_{\boldsymbol{x}}(\boldsymbol{x}_t+\varepsilon\boldsymbol{u}_b)-\bar{\phi}_{\boldsymbol{x}})\boldsymbol{u}_b \tag{11}$$

其中，$\boldsymbol{u}_b$为1维的相互独立的均匀分布张量，$\varepsilon$为一个正的小值。$\hat{\nabla}S_{\boldsymbol{x}}(\boldsymbol{x}_t,\varepsilon)$在决策边界附近且$\boldsymbol{\delta}\to 0^+$时是渐进无偏的，并通过与查询得到的符号函数平均值$\bar{\phi}_{\boldsymbol{x}}$相减使得方差降低，从而提高替代梯度$\hat{\nabla}S_{\boldsymbol{x}}(\boldsymbol{x}_t,\varepsilon)$的估计精度。

(3)二分搜索步长至攻击成功，并利用步长与梯度方向更新信号样本：在第$t$轮迭代时，从$\boldsymbol{x}_t$开始运算，此时$\boldsymbol{x}_t$满足$\phi_{\boldsymbol{x}}(\boldsymbol{x}_t)=1$，采用的搜索步长尺寸初始化为$\xi_t:=\|\boldsymbol{x}_t-\boldsymbol{x}\|_2/\sqrt{t}$。将步长与梯度方向相乘后与信号样本相叠加得到更新后的信号样本$\tilde{\boldsymbol{x}}_{t+1}$

$$\tilde{\boldsymbol{x}}_{t+1}=\boldsymbol{x}_t+\xi_t\frac{\hat{\nabla}S(\boldsymbol{x}_t,\boldsymbol{\delta})}{\left\|\hat{\nabla}S(\boldsymbol{x}_t,\boldsymbol{\delta})\right\|_2} \tag{12}$$

如果满足$\phi_{\boldsymbol{x}}(\tilde{\boldsymbol{x}}_{t+1})=-1$，将搜索步长缩减为原始的1/2，直至$\phi_{\boldsymbol{x}}(\tilde{\boldsymbol{x}}_{t+1})=1$，取当前步长，并使用式(12)对信号样本进行更新。

(4)将更新后的信号样本投影到决策边界附近且保持攻击成功：更新后的对抗样本$\tilde{\boldsymbol{x}}_{t+1}$需投影到决策边界附近，以缩小与原始信号样本$\boldsymbol{x}$的距离，并提升下一轮迭代过程中梯度估计的准确性。具体的投影方法是将$\tilde{\boldsymbol{x}}_{t+1}$与$\boldsymbol{x}$按照比例$\beta_t$相叠加：$\boldsymbol{x}_{t+1}=\beta_t\boldsymbol{x}+(1-\beta_t)\tilde{\boldsymbol{x}}_{t+1}$。式中$\beta_t$通过二分搜索获得，具体过程如下：将$\beta_t$初始值设置为0.5，并在[0,1]内进行二分搜索，搜索方法与步骤2相同。当相邻两次$\beta_t$的搜索值相差小于预设阈值$\theta=0.001$且$\phi_{\boldsymbol{x}}(\boldsymbol{x}_0)=1$时，完成搜索，随后利用$\beta_t$投影得到信号对抗样本$\boldsymbol{x}_{t+1}$，该样本$\boldsymbol{x}_{t+1}$在决策边界$S_{\boldsymbol{x}}$附近，且满足$\phi_{\boldsymbol{x}}(\boldsymbol{x}_{t+1})=1$。

(5)重复步骤(2)，(3)和(4)，经多轮迭代后生成扰动能量较低的信号对抗样本。

方法流程图如图3所示。

### 2.3 基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法

2.2节根据信号检测网络对抗攻击模型，参考HSJA对抗攻击算法，设定查询决策边界，提出了基于决策边界的信号检测网络黑盒查询对抗攻击方法。通过该方法，可以利用多轮查询迭代有效生成信号对抗样本。然而，在梯度查询的过程中，扰动采样空间的选择对梯度查询的有效性至关重要。在图像查询攻击方法的研究中，为了提高图像分类任务黑盒查询攻击梯度估计的有效性，Li等人[31]根据图像的特点，分别从空间、频域和主成分3个角度选取不同的梯度估计采样子空间，提高梯度估计的有效性，并取得了比全空间采样更好的攻击效果。

对信号对抗攻击而言，信号的分布与图像并不

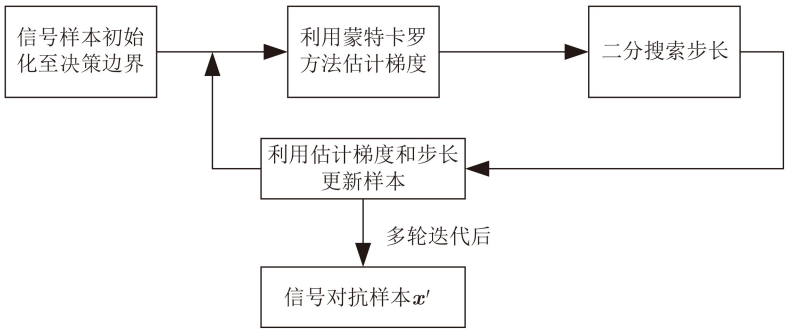



图 3 信号检测网络黑盒对抗攻击方法示意图

相同，无法直接借鉴图像的采样子空间来提高信号对抗攻击的梯度估计有效性。受文献[32–35]启发，在针对信号调制识别的对抗攻击研究中，郭宇琦等人[36,37]分别选取信号样本中数值最大和最小的$t$个点，利用压缩感知技术，通过多轮查询样本识别准确率来重构准确率的梯度信息，生成信号对抗样本。受此启发，本文对信号检测网络的白盒对抗攻击结果中对抗扰动大小与信号幅度对应比例进行了统计，得到数据如表1所示。

表1统计了快速梯度法(Fast Gradient Method, FGM)和投影梯度下降法(Projected Gradient Descent, PGD)两种对抗攻击方法生成的对抗扰动幅度与信号幅度对应位置比例。参数$p$表示信号样本或者对抗扰动中幅度最大的前$pN$个点的选取比例($N$为信号总点数)。对于每个$p$值，本文计算了对抗扰动幅度最大的前$pN$个点与信号幅度最大的前$pN$个点的位置交集，将交集数量除以选取总点数$pN$，得到重叠比例，表中Max,Min和Avg分别表示该比例的最大值、最小值和平均值。最后一项为Avg与$p$的比值，若其显著大于1，说明扰动集中于信号高幅度区域；若接近1，则二者无明显关联。

由表1中可以看出，信号对抗扰动较大的分量与信号幅度较大的分量具有较强的对应关系，而对抗扰动是梯度方向的单步或多轮迭代所得到的，因此在信号对抗攻击中，信号扰动幅度较高的分量具有一定的梯度方向倾向性，即信号对抗扰动的梯度方向更集中于幅度较大的信号分量。也就是说，通过对较大幅度分量进行查询并估计梯度，可能会更加有效。

**表 1 不同攻击方法下对抗扰动幅度与信号幅度对应位置比例**

| 对抗攻击方法 | $p$ | Max | Min | Avg | Avg/$p$ |
|---|---|---|---|---|---|
| FGM | 0.02 | 0.50 | 0.06 | 0.35 | 17.50 |
| | 0.04 | 0.55 | 0.20 | 0.43 | 10.75 |
| | 0.10 | 0.62 | 0.40 | 0.54 | 5.40 |
| PGD | 0.02 | 0.49 | 0.07 | 0.35 | 17.50 |
| | 0.04 | 0.54 | 0.24 | 0.43 | 10.75 |
| | 0.10 | 0.61 | 0.39 | 0.53 | 5.30 |

根据上述规律，信号检测网络对抗攻击方法在通过查询求解近似梯度的过程中，可以将全空间采样改为按照一定比例$p$选取较大幅度值的信号分量进行采样，在样本更新时，仅对选定的稀疏分量进行更新，从而达到提升梯度查询有效性的目的。基于稀疏子空间采样查询的梯度方向估计算法和基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法如算法1和算法2所示，通过上述算法生成稀疏子空间采样条件下的信号检测对抗样本。

## 3 实验结果

为了验证本文信号对抗攻击方法的有效性，本章开展对抗攻击实验并分析实验结果。首先使用原始信号样本对信号检测网络进行训练，并统计信号

**算法 1 基于稀疏子空间采样查询的梯度方向估计**

输入：信号长度$n$，信号采样分量比例$p$，决策边界附近的信号数据$\boldsymbol{x}_0\in\mathbb{R}^{1\times n}$，信号采样稀疏子空间$\boldsymbol{W}\in\mathbb{R}^{1\times np}$，稀疏采样的数量$B$，符号函数$\phi_{\boldsymbol{x}}(\cdot)$

输出：估计梯度$\hat{\nabla}S_{\boldsymbol{x}}$

(1) 从信号采样稀疏子空间$\boldsymbol{W}$中采样$B$个随机均匀分布噪声，组成向量组$\boldsymbol{V}\in\mathbb{R}^{B\times np}$。

(2) 在信号幅度较小的$1-p$个位置上补0，将向量组$\boldsymbol{V}$扩充为稀疏采样扰动向量组$\boldsymbol{U}\in\mathbb{R}^{B\times n}$。

(3) 将稀疏采样扰动向量$\boldsymbol{u}_b{}^{1\times n}\in\boldsymbol{U}$与信号数据$\boldsymbol{x}_0$相叠加，利用蒙特卡罗方法估计梯度：

$\hat{\nabla}S_{\boldsymbol{x}}(\boldsymbol{x}_t,\varepsilon):=\dfrac{1}{B-1}\sum_{b=1}^{B}(\phi_{\boldsymbol{x}}(\boldsymbol{x}_t+\varepsilon\boldsymbol{u}_b)-\bar{\phi}_{\boldsymbol{x}})\boldsymbol{u}_b$。

(4) 返回估计梯度$\hat{\nabla}S_{\boldsymbol{x}}$。

检测网络的指标。随后，选择信号目标数量消失比例0.3作为决策边界，分别统计在全空间和稀疏子空间采样所得到的信号对抗样本攻击结果。最后，为直观展示信号对抗样本的时域特征，绘制信号对抗样本与干净样本波形图进行对比分析。

### 3.1 实验设置

采用的信号宽带检测数据集为6 400 kHz中频采样文件，每个文件采样时长为200 ms，并含有多个载波。数据集以16位整型的2进制的形式存储，单个信号文件共包含16×1 280 000个二进制数据，经过转换后为1 280 000个16位整型数据。

信号检测网络模型由时频图转换模块和YOLOv5目标检测网络级联组成。时频图转换模块在处理信号数据时，单一信号文件从16位整型2进制数据读取为整型数据，随后将读取后的信号数据进行归一化处理，设定快速傅里叶变换点数(Number of points for Fast Fourier Transform, NFFT)为2 048，采样率(sampling Frequency, Fs)为6 400 000，重叠点数为48，采用Blackman窗函数来计算时频图，设定生成的时频图为灰度图，灰度值范围设定为0～255。数据集中的标签原本为信号的频率信息、时间信息和类别信息，将标签的时频信息转换为信号目标在时频图中的坐标信息。随后我们利用生成的时频图和标签构建信号时频图数据集，并对YOLOv5网络进行训练，经训练后的YOLOv5检测网络在信号检测数据集上的mAP可以达到0.872，召回率为0.813，精确率为0.953。

在对抗攻击实验中，本文以使得一定比例$\alpha$的信号目标消失作为决策边界，设单个干净样本的召回率$R_{\mathrm{beg}}$、单个对抗样本的召回率为$R_{\mathrm{adv}}$，在此我们给出单个信号样本攻击成功定义如式(13)

$$S\left(R_{\mathrm{adv}}\right):=\begin{cases}1, & \alpha \le 1-R_{\mathrm{adv}}/R_{\mathrm{beg}}\\ 0, & \alpha > 1-R_{\mathrm{adv}}/R_{\mathrm{beg}}\end{cases} \tag{13}$$

用1减去单个样本降低的召回率与干净样本召回率的比值，当大于等于预设的比例$\alpha$时，则认为对该样本攻击成功。

### 3.2 对抗样本攻击结果

在对抗攻击实验中，选择信号目标消失比例0.3作为决策边界，分别在全空间和稀疏子空间下进行扰动采样，其中稀疏子空间稀疏度为0.1，即对抗扰动仅添加在幅值最大的10%的分量上，共迭代21轮，每轮梯度查询次数设置为100次，梯度查询总计2 100次。在生成信号对抗样本后，根据对抗扰动的大小，向干净信号样本添加同等2范数强度的噪声扰动，生成随机噪声干扰样本。分别统计4类信号样本的mAP值、精确率、召回率和攻击成功率如表2所示，各列对比结果中最优的数值通过加粗标识。

**算法 2　基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法**

输入：信号检测网络$C(\cdot)$，信号样本$\boldsymbol{x}$，迭代轮数$T$。

输出：信号对抗样本$\boldsymbol{x}'$。

(1) 定义决策函数$S_{\boldsymbol{x}}(\boldsymbol{x}'):=C(\boldsymbol{x})\times\alpha-C(\boldsymbol{x}')$、决策边界$\mathrm{bd}(S_{\boldsymbol{x}}):=\mathrm{bd}\left\{\boldsymbol{x}'\,|S_{\boldsymbol{x}}(\boldsymbol{x}')=0\right\}$和查询符号函数

$\phi_{\boldsymbol{x}}(\boldsymbol{x}'):=\mathrm{sign}(S_{\boldsymbol{x}}(\boldsymbol{x}'))=\begin{cases}1, & S_{\boldsymbol{x}}(\boldsymbol{x}')\ge 0\\ -1, & \text{其他}\end{cases}$。

(2) 对信号样本$\boldsymbol{x}$进行初始化操作，得到信号样本$\boldsymbol{x}_0$在决策边界$\boldsymbol{S}_x$附近且满足$\phi_{\boldsymbol{x}}(\boldsymbol{x}_0)=1$。

(3) for $t$ 从 0～$T-1$:

使用算法1，估计梯度方向$\hat{\nabla}S_{\boldsymbol{x}}(\boldsymbol{x}_t)$；

搜索步长$\xi_t$，并更新信号样本$\tilde{\boldsymbol{x}}_{t+1}=\boldsymbol{x}_t+\xi_t\dfrac{\hat{\nabla}S(\boldsymbol{x}_t,\boldsymbol{\delta})}{\left\|\hat{\nabla}S(\boldsymbol{x}_t,\boldsymbol{\delta})\right\|_2}$；

搜索投影比例$\beta_t$，将$\tilde{\boldsymbol{x}}_{t+1}$投影到决策边界得到$\boldsymbol{x}_{t+1}=\beta_t\boldsymbol{x}+(1-\beta_t)\tilde{\boldsymbol{x}}_{t+1}$，$\boldsymbol{x}_{t+1}$满足$\phi_{\boldsymbol{x}}(\boldsymbol{x}_{t+1})=1$；

end for

(4) 返回信号对抗样本$\boldsymbol{x}'=\boldsymbol{x}_T$。

**表 2　0.3决策边界下不同样本攻击结果**

| 样本类别 | mAP | $P$ | $R$ | 攻击成功率(%) |
|---|---|---|---|---|
| 干净样本 | 0.872 | 0.953 | 0.813 | / |
| 全空间采样对抗样本 | 0.519 | 0.708 | 0.508 | 48.0 |
| 全空间采样下同等强度随机噪声干扰样本 | 0.549 | 0.736 | 0.539 | 40.8 |
| 稀疏子空间采样对抗样本 | **0.492** | **0.664** | **0.478** | **49.2** |
| 稀疏子空间采样下同等强度随机噪声干扰样本 | 0.555 | 0.73 | 0.55 | 36.6 |

由表2可以看出：同干净样本相比，两种查询空间生成的对抗样本，均可以使得信号检测网络的mAP值、精确率和召回率得到显著下降，其中稀疏子空间采样生成的对抗样本相比于干净样本使得mAP值降低了43.6%、召回率降低41.2%；所生成对抗样本相较于同等强度的随机噪声干扰呈现出更好的攻击效果，且全空间查询的攻击成功率相较于同等强度随机噪声提升了17.6%，稀疏子空间查询的攻击成功率相较于同等强度随机噪声提升了34.4%。在相同的查询次数下，稀疏子空间各项指标均优于全空间，呈现出更好的攻击效果，其中攻击成功率提升了2.5%。

### 3.3 对抗扰动能量比例分布及迭代下降过程

本文统计了不同采样空间扰动的平均能量比，其中全空间采样的对抗扰动平均能量比为5.18%，而稀疏子空间采样对抗扰动平均能量比为5.00%，比全空间采样方法降低了3.47%。

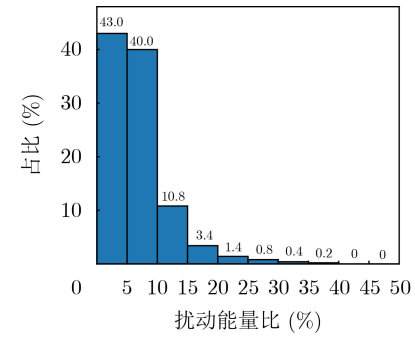
为研究对抗扰动能量分布，绘制全空间查询和稀疏子空间查询的扰动能量比例分布直方图如图4所示。图4(a)和图4(b)为扰动能量比50%以下的分布直方图，图4(c)和图4(d)为扰动能量比10%以下的分布直方图。在21轮迭代后，根据统计结果，两种查询空间的扰动能量比例均主要集中在10%以下，其中全空间占比83%，稀疏子空间占比82.6%，两者基本一致。在扰动能量比低于10%的分布中，稀疏子空间主要集中在3%～4%，全空间主要集中在4%～5%。在扰动能量比低于5%的数据分布中，稀疏子空间占比44.4%略高于全空间占比43%。其中稀疏子空间4%以下占比31.8%，而全空间4%以下占比26.4%。两种方法下的对抗扰动能量分布基本一致。

为研究在两种采样空间下，扰动能量随迭代轮数下降情况，本文选择了经过21轮迭代后扰动能量比在10%以下的信号对抗样本，并统计每轮迭代后信号扰动样本的扰动能量比平均值，绘制扰动能量比平均值随迭代轮数变换折线图。

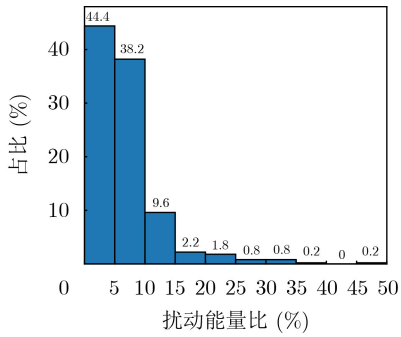
图5中蓝色曲线对应全空间查询，红色虚线对应稀疏子空间查询。根据图中扰动能量比平均值随迭代轮数下降情况，我们可以得到如下结论：本文提出的信号检测网络对抗攻击方法在两种采样空间下，经过多轮迭代均能有效降低信号扰动强度。在迭代开始时，两种空间的扰动能量比较接近，当查询轮数逐渐增加时，稀疏子空间查询扰动能量略低于全空间查询，且差值逐渐增大。

### 3.4 信号样本波形图对比

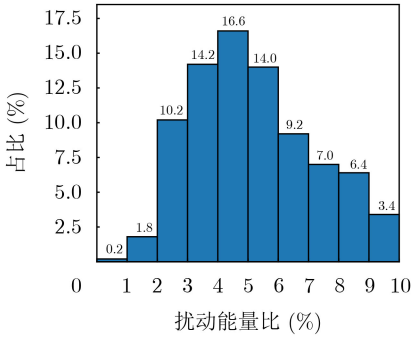
为了更直观地展示信号对抗样本的时域特征，本文随机选取了100个连续采样点进行波形对比分析，结果如图6所示。图6(a)展示了全空间查询生成的对抗样本与原始信号的波形对比，而图6(b)则呈现了稀疏子空间查询生成的对抗样本与原始信号



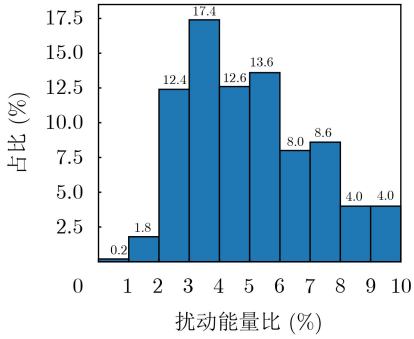

(a) 全空间查询扰动能量比50%以下的分布直方图




(b) 稀疏子空间查询扰动能量比50%以下的分布直方图




(c) 全空间查询扰动能量比10%以下的分布直方图




(d) 稀疏子空间查询扰动能量比10%以下的分布直方图

图 4 不同查询空间对抗样本扰动能量比分布直方图

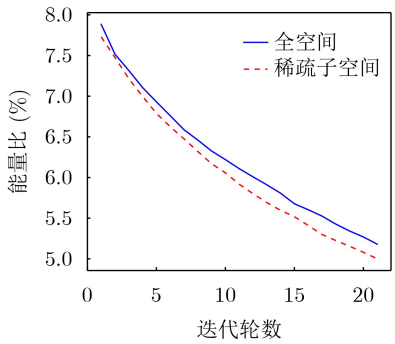
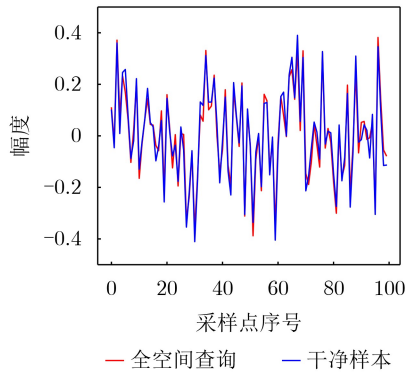


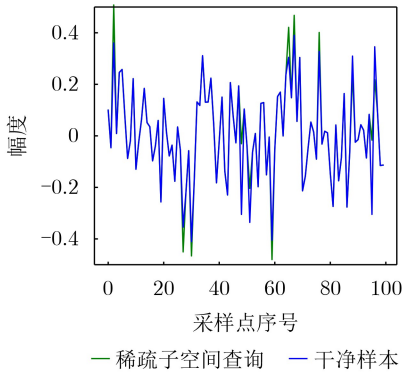

图 5 不同查询空间扰动能量比平均值随迭代轮数下降曲线

的波形对比。其中，蓝色曲线代表原始信号样本，其他颜色曲线分别对应不同方法生成的对抗样本。图6(b)可以看出对抗扰动都施加在了信号幅度值较大的位置，最终生成的对抗样本形态上与原始信号非常接近，且施加扰动的位置也非常少，直接体现了对抗扰动的不易察觉性。

从图6的实验结果可以看出，本文提出的两种方法均能在有效生成对抗样本的同时，较好地保持原始信号的主体波形特征。具体而言，全空间查询方法在每个采样点上都引入了扰动，而稀疏子空间查询方法则仅对信号幅值较大的分量进行优化调整，其余分量保持不变。这一策略使得稀疏子空间查询方法具有显著的对抗扰动0范数控制特性，即在保证攻击效果的同时最小化扰动分量的数量，这一重要特性在全空间采样方法(以优化扰动2范数为目标)中难以实现。此外，由于稀疏采样在梯度查询过程中仅对采样到的10%信号分量施加随机扰动并进行梯度计算，其余90%的信号扰动分量无需进行计算，较全空间查询方法显著降低了计算开销。




(a) 信号样本与全空间查询对抗样本波形图




(b) 信号样本与稀疏子空间查询对抗样本波形图

图 6 信号样本与信号对抗样本波形图

## 4　结论

本文提出了一种基于稀疏子空间采样的信号检测网络黑盒查询对抗攻击方法。本方法将信号样本检测数量消失比例作为判断攻击是否成功的约束条件，构造信号检测网络对抗样本攻击模型，参考HSJA算法设计基于决策边界的信号检测网络黑盒查询对抗攻击方法，求解该模型生成信号对抗样本，并根据信号对抗扰动特点构建稀疏采样子空间进行查询攻击，进一步改善查询效率。实验结果表明，在信号目标数量消失比例0.3的决策边界下，稀疏子空间采样黑盒对抗攻击方法使得信号检测网络mAP值降低了43.6%、召回率降低了41.2%。与全空间采样方法相比，稀疏子空间采样方法攻击成功率提升了2.5%，且对抗扰动平均能量比降低了3.47%。稀疏子空间采样攻击方法可以使得信号检测网络性能明显下降，相较于全空间采样具有攻击成功率高、扰动强度更小等优势。

## 参 考 文 献

[1] REN Shaoqing, HE Kaiming, GIRSHICK R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

[2] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: Unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779–788. doi: 10.1109/CVPR.2016.91.

[3] WANG C Y, BOCHKOVSKIY A, and LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 7464–7475. doi: 10.1109/CVPR52729.2023.00721.

[4] ZHA Xiong, PENG Hua, QIN Xin, *et al*. A deep learning framework for signal detection and modulation classification[J]. *Sensors*, 2019, 19(18): 4042. doi: 10.3390/

s19184042.

[5] VAGOLLARI A, HIRSCHBECK M, and GERSTACKER W. An end-to-end deep learning framework for wideband signal recognition[J]. *IEEE Access*, 2023, 11: 52899–52922. doi: 10.1109/ACCESS.2023.3280454.

[6] LI Qing, ZHOU Xin, MENG Xiandong, *et al*. Lightweight RadioYOLO for radio signal detection[C]. 2022 International Conference on Algorithms, Data Mining, and Information Technology, Xi'an, China, 2022: 117–123. doi: 10.1109/ADMIT57209.2022.00027.

[7] 耿频永, 曹叶文, 赵晓蕾, 等. 基于频率敏感注意力的短波宽带特定信号检测[J]. 数据采集与处理, 2023, 38(1): 63–73. doi: 10.16337/j.1004-9037.2023.01.004.
GENG Pinyong, CAO Yewen, ZHAO Xiaolei, *et al*. Shortwave wideband specific signal detection based on frequency-sensitive attention[J]. *Journal of Data Acquisition and Processing*, 2023, 38(1): 63–73. doi: 10.16337/j.1004-9037.2023.01.004.

[8] 李润东. 基于深度学习的通信信号智能盲检测与识别技术研究[D]. [博士论文], 电子科技大学, 2021.
LI Rundong. Research on intelligent blind detection and recognition of communication signals based on deep learning[D]. [Ph. D. dissertation], University of Electronic Science and Technology of China, 2021.

[9] XIE Cihang, WANG Jianyu, ZHANG Zhishuai, *et al*. Adversarial examples for semantic segmentation and object detection[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 1378–1387. doi: 10.1109/ICCV.2017.153.

[10] WANG Derui, LI Chaoran, WEN Sheng, *et al*. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples[J]. *IEEE Transactions on Cybernetics*, 2022, 52(8): 7427–7440. doi: 10.1109/TCYB.2020.3041481.

[11] ZHU Hegui, SUI Xiaoyan, REN Yuchen, *et al*. Boosting transferability of targeted adversarial examples with non-robust feature alignment[J]. *Expert Systems with Applications*, 2023, 227: 120248. doi: 10.1016/j.eswa.2023.120248.

[12] CHOW K H, LIU Ling, LOPER M, *et al*. Adversarial objectness gradient attacks in real-time object detection systems[C]. 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, Atlanta, USA, 2020: 263–272. doi: 10.1109/TPS-ISA50397.2020.00042.

[13] SADEGHI M and LARSSON E G. Adversarial attacks on deep-learning based radio signal classification[J]. *IEEE Wireless Communications Letters*, 2019, 8(1): 213–216. doi: 10.1109/LWC.2018.2867459.

[14] ZHANG Sicheng, FU Jiangzhi, YU Jiarun, *et al*. Channel-robust class-universal spectrum-focused frequency adversarial attacks on modulated classification models[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2024, 10(4): 1280–1293. doi: 10.1109/TCCN.2024.3382126.

[15] TIAN Qiao, ZHANG Sicheng, MAO Shiwen, *et al*. Adversarial attacks and defenses for digital communication signals identification[J]. *Digital Communications and Networks*, 2024, 10(3): 756–764. doi: 10.1016/j.dcan.2022.10.010.

[16] LI Dongyang, WANG Linyuan, XIONG Guangwei, *et al*. Signal adversarial examples generation for signal detection network via white-box attack[J]. *Electronics Letters*, 2025, 61(1): e70348. doi: 10.1049.ell2.70348.

[17] SHIN Y, NAM S W, AN C K, *et al*. Design of a time-frequency domain matched filter for detection of non-stationary signals[C]. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, 2001: 3585–3588. doi: 10.1109/ICASSP.2001.940617.

[18] DIAMANT R. Closed form analysis of the normalized matched filter with a test case for detection of underwater acoustic signals[J]. *IEEE Access*, 2016, 4: 8225–8235. doi: 10.1109/ACCESS.2016.2630498.

[19] GARDNER W A. Exploitation of spectral redundancy in cyclostationary signals[J]. *IEEE Signal Processing Magazine*, 1991, 8(2): 14–36. doi: 10.1109/79.81007.

[20] ZENG Yonghong and LIANG Yingchang. Maximum-minimum eigenvalue detection for cognitive radio[C]. The 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, Greece, 2007: 1–5. doi: 10.1109/PIMRC.2007.4394211.

[21] URKOWITZ H. Energy detection of unknown deterministic signals[J]. *Proceedings of the IEEE*, 1967, 55(4): 523–531. doi: 10.1109/PROC.1967.5573.

[22] PAPERNOT N, MCDANIEL P, and GOODFELLOW I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples[EB/OL]. https://arxiv.org/abs/1605.07277, 2016.

[23] LIU Yanpei, CHEN Xinyun, LIU Chang, *et al*. Delving into transferable adversarial examples and black-box attacks[C]. The 5th International Conference on Learning Representations, Toulon, France, 2017.

[24] WANG Zhibo, YANG Hongshan, FENG Yunhe, *et al*. Towards transferable targeted adversarial examples[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 20534–20543. doi: 10.1109/CVPR52729.2023.01967.

[25] CHEN Pinyun, ZHANG Huan, SHARMA Y, *et al*. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]. The 10th ACM Workshop on Artificial Intelligence and Security, Dallas, USA, 2017: 15–26. doi: 10.1145/3128572.3140448.

[26] GUO Chuan, GARDNER J R, YOU Yurong, *et al*. Simple black-box adversarial attacks[C]. The 36th International Conference on Machine Learning, Long Beach, USA, 2019: 2484–2493.

[27] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, *et al*. Square attack: A query-efficient black-box adversarial attack via random search[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 484–501. doi: 10.1007/978-3-030-58592-1_29.

[28] BRENDEL W, RAUBER J, and BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]. The 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

[29] CHENG Minhao, SINGH S, CHEN P H, *et al*. Sign-OPT: A query-efficient hard-label adversarial attack[C]. The 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020.

[30] CHEN Jianbo, JORDAN M I, and WAINWRIGHT M J. HopSkipJumpAttack: A query-efficient decision-based attack[C]. Proceedings of 2020 IEEE Symposium on Security and Privacy, San Francisco, USA, 2020: 1277–1294. doi: 10.1109/SP40000.2020.00045.

[31] LI Huichen, XU Xiaojun, ZHANG Xiaolu, *et al*. QEBA: Query-efficient boundary-based blackbox attack[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 1221–1230. doi: 10.1109/CVPR42600.2020.00130.

[32] SHAMIR A, SAFRAN I, RONEN E, *et al*. A simple explanation for the existence of adversarial examples with small hamming distance[EB/OL]. https://arxiv.org/abs/1901.10861, 2019.

[33] CAI Hanqin, LOU Yuchen, MCKENZIE D, *et al*. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization[C]. The 38th International Conference on Machine Learning, Graz, Austria, 2021: 1193–1203.

[34] CAI Hanqin, MCKENZIE D, YIN Wotao, *et al*. A one-bit, comparison-based gradient estimator[J]. *Applied and Computational Harmonic Analysis*, 2022, 60: 242–266. doi: 10.1016/j.acha.2022.03.003.

[35] 武越, 苑咏哲, 岳铭煜, 等. 点云配准中多维度信息融合的特征挖掘方法[J]. 计算机研究与发展, 2022, 59(8): 1732–1741. doi: 10.7544/issn1000-1239.20220042.
WU Yue, YUAN Yongzhe, YUE Mingyu, *et al*. Feature mining method of multi-dimensional information fusion in point cloud registration[J]. *Journal of Computer Research and Development*, 2022, 59(8): 1732–1741. doi: 10.7544/issn1000-1239.20220042.

[36] 郭宇琦, 李东阳, 尹志宁, 等. 采用1-bit压缩感知的信号识别网络黑盒对抗攻击方法[J]. 信息工程大学学报, 2024, 25(5): 593–600. doi: 10.3969/j.issn.1671-0673.2024.05.014.
GUO Yuqi, LI Dongyang, YIN Zhining, *et al*. Black-box adversarial attacks on signal recognition networks using 1-bit compressed sensing[J]. *Journal of Information Engineering University*, 2024, 25(5): 593–600. doi: 10.3969/j.issn.1671-0673.2024.05.014.

[37] 郭宇琦. 无线I/Q信号调制识别神经网络对抗样本生成方法[D]. [硕士论文], 战略支援部队信息工程大学, 2023.
GUO Yuqi. Research on wireless I/Q signal adversarial examples in modulation recognition neural networks[D]. [Master dissertation], PLA Strategy Support Force Information Engineering University, 2023.

李东阳：男，助理工程师，研究方向为智能信号处理和人工智能安全.
王林元：男，副教授，研究方向为人工智能基础理论、稀疏优化理论、脑机交互混合智能等.
彭进先：男，工程师，研究方向为复杂电磁环境构设、电子对抗等.
马德魁：男，助教，研究方向为人工智能基础理论、智能信息处理等.
闫　镔：男，教授。研究方向为智能信息处理.

责任编辑：余　蓉

# A Black-Box Query Adversarial Attack Method for Signal Detection Networks Based on Sparse Subspace Sampling

LI Dongyang[①②]　WANG Linyuan[①]　PENG Jinxian[②]　MA Dekui[①]　YAN Bin[①]

[①](*Henan Key Laboratory of Imaging and Intelligent Processing, Information Engineering University, Zhengzhou* 450001, *China*)

[②](63611 *Troops, People's Liberation Army of China, Kuerle* 841000, *China*)

**Abstract**:

**Objective**　The application of deep neural networks to signal detection has raised concerns regarding their vulnerability to adversarial example attacks. In black-box attack scenarios, where internal model information is

inaccessible, this paper proposes a black-box query adversarial attack method based on sparse subspace sampling. The method offers an approach to evaluate the robustness of signal detection networks under black-box conditions, providing theoretical support and practical guidance for improving the reliability and robustness of these networks.

**Methods** This study combines the characteristics of signal detection networks with the attack objective of reducing the recall rate of signal targets. The disappearance ratio of detected signal targets is used as the constraint for determining attack success, forming an adversarial attack model for signal detection networks. Based on the HopSkipJumpAttack (HSJA) algorithm, a black-box query adversarial attack method for signal detection networks is designed, which generates adversarial examples by approaching the model's decision boundary. To further improve query efficiency, a sparse subspace query adversarial attack method is proposed. This approach constructs sparse subspace sampling based on the characteristics of signal adversarial perturbations. Specifically, during the generation of adversarial examples, signal components with large amplitudes are selected in proportion, and only these components are perturbed.

**Results and Discussions** Experimental results show that under a decision boundary condition with a signal target disappearance ratio of 0.3, the proposed sparse subspace sampling black-box adversarial attack method reduces the mean Average Precision (mAP) by 43.6% and the recall rate by 41.2%. Under the same number of queries, all performance metrics for the sparse subspace sampling method exceed those of the full-space sampling approach, demonstrating improved attack effectiveness, with the success rate increasing by 2.5% (Table 2). In terms of signal perturbation intensity, the proposed method effectively reduces perturbation intensity through iterative optimization under both sampling spaces. At the beginning of the iterations, the perturbation energies for the two spaces are similar. As the number of query rounds increases, the perturbation energy required for sparse subspace sampling becomes slightly lower than that of full-space sampling, and the difference continues to widen. The average adversarial perturbation energy ratio for full-space sampling is 5.18%, whereas sparse subspace sampling achieves 5.00%, reflecting a 3.47% reduction relative to full-space sampling (Fig. 4). For waveform perturbations, both sampling strategies proposed in this study can generate effective adversarial examples while preserving the primary waveform characteristics of the original signal. Specifically, the full-space query method applies perturbations to every sampling point, whereas the sparse subspace query method selectively perturbs only the large-amplitude signal components, leaving other components unchanged (Fig. 5). This selective approach provides the sparse subspace method with a notable l0-norm control property for adversarial perturbations, minimizing the number of perturbed components without compromising attack performance. In contrast, the full-space sampling method focuses on optimizing the l2-norm of perturbations, without achieving this selective control.

**Conclusions** This study proposes a black-box query adversarial attack method for signal detection networks based on sparse subspace sampling. The disappearance ratio of detected signal targets is used as the success criterion for attacks, and an adversarial example generation model for signal detection networks is established. Drawing on the HSJA algorithm, a decision-boundary-based black-box query attack method is designed to generate adversarial signal examples. To further enhance query efficiency, a sparse subspace sampling strategy is constructed based on the characteristics of signal adversarial perturbations. Experimental results show that under a decision boundary with a target disappearance ratio of 0.3, the proposed sparse subspace sampling black-box attack method reduces the mAP of the signal detection network by 43.6% and the recall rate by 41.2%. Compared with full-space sampling, sparse subspace sampling increases the attack success rate by 2.5% and reduces the average perturbation energy ratio by 3.47%. The sparse subspace method significantly degrades signal detection network performance while achieving superior attack efficiency and lower perturbation intensity relative to full-space sampling. Furthermore, the full-space query method introduces perturbations at all sampling points, whereas the sparse subspace method selectively perturbs only the high-amplitude signal components, leaving other components unchanged. This approach enforces $l_0$-norm sparsity constraints, minimizing the number of perturbed components without compromising attack effectiveness. The proposed method provides a practical solution for evaluating the robustness of signal detection networks under black-box conditions and offers theoretical support for improving the reliability of these networks against adversarial threats.

**Key words**: Signal detection network; Signal adversarial examples; Black-box query attack; Sparse subspace sampling