

深度学习软测量模型对抗鲁棒性增强: 一种对抗训练与攻击检测融合的对抗防御方法

郭润元¹ 李安伦¹ 刘涵¹ 张友民² 刘丁¹

摘 要 深度学习软测量 (DLSS) 已成为解决复杂工业过程关键变量测量难题的有效方法, 但其极易遭受肉眼难以察觉的对抗攻击而输出虚假预测结果, 进而危害生产安全. 现有对抗攻击检测方法对微小扰动对抗样本 (SPAS) 普遍难以检出, 且对抗训练防御方法仍广泛存在对抗鲁棒过拟合现象. 针对上述问题, 提出一种对抗训练与攻击检测融合 (ATADI) 的对抗防御方法: 首先, 理论证明仅采用 SPAS 进行对抗训练更有助于缓解对抗鲁棒过拟合. 其次, 在此基础上提出一种基于 SPAS 的特征锚定式对抗训练 (FAAT) 方案, 即 SPAS-FAAT. ATADI 中的攻击检测器在防御过程中仅检测较大扰动对抗样本 (LPAS) 以规避其难以检出 SPAS 的不足, SPAS 则被输入 SPAS-FAAT 后的 DLSS 以获得最终软测量结果. 最后, 在转子热变形软测量对抗攻防案例上进行了实验. 结构消融实验结果显示, 将对抗训练与攻击检测相融合可作为 DLSS 对抗防御的有效手段, 各类攻击产生的 LPAS 均能以不低于 98.42% 的准确率被检出, 且 SPAS-FAAT 后的 DLSS 在 SPAS 与正常样本上的软测量结果均满足精度要求. ATADI 方法显著增强了 DLSS 的对抗鲁棒性, 为对抗防御方法的研究提供了新思路.

关键词 软测量; 深度学习; 对抗防御; 对抗训练; 对抗攻击检测

引用格式 郭润元, 李安伦, 刘涵, 张友民, 刘丁. 深度学习软测量模型对抗鲁棒性增强: 一种对抗训练与攻击检测融合的对抗防御方法. 自动化学报, 2026, 52(9): 2068–2081

DOI 10.16383/j.aas.c250725

CSTR 32138.14.j.aas.c250725

Enhancing Adversarial Robustness of Deep Learning-based Soft Sensors: A Defense Method Integrating Adversarial Training and Attack Detection

GUO Run-Yuan¹ LI An-Lun¹ LIU Han¹ ZHANG You-Min² LIU Ding¹

Abstract Deep learning-based soft sensors (DLSS) effectively measure critical variables in complex industrial processes. However, DLSS models are highly vulnerable to imperceptible adversarial attacks that generate misleading outputs and endanger production safety. Existing adversarial detectors poorly identify small perturbation adversarial samples (SPAS), and adversarial training defense methods still widely suffer from adversarial robustness overfitting. To tackle these issues, an adversarial training and attack detection integration (ATADI) defense method is proposed. First, it theoretically proves that SPAS-only adversarial training better mitigates such overfitting. Second, on this basis, a SPAS-based feature-anchored adversarial training (FAAT) scheme, namely SPAS-FAAT, is proposed. In ATADI, the detector only targets large perturbation adversarial samples (LPAS) to avoid SPAS detection difficulties during defense, while SPAS are directly fed into SPAS-FAAT-enhanced DLSS to obtain the final soft sensing results. Finally, experiments are conducted on an adversarial attack-defense case of rotor thermal deformation soft sensing: Structural ablation results verify that the integration of adversarial training and attack detection serves as an effective defense means for DLSS. LPAS from various adversarial attacks are detected with over 98.42% accuracy, and SPAS-FAAT-enhanced DLSS achieves satisfactory accuracy in the soft sensing results for both SPAS and normal samples. The ATADI method significantly improves DLSS adversarial robustness, providing a new avenue for adversarial defense method research.

Keywords soft sensors; deep learning; adversarial defense; adversarial training; adversarial attack detection

Citation Guo Run-Yuan, Li An-Lun, Liu Han, Zhang You-Min, Liu Ding. Enhancing adversarial robustness of deep learning-based soft sensors: A defense method integrating adversarial training and attack detection. *Acta Automatica Sinica*, 2026, 52(9): 2068–2081

收稿日期 2025-12-16 录用日期 2026-04-20

Manuscript received December 16, 2025; accepted April 20, 2026

国家自然科学基金 (62127809, 62403377, 92270117), 陕西省自然科学基金基础研究计划青年项目 (2024JC-YBQN-0697), 西安自动化学会青年人才托举计划项目 (XAAA-QNRC-2025-10) 资助

Supported by National Natural Science Foundation of China (62127809, 62403377, 92270117), Youth Program of the Natural Science Basic Research Program of Shaanxi Province (2024JC-YBQN-0697), and Xi'an Association of Automation Young

Talent Nurturing Program (XAAA-QNRC-2025-10)

本文责任编辑 许斌

Recommended by Associate Editor XU Bin

1. 西安理工大学自动化与信息工程学院 西安 710048 中国 2. 康考迪亚大学机械、工业与航空工程系 蒙特利尔 H3G1M8 加拿大

1. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China 2. Department of Mechanical, Industrial and Aerospace Engineering, Concordia University, Montreal H3G1M8, Canada

复杂工业过程中通常存在一大类与产品质量密切相关、需加以严格控制的关键过程变量, 由于硬件传感器的技术制约或经济成本等原因, 当前难以对这些变量实时准确地测量^[1]. 为解决该问题, 软测量技术应运而生, 通过选用与难测关键变量 (又称“主导变量”) 密切相关的易测过程变量 (又称“辅助变量”) 作为输入, 从而建立辅助变量与主导变量之间的某种数学模型, 最终实现对主导变量的在线测量^[2]. 建立软测量模型的主要方法有机理知识驱动和数据驱动两种^[3]. 由于在复杂工业过程中建立精确的机理知识模型较为困难, 近年来数据驱动的软测量建模方法成为主流. 其中, 基于深层神经网络建立的深度学习软测量 (deep learning-based soft sensors, DLSS) 模型已被证实能够凭借其自身强大的数据时空特征提取能力而实现更高精度的主导变量预测^[4]. 然而, DLSS 存在对抗鲁棒性缺失的瓶颈, 严重阻碍了其广泛部署和应用.

软测量模型的对抗鲁棒性是指模型防御对抗攻击的能力. 对抗攻击现象最早在使用神经网络模型进行图像分类任务时被发现, 即通过向干净的图像样本中添加精心设计且人类肉眼难以察觉的微小扰动以产生对抗样本, 并将该对抗样本输入神经网络, 即可开展攻击, 从而使模型输出错误的分类结果^[5]. 除了分类识别任务外, 对抗攻击在回归预测类任务中也大量存在, 这对 DLSS 的可靠部署及应用造成了严重的威胁^[6]. 一方面, 由于神经网络模型已被证实容易遭受对抗攻击的威胁, 而 DLSS 基于深度神经网络构建, 因此 DLSS 天然地容易受到对抗攻击的影响^[7]. 另一方面, 由于软测量模型在工业控制系统中进行部署, 越来越多的工业控制系统与企业、公共网络高度互联, 工业生产与外界不再物理隔离, 这导致攻击者容易利用各类软件和硬件漏洞入侵工业控制系统. 入侵后, 攻击者能够隐蔽地进行横向渗透并进一步劫持和篡改传输链路上的数据, 从而开展针对 DLSS 的对抗攻击^[8]. 例如, 文献 [9] 提出一种针对 DLSS 的迭代型直接攻击输出 (iterative directly attack output, IDAO) 方法, 结果显示 IDAO 能够对合成氨过程中转化炉内氧含量的 DLSS 模型实施对抗攻击, 攻击结果使 DLSS 给出虚假的氧含量输出值, 诱使技术人员误判过程状态并执行错误操作, 危害工业生产安全. 文献 [10] 提出一种平移镜像输出攻击 (translation mirror output attack, TMOA) 方法, 并使用 TMOA 对硫回收装置中 SO_2 浓度预测的 DLSS 模型成功实施了对抗攻击. 事实上, 尽管目前并非所有部署 DLSS 的控制系统都有遭受对抗攻击的公开报道, 但随着近两年不断有新的对抗攻击方法被提出, 且对抗防

御方法的研究自身具有显著的预防性特征, 即必须在攻击发生前构建全面的防御框架, 而非事后补救, 故急需开展针对对抗攻击的对抗防御方法研究, 以提升 DLSS 的对抗鲁棒性^[11].

现有的对抗防御方法主要分为主动式防御与反应式防御两种. 主动式防御通过增强 DLSS 自身的对抗鲁棒性来提升其防御对抗攻击的能力, 其中最具代表性的方法是对抗训练^[12]. 对抗训练通过将对抗样本纳入 DLSS 的训练过程, 显著改善了 DLSS 在面对对抗攻击时主导变量预测精度下降的问题. DLSS 领域的对抗训练在形式上被认作一种内部最大化-外部最小化的数学优化问题. 其中, 内部最大化指针对给定的正常输入样本构造出更具攻击性的对抗样本的过程; 而外部最小化指在得到上述对抗样本后, 以正常样本和对抗样本共同作为输入, 将 DLSS 的训练损失最小化的模型训练过程^[13]. 因此, 要想实现更优的对抗训练效果, 就需要使内部最大化和外部最小化问题均实现更优的求解.

针对内部最大化问题的优化求解, 文献 [14] 提出一种知识引导型对抗网络 (knowledge-guided adversarial attack, KGAA) 方法. KGAA 将测量环节的机理知识加入目标函数, 并添加新的约束条件, 通过定义障碍函数以重构优化问题, 从而克服攻击 DLSS 时优化问题的病态, 成功得到攻击性更强的对抗样本. 文献 [15] 提出一种基于生成式模型的知识辅助生成式对抗网络 (knowledge-aided generative adversarial network, KAGAN), 该网络通过直接生成对抗样本从而避免了对迁移梯度的不准确估计, 为生成强攻击性对抗样本提供了便利. 与内部最大化问题所得到的广泛研究不同, 当前对外部最小化问题优化求解的研究仍较为有限, 文献 [16] 在 DLSS 的最后一个隐藏层中使用标准对抗训练 (standard adversarial training, SAT) 方案开展防御, 一定程度上提升了模型对噪声和攻击的鲁棒性. 文献 [17] 将对抗训练过程创新性地看作领域自适应过程, 将对抗样本所在域看作目标域, 将原正常样本所在域看作源域, 提出一种域自适应对抗训练 (domain-adaptive adversarial training, DAAT) 方法, 有效提升了 DLSS 的对抗鲁棒性. 然而, 当前以文献 [16] 和文献 [17] 为代表的对抗训练方法均存在较严重的对抗鲁棒过拟合, 表现为: 虽然将对抗样本输入已完成对抗训练的 DLSS 时, 主导变量的预测精度得到显著提升, 即 DLSS 能够抵御对抗攻击的威胁, 但若此时将原正常样本输入该 DLSS, 模型对其主导变量预测的精度反而大幅下降, 这被称作对抗鲁棒过拟合引起的自然预测精度下降现象^[18]. 事实上, 当前复杂工业过程对控制品质的要

求不断提高,要求软测量模型必须实现对主导变量的高精度预测.因此,面对对抗鲁棒过拟合问题,需要进一步研究新的对抗训练方法,避免 DLSS 对抗样本过度学习,从而训练出具备优良对抗鲁棒性的 DLSS.

除主动式防御之外,反应式防御同样是重要的对抗防御方法.该类方法旨在在 DLSS 接收输入样本前及时识别并干预潜在对抗样本,从而避免 DLSS 被对抗样本攻击^[19].反应式防御的主要形式为对抗攻击检测,当前对抗攻击检测方法按对抗样本是否参与训练,可分为无监督检测与有监督检测两大类.无监督检测方法通过挖掘正常样本在特征空间或误差空间中的内在结构与分布规律,从而识别偏离这些统计特征的异常输入^[20].文献^[21]利用自编码器对输入样本进行重构,通过计算原始样本与重构样本之间的重构误差,检测潜在的对抗攻击.然而,对抗攻击与其他过程异常最大的不同在于其天然具有极强的隐蔽性,即在生成对抗样本时所添加的扰动相比一般过程异常较为微小,且不断有能够生成更微小扰动的强隐蔽性对抗攻击方法被提出.如文献^[15]提出的 KAGAN 和文献^[17]提出的融合历史梯度的对抗攻击方法 (historical gradient-based adversarial attack, HGAA),它们添加的扰动相比 IDAO 和 KGAA 来说更为微小,导致常规的无监督检测方法在面临这类强隐蔽性对抗攻击时难以发挥效用.与无监督方法不同,有监督检测方法通过利用正常样本与对抗样本在数据流形上的差异来检测对抗样本.文献^[22]以主网络的中间层特征为输入,通过在深度神经网络中增添小型轻量级的多层感知机检测器,对输入样本的特征分布进行判别,从而区分正常样本与对抗样本.然而,该类方法往往仅能检测参与训练的攻击类型,在面对未参与训练的未知对抗攻击类型时,其检测性能会严重下降,导致此类方法的实用性降低.文献^[23]针对此问题提出一种基于双向一致性判别 (bidirectional consistency discrimination, BCD) 的对抗攻击检测方法.该方法通过对待检测样本的特征值进行前向预测并反向重构历史样本的特征值,从而计算出待检测样本在辅助变量时间序列中的双向一致性.通过一致性残差矩阵的范数量化数据一致性程度,并结合阈值判别实现对抗攻击的有效检测.由于判别过程仅依赖辅助变量序列,因此该方法对未知攻击也能有效检出.需要说明的是,对抗攻击在产生对抗样本时虽然添加的是肉眼难以察觉的微小扰动,但这些满足范数约束的扰动相对而言会有数值大小之分.在生成的对抗样本中,将所添加扰动相对较大者统称为较大扰动对抗样本 (large perturbation

adversarial samples, LPAS), 剩余的样本则为微小扰动对抗样本 (small perturbation adversarial samples, SPAS).由文献^[23]的实验结果可知,BCD 检测方法仅对普通对抗攻击方法有效,对强隐蔽性对抗攻击 (如 HGAA 和 KAGAN) 来说,由于其生成对抗样本时所添加的扰动更为微小,导致 LPAS 和 SPAS 与正常样本的数值差距也更加微小,BCD 检测方法对 LPAS 尚能检出,但对 SPAS 难以检出.由此可见,无论是有监督还是无监督检测,当前检测方法尚不足以对强隐蔽性对抗攻击开展有效防御.

针对上述攻击检测与对抗训练方法各自的不足,一个自然的想法是将对抗训练与攻击检测融合,将容易检测的 LPAS 交由检测器检出,而将检测器难以检出的 SPAS 交由对抗训练方法进行处理.由此,攻击检测的目标从“检出全部对抗样本”转变为“精确检出 LPAS”,对抗训练的目标也由“提升模型对所有对抗样本的对抗鲁棒性”转变为“针对性地提升模型对 SPAS 的鲁棒性”.目前,DLSS 对抗防御领域尚未出现将对抗训练与攻击检测融合的研究.在图像分类领域,虽然有将检测与对抗训练相结合的研究,但其仅将检测与训练机械式地拼接并面向全体对抗样本^[24],而非本文所提出的将检测和训练步骤分别与 LPAS 和 SPAS 适配的方案.同时,已有研究表明参与对抗训练的对抗样本数量的减少能够缓解对抗鲁棒过拟合^[25],故使用对抗样本中的 SPAS 开展对抗训练,已初步具备理论基础.然而,在 DLSS 对抗防御中舍弃 LPAS,仅选取 SPAS 参与对抗训练的必要性 and 优势仍需进一步证明.

综上,针对 DLSS 在开展对抗防御时面临的对抗鲁棒过拟合、SPAS 难以检出的问题,本文提出一种对抗训练与攻击检测融合 (adversarial training and attack detection integration, ATADI) 的对抗防御方法.主要贡献如下:

- 1) 对选取 SPAS 开展对抗训练的必要性 and 优势给出证明,并提出基于 SPAS 进行的特征锚定式对抗训练 (feature-anchored adversarial training, FAAT) 方案,命名为 SPAS-FAAT.

- 2) 在 DLSS 对抗防御领域,首次提出将对抗训练与攻击检测融合以开展防御.ATADI 巧妙规避了攻击检测与对抗训练各自的不足,LPAS 将交予 ATADI 中的攻击检测模块直接检出,无法检出的 SPAS 将输入 SPAS-FAAT 后的 DLSS,以得到符合精度要求的主导变量预测结果.

- 3) 在空气预热器 (后文均简称“空预器”) 转子热变形软测量对抗攻防案例上,对 ATADI 对抗防御的有效性进行充分验证,显示出该方法在提升

DLSS 对抗鲁棒性上的显著优势.

1 预备知识

1.1 DLSS 的对抗攻击

以攻击者对待攻击 DLSS 模型信息的掌握程度为依据, DLSS 对抗攻击可划分为白盒攻击与黑盒攻击两类. 白盒攻击假设攻击者能够掌握待攻击 DLSS 的模型架构、训练参数及训练数据等内部信息. 对实际复杂工业过程来说, 攻击者通常难以获得 DLSS 的完整内部信息, 因此白盒攻击更多作为一种理想化的攻击方法. 黑盒攻击则假设攻击者对待攻击模型的内部结构、参数等信息一无所知或仅掌握极少信息, 更贴近实际工业场景. 通常, DLSS 对抗攻防领域的相关研究均假设攻击者能够成功劫持传输链路上的数据, 否则, 对抗攻击与防御方法的研究就无从谈起. 因此, 黑盒攻击者通常通过劫持待攻击 DLSS 模型的输入-输出数据, 构建近似刻画待攻击 DLSS 模型输入-输出关系的代理 DLSS 模型, 并在代理 DLSS 模型上生成具有迁移性的对抗样本, 进而输入原待攻击 DLSS 模型以产生虚假输出, 实现黑盒对抗攻击.

在各类针对 DLSS 的攻击方法中, KGAA 被认为是最具备切实可行性的对抗攻击方法之一, 故以其为例, 简述对抗样本的生成过程^[13]. KGAA 方法将测量环节的机理知识加入目标函数, 并添加新的约束条件, 通过定义障碍函数 $G_{\theta}(\mathbf{x} + \delta, r)$ 解决约束非线性问题. 障碍函数表示如下:

$$G_{\theta}(\mathbf{x} + \delta, r) = -L(y_{\text{KDSS}}, f_{\theta}(\mathbf{x} + \delta)) - rD(\mathbf{x} + \delta) \quad (1)$$

$$D(\mathbf{x} + \delta) = \frac{1}{f(\mathbf{x} + \delta) - y_{\text{KDSS}}} \quad (2)$$

其中, f 表示 DLSS 模型; θ 表示待优化的 DLSS 参数; δ 表示对抗攻击产生的不可察觉扰动; $\mathbf{x}$ 表示输入样本; $\mathbf{x} + \delta$ 表示生成的对抗样本; $f_{\theta}(\mathbf{x} + \delta)$ 为被攻击的 DLSS 模型的虚假输出; y_{KDSS} 为知识驱动软测量模型对输入 $\mathbf{x}$ 的预测值; L 表示损失函数, 本工作中选择均方误差作为软测量场景下的损失函数; r 为较小的正数. 式 (2) 为障碍项的定义, 当 $\mathbf{x}$ 逼近可行域边界时, $G_{\theta}(\mathbf{x} + \delta, r)$ 逼近 $-\infty$.

求解如下优化问题即可得到能够有效攻击 DLSS 的对抗样本^[14]:

$$\begin{aligned} & \max_{\delta} G_{\theta}(\mathbf{x} + \delta, r) \\ & \text{s.t. } \mathbf{x} \in \mathcal{S}, \|\delta\|_{\infty} \leq \varepsilon \end{aligned} \quad (3)$$

其中, $\mathcal{S}$ 代表可行域; ε 为扰动阈值, 用于约束扰动强度. KGAA 攻击方法克服了攻击 DLSS 时优化问

题的病态, 成功得到了具备较强攻击性的对抗样本.

1.2 DLSS 的对抗训练防御

在 DLSS 模型训练过程中, 模型参数的更新仅依赖正常样本的主导变量预测误差, 其优化目标旨在最小化模型输出与实际测量值之间的损失, 如下式所示^[26]:

$$\theta^* = \arg \min_{\theta} E_{(x, y)} [L(f_{\theta}(\mathbf{x}), y)] \quad (4)$$

其中 y 是对应主导变量的实际测量值, θ^* 是训练得到的 DLSS 最优参数, E 表示数学期望, (x, y) 表示从数据分布中采样的数据及标签对.

为提升模型的对抗鲁棒性, SAT 在 DLSS 训练过程中引入精心设计的对抗样本, 通过内部最大化-外部最小化的双层优化, 提升模型在对抗攻击下的预测精度, 其典型优化目标可表示为下式^[26]:

$$\theta^{\text{adv}} = \arg \min_{\theta} E_{(x, y)} \left[\max_{\|\delta\|_{\infty} \leq \varepsilon} L(f_{\theta}(\mathbf{x} + \delta), y) \right] \quad (5)$$

其中 θ^{adv} 是通过对抗训练得到的最优 DLSS 参数.

2 本文方法

本节介绍所提出的 ATADI 对抗防御方法, 主要包含三个部分: 第一部分介绍 ATADI 对抗防御框架与流程; 第二部分对 FAAT 方法进行说明; 第三部分对选取 SPAS 开展对抗训练的必要性 and 优势给出证明.

2.1 ATADI 对抗防御框架与流程

ATADI 对抗防御框架由攻击检测器与 SPAS-FAAT 后的对抗鲁棒 DLSS 组成, 如图 1 所示. 其中, 攻击检测器采用基于 BCD 的对抗攻击检测方法构建^[23], 可以直接检出扰动较大的对抗样本. 一般将检出的对抗样本进行净化后再输入 DLSS 完成最终预测, 由于对抗攻击消除不属于本文的研究内容, 故不做赘述. 对于检测器无法识别的 SPAS, 则直接输入 SPAS-FAAT 后的 DLSS, 获得主导变量预测结果. SPAS-FAAT 方法构建对抗鲁棒 DLSS 的具体步骤如下: 首先采用 BCD 检测方法进行对抗攻击检测实验, 得到检测器的漏检样本集合; 然后取该集合中的最大扰动幅值作为 SPAS 与 LPAS 的划分阈值, 称该阈值为该类攻击下的标准阈值; 最后将扰动小于该阈值的 SPAS 与全部正常样本共同用于 SPAS-FAAT, 在对抗训练后即可得到具备对抗鲁棒性的 DLSS.

2.2 FAAT 方法

如图 2 所示, DLSS 模型在结构上可拆分为前

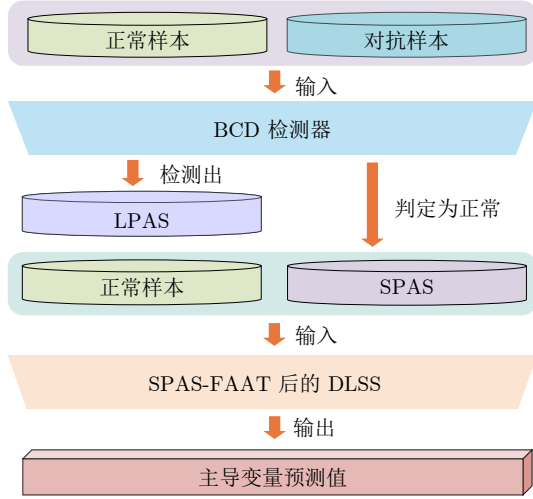


图 1 ATADI 对抗防御框架图

Fig.1 ATADI adversarial defense framework diagram

后两段. 将网络中参数量最大的层及其之前的网络结构定义为 DLSS 模型的前段 F , 参数为 θ^f ; 其余层结构为 DLSS 模型的后段 B , 参数为 θ^b . 前段的功能为将输入样本 $\mathbf{x}$ 映射为隐层特征分布, 后段利用隐层样本预测主导变量.

本文提出的 FAAT 方法, 其具体训练过程分为 3 步:

第 1 步. 利用正常样本对 DLSS 模型进行训练. DLSS 模型训练的目标函数为最小化正常样本的主导变量预测损失 L_p , 公式如下所示:

$$L_p = \sum_{\mathbf{x}_i^{\text{orig}}, y_i^{\text{orig}} \in D^{\text{orig}}} \text{MSE}(B(F(\mathbf{x}_i^{\text{orig}})), y_i^{\text{orig}}) \quad (6)$$

其中, D^{orig} 为原始样本集, 由原始输入样本与对应标签构成; $\mathbf{x}_i^{\text{orig}}$ 为第 i 个原始输入样本; y_i^{orig} 为第 i 个原始输入样本的标签.

训练完成后, 提取 DLSS 模型的前段 F , 并保持其网络结构与参数不变.

第 2 步. 在结构与参数固定的 DLSS 前段 F 之后连接一个新的网络参数随机初始化的神经网络模块, 称为特征锚定器 A . 将正常样本输入 DLSS 模型前段以获得其隐层样本, 随后由特征锚定器 A 对该样本进行重构以恢复原始输入样本. 在此过程中, 仅优化特征锚定器 A 的参数 θ^a , 其目标为最小化重构损失 L_r , 如下所示:

$$\arg \min_{\theta^a} L_r \quad (7)$$

$$L_r = \sum_{\mathbf{x}_i^{\text{orig}} \in D^{\text{orig}}} \text{MSE}(A(F(\mathbf{x}_i^{\text{orig}})), \mathbf{x}_i^{\text{orig}}) \quad (8)$$

特征锚定器 A 在 DLSS 模型前段 F 结构和参

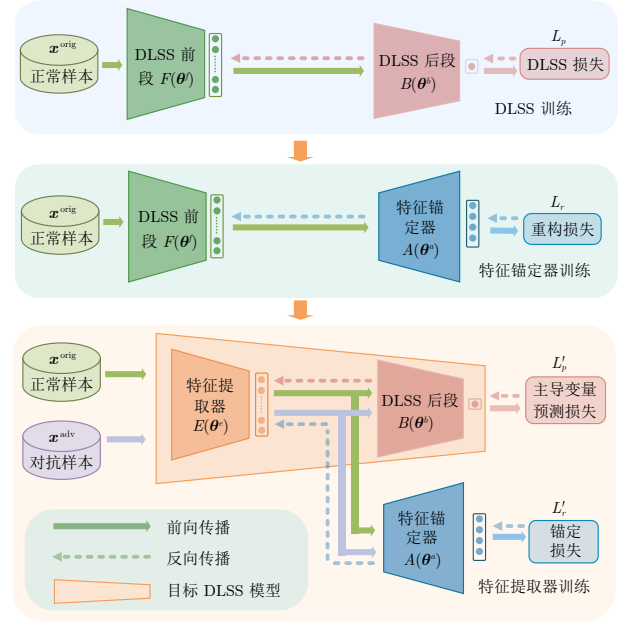


图 2 FAAT 过程示意图

Fig.2 FAAT process diagram

数固定的条件下, 通过重构前段 F 输出以学习正常样本的隐层特征分布. 由于该分布为参数与结构固定的 DLSS 模型后段 B 在达到最高主导变量预测精度时的隐层特征分布, 因此可作为后续对抗训练阶段稳定且可靠的参考分布. 在对抗训练过程中, 可对隐层特征空间施加强约束, 迫使 DLSS 前段在面对对抗样本时提取的隐层样本与正常样本的特征分布接近.

第 3 步. 构建由参数随机初始化的 DLSS 模型前段, 即特征提取器 E 、结构与参数均固定的特征锚定器 A 以及 DLSS 模型后段 B 组成的对抗训练框架. 在训练过程中, 将同一批次的正常样本 $\mathbf{x}^{\text{orig}}$ 与对应的对抗样本 $\mathbf{x}^{\text{adv}}$ 同时输入至特征提取器 E , 将所得隐层样本分别输入至 DLSS 模型后段 B 与特征锚定器 A , 计算主导变量预测损失 L'_p 和锚定损失 L'_r , 如式 (9) 和 (10) 所示.

$$L'_p = \sum_{\mathbf{x}_i^{\text{orig}}, y_i^{\text{orig}} \in D^{\text{orig}}} \text{MSE}(B(E(\mathbf{x}_i^{\text{orig}})), y_i^{\text{orig}}) + \sum_{\mathbf{x}_i^{\text{adv}} \in D^{\text{adv}}, y_i^{\text{orig}} \in D^{\text{orig}}} \text{MSE}(B(E(\mathbf{x}_i^{\text{adv}})), y_i^{\text{orig}}) \quad (9)$$

$$L'_r = \sum_{\mathbf{x}_i^{\text{orig}} \in D^{\text{orig}}} \text{MSE}(A(E(\mathbf{x}_i^{\text{orig}})), \mathbf{x}_i^{\text{orig}}) + \sum_{\mathbf{x}_i^{\text{orig}} \in D^{\text{orig}}, \mathbf{x}_i^{\text{adv}} \in D^{\text{adv}}} \text{MSE}(A(E(\mathbf{x}_i^{\text{adv}})), \mathbf{x}_i^{\text{orig}}) \quad (10)$$

其中, D^{adv} 为对抗样本集; L'_p 用于约束模型在对抗

扰动下仍能提取出与对应正常样本接近的隐层特征分布, 从而提升模型对抗鲁棒性; L'_r 作为强约束项, 用于约束特征提取器 E 对正常样本与对抗样本的隐层特征提取过程, 使提取的隐层特征分布不会偏离未进行对抗训练前 DLSS 模型前段 F 所提取的正常样本隐层特征分布, 从而缓解对抗鲁棒过拟合现象. 两种损失函数共同构成总损失, 用于联合优化特征提取器参数 θ^e , 如式 (11) 所示, 其中 λ 为平衡系数.

$$\arg \min_{\theta^e} (L'_p + \lambda L'_r) \quad (11)$$

在完成特征提取器训练后, 将其与参数和结构固定的 DLSS 模型后段进行拼接, 得到最终的对抗鲁棒 DLSS 模型.

2.3 证明过程

由于对抗样本与正常样本在输入空间和特征空间中的分布存在差异, 引入对抗样本参与训练意味着模型需要在优化过程中同时平衡两类样本的损失优化. 因此, 模型参数的最优解不再对应于仅正常样本参与训练得到的最优参数 θ^* , 而是偏移至对抗训练后能同时拟合两类样本分布的最优参数 θ^{adv} . 该参数偏移改变了模型对正常样本的拟合方向, 进而导致对抗训练后的 DLSS 对正常样本的主导变量预测精度下降. 为分析对抗训练引起的参数偏移如何影响模型输出, 从而揭示参与对抗训练的对抗样本扰动大小与对抗训练后 DLSS 自然预测精度之间的关系, 下面从模型函数的参数依赖关系入手进行推导.

证明. DLSS 的表达式 $f_{\theta}(x)$ 是关于输入 x 的函数, 但同时也依赖模型参数 θ , 因此可以表示为 $f(\theta, x)$. 若将该双变量函数中的 x 固定, 则可以表示为一个关于模型参数 θ 的函数 $g(\theta) = f(\theta, x)$, 此式表示对相同的输入 x , 模型参数 θ 的变化会影响模型的输出结果. 由于 DLSS 领域的对抗攻击具有较强的隐蔽性, 其扰动通常较小, 因此对抗训练得到的模型参数相较于正常训练得到的模型参数, 发生的偏移属于小参数偏移, 并且在训练时采用平滑激活函数和权重衰减的优化策略, 函数可微且 Hessian 有界. 当 DLSS 参数为 θ^{adv} 时, $f_{\theta}(x)$ 可在 θ^* 处作一阶泰勒展开, 得到:

$$f_{\theta^{\text{adv}}}(x) = g(\theta^{\text{adv}}) \approx g(\theta^*) + \nabla_{\theta} g^T(\theta^*)(\theta^{\text{adv}} - \theta^*) \quad (12)$$

其中, $\nabla_{\theta} g^T(\theta^*) = \nabla_{\theta} f_{\theta^*}^T(x)$, 代入式 (12) 可得:

$$f_{\theta^{\text{adv}}}(x) \approx f_{\theta^*}(x) + \nabla_{\theta} f_{\theta^*}^T(x)(\theta^{\text{adv}} - \theta^*) \quad (13)$$

因此, 对抗训练后的 DLSS 相较于正常训练的 DLSS 在正常样本上的主导变量预测偏差为:

$$|f_{\theta^{\text{adv}}}(x) - f_{\theta^*}(x)| \approx |\nabla_{\theta} f_{\theta^*}^T(x)(\theta^{\text{adv}} - \theta^*)| \approx |\nabla_{\theta} f_{\theta^*}^T(x)| \cdot |\Delta\theta| \quad (14)$$

当 DLSS 模型训练时, 参数更新如式 (15) 所示. 其中, θ^0 为 DLSS 模型初始参数, α 为学习率.

$$\theta^* = \theta^0 - \alpha E_{(x, y)} [\nabla_{\theta} L(f_{\theta}(x), y)] \quad (15)$$

当 DLSS 模型对抗训练时, 模型参数更新为:

$$\theta^{\text{adv}} = \theta^0 - \alpha E_{(x, y)} [\nabla_{\theta} L(f_{\theta}(x + \delta), y)] \quad (16)$$

式 (16) 与式 (15) 相减可得到:

$$\Delta\theta = \theta^{\text{adv}} - \theta^* = -\alpha E_{(x, y)} [\nabla_{\theta} L(f_{\theta}(x + \delta), y) - \nabla_{\theta} L(f_{\theta}(x), y)] \quad (17)$$

由式 (17) 可得:

$$|\Delta\theta| \propto E_{(x, y)} [|\nabla_{\theta} L(f_{\theta}(x + \delta), y) - \nabla_{\theta} L(f_{\theta}(x), y)|] \quad (18)$$

假设损失函数可微, 对扰动点做一阶展开:

$$\nabla_{\theta} L(f_{\theta}(x + \delta), y) \approx \nabla_{\theta} L(f_{\theta}(x), y) + \nabla_x \nabla_{\theta} L(f_{\theta}(x), y) \cdot \delta \quad (19)$$

由式 (19) 可得:

$$|\nabla_{\theta} L(f_{\theta}(x + \delta), y) - \nabla_{\theta} L(f_{\theta}(x), y)| \approx |\nabla_x \nabla_{\theta} L(f_{\theta}(x), y)| \cdot \delta \quad (20)$$

由 (20) 可得, 当 $|\delta|$ 增大时, $|\Delta\theta|$ 增大, $|f_{\theta^{\text{adv}}}(x) - f_{\theta^*}(x)|$ 增大. $\square$

结合上述推导, 可以得出结论: 参与对抗训练的对抗样本扰动越小, 对抗训练后的模型在正常样本上的预测精度下降越少. 故本文基于第 2.1 节所述的阈值划分方式, 仅将扰动小于标准阈值的对抗样本 (即 SPAS) 用于对抗训练. 剩余的对抗样本 LPAS 由于扰动较大, 尝试直接对其进行检出, 而不参加对抗训练, 以免导致模型经训练后在正常样本上的预测精度下降. 因此, 第 2.1 节所述的 ATADI 防御方法贴合所证明的理论设计的对抗防御框架, 所述的阈值划分方式则为开展 ATADI 防御做好了数据准备.

3 实验验证与分析

3.1 案例介绍

空预器是电站锅炉的关键设备, 用于回收排烟

余热以预热助燃空气,从而保障锅炉出力并提高燃烧效率.然而在实际运行中,空预器转子因受热会产生“蘑菇状”变形,使顶部扇形板与转子之间出现较大漏风间隙,造成助燃风流失,导致机组降负荷运行,严重影响运行经济性.为解决这一问题,目前普遍采用硬件间隙传感器对漏风间隙进行非接触式测量,并在与间隙设定值比较后通过调节扇形板高度来控制漏风.但该硬件传感器在高温高粉尘环境下易损坏,且只能在停炉检修时更换.为此,当前通常通过构建能够准确预测转子热变形的 DLSS 模型,并基于预测得到的热变形量,进一步推算关键运行间隙,使间隙控制系统在硬件传感器失效条件下仍能够实现自动调节,从而有效抑制漏风现象^[27].

尽管现有转子热变形 DLSS 在测量精度上已满足要求,但此类智能监测模型正在遭受对抗攻击的严重威胁.虽然当前电站的空预器间隙控制系统尚未有遭受对抗攻击的公开报道,但该类电站所在的电力系统中已出现数据劫持与虚假数据注入攻击的真实安全事件^[28-29].虚假数据注入攻击与对抗攻击同属数据欺骗攻击范畴,均通过篡改传感器数据误导系统决策.当前工业控制系统与电站企业网、互联网的边界日益模糊,系统攻击面显著扩大.随着 DLSS 在空预器间隙控制等关键控制回路中的部署,具有实施虚假数据注入攻击能力的攻击者在掌握对抗攻击理论后,完全可能将攻击手段从虚假数据注入攻击升级为对抗攻击,故必须在 DLSS 的构建阶

段提升其对抗鲁棒性.

具体到本案例,攻击者的能力边界设定如下:基于 DLSS 黑盒对抗攻击的普遍假设,攻击者无法获取部署于工业服务器上的转子热变形 DLSS 的模型架构、训练参数及训练数据.但攻击者具有工控网络渗透能力和对抗攻击知识储备,通过劫持在工业服务器、分散控制系统及可编程逻辑控制器通信链路中长期运行的温度-热变形对应数据,构建代理 DLSS 模型.在此基础上生成具有迁移性的对抗样本,并利用 Modbus 或 OPC 协议漏洞将对抗样本注入传感器数据链路.攻击者通过篡改空气、烟气进出口温度测量值的方式,导致 DLSS 输出虚假的转子热变形预测结果,进而通过漏风间隙控制柜生成错误的执行指令,驱动间隙调节执行器对扇形板进行误调节.如图 3 所示,若扇形板被误调至上极限位置,将导致顶部间隙过大从而显著增加漏风,使间隙控制失效并造成重大经济损失;若被误调至下极限位置,则会导致扇形板与转子激烈摩擦,造成设备机械损毁甚至引发机组非计划停机.类似的机械事故在工业现场已有先例:当传统硬件传感器发生漂移故障或早期浅层软测量模型出现系统性偏差时,均导致扇形板过度调节引发设备卡死或漏风率激增^[30].对抗攻击作为更具隐蔽性的数据欺骗手段,其作用于控制回路后的物理影响与传统故障在机理上等效,且因其自身扰动微小,更难以被及时检测.因此,针对空预器转子热变形 DLSS 开展预

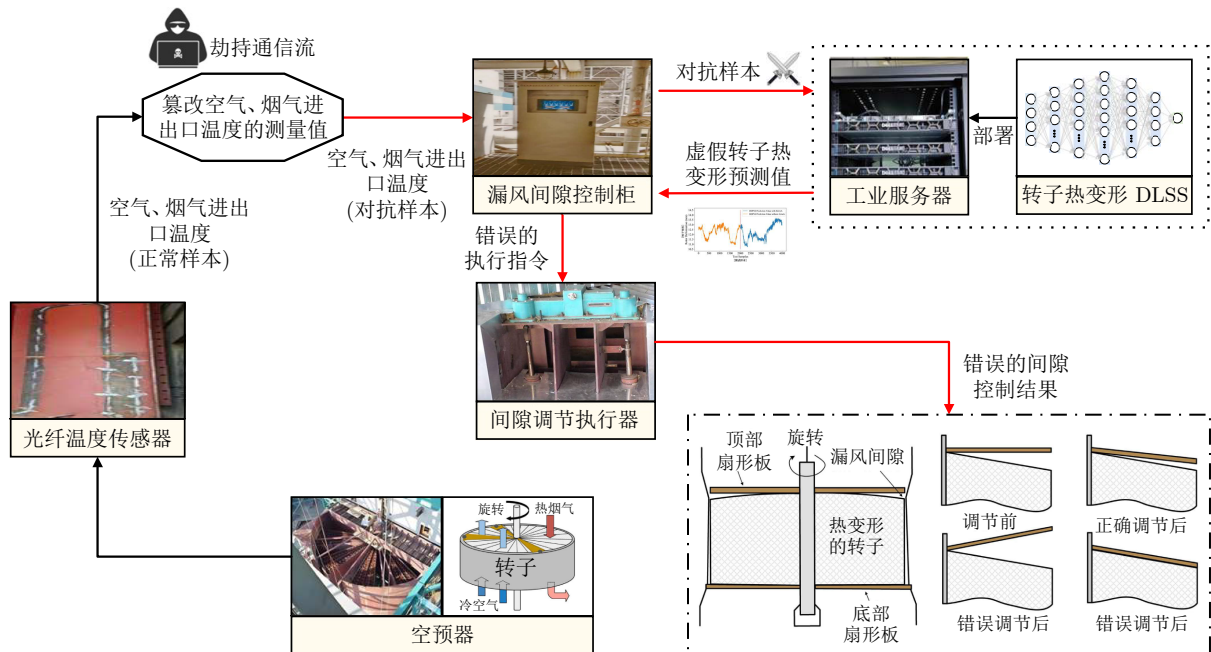


图 3 转子热变形 DLSS 的对抗攻击过程及攻击结果示意图

Fig. 3 Adversarial attack process and attack results diagram of rotor thermal deformation DLSS

防性的对抗防御研究, 对保障电站关键基础设施的安全、经济运行具有重要意义. 第 4 节将在此案例上对所提出的 ATADI 对抗防御方法的有效性进行验证.

3.2 参数设置与指标选定

首先, 构建转子热变形 DLSS 模型, 模型的 4 个辅助变量为入口烟气温度、出口空气温度、入口空气温度和出口烟气温度, 主导变量为转子热变形值. DLSS 采用全连接前馈神经网络, 网络结构为 4-16-32-64-32-4-1, 且其预测精度满足工程实际需求. 隐藏层神经元的激活函数均采用平滑激活函数 Sigmoid. 模型的损失函数为预测值和实际测量值的均方误差, 使用 Adam 优化器训练. 本文所有实验均在搭载 NVIDIA GeForce RTX 3060 Ti 显卡的计算机上完成. 总样本数为 23 000, 涵盖时间范围为从 0 到 23 000 的序列数据, DLSS 训练、对抗样本生成、防御框构建与测试样本选取均按时间顺序进行, 相互之间没有重叠. DLSS 训练集由前 8 000 条样本构成, 选取时间区间 8 000 到 10 000 的样本作为 DLSS 测试集. 对时间区间 8 000 到 23 000 的样本分别施加四种攻击机制与隐蔽性强弱不同的对抗攻击, 包括 IDAO^[9]、KGAA^[14]、KAGAN^[15] 与 HGAA^[17], 以生成四种对抗样本. 如前文第 2.1 节所述, 针对上述对抗样本, 采用基于 BCD 方法构建的检测器在时间区间 8 000 到 18 000 内的各类对抗样本上进行检测实验, 将所有漏检对抗样本的最大扰动幅值作为 LPAS 与 SPAS 的划分阈值, 并将其定义为该攻击类型下的标准阈值, 检测器的具体参数参照文献 [23] 确定. 据此, 四种对抗攻击对应的标准阈值得以确定, IDAO 为 0.040、KGAA 为 0.035、KAGAN 为 0.032、HGAA 为 0.030. 以 IDAO 为例, 将 IDAO 攻击下生成的对抗样本中扰动幅值大于 0.040 的样本标记为 LPAS, 将扰动幅值小于或等于 0.040 的样本标记为 SPAS.

本文亦对 KGAA 攻击下标准阈值的取值进行了敏感度分析, 以探讨标准阈值取值的变化对 BCD 检测性能和对抗训练后 DLSS 预测精度的影响. 需要说明的是, 如第 2.1 节中的标准阈值确定方法所述, 对 LPAS 与 SPAS 划分的标准阈值并非传统意义上的可调超参数, 而是由 BCD 检测器在特定类型攻击下的检测结果所唯一确定的最优判别边界. 如果 BCD 方法中用于判断当前样本是否为对抗样本的判别阈值发生改变, 则在特定类型攻击下的标准阈值亦会发生变化. 对标准阈值进行灵敏度分析, 即对标准阈值尝试取更大或更小的值并观察其对防御方法性能的影响. 但是, 更小的标准阈值

意味着 BCD 方法中的判别阈值变得更小, 由于更小的判别阈值会将更多正常样本误检为对抗样本, 故 BCD 方法在参数设置时并未考虑将判别阈值设定过小. 在本文的分析中, 仅考虑 BCD 方法的判别阈值增大的情况, 因此仅对标准阈值尝试取更大的值进行灵敏度分析. 由于 BCD 检测器对 LPAS 的高召回率 (Recall) 是 ATADI 得以发挥效用的前提, 因此, 以标准阈值 0.032 为基准、0.004 为步长, 总共选取 7 个阈值进行实验, 得出在这 7 个阈值下的召回率取值, 绘制如图 4 所示的折线图.

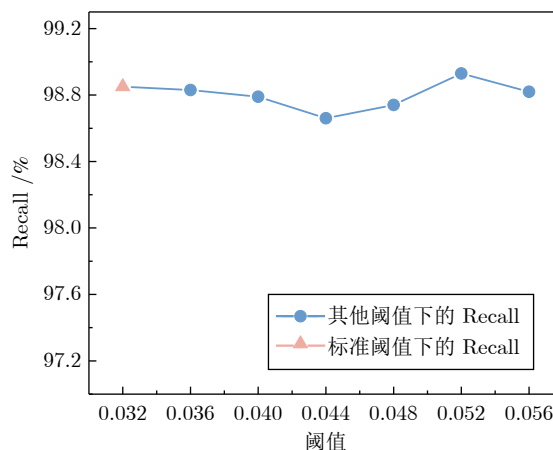


图 4 ATADI 对抗防御方法在不同阈值下的召回率折线图

Fig. 4 Line plot of Recall for ATADI adversarial defense method under different thresholds

同理, 绘制出如图 5 所示的 DLSS 预测精度折线图. 由图 4 可知, 若取标准阈值大于 0.032, 无论比 0.032 大多少, 召回率均在标准阈值所对应的召回率值 98.95% 上下浮动, 总体变化不大. 观察图 5 可知, 若取标准阈值大于 0.032, 随着阈值的取值越来越大, 将导致平均绝对误差 (mean absolute error, MAE) 不断增高. 这是因为阈值越大, 就有越多原本被划分为 LPAS 的对抗样本被误分为 SPAS 并用于对抗训练, 使用这些扰动较大的样本进行对抗训练, 降低了对抗鲁棒性, 致使 DLSS 的预测精度下降. 综上分析, 标准阈值的取值变大虽然不会明显影响 BCD 检测模型的召回率, 但会造成 DLSS 的预测精度下降, 因此, 依据第 2.1 节中所述方法确定的标准阈值能够作为最优选择, 以确保 ATADI 的防御效果. 将每类对抗样本中时间区间在 8 000 到 18 000 的 SPAS 与其对应的正常样本作为训练集, 经 SPAS-FAAT 后分别得到针对四种攻击类型的对抗鲁棒 DLSS 模型. 具体来说, 将通道序列 4-16-32-64 视为 DLSS 模型的前段, 将通道序列 32-4-1 视为后段. 构建一个结构为 64-32-16-8-4 的五层

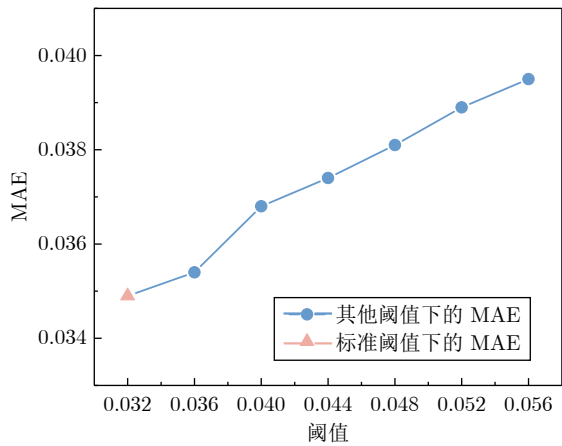


图 5 ATADI 对抗防御方法在不同阈值下的主导变量预测误差折线图

Fig. 5 Line plot of prediction error of dominant variable for ATADI adversarial defense method under different thresholds

全连接网络作为特征锚定器, 隐藏层均采用 ReLU 激活, 输出层为线性映射以生成 4 维重构结果. 令 DLSS 模型前段随机初始化, 作为可训练的特征提取器, 并固定特征锚定器与 DLSS 模型后段的结构与参数. 在 SPAS-FAAT 过程中, 仅优化特征提取器的参数, 并同样采用 Adam 优化器, 平衡系数 λ 取值为 0.6. 训练完成后, 将优化得到的特征提取器与固定的 DLSS 模型后段拼接, 构成最终的对抗鲁棒 DLSS 模型. 将训练得到的四种对抗鲁棒 DLSS 模型分别与检测器结合, 构成针对四类对抗攻击的 ATADI 防御方法. 进一步, 将时间区间在 18 000 到 23 000 的正常样本与其对应的四类对抗样本作为测试集, 对 ATADI 对抗防御方法的有效性进行验证和评估. 鉴于在该防御方法中检测器的任务转变为仅识别 LPAS, 本文首先以准确率 (Accuracy)、精确率 (Precision) 和召回率作为评价指标, 对检测器识别 LPAS 的性能进行了评估. 指标的公式如下:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (21)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (22)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (23)$$

其中, TP 是正确分类为 LPAS 的个数, TN 是正确分类为正常样本的个数, FP 是错误分类为 LPAS 的个数, FN 是错误分类为正常样本的个数. Accuracy 反映检测器对可识别 LPAS 的整体正确率, 用于衡量整体性能; Precision 表示被检测器预测为对

抗样本的样本中真正为对抗样本的比例, 用于反映误报情况; Recall 衡量检测器成功检出的对抗样本占全部对抗样本的比例, 用于反映漏检情况.

随后, 比较 ATADI 对抗防御框架下, SPAS-FAAT 前后 DLSS 在正常样本与 SPAS 上的主导变量预测精度, 以验证在 ATADI 对抗防御中的 SPAS-FAAT 方法在提升模型对抗鲁棒性方面的有效性. 具体而言, 采用 DLSS 预测值与实际测量值之间的 MAE 作为评价指标, 其计算公式如下:

$$\text{MAE} = \frac{1}{z} \sum_{l=1}^z |y_{\text{pred}}^{(l)} - y_{\text{measured}}^{(l)}| \quad (24)$$

其中, y_{pred} 为 DLSS 的预测值, y_{measured} 为对应的实际测量值, z 为样本数量. 若 SPAS-FAAT 得到的 DLSS 模型在正常样本上的预测精度得以保持, 同时在对抗样本上的 MAE 低于 SAT 后的 DLSS 模型, 说明该方法在提升 DLSS 对抗鲁棒性方面具有优势.

4 ATADI 对抗防御方法的实验结果与分析

4.1 对 LPAS 进行检测的实验结果与分析

对 ATADI 框架下 BCD 检测器的检测性能进行评估. 将包含四种攻击方法所生成 LPAS 的测试集样本输入 BCD 检测器进行测试, 实验结果如表 1 所示.

表 1 BCD 检测器对四类攻击所产生 LPAS 的检测结果 (%)

Table 1 Detection results of the BCD detector for LPAS generated by four types of attacks (%)

攻击方法	Accuracy	Precision	Recall
IDAO	99.63	98.95	99.45
HGAA	98.76	97.55	98.30
KGAA	99.21	98.40	98.95
KAGAN	98.42	97.10	97.95

由上表可以看出, BCD 检测器在四种不同的对抗攻击下均表现出较为稳定且可靠的检测能力: 从整体检测的 Accuracy 来看, 四种攻击下的 Accuracy 均保持在 98.42% 以上, 说明虽然四种攻击的产生机制不同, 但无论对于哪种攻击, 检测器均能够维持较高的总体判别准确性. 四类攻击的 Recall 均处于 97.95% 到 99.45% 区间内, 尤其在 IDAO 攻击下达到 99.45%. 这说明所提方法漏检率较低, 能够有效识别多数 LPAS. 虽然检测器对 LPAS 仍

有约 1.34% 的漏检, 但这部分漏检的 LPAS 样本通常是所有 LPAS 中扰动幅值较小的部分, 并且由于空预器间隙控制系统的调节策略中引入了包含两个调节周期的滞后启动机制, 1.34% 左右的漏检样本难以在连续周期内持续触发有效的调节指令, 因此一般不会对实际间隙调节过程造成显著影响. Precision 稳定保持在 97.10% 以上, 表明在对抗样本能够被准确检测的前提下, 并未引入过多误报, 检测方法在召回率与精确率之间取得了较为合理的平衡. 综上所述, 在 ATADI 对抗防御框架下, BCD 检测器能够有效检出 LPAS, 为后续开展 SPAS-FAAT 对抗训练防御奠定了基础.

4.2 对 SPAS 开展防御的实验结果与分析

本节对 ATADI 对抗防御框架中 SPAS-FAAT 对抗训练的有效性进行测试, 并与标准对抗训练方法 SAT 进行了对比. 经 SPAS-FAAT 后的 DLSS 表示为 $DLSS_{SPAS-FAAT}$, 经 SAT 后的 DLSS 表示为 $DLSS_{SAT}$. 实验中 $DLSS_{SAT}$ 采用标准对抗训练策略, 其优化目标仅由软测量任务损失构成, 即通过同时最小化正常样本与其对应对抗样本的预测误差来完成训练. 相比之下, $DLSS_{SPAS-FAAT}$ 在任务损失的基础上引入了特征锚定机制, 在训练过程中显式约束对抗样本与其对应正常样本在隐层特征空间中的特征一致性. 因此, $DLSS_{SPAS-FAAT}$ 的损失函数由主导变量预测损失和特征锚定损失两部分组成. 为保证对比结果的公平性, $DLSS_{SPAS-FAAT}$ 与 $DLSS_{SAT}$ 在训练过程中采用了完全一致的训练轮数、优化器配置, 且两者在针对特定对抗攻击进行对抗训练时, 采用相同的攻击方法生成对抗样本, 对抗扰动设置在 0.08 以内, 使温度的变化处于允许的正常波动范围. $DLSS_{SAT}$ 的对抗训练耗时为 12.6 s, $DLSS_{SPAS-FAAT}$ 的对抗训练耗时为 18.3 s. 转子热变形 DLSS 的对抗训练均为离线进行, 其在线测量时神经网络前向计算耗时均在 2.8 ms 内, 实时性完全满足实际需求.

针对 ATADI 对抗防御方法, 本文测试了 SPAS-FAAT 策略在增强 DLSS 模型对 SPAS 对抗鲁棒性方面的性能, 并与 SAT 进行了对比分析. 图 6 和图 7 分别展示了在 KGAA 与 KAGAN 对抗攻击下, $DLSS_{SPAS-FAAT}$ 与 $DLSS_{SAT}$ 对 SPAS 的主导变量预测曲线对比.

可以观察到, $DLSS_{SAT}$ 预测曲线整体偏离了实际测量值曲线, 表现出对 SPAS 的脆弱性, 而 $DLSS_{SPAS-FAAT}$ 预测曲线与实际测量值曲线高度接近, 表明其对 SPAS 具备更强的防御能力. 定量结

果如表 2 所示, 测试样本包括 x_{SPAS}^{IDAO} 、 x_{SPAS}^{KGAA} 、 x_{SPAS}^{HGAA} 和 x_{SPAS}^{KAGAN} , 分别表示由 IDAO、KGAA、HGAA 和

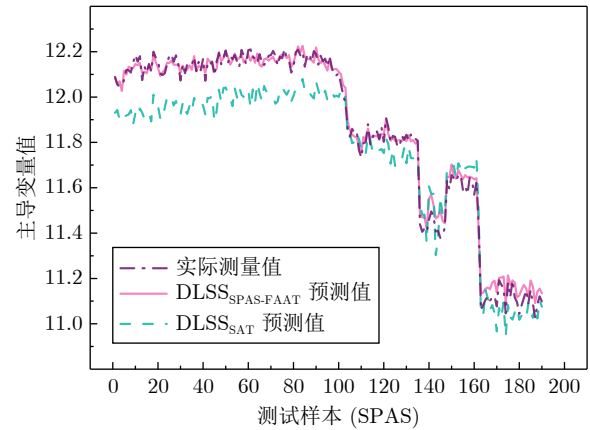


图 6 KGAA 攻击下实施对抗防御后的 DLSS 对抗鲁棒性预测曲线 (输入 SPAS)

Fig.6 Adversarial robustness prediction curves of the DLSS under KGAA attacks after implementing the adversarial defense (SPAS input)

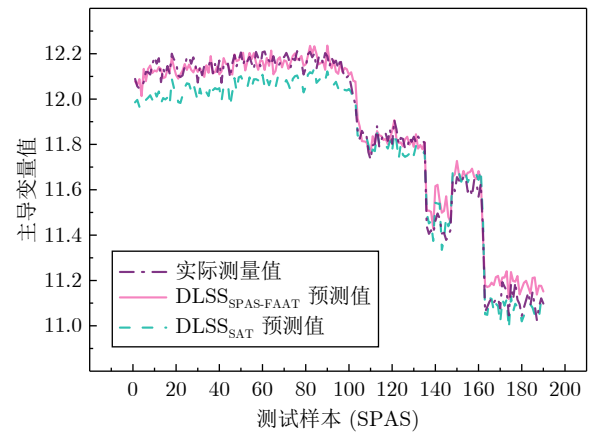


图 7 KAGAN 攻击下实施对抗防御后的 DLSS 对抗鲁棒性预测曲线 (输入 SPAS)

Fig.7 Adversarial robustness prediction curves of the DLSS under KAGAN attacks after implementing the adversarial defense (SPAS input)

表 2 四种不同对抗攻击下的 DLSS 对抗鲁棒性测试指标 (输入 SPAS)

Table 2 DLSS adversarial robustness test metrics under four different adversarial attacks (SPAS input)

测试样本	MAE		
	$DLSS_{SPAS-FAAT}$	$DLSS_{Original}$	$DLSS_{SAT}$
x_{SPAS}^{IDAO}	0.0282	0.0421	0.0391
x_{SPAS}^{KGAA}	0.0197	0.0415	0.0310
x_{SPAS}^{HGAA}	0.0136	0.0446	0.0267
x_{SPAS}^{KAGAN}	0.0263	0.0375	0.0287

KAGAN 攻击生成的 SPAS 样本. $DLSS_{SPAS-FAAT}$ 在四类典型对抗攻击生成的 SPAS 下的误差指标均显著低于 $DLSS_{SAT}$ 与原始深度学习软测量模型 $DLSS_{Original}$, 表明无论是对基础 DLSS 对抗攻击方法, 如 KGAA 和 IDAO, 还是对强隐蔽性 DLSS 对抗攻击方法, 如 HGAA 和 KAGAN, 所提出的对抗训练方法均能够显著提升 DLSS 对 SPAS 的对抗鲁棒性. 从不同攻击类型的对比结果看, $DLSS_{SPAS-FAAT}$ 在 KAGAN 与 HGAA 这两种攻击下对 DLSS 的鲁棒性提升尤为明显. 以 HGAA 为例, 其误差指标仅为 0.0136, 较原始 DLSS 降低约 69.5%, 较 SAT 防御方法降低约 49.1%, 说明在强隐蔽性对抗攻击下, SPAS-FAAT 能够更有效地抑制对抗扰动对模型预测的影响.

总体而言, 图 6、图 7 与表 2 的实验结果验证了在 ATADI 对抗防御框架下, SPAS-FAAT 方法在不同扰动强度与攻击模式下均能稳定降低 DLSS 对 SPAS 的预测误差, 使训练后的模型在对抗鲁棒性上显著优于传统 SAT 方法下得到的模型.

4.3 缓解对抗鲁棒过拟合的实验结果与分析

在对抗鲁棒性提升的基础上, 还应确保对抗训练后的 DLSS 不存在严重的对抗鲁棒过拟合. 因此, 本节进一步进行实验, 观察经 SPAS-FAAT 后, DLSS 模型在正常样本作为输入时其预测精度的变化情况, 并与传统 SAT 方法进行对比. 相关实验配置与第 4.2 节保持一致, 图 8 与图 9 分别展示了在 KGAA 与 KAGAN 对抗攻击下, $DLSS_{SPAS-FAAT}$ 与 $DLSS_{SAT}$ 以正常样本作为输入时的主导变量预测曲线.

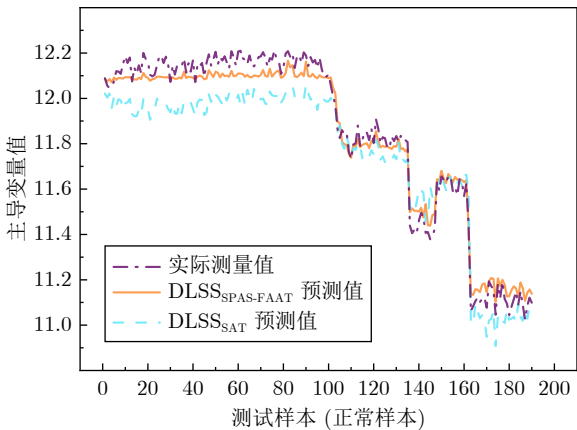


图 8 KGAA 攻击下实施对抗防御后的 DLSS 对抗鲁棒性预测曲线 (输入正常样本)

Fig.8 Adversarial robustness prediction curves of the DLSS under KGAA attacks after implementing the adversarial defense (normal sample input)

由图 8 和图 9 可知, 当正常样本作为输入时, $DLSS_{SAT}$ 的预测曲线在整体幅值和局部变化趋势上相较实际测量值均出现明显偏移, 表明传统 SAT 方法在提升对抗鲁棒性的同时产生了较为严重的对抗鲁棒过拟合问题. 相比之下, $DLSS_{SPAS-FAAT}$ 的预测曲线在整体趋势和数值水平上均与实际测量值保持较好的一致性, 未出现明显的数值偏移. 这说明 SPAS-FAAT 能抑制正常样本预测精度的退化, 并有效缓解了对抗鲁棒过拟合. 除了将 KGAA 和 KAGAN 的测试指标结果列出之外, 表 3 亦将 IDAO 和 HGAA 的测试指标结果一同列出, 测试样本包括 x_{normal}^{IDAO} 、 x_{normal}^{KGAA} 、 x_{normal}^{HGAA} 和 x_{normal}^{KAGAN} , 均为正常样本, 分别对应 IDAO、KGAA、HGAA 和 KAGAN 攻击类型下的实验设置. 观察表 3 实验结果可知, 无论攻击方法如何, $DLSS_{SAT}$ 的 MAE 相比 $DLSS_{Original}$ 进一步增大, 即 $DLSS_{SAT}$ 在正常样本上的预测精度下降, 而 $DLSS_{SPAS-FAAT}$ 在正常样本上的预测精度相比 $DLSS_{Original}$ 仅有微小变化. 这从定量的角度说明了无论面对基础 DLSS 对抗攻

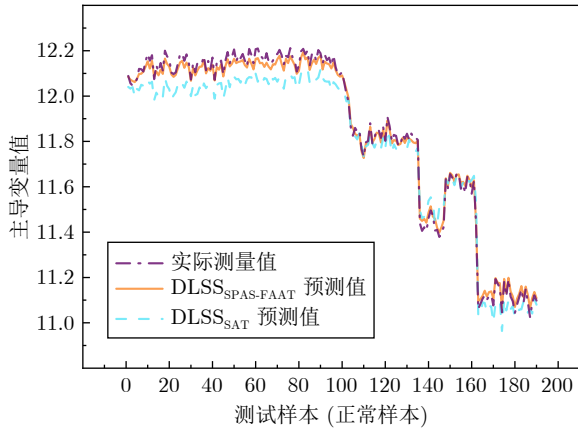


图 9 KAGAN 攻击下实施对抗防御后的 DLSS 对抗鲁棒性预测曲线 (输入正常样本)

Fig.9 Adversarial robustness prediction curves of the DLSS under KAGAN attacks after implementing the adversarial defense (normal sample input)

表 3 四种不同对抗攻击下的 DLSS 对抗鲁棒性测试指标 (输入正常样本)

Table 3 DLSS adversarial robustness test metrics under four different adversarial attacks (normal sample input)

测试样本	MAE		
	$DLSS_{SPAS-FAAT}$	$DLSS_{Original}$	$DLSS_{SAT}$
x_{normal}^{IDAO}	0.0386	0.0391	0.0410
x_{normal}^{KGAA}	0.0349	0.0341	0.0361
x_{normal}^{HGAA}	0.0338	0.0329	0.0358
x_{normal}^{KAGAN}	0.0328	0.0326	0.0349

击方法还是强隐蔽性对抗攻击方法, SPAS-FAAT 均能有效缓解对抗鲁棒过拟合现象。

结合第 4.1 ~ 4.3 节的实验结果与分析可知, ATADI 方法中的检测器能够有效检出 LPAS。同时, 该方法中的 SPAS-FAAT 能够使 DLSS 对 SPAS 具备对抗鲁棒性, 且仅损失 DLSS 在正常样本上一小部分的预测精度。综上所述, 在使用 ATADI 方法开展对抗防御时, 对抗鲁棒过拟合问题得到有效缓解, 并且 DLSS 的对抗鲁棒性得到了有效提升。

4.4 结构消融实验的实验结果与分析

第 4.1 ~ 4.3 节从功能上验证了 ATADI 的有效性, 本节将通过结构消融实验从结构上进一步验证 ATADI 防御中攻击检测与对抗训练两个环节各自发挥的作用。因此, 结构消融实验应包含仅开展攻击检测的 ATADI 和仅开展对抗训练的 ATADI 的对比, 实验均采用 KGAA 作为攻击方法。首先, 对仅开展攻击检测的 ATADI 进行实验。大多数 LPAS 被检测器检出, SPAS 和极少数未被检测出的 LPAS 则输入未对抗训练的 DLSS 模型, 预测误差为 0.041 5。当对仅开展对抗训练的 ATADI 进行实验时, 因为缺少检测器对 LPAS 的检出, 全部对抗样本输入 SPAS-FAAT 后的 DLSS 模型, 预测误差为 0.063 4。而在开展完整的 ATADI 防御时, LPAS 被检测器检出, 而 SPAS 则输入 SPAS-FAAT 后的 DLSS 模型, 预测误差为 0.019 7, 远小于仅开展攻击检测的 ATADI 和仅开展对抗训练的 ATADI 的预测误差。因此, 可以得出结论: 所提 ATADI 对抗防御方法中的攻击检测方法 with 对抗训练方法在单独使用时均难以取得理想的防御效果, 而二者结合使用时能够显著提升 DLSS 整体防御性能。

对于 ATADI 防御框架中 SPAS-FAAT 方法, 由第 2.3 节可知, 其在训练时仅采取 SPAS 即可有效提升 DLSS 对抗鲁棒性。然而 FAAT 方法在 SPAS-FAAT 中发挥的作用仍有待验证。因此, 本文通过将 SPAS-FAAT 方法中的 FAAT 替换为 SAT 与 DAAT, 以验证 FAAT 在 SPAS-FAAT 方法中的作用。实验结果如表 4 所示。

在本文防御框架下, 对抗训练环节采用 SPAS 进行 SAT 与 DAAT 后的 DLSS 模型分别表示为 $DLSS_{SPAS-SAT}$ 和 $DLSS_{SPAS-DAAT}$ 。结果显示, $DLSS_{SPAS-FAAT}$ 在正常样本与 SPAS 上的预测精度明显高于 $DLSS_{SPAS-SAT}$, 并优于 $DLSS_{SPAS-DAAT}$, 即使用 FAAT 进行对抗训练后得到的 DLSS 在正常样本和 SPAS 作为输入时均实现了更优的指标。

表 4 采用不同对抗训练方法的 DLSS 对抗鲁棒性测试结果

Table 4 Adversarial robustness test results of DLSS with different adversarial training methods

测试样本	MAE		
	$DLSS_{SPAS-SAT}$	$DLSS_{SPAS-DAAT}$	$DLSS_{SPAS-FAAT}$
SPAS	0.023 8	0.021 7	0.019 7
正常样本	0.036 1	0.035 1	0.034 9

这是因为相比于 SAT 与 DAAT, FAAT 引入了特征锚定机制, 在训练过程中显式约束对抗样本与其对应的正常样本在隐层特征空间中的特征一致性, 从而使 DLSS 提升对 SPAS 防御能力的同时, 维持对正常样本的预测精度。实验结果显示 FAAT 在提升对抗鲁棒性方面也发挥了重要作用。

5 结束语

本文提出一种 ATADI 对抗防御方法, 以应对 DLSS 在开展对抗防御时面临的对抗鲁棒过拟合、SPAS 难以检出的问题。实验结果显示, 将对抗训练与攻击检测进行融合以开展防御, 这巧妙地规避了各自的不足, 有效提升了 DLSS 的对抗防御能力。在各类攻击下, ATADI 框架中的检测器均能以不低于 98.42% 的准确率检出 LPAS。相比于其他对抗训练方法, SPAS-FAAT 后的 DLSS 更具对抗鲁棒性, 难以检出的 SPAS 输入该 DLSS 后能够取得更优的软测量结果, 表现为 DLSS 不仅在 SPAS 上实现了高精度预测, 而且在正常样本上的主导变量预测精度也能够符合实际要求, 有效克服了对抗鲁棒过拟合问题。需要指出的是, 本文提出的 ATADI 方法作为一种新颖的框架, 在面对未知类型对抗攻击时的泛化性仍有待研究。同时, 研究对抗样本的净化方法以进一步处理检出的 LPAS, 有助于完善 ATADI 对抗防御框架, 这将作为今后的研究方向。

附录 A

表 A1 缩写注解表

Table A1 List of abbreviations

缩写	含义
DLSS	深度学习软测量
SPAS	微小扰动对抗样本
LPAS	较大扰动对抗样本
BCD	双向一致性判别
FAAT	特征锚定式对抗训练

表 A1 缩写注解表 (续表)

Table A1 List of abbreviations (continued table)

缩写	含义
ATADI	对抗训练与攻击检测融合
SAT	标准对抗训练
DAAT	域自适应对抗训练
KGAA	知识引导型对抗攻击
SPAS-FAAT	基于微小扰动对抗样本特征锚定式对抗训练
DLSS _{Original}	原始深度学习软测量模型
DLSS _{SAT}	标准对抗训练后得到的深度学习软测量模型
DLSS _{DAAT}	域自适应对抗训练后得到的深度学习软测量模型
DLSS _{SPAS-FAAT}	基于微小扰动对抗样本特征锚定式对抗训练后得到的深度学习软测量模型

参考文献

1 Guo R Y, Liu H, Xie G, Zhang Y M, Liu D. A self-interpretable soft sensor based on deep learning and multiple attention mechanism: From data selection to sensor modeling. *IEEE Transactions on Industrial Informatics*, 2023, **19**(5): 6859–6871

2 Shi X D, Li R H, Morales H, Huang W, Xiong W L. Semi-supervised probabilistic learning network for soft sensor modeling with partially labeled data. *IEEE Transactions on Automation Science and Engineering*, 2025, **22**: 16309–16321

3 Qiao Jing-Hui, Chai Tian-You. Data and model-based soft measurement model of cement raw meal decomposition ratio. *Acta Automatica Sinica*, 2019, **45**(8): 1564–1578 (乔景慧, 柴天佑. 数据与模型驱动的水泥生料分解率软测量模型. *自动化学报*, 2019, **45**(8): 1564–1578)

4 Meng Xi, Zhang Yan, Qiao Jun-Fei. A soft-sensing method for effluent nitrogen-containing parameters based on dynamic fuzzy neural networks. *Control Theory & Applications*, 2024, **41**(12): 2383–2392 (蒙西, 张寅, 乔俊飞. 基于动态模糊神经网络的出水含氮参数软测量方法. *控制理论与应用*, 2024, **41**(12): 2383–2392)

5 Zühlke M M, Kudenko D. Adversarial robustness of neural networks from the perspective of Lipschitz calculus: A survey. *ACM Computing Surveys*, 2025, **57**(6): 1–41

6 Wu T, Wang X C, Qiao S J, Zhang Y M, Xian X P, Liu Y B, et al. Small perturbations are enough: Adversarial attacks on time series prediction. *Information Sciences*, 2022, **587**: 794–812

7 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv: 1412.6572, 2014.

8 Yin Z Q, Zhuo Y, Ge Z Q. Transfer adversarial attacks across industrial intelligent systems. *Reliability Engineering & System Safety*, 2023, **237**: Article No. 109299

9 Kong X Y, Ge Z Q. Adversarial attacks on neural-network-based soft sensors: Directly attack output. *IEEE Transactions on Industrial Informatics*, 2022, **18**(4): 2443–2451

10 Chen L, Zhu Q X, He Y L. Adversarial attacks for neural network-based industrial soft sensors: Mirror output attack and translation mirror output attack. *IEEE Transactions on Industrial Informatics*, 2024, **20**(2): 2378–2386

11 Jiang Q C, Fan S H, Zhu Z Y, Hou Z X, Zhong W M, Tan L, et al. Adversarial attacks on industrial soft sensors: Multi-target attacks based on diffusion models. *Information Sciences*, 2026, **725**: Article No. 122732

12 Zhou X L, Wu O, Yang N. Adversarial training with anti-adversaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 10210–10227

13 Zhao S J, Wang X Z, Wei X X. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis and Machine Intelli-*

gence, 2024, **46**(12): 9338–9352

14 Guo R Y, Liu H, Liu D. When deep learning-based soft sensors encounter reliability challenges: A practical knowledge-guided adversarial attack and its defense. *IEEE Transactions on Industrial Informatics*, 2024, **20**(2): 2702–2714

15 Guo R Y, Chen Q Y, Tong S, Liu H. Knowledge-aided generative adversarial network: A transfer gradient-less adversarial attack for deep learning-based soft sensors. In: Proceedings of the 14th Asian Control Conference (ASCC). Dalian, China: IEEE, 2024. 1254–1259

16 Xie Y F, Wang J, Xie S W, Chen X F. Adversarial training-based deep layer-wise probabilistic network for enhancing soft sensor modeling of industrial processes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **54**(2): 972–984

17 Guo R Y, Chen Q Y, Liu H, Wang W Q. Adversarial robustness enhancement for deep learning-based soft sensors: An adversarial training strategy using historical gradients and domain adaptation. *Sensors*, 2024, **24**(12): Article No. 3909

18 Liu C, Huang Z C, Salzmann M, Zhang T, Stüsstrunk S. On the impact of hard adversarial instances on overfitting in adversarial training. *Journal of Machine Learning Research*, 2024, **25**(356): 1–46

19 Chen Y, Li X D, Hu P, Peng D Z, Wang X. Diffilter: Defending against adversarial perturbations with diffusion filter. *IEEE Transactions on Information Forensics and Security*, 2024, **19**: 6779–6794

20 Liu H, Zhao B, Zhang K H, Liu P. Nowhere to hide: A lightweight unsupervised detector against adversarial examples. arXiv preprint arXiv: 2210.08579, 2022.

21 Liu H, Zhao B, Guo J B, Zhang K H, Liu P. A lightweight unsupervised adversarial detector based on autoencoder and isolation forest. *Pattern Recognition*, 2024, **147**: Article No. 110127

22 Mumcu F, Yilmaz Y. Universal and efficient detection of adversarial data through nonuniform impact on network layers. arXiv preprint arXiv: 2506.20816, 2025.

23 Guo R Y, Li A L, Liu H. An adversarial attack detection method based on bidirectional consistency discrimination for deep learning-based soft sensors. In: Proceedings of the 2025 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes. Urumqi, China: IEEE, 2025. 1–6

24 Yin X W, Kolouri S, Rohde G K. GAT: Generative adversarial training for adversarial example detection and robust classification. arXiv preprint arXiv: 1905.11475, 2022.

25 Chen E C, Lee C R. Data filtering for efficient adversarial training. *Pattern Recognition*, 2024, **151**: Article No. 110394

26 Hong S, An N, Cho H, Lim J, Han I S, Moon I, et al. A dynamic soft sensor based on hybrid neural networks to improve early off-spec detection. *Engineering With Computers*, 2023, **39**(4): 3011–3021

27 Guo R Y, Liu H. Semisupervised dynamic soft sensor based on complementary ensemble empirical mode decomposition and deep learning. *Measurement*, 2021, **183**: Article No. 109788

28 Wu Ying-Jun, Ru Ying-Tao, Liu Jin-Tao, Shi Zhan-Yu, Gu Song, Ni Ming. False data injection attack detection for automatic generation control system based on set-membership filtering. *Automation of Electric Power Systems*, 2022, **46**(1): 33–41 (吴英俊, 汝英涛, 刘锦涛, 施展宇, 顾松, 倪明. 基于集员滤波的自动发电控制系统虚假数据注入攻击检测. *电力系统自动化*, 2022, **46**(1): 33–41)

29 Guo W, Liu F, Wang Y, Sidorov D, Wu J. Adaptive event-triggered sliding mode load frequency control for cyber-physical power systems under false data injection attacks. *IEEE Transactions on Industrial Informatics*, 2024, **21**(4): 2947–2956

30 Liu Ding, Liu Han. *Control Methods and Applications of Air Preheater for Power Plant Boilers*. Beijing: Science Press, 2016. (刘丁, 刘涵. 电站锅炉空气预热器控制方法及应用. 北京: 科学出版社, 2016.)



郭润元 西安理工大学自动化与信息工程学院讲师. 2023 年获得西安理工大学博士学位. 主要研究方向为软测量, 工业过程控制, 可信人工智能.

E-mail: guorunyuan@xaut.edu.cn

(GUO Run-Yuan Lecturer at the School of Automation and Information Engineering, Xi'an University of Technology. He received his Ph.D. degree from Xi'an University of Technology in 2023. His research interests include soft sensors, industrial process control, and trustworthy artificial intelligence.)



李安伦 西安理工大学自动化与信息工程学院硕士研究生. 主要研究方向为深度学习软测量, 软测量对抗攻击与防御.

E-mail: 2230320161@stu.xaut.edu.cn

(LI An-Lun Master student at the School of Automation and Information Engineering, Xi'an University of Technology. His research interests include deep learning-based soft sensors and adversarial attack and defense for soft sensors.)



刘 涵 西安理工大学自动化与信息工程学院教授. 2003 年获得西安理工大学博士学位. 主要研究方向为工业人工智能, 工业过程智能监测与诊断. 本文通信作者.

E-mail: liuhan@xaut.edu.cn

(LIU Han Professor at the School of Automation and Information Engineering, Xi'an University of Technology. He received his Ph.D. de-

gree from Xi'an University of Technology in 2003. His research interests include industrial artificial intelligence, intelligent monitoring and diagnosis for industrial processes. Corresponding author of this paper.)



张友民 加拿大康考迪亚大学机械、工业与航空工程系教授, IEEE Fellow, CSME Fellow. 1995 年获得西北工业大学博士学位. 主要研究方向为故障检测与诊断, 多智能体系统容错协同控制.

E-mail: ymzhang@encs.concordia.ca

(ZHANG You-Min Professor in the Department of Mechanical, Industrial and Aerospace Engineering, Concordia University, Canada, IEEE Fellow, CSME Fellow. He received his Ph.D. degree from Northwestern Polytechnical University in 1995. His research interests include fault detection and diagnosis, and fault-tolerant cooperative control of multi-agent systems.)



刘 丁 西安理工大学自动化与信息工程学院教授. 1997 年获得西安交通大学博士学位. 主要研究方向为复杂工业过程信息测量、建模和智能优化控制, 智能机器人及控制.

E-mail: liud@xaut.edu.cn

(LIU Ding Professor at the School of Automation and Information Engineering, Xi'an University of Technology. He received his Ph.D. degree from Xi'an Jiaotong University in 1997. His research interests include information measurement, modeling and intelligent optimal control of complex industrial processes, and intelligent robots and their control.)