控制与决策
Control and Decision

网络攻击下信息物理系统安全防护研究综述

芦安洋, 张佳楠, 王庆杰, 纪寒康, 尹利榜, 朱立秋, 孙秉旭

引用本文:
芦安洋, 张佳楠, 王庆杰, 等. 网络攻击下信息物理系统安全防护研究综述[J]. 控制与决策, 2026, 41(1): 1-18.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0866

网络攻击下信息物理系统安全防护研究综述

芦安洋[1,2†]，张佳楠[1]，王庆杰[1]，纪寒康[1]，尹利榜[1]，朱立秋[1]，孙秉旭[1]

(1. 东北大学 信息科学与工程学院， 沈阳 110819；
2. 东北大学 流程工业综合自动化国家重点实验室， 沈阳 110819)

摘　要: 信息物理系统作为工业 4.0、智能电网等领域的核心, 其信息层与物理层深度耦合的特性在带来高效能的同时, 也引入了严峻的安全风险. 网络攻击尤其威胁着 CPS 的正常运行, 可能导致严重的物理层破坏和系统瘫痪. 鉴于此, 从防护者视角出发, 系统总结网络攻击下 CPS 安全防护的研究进展, 并深入分析安全状态估计、安全控制与攻击检测 3 大关键技术. 首先, 梳理典型网络攻击模型, 涵盖攻击类型与攻击位置, 并重点阐述安全状态估计的研究现状; 然后, 针对传感器数据篡改问题, 提出基于遍历搜索、凸优化及人工智能的方法, 以解决稀疏攻击下的状态重构难题; 接着, 分别探讨针对信息层网络攻击的安全控制策略以及针对物理层干扰的避碰避障控制方法, 系统总结基于模型的攻击检测方法与基于数据驱动的攻击检测方法; 最后, 总结当前研究存在的问题与挑战, 并对未来研究方向进行展望.

关键词: 信息物理系统；网络攻击；安全防护；安全状态估计；安全控制；攻击检测

中图分类号: TP273　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0866

引用格式: 芦安洋, 张佳楠, 王庆杰, 等. 网络攻击下信息物理系统安全防护研究综述 [J]. 控制与决策, 2026, 41(1): 1-18.

A survey on secure protection of cyber-physical systems under cyber attacks

LU An-yang[1,2†]， *ZHANG Jia-nan*[1]， *WANG Qing-jie*[1]， *JI Han-kang*[1]， *YIN Li-bang*[1]， *ZHU Li-qiu*[1]， *SUN Bing-xu*[1]

(1. College of Information Science and Engineering，Northeastern University，Shenyang 110819，China；2. State Key Laboratory of Synthetical Automation for Process Industries， Northeastern University， Shenyang 110819，China)

Abstract: Cyber-physical systems (CPS), serving as the cornerstone of Industry 4.0 and smart grids, feature deep integration between cyber and physical layers. While this integration enables high efficiency, it simultaneously introduces significant security vulnerabilities. Cyberattacks pose severe threats to CPS operations, potentially causing physical damage and system failures. This paper systematically reviews recent advances in CPS security from a defensive perspective, with focus on three key technologies: secure state estimation, security control, and attack detection. First, we categorize typical attack models by attack type and location. Next, the state-of-the-art in secure state estimation are summarized. To address sensor data tampering, researchers have developed methods based on exhaustive search, convex optimization, and artificial intelligence to reconstruct system states under sparse attacks. Then, security control strategies against cyber-layer attacks and collision/obstacle avoidance methods for physical-layer disturbances are examined. Subsequently, model-based and data-driven attack detection approaches are systematically summarized. Finally, the paper summarizes the current limitations and challenges in the field and outlines potential avenues for future research.

Keywords: cyber-physical systems； cyber attacks； secure protection； secure state estimation； secure control； attack detection

收稿日期: 2025-08-23；**录用日期:** 2025-12-04.

基金项目: 国家自然科学基金项目 (62522309, 62473088)；辽宁省自然科学基金项目 (2025JH6/101000012)；广东省基础与应用基础研究基金项目 (2023A1515140007).

†通信作者. E-mail: luanyang@ise.neu.edu.cn.

0 引 言

信息物理系统 (cyber-physical systems, CPS) 作为计算、通信与物理过程深度融合的智能系统, 通过信息空间的智能决策深度驱动并反馈物理实体行为, 已成为实现工业 4.0、智慧城市、智能电网、无人驾驶等领域的核心技术[1-2]. 在智能电网中, CPS 实现发电、输电、用电的实时协调优化; 在智能制造中, CPS 构建了柔性生产线与智能工厂的神经中枢; 在智能交通与无人系统 (如自动驾驶汽车、无人机集群) 中, CPS 是实现环境感知、智能决策与精准控制的关键[3]. 然而, CPS 信息层与物理层的高度耦合在带来巨大效能的同时, 也引入了前所未有的安全脆弱性. 因此, 深入研究 CPS 在面对日益严峻网络威胁时的安全性, 具有重要的理论价值.

CPS 的安全性分析主要包含针对信息层的网络安全和针对物理层的安全两个层面. 在信息安全层面 (如传感器、控制器、执行器、网络通信), 系统主要面临诸如拒绝服务攻击 (denial of service, DoS)、重放攻击及虚假数据注入攻击 (false data injection, FDI) 等网络威胁; 在物理安全层面 (如传感器/执行器损坏、外部物理力干扰), 核心目标在于防范因异常操作或外部干扰引发的硬件设备故障或安全事故, 其中避碰与避障技术是保障物理安全的关键研究方向. CPS 的安全攸关全局, 一旦遭受成功攻击, 将直接引发系统崩溃, 并造成难以承受的后果. 例如, 2010 年, 伊朗纳坦兹核设施遭遇网络攻击, 其工业控制系统被恶意代码渗透, 致使大量离心机在运行中出现故障并报废[4]; 2014 年, 在德国某钢厂, 攻击者由办公网入侵生产网, 导致高炉无法按规程停机并造成“大规模物理损害”; 2015 年乌克兰电网攻击使得 3 家地区配电公司遭到协调入侵, 多个变电站被断开, 致使用户停电数小时[5]. 因此, CPS 安全防护的核心挑战在于如何有效应对信息层网络攻击与物理层干扰/攻击的协同威胁, 确保系统在恶意环境下的功能安全、信息安全以及物理安全.

针对上述严峻挑战, 研究人员从防护者视角出发做出了大量努力, 包括安全状态估计、安全控制和攻击检测[6]. 首先, 安全状态估计是 CPS 安全防护的重要手段[7], 当部分传感器数据被攻击者篡改 (如 FDI 攻击) 时, 安全状态估计的目标是在此条件下仍能准确地重构系统内部的关键运行状态 (如电网的电压相角、无人车的位置速度). 其核心挑战在于区分恶意篡改数据与正常噪声扰动, 主要方法包括基于优化的鲁棒估计 (如稀疏恢复)、基于观测器的鲁棒设计 (如未知输入观测器、滑模观测器) 以及数据驱动方法等. 其次, 安全控制是保障 CPS 在遭受攻击下仍能维持稳定运行并满足关键性能与安全约束的核心手段[8]. 针对信息层网络攻击的安全控制, 研究人员主要解决控制指令在网络传输或生成过程中被篡改、延迟或阻断的问题. 核心策略包括鲁棒/弹性控制 (如H_∞ 控制、滑模控制)、基于事件触发的通信优化以及安全控制重构/容错控制等. 针对物理层干扰的安全控制, 研究人员主要解决外部物理干扰或攻击引发的物理效应导致系统违反物理安全约束 (如与障碍物或其他智能体发生碰撞) 的问题, 其核心目标是在存在干扰下确保物理安全, 关键技术包括基于控制障碍函数的安全关键控制、融合安全约束的模型预测控制、鲁棒/自适应避障算法等. 最后, 攻击检测是主动发现和识别恶意活动的哨兵[9], 其目标是在复杂的系统噪声和扰动背景下, 快速、准确地识别出网络攻击或异常行为并尽可能定位攻击源. 主要技术路线包括基于模型的残差分析 (如观测器、卡方检测)、基于数据驱动的机器学习方法 (如异常检测、分类模型) 以及信号处理技术 (如水印) 等.

基于上述讨论, 本文旨在系统梳理网络攻击下信息物理系统安全防护领域的最新研究进展, 重点围绕安全状态估计、安全控制、攻击检测 3 个核心角度展开深入分析. 首先, 分析 CPS 面临的典型网络攻击模型与安全威胁场景, 详细阐述安全状态估计的主要方法、原理与研究现状; 然后, 深入探讨针对信息层网络攻击的安全控制策略以及针对物理层干扰的避碰避障控制方法, 并系统评述攻击检测技术的分类与研究现状; 最后, 总结当前研究面临的关键挑战并展望未来发展方向. 通过对现有研究的系统归纳、比较与前瞻性分析, 本文期望能为相关研究者和实践者提供清晰的技术脉络与未来方向的思考.

1 典型网络攻击介绍

CPS 的开放互联特性使其面临多样化的网络攻击威胁, 如图 1 所示. 这些攻击不仅类型各异, 其发生的网络位置也深刻影响着攻击的破坏机制与防御

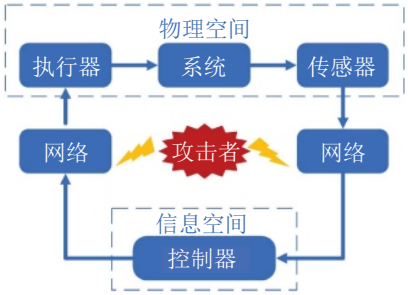

图1 网络攻击下的信息物理系统

难度. 本节将从攻击类型 (如 DoS 攻击、FDI 攻击、重放攻击) 和攻击位置 (如传感器-控制器通道、控制器-执行器通道、通信网络、智能体节点) 两个关键维度, 系统梳理 CPS 面临的典型网络攻击模式, 为后续安全防护策略的讨论奠定基础[10].

1.1　攻击类型

根据攻击机制与目标, CPS 面临的主要攻击类型包括 DoS 攻击、FDI 攻击和重放攻击.

1) DoS 攻击通过耗尽通信信道或节点资源 (如带宽、计算能力), 阻断智能体间的实时信息交换. 其核心是破坏可用性, 导致控制回路中断或状态估计失效[11]. 文献 [10] 总结的 DoS 攻击典型形式包括随机 DoS 攻击和频率-持续时间约束 DoS 攻击.

随机 DoS 攻击指数据包丢失服从概率模型 (如伯努利分布或 Markov 过程). 令$y_i(t)$表示在时间t来自通信信道i的初始数据, $\tilde{y}_i(t)$表示时间t通信信道i传输到控制器的数据, $y_i(t)$与$\tilde{y}_i(t)$之间的关系描述如下:

$$\tilde{y}_i(t) = \alpha_i(t) y_i(t),$$

其中随机变量$\alpha_i(t) = \{0, 1\}$描述攻击事件. 具体地, 如果通信信道在时间t受到攻击, 则$\alpha_i(t) = 0$; 否则, $\alpha_i(t) = 1$. 假设 DoS 攻击造成的数据丢失服从伯努利分布, 则随机变量$\alpha_i(t)$满足

$$\Pr\{\alpha_i(t) = 1\} = \bar{\alpha}_i,\ \Pr\{\alpha_i(t) = 0\} = 1 - \bar{\alpha}_i,$$

其中$\bar{\alpha}_i \in (0, 1)$.

频率-持续时间约束 DoS 攻击指攻击者受固有能量的限制, 并不能持续发出攻击. 文献 [12] 通过对 DoS 攻击的频率和持续时间进行限制, 研究了 CPS 在遭受 DoS 攻击时, 具有多传输信道的输入状态稳定控制问题. 特别地, 令$\{h_n^i\}_{n\in\mathcal{N}}(h_0^i \geqslant 0)$表示攻击开/关转换序列. 定义$\mathcal{A}_n^i = [h_n^i, h_n^i + T_n^i)$是通信信道$i$在第$n$个 DoS 攻击持续时间, 这表示在该时间间隔内不能通信, 其长度为$T_n^i \geqslant 0$. 为了分析系统可以容忍的 DoS 攻击的影响, 在文献 [12] 中引入两个概念, 即 DoS 攻击频率和持续时间.

假设 1 (DoS 攻击频率)　对于$t_2 \geqslant t_1 \geqslant 0$, 存在常数$\eta^{1i}$和$T^i$使得下式成立:

$$\mathbb{N}^i(t_1, t_2) \leqslant \eta^{1i} + \frac{t_2 - t_1}{T^i},$$

其中$\mathbb{N}^i(t_1, t_2)$为在间隔$[t_1, t_2)$上发生的 DoS 攻击的次数.

假设 2 (DoS 攻击持续时间)　对于$t_2 \geqslant t_1 \geqslant 0$, 存在常数$\eta^{2i}$和$\mu^i$使得下式成立:

$$|\Xi^i(t_1, t_2)| \leqslant \eta^{2i} + \frac{t_2 - t_1}{\mu^i},$$

其中$|\Xi^i(t_1, t_2)| = \bigcup\limits_{n\in N} \mathcal{A}_n^i \bigcap [t_1, t_2]$.

2) FDI 攻击通过篡改传输数据破坏传输完整性, 相比于 DoS 攻击更具有破坏性. 根据 FDI 攻击的主要策略, 将其分为加性攻击、乘性攻击和替换攻击.

以通信信道$y_i(t)$为例, 加性攻击是指向原始信号$y_i(t)$注入虚假数据$a_i(t)$, 即$\tilde{y}_i(t) = y_i(t) + a_i(t)$[13]. 乘性攻击是指缩放原始信号 $\tilde{y}_i(t) = c_i(t) y_i(t)$, 其中$c_i(t)$为缩放率, 该式可转化为加性形式. 替换攻击是指直接将原始信号替换为伪造信号 $\tilde{y}_i(t) = r_i(t)$, 其中$r_i(t)$可以是任意攻击信号, 隐蔽性极强. 此类攻击常见于传感器- 控制器通道 (如篡改 GPS 数据误导无人车定位) 或控制器-执行器通道 (如篡改断路器指令引发电网瘫痪).

3) 重放攻击通过截获并重复发送历史合法数据包, 破坏信息的新鲜性[14], 其本质是利用协议漏洞绕过认证机制, 无需破解加密内容. 例如, 重放过去的传感器数据可使控制器基于过时状态决策, 导致无人系统无法响应环境变化. 对此, 学者们进行了大量研究, 文献 [15] 研究了针对重放攻击的离散时间隐马尔可夫跳跃系统的静态输出反馈安全控制问题; 文献 [16] 基于奇偶空间方法, 从新的角度研究了重放攻击检测问题.

注 1　FDI 攻击: 当面对受保护的测量值、随机水印或在具备足够传感冗余并采用鲁棒估计方法时, 其攻击的隐蔽性显著下降, 从而导致失效. 重放攻击: 在系统引入时间戳或动态水印等防护措施后, 该攻击的有效性将被消除. DoS 攻击: 当丢包/信道占用低于容忍上界, 或当系统采用事件触发机制且满足平均驻留时间条件时, 其破坏效果显著降低.

表 1 详细对比了不同攻击模型的优劣势及其典型适用场景.

表1　不同网络攻击模型对比

攻击类型	隐蔽性	破坏力	实施难度	典型适用场景	主要优势
DoS攻击	低	中等	低	带宽受限系统	实施简单
FDI攻击	高	高	中高	状态估计系统	隐蔽性强
重放攻击	中	中	低	认证薄弱系统	绕过加密

1.2　攻击位置

根据攻击在 CPS 信息流中发生的关键位置, 可将其分为针对传感器-控制器通道、控制器-执行器通道、通信网络本身以及智能体节点的攻击.

1) 传感器-控制器通道攻击: 此类攻击发生在传感器节点或其向控制器传输感知数据的通道上[17],

攻击者篡改传感器测量值 (如温度、位置) 或阻断其传输. 模型描述为 $\tilde{y}_i(t)=y_i(t)+\alpha_i(t)y_i(t)$, 其中 $y_i(t)$为注入的虚假信号. 此类攻击直接影响状态估计准确性, 是 FDI 攻击的高发区域.

2) 控制器-执行器通道攻击: 此类攻击发生在控制器节点或其向执行器发送控制指令的通道上, 攻击者通过篡改或阻断控制指令传输, 导致执行器接收错误命令[18]. 例如在工业控制中, 篡改机械臂运动指令可能引发设备碰撞事故, 模型为 $\tilde{u}_i(t)=u_i(t)+\alpha_i(t)u_i(t)$.

3) 通信网络攻击: 此类攻击发生在分布式 CPS 的信息层, 具体位于构成该层的多智能体网络内相邻智能体的通信通道上, 会显著干扰甚至破坏智能体间的互联协作行为[19], 其呈现形式可分为 DoS 攻击和 FDI 攻击. 第 1.1 节对 DoS 攻击和 FDI 攻击的具体形式进行了详细描述.

4) 智能体攻击: 此类攻击发生在分布式 CPS 信息层的多智能体上. 通常, 攻击者劫持多个智能体节点将其变为恶意节点或拜占庭节点[20-21]. 具体地, 恶意节点是指向所有邻居发送相同但虚假信息的叛变节点, 拜占庭节点是指可对不同邻居发送不一致信息 ($x_j^i(t)\neq x_j^k(t)$) 的叛变节点. 根据破坏范围可分为 F-全局攻击 (全网至多 F 个节点被攻击) 和 F-局部攻击 (每个正常节点的邻居中至多F个恶意节点). 此类攻击利用分布式协作机制扩散破坏, 对无人机集群、智能电网等场景构成严重威胁, 需通过加权中值一致性算法等技术隔离恶意节点.

2 安全状态估计

为获取受攻击干扰的 CPS 实时动态, 利用可能被破坏的传感器数据重建系统状态至关重要, 如图 2 所示. 安全状态估计因其能够实时识别攻击并重构真实系统状态而备受重视. 具体地, 给定一组测量值 $\mathcal{Y}$, 防守方的目标是在恶意攻击的干扰下估计出系统状态x, 使得

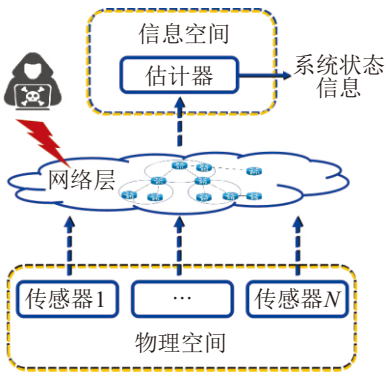

图2 安全状态估计基本结构

$$||\hat{x}-x||<\varepsilon.$$

其中: $\hat{x}$为状态估计值; $\varepsilon\geqslant 0$为估计误差的界, 当 $\varepsilon=0$时状态被准确重构. 然而, 与传统控制系统不同, CPS 中物理与网络组件的紧密集成以及各种恶意攻击的存在, 使得安全状态估计面临核心挑战: 如何准确识别并排除受攻击的数据源. 稀疏攻击下被攻击信道是未知的, 需要从多种可能性中找到正确的被攻击信道集合.

综上, 该问题本质上是一个组合优化问题, 需要从潜在受攻击通道的所有可能组合中寻找正确的受攻击数据源集合, 即对C_n^s种可能的组合进行搜索. 针对较大的被攻击信道数s和通道数n, 其难点在于解决由此带来的高计算复杂度. 基于此, 将现有的安全状态估计方法分为 3 类: 基于遍历搜索的方法、基于凸优化的方法和基于人工智能的方法.

2.1 基于遍历搜索的方法

基于遍历搜索的方法是解决安全状态估计问题的一类直观且理论上完备的策略. 其核心思想在于: 既然攻击是稀疏的 (最多s个位置被攻击), 那么理论上存在一个未被攻击或攻击影响最小的传感器通道子集. 遍历搜索法就是系统地枚举所有可能的s-稀疏攻击模式组合, 对每种组合假设其是被攻击的集合, 然后基于剩余的 (未受攻击的) 测量数据进行状态估计, 最终基于最小残差准则从所有候选估计结果中选择一个最优的作为最终的状态估计值. 具体描述如下:

$$\{\hat{x},\hat{A}\}=\underset{\hat{x}\in R,\hat{A}\in K_s}{\arg\min}\ ||\mathcal{Y}_{\hat{A}}-\mathcal{O}_{\hat{A}}\hat{x}||^2.$$

其中: $\hat{A}$为受攻击通道集合的估计值; K_s为s个精确攻击下所有可能的受攻击通道组合 (规模为C_n^s); $\mathcal{Y}_{\hat{A}}$和 $\mathcal{O}_{\hat{A}}$ 分别为观测向量 $\mathcal{Y}$ 和观测矩阵 $\mathcal{O}$ 剔除$\hat{A}$对应行后的子集. 当系统规模扩大 (n和s 增大) 时, 计算复杂度因组合爆炸 (C_n^s 级增长) 而难以应用于大规模 CPS. 因此为缓解计算负担, 当前研究聚焦于通过理论方法降低计算复杂度或搜索次数[22-24].

减少候选组合的数量是降低搜索次数的直接方法之一. 目前, 如何优化候选项以在减少搜索次数的同时保证状态估计的准确性, 已得到广泛研究. 基于可满足性模理论的方法可以在线排除不满足条件的候选项, 从而减少搜索次数. 文献 [22] 设计了一种基于可满足性模理论的理论求解器, 该求解器检查当前候选项是否可行, 并在不可行时提供冲突的原因、证书或反例. 每个证书都会导致学习新的约束, 使用这些约束来修剪搜索空间. 同样, 文献 [25] 也提出了

一种基于可满足性模理论的候选项优化方法, 该方法是文献 [22] 的扩展, 在系统无扰动情况下, 状态估计能够获得更好的性能. 另外, 基于系统 2s 稀疏可观的假设, 集合覆盖方法可用于减少搜索空间. 文献 [23] 提出一种基于集合覆盖技术的状态估计方法, 并从理论上证明了该方法至少可以将候选组合缩减为原来的一半.

文献 [24] 基于集合论视角, 运用受限集合划分技术有效降低了搜索次数. 文献 [26] 基于传感器类型的数量通常远少于传感器数量的事实, 开发了一种通过等价传感器的快速状态估计算法, 该方法通过验证传感器类型的测量数据的相似度, 提取攻击位置信息以排除一些不匹配的搜索候选. 文献 [27] 发展了一种基于正交投影的安全状态估计方案, 该方案在安全状态估计策略设计基础上, 结合扰动解耦方法实现了受稀疏攻击及干扰影响下系统状态的准确重建.

此外, 遍历搜索方法的一个关键工作是候选组合正确性判断. 在固定通道攻击的情况下, 切换机制被广泛用于降低计算复杂度. 文献 [28] 提出了一种基于自适应切换机制的安全状态估计方法, 通过切换函数矩阵, 自适应地截断受攻击通道, 最终切换到合适的模式并保持不变. 文献 [29] 提出了切换投影梯度下降算法, 用于解决高计算复杂度问题, 采用测量数据的预处理方法提高算法的收敛速度, 提出了一个新的投影算子, 直接基于所获得的估计来减少搜索时间. 文献 [30] 构建一组具有自适应切换机制的模糊状态估计器, 用于估计稀疏传感器攻击下 T-S 模糊系统的状态.

遍历搜索的优势在于其高精度的状态估计, 且其方法主要用于集中式安全状态估计. 虽然现有方法在一定程度上减少了搜索频率, 但随着 CPS 规模的增大, 传感器数量的增多, 其计算复杂度仍然会显著增加.

2.2　基于凸优化的方法

仅从理论分析的角度, 遍历搜索必然可以找到真实状态和被攻击通道集合的可靠估计, 但对一个已被证明是 NP 难的问题[31], 使用暴力解法会使得计算复杂度急剧增加: 对于有n个传感器信道的系统, 假设已知有s个信道被攻击, 则暴力搜索的次数为C_n^s. 可以看出, 如果n很大 (例如$n>100$, 这在智能电网的建模中很常见), 则遍历搜索的实用性大大降低, 所以为了避免暴力搜索, 研究者将解决优化问题相关的技术引入安全状态估计中. 回顾安全状态估计的核心目标, 是在传感器网络可能遭受恶意攻击 (如虚假数据注入、数据篡改) 时, 仍能准确重构系统真实状态, 这一问题天然具备优化问题的数学形式: 已知受污染的观测值, 需寻找估计值以最小化残差范数, 同时满足系统动态约束和安全性要求. 凸优化因其理论完备性与计算高效性, 已成为解决状态估计问题的首选框架.

凸优化问题是在目标函数与不等式约束函数均为凸函数的条件下, 对目标函数值进行最小化. 一个典型凸优化问题的标准形式为

$$\begin{aligned}&\min_x\ f_0(x).\\&\text{s.t. } f_i(x)\leqslant 0,\ i=1,2,\ldots,n;\\&\quad a_j^{\mathrm{T}}x=b_j,\ j=1,2,\ldots,p.\end{aligned}$$

其中: $f_0(x)$为目标函数, $f_i(x)$为凸约束函数, 可行域$\mathcal{X}=\{x|f_i(x)\leqslant 0,\ i=1,2,\ldots,n,\ \ a_j^{\mathrm{T}}x=b_j,\ j=1,2,\ldots,p\}$为凸集.

凸优化方法在优化问题中的优势显而易见: 首先相较于非凸优化, 凸优化存在且仅存在一个最优解, 不必寻找多个局部最优解; 而且相较于非凸优化, 凸优化问题可利用多种高效算法进行解决, 例如线性规划、梯度下降[32] 等. 上述特性也在安全状态估计领域赋予了凸优化得天独厚的优势: 当系统面临传感器篡改、通信干扰等威胁时, 基于凸优化的估计框架能够将复杂的防御问题转化为具有严格数学保障的可计算模型; 相较于非凸优化方法可能陷入多个局部最优解的困境 (例如某些基于神经网络的检测器需反复调整初始值以避免误收敛), 凸优化凭借其唯一全局最优解的特性, 从根本上消除了结果的不确定性. 这一优势在关键基础设施 (如电网、交通系统) 中尤为重要 —— 防御策略的可靠性直接关系到系统安全, 而同等条件下非凸算法常因局部最优解产生灾难性偏离.

另外, 计算效率的显著提升是凸优化方法的另一个核心竞争力: 虽然暴力搜索在理论分析中往往是可行的, 但在实际应用中, 系统都是在高维状态空间 ($n>100$) 下建模的, 例如智能电网、社会网络等. 面对指数级增长的计算复杂度, 暴力搜索算法必然失效. 而将估计问题转化为凸优化问题后, 利用凸优化的高效算法可快速简便地进行安全状态估计, 部分情况下甚至可以在多项式时间内完成任务[31]. 此外, 研究者也因地制宜地对一般的高效算法进行了改进[33], 使其对状态估计这一应用场景的适用性更强.

除了一般的估计问题外, 凸优化也常被用于进

行系统脆弱性分析, 即站在攻击者的角度, 找出一种攻击策略, 其对系统具有最强破坏力的同时, 不会触发系统的攻击检测器. 攻击策略的破坏能力可由不可靠的估计值与真实值之间的差的范数进行量化, 这自然地形成了一个最大化问题 (或等效的最小化负偏差范数问题). 在远程状态估计中, 攻击检测器通常是卡方检测器或 K-L 散度 (Kullback-Leibler divergence) 检测器, 这些检测器均可视为特殊的凸约束函数: 卡方检测器通过一个特殊的二次型与阈值进行对比以检测, 这定义了一个凸的椭球约束集; 而 K-L 散度自身就是一个凸函数, 即使约束可能不是凸集, 但系统带有高斯噪声这一常见的假设的前提下, 约束也可被松弛或等价转化为凸约束. 由此可见, 凸优化非常适用于脆弱性分析, 这种建模允许研究者利用凸优化的强大理论和高效算法 (内点法、梯度投影等), 精确计算系统在最坏隐蔽攻击下的最大可能状态偏差 (破坏性量化), 识别系统最薄弱的环节 (哪些传感器或通信链路被攻击影响最大) 和最优攻击模式 (攻击信号的特征). 相关的研究成果也非常丰硕, 例如文献 [34-35] 等.

虽然凸优化方法有许多优势, 但其也有一定的局限性. 文献 [36] 引入了凸优化方法以替代暴力搜索, 降低了复杂度, 但也不可避免地引入新的限制条件. 尽管由于凸优化对凸函数的强制要求, 引入额外条件是合情合理的, 但这也意味着凸优化在安全状态估计中不是万能的, 毕竟并不是所有的 NP 难问题都可利用凸优化解决.

鉴于凸优化的缺陷, 研究者也开发了一些额外的安全状态估计技术, 包括饱和自适应技术[37]、状态分解技术[38] 等. 自适应饱和技术一般用于分布式估计, 为了实现更好的共识-更新的估计策略, 引入饱和自适应参数降低攻击带来的影响, 将估计值限制在一定范围内, 使得所有节点的估计值最终都收敛到一个范围内以得到可靠估计值. 此外, 文献 [32] 将饱和自适应技术引入到集中式框架下, 在梯度下降算法中加入饱和自适应参数以避免暴力搜索; 而状态分解技术是最近比较热门的新兴技术, 将状态分解为若干子状态分别估计, 继而完成状态重构. 利用状态分解, 可将部分 NP 难的状态估计问题在多项式时间内解决, 由于其在复杂度方面的巨大优势, 状态分解在近年受到了越来越多的关注.

值得注意的是, 凸优化方法并没有因上述新技术的兴起而受到冷落, 而是在积极优化自身的同时与新技术结合, 取长补短. 例如文献 [39] 提出了在一定较为宽松的限制条件下将估计问题转化为凸优化问题的方法; 文献 [23] 放弃寻找凸约束函数, 直接利用梯度下降设计安全状态估计算法; 文献 [37] 完全放弃凸优化, 改用引入多数投票以获得可靠的估计值. 对于可以使用凸优化的场景, 文献 [38] 也引入了状态分解以进一步降低复杂度. 此外, 还有针对凸优化算法创新的研究, 例如文献 [32] 在算法中引入饱和自适应函数, 中和攻击带来的估计偏差; 文献 [40] 在梯度下降算法中引入了额外的恶意攻击检测器, 检测器触发时, 将已被确认的攻击节点排除在梯度更新外, 防止信息污染.

综上所述, 凸优化在安全状态估计中占据了重要地位, 其通过全局最优解的唯一性、高效求解能力 (梯度下降/内点法等) 以及灵活的问题建模能力 (如将卡方检测转化为椭球约束), 为抵御虚假数据注入、拜占庭攻击等威胁提供了坚实的数学框架. 研究证明, 在满足一般的稀疏可观测性假设下, 基于凸优化的方案能严格保证估计误差有界性与算法实时性, 成为电力网络、无人集群等关键系统的首选防御范式. 但由于安全状态估计的 NP 难特性, 该方法仍有相当大的改进空间, 需进行进一步研究.

2.3 基于人工智能的方法

传统方法 (如基于残差分析、观测器设计) 在面对复杂攻击 (尤其是隐蔽攻击、协同攻击)、系统非线性、模型不确定性等条件下, 解决安全状态估计问题时常表现出局限性, 而基于人工智能的方法 (尤其是机器学习、深度学习) 在此类情景有着独特优势.

在基于机器学习的方法中, 文献 [41] 提出了一种强化学习算法以评估攻击者和防御者策略对状态估计的影响, 该算法设计了两种场景: 可靠通道 (即丢包的原因只是 DoS 攻击) 和不可靠通道 (丢包的原因不完全来自 DoS 攻击), 旨在使传感器和攻击者能够动态学习和调整策略. 文献 [42] 提出了一种基于机器学习的状态估计器, 添加了第 2 层安全性, 并将预测的系统状态与传统状态估计器估计的系统状态进行比较, 即使网络攻击绕过传统的不良数据检测, 也可以检测到网络攻击. 类似地, 文献 [43] 也开发了一个机器学习和状态估计器的协作框架, 该方法开发多目标多变量线性回归模型, 用于数据驱动的智能电网的测量估计, 通过分析基于物理的状态估计和基于机器学习的测量估计过程的残差空间以检测参数攻击.

在基于深度学习的方法中, 文献 [44] 提出了一种基于进化算法和深度集成学习技术的新型分数阶扩展卡尔曼滤波, 用于信息物理电力系统的状态估

计问题. 通过集成 ResNet 用于预测电力系统实时状态, 进而进行攻击检测与安全状态估计. 文献 [45] 将基于表示学习的卷积神经网络用于电力 CPS 的智能攻击定位和系统恢复, 该方法将多重网络攻击位置检测问题表述为多标签分类问题, 基于表示学习的卷积神经网络最初用作分类器, 进而对受攻击的部分进行状态恢复. 文献 [46] 提出了一种基于注意力的时间卷积自编码器, 该方法结合注意力机制和时间卷积网络的优点来捕获时空信息, 此模型标识 FDI 攻击位置, 并将相应的测量值替换为重建的值. 文献 [47] 创新性地开发了一种基于区间状态估计的防御机制, 在此机制中设计了一个典型的深度学习模型, 即堆叠式自编码器, 以帮助正确提取电力负载数据中的非线性和非平稳特征, 提高了电力负载状态预测的准确性. 但基于人工智能的方法也存在一定的局限性: 1) 通常需要大量高质量标注数据, 而在实际 CPS 中获取攻击样本成本高且存在安全风险; 2) 当系统受实时算力与可解释性约束时, 该方法存在泛化不稳、误报或漏报偏高的问题.

3 安全控制

3.1 针对信息层网络攻击的安全控制

CPS 的信息层是实现感知、决策与控制的关键, 但也是网络攻击的主要目标. 信息层网络攻击 (如 DoS 攻击、FDI 攻击、重放攻击等) 旨在破坏控制信息的完整性、可用性和及时性, 导致控制信号被篡改、延迟或丢失. 这些攻击会严重破坏控制回路的稳定性和安全性. 针对这些攻击的安全控制, 其核心目标是设计能够在攻击存在下维持系统性能 (如稳定性、关键性能指标) 的控制器, 即弹性控制. 本节聚焦于信息层攻击, 综述解决控制信号异常问题的弹性控制策略, 主要方法包括事件触发控制、最优控制、基于观测器的弹性控制等.

1) 事件触发控制: 摒弃传统的周期性采样和通信, 仅在系统状态 (或估计状态) 满足预先设计的触发条件时, 才进行传感器测量传输或控制指令更新. 常用于解决 DoS 攻击下控制信号丢失的问题和提高资源效率. 事件触发控制可以减少不必要的数据传输, 从而减少网络通信量和控制器计算负担. 以文献 [48] 为例, 控制信号可以表示为

$$u(t) = Kx(t_k h),\ \forall t \in [t_k h, t_{k+1} h].$$

其中: $t_k h$为最近一次触发时刻, $t_{k+1} h$为下一次触发时刻. 是否触发可以用$e(t) > \epsilon$表示, $e(t)$表示误差, ϵ表示阈值, 即当误差大于阈值ϵ时触发, 发送控制信号, 否则采用上一次最新触发的信号. 文献 [49] 研究了一种针对易受 DoS 攻击的远程电机系统的事件触发滑模预测控制方法, 该方法采用滑模策略, 利用事件触发抵消传感器信号的中断. 针对面临资源限制和欺骗性攻击的工业 CPS, 文献 [50] 设计了一种安全的事件触发控制策略, 该策略采用神经网络学习的近似算法来估计不确定的非线性, 利用 Nussbaum 型函数解决时变攻击注入信号的未知符号问题, 然后开发了一种事件触发机制以减少通信负载.

2) 最优控制: 最优控制是一种数学框架, 旨在为动态系统建立能在特定时间段内优化其性能的控制策略. 该框架的核心在于寻找一个控制函数, 当该控制函数作用于行为可由微分方程描述的系统时, 能够最小化或最大化某个代价或性能指标. 标准的最优控制问题可以表述如下:

$$J = \int_{t_0}^{t_f} L(x(t), u(t), t)\mathrm{d}t.$$

其中: L为代价函数, t_0为初始时间, t_f为最终时间. 文献 [51] 研究了一种基于零和博弈的最优控制策略, 专门针对受执行器 FDI 攻击影响的 CPS. 该策略在无限时域二次成本框架内使用动态规划方法, 从而产生最佳的防御和攻击策略, 增强 CPS 对复杂网络操纵的鲁棒性. 文献 [52] 提出了一种无模型 Q-学习算法, 用于解决暴露于 DoS 攻击和 FDI 攻击的 CPS 中的最优控制问题.

3) 基于观测器的弹性控制: 设计鲁棒状态观测器 (如龙伯格观测器、滑模观测器、卡尔曼滤波器等), 从被攻击篡改或丢失的测量信号中尽可能准确地估计真实的系统状态$x(t)$或攻击信号. 控制器则基于估计的状态或补偿后的信号进行设计, 常用于解决状态信息被篡改导致控制信号计算错误的问题.

以离散信息物理系统为例, 具体描述如下:

$$\begin{cases} x(t+1) = Ax(t) + Bu(t), \\ y_i(t) = Cx(t). \end{cases}$$

其中: $x(t)$为系统状态, $y_i(t)$为测量输出, $u(t)$为控制输入, $i \in \{1, 2, \ldots, n\}$, n为传输通道的数量. 由于基于观测器的弹性控制主要目标是确定基于弹性观测器控制的稳定性条件, 使得闭环系统在 DoS 攻击不存在时保持稳定. 首先, 由于无法直接测量系统状态, 引入观测器进行状态估计. 针对离散 CPS 的动态观测器由下式给出:

$$\hat{x}(t+1) = A\hat{x}(t) + Bu(t) + \sum_{i=1}^{n} L_i(y_i(t) - C\hat{x}_i(t)).$$

其中: $\hat{x}(t)$为系统状态的估计值, L_i为观测器增益. 此外, 基于观测器的控制器形式如下:

$$u(t) = K\hat{x}(t),$$

其中K为控制器增益矩阵. 这种方法的特点在于通过状态或攻击信号的重构与补偿, 在模型驱动框架下实现受攻击系统的稳定运行.

近年来, 使用基于观测器的弹性控制解决 CPS 安全问题引起了人们的关注. 文献 [53] 研究了执行器与传感器遭受攻击时, 连续时间 CPS 基于事件触发机制的安全观测器控制问题. 通过将原始被控对象增广为离散时间系统, 提出离散时间安全观测器以实现状态与攻击信号的联合估计. 此外, 该方法创新性地设计了基于安全观测器的控制器, 通过执行器攻击估计值实现对攻击的中和补偿. 文献 [54] 研究了 DoS 攻击下具有多重传输 CPS 的基于观测器的弹性控制问题. 区别于采用现有静态输出反馈控制器, 本文采用基于观测器的控制器. 虽然这些贡献都大大推进了 CPS 在网络攻击下的应用, 但没有考虑 CPS 在实际环境中遭受避碰避障时的可适应性.

上述方法的有效性已在多种真实工业场景与实验平台中得到初步验证. 例如, 在无人机路径规划的实例中, 可将最优控制理论与凸优化方法应用于无攻击干扰下的实时路径规划问题. 该系统要求无人机基于机载传感器实时感知周围障碍物, 并将自身状态、与障碍物的距离及与终点的距离作为约束条件, 构建并求解一个凸优化问题以生成最优避障路径. 实验结果表明, 该控制策略不仅能有效避开静态与动态障碍物, 而且其规划出的路径长度相较于传统方法显著缩短, 提升了巡航效率. 这一实例验证了上述策略在计算效率与安全性能方面均能满足实际要求.

除上述安全控制方式外, CPS 在对抗网络攻击的安全控制领域仍有多种有效策略值得关注与研究, 例如滑模控制、自适应控制、模糊控制等控制方法[55-57]. 此外, 指令控制因其更加符合实际需求引起了广泛关注[58-59]. 值得注意的是, 单一控制策略往往难以全面应对复杂多变的网络攻击环境. 因此, 将多种控制方法进行协同设计与深度融合能够得到更好的控制效果[60].

3.2 针对物理层威胁的避碰避障

3.2.1 无攻击干扰下的避碰避障问题

在保障 CPS 物理层安全的框架下, 避碰避障问题是确保 CPS 在复杂环境中安全运行的核心挑战. 尽管第 3.1 节所述的安全控制策略能有效抵御信息层网络攻击 (如 DoS、FDI 攻击), 物理层的安全性仍需通过专门的避碰避障机制实现. 在多智能体 CPS 的物理层安全控制中, 无攻击干扰下的避碰避障是保障系统自主运行的核心能力, 其目标是在不存在外部恶意攻击的环境中, 实现多智能体高精度轨迹跟踪与协同优化的同时, 严格规避静态/动态障碍物及智能体间碰撞.

避碰避障问题需满足双重约束[61-62], 即安全距离约束

$$\begin{cases} \|q_i(t) - q_j(t)\| > r_{ij},\ \forall t \geqslant 0,\ i \neq j; \\ \|q_i(t) - \mu_l(t)\|_{D_l} > \rho_{il},\ \forall t \geqslant 0,\ l \in \mathcal{M}. \end{cases}$$

其中: $r_{ij} = \max(r_i, r_j)$为智能体间最小安全距离, ρ_{il}为智能体i与障碍物l在椭圆坐标系下的安全半径, μ_l 为障碍物中心, $D_l = D_l^{\mathrm{T}} > 0$ 为椭圆形状矩阵.

连通性保持为$\|q_i(t) - q_j(t)\| < R_c,\ \forall j \in \mathcal{N}_i$, $t \geqslant 0$, 其中R_c为通信半径, 确保拓扑连通性.

上述无攻击干扰下的避碰避障问题建模, 本质上构建了一个兼具运动学约束与安全要求的非线性控制问题. 为应对这一问题, 现有研究聚焦于以下方法: 人工势场法、障碍李雅普诺夫函数法以及数据驱动方法. 下文将系统分析这 3 种方法, 揭示其理论特性.

1) 人工势场法.

人工势场法 (artificial potential field, APF) 是解决 CPS 避碰避障问题的经典方法, 其核心思想是通过构建虚拟势场来引导智能体运动, 在无人机编队、移动机器人等领域已得到广泛应用. APF 通过吸引势场和排斥势场的叠加生成控制指令, 有

$$\boldsymbol{u}_i = -\nabla\phi_{\mathrm{att}}(\boldsymbol{q}_i - \boldsymbol{q}_{\mathrm{goal}}) - \sum_{j\in\mathcal{N}_i}\nabla\phi_{\mathrm{rep}}(\boldsymbol{q}_{ij}) - \sum_{k}\nabla\psi_{\mathrm{obs}}(\boldsymbol{q}_{ik}).$$

其中: 吸引势场为$\phi_{\mathrm{att}}(\boldsymbol{q})$, 其梯度方向始终指向目标点, 系数$k_{\mathrm{att}}$ 控制收敛速度; 排斥势场为$\psi_{\mathrm{obs}}(\boldsymbol{q}_{ik})$, 当智能体间距$\|\boldsymbol{q}_{ij}\|$小于安全距离$d_{\mathrm{safe}}$时激活, 产生径向排斥力.

近期, 研究者通过结构改进和算法融合两个维度对传统 APF 进行增强. 文献 [63] 通过引入排斥性 APF 梯度的时间导数作为阻尼项, 在实现无碰撞编队控制的同时减少相邻避碰振荡. 文献 [64] 提出了一种自适应 APF 方法, 用于自动驾驶汽车的防撞, 确保模型预测控制策略的实施路径平稳. 值得注意的是, 近年来 APF 取得了新的进展. 面对未知环境, 文献 [65] 采用强化学习求解 APF 的局部最优. 文献 [65] 面对动态障碍时将人工势场与强化学习相结合, 实现了复杂环境下的安全动态避障.

2) 障碍-李雅普诺夫函数法.

障碍-李雅普诺夫函数法是解决 CPS 避碰避障问题的核心工具, 其核心思想是通过构造具有排斥特性的函数, 将安全约束嵌入控制器设计. 如文献 [66] 所述, 该方法能严格保证安全性, 同时保持跟踪性能.

障碍函数法的核心在于构造具有定向排斥特性的势场函数. 对于椭圆障碍物, 其障碍函数可构造为

$$B(q)=\begin{cases}\dfrac{\eta(1+\cos\pi\xi)}{(\|q-o\|_D^2-r^2)^{k+1}}, & \|q-o\|_D<R;\\ 0, & \text{otherwise.}\end{cases}$$

参数设计准则如下: 形状矩阵D反映椭圆几何特征, $\xi=(\|q-o\|_D^2-r^2)/(R^2-r^2)$实现过渡光滑化, 增益系数$\eta$和指数$k$共同调节障碍强度.

上述静态障碍函数虽能确保避碰, 但会因突变斥力导致系统动态特性恶化. 为此, 可以通过积分结构消除传统加性李雅普诺夫障碍函数在障碍区的动态失配问题, 构造如下复合李雅普诺夫障碍函数, 进而通过该函数分析系统的稳定性:

$$V=\underbrace{\frac{1}{2}\Big(\int_0^t B(q(\tau))\mathrm{d}\tau+1\Big)}_{\text{障碍积累项}}\underbrace{\|z_1\|^2}_{\text{Lyapunov 项}}.$$

近年来, 障碍-李雅普诺夫函数法有了新进展. 文献 [66] 通过将障碍函数积分项与传统李雅普诺夫函数相乘, 解决了动态障碍环境下跟踪控制与避障的冲突问题. 文献 [66] 开发了分布式双重障碍函数框架, 结合命令调节器实现多智能体系统的协同避障与轨迹跟踪.

3) 深度强化学习方法.

传统方法 (如人工势场法、障碍函数法) 依赖于精确的环境建模, 难以应对动态、未知的复杂环境. 近年来, 深度强化学习 (deep reinforcement learning, DRL) 因其无模型学习和自适应优化的能力, 成为避碰避障的重要研究方向, 能够通过学习智能体的交互经验, 动态优化策略. 相较于传统方法, DRL 的优势在于: 无需精确环境建模, 通过试错学习直接生成避障策略; 可处理高维状态空间 (如 LiDAR 点云、视觉输入); 支持多智能体协同优化.

文献 [67] 提出的多调节器辅助强化学习, 通过课程学习和人工势场引导解决多障碍物场景下的协同包围问题. 为了处理障碍物数量未知的复杂环境, 文献 [68] 采用长短期记忆网络对动态障碍物进行编码, 从而提高了避障效率. 针对两个受安全约束的动态系统, 文献 [68] 为了进一步保证系统的安全, 结合控制障碍函数和离线学习技术, 提出了安全感知的追逃策略.

3.2.2　外部攻击干扰下的避碰避障问题

外部攻击干扰下的避碰避障问题是一个在自动驾驶、无人机、机器人系统等领域的关键挑战, 涉及在恶意攻击 (如 DoS 攻击或 FDI 攻击) 下, 确保智能体避免相互碰撞以及避开障碍物, 是近年来逐渐受到关注的一个话题.

除了上述 4 种分类方法, 按照智能体感知环境的方式, 避碰避障方法可以分为基于通信避障以及基于传感器 (如车载摄像头或者雷达) 避障. 大多数避碰避障方法是基于传感器的, 这也是最直接有效的方法. 目前, 大多数文献均假设智能体的感知范围是圆形或者球形, 只要其他物体处于智能体的感知范围之内, 智能体就能够获得二者的距离信息, 从而避免碰撞. 但是在实际应用中, 由于障碍物的视线遮挡, 智能体无法知晓处于障碍物另一端的智能体的距离信息, 这使得智能体的感知范围并非是规则形状, 此时智能体之间的通信就显得尤为重要. 在实际应用中, 外部的攻击干扰主要集中于通信方面, 如图 3 所示. 因此, 如何在攻击干扰下确保通信的可靠性, 并将结果用于智能体避碰避障, 是需要解决的一个问题.

在 DoS 攻击方面, 文献 [69] 应用障碍李雅普诺夫函数, 研究了在异步 DoS 攻击下的不确定非线性多智能体系统的无碰撞自适应模糊安全编队控制问题. 智能体之间通过相互通信来估计领导者的状态, 并用于避碰避障中. 文献 [70] 提出了一种基于障碍函数的新型弹性避碰策略, 能够在异步 DoS 攻击下根据智能体的实时状态生成避碰最优共识命令. 文献 [71] 利用人工势场法设计具有连续偏导数的势函数, 并将其融入弹性分布式编队估计器中, 提出了在间歇执行器故障和同步 DoS 攻击下, 针对具有未知非线性动力学的欠驱动无人水面艇的无碰撞分布式模糊自适应弹性编队控制方案. 在文献 [72] 中, DoS 攻击会增加通信网络的服务时间, 并导致额外的传输延迟, 从而增加编队中车辆发生追尾碰撞的风险. 该文章提出了一种在 DoS 攻击和外部干扰下的车辆编队控制方法, 保证在攻击持续时间的严格上限内, 在所有空间变化的 DoS 攻击下, 既能跟踪期望的间距策略, 又能跟踪空间变化的参考速度.

在隐秘攻击和 FDI 攻击方面, 文献 [73] 为联网自动驾驶车辆提供了一种异常检测方案, 利用半定规划形式的风险评估工具, 量化隐秘攻击的潜在影响, 当风险值达到一定条件时, 认为系统受到了网络攻击. 文献 [74] 在此基础上结合状态估计器以及攻

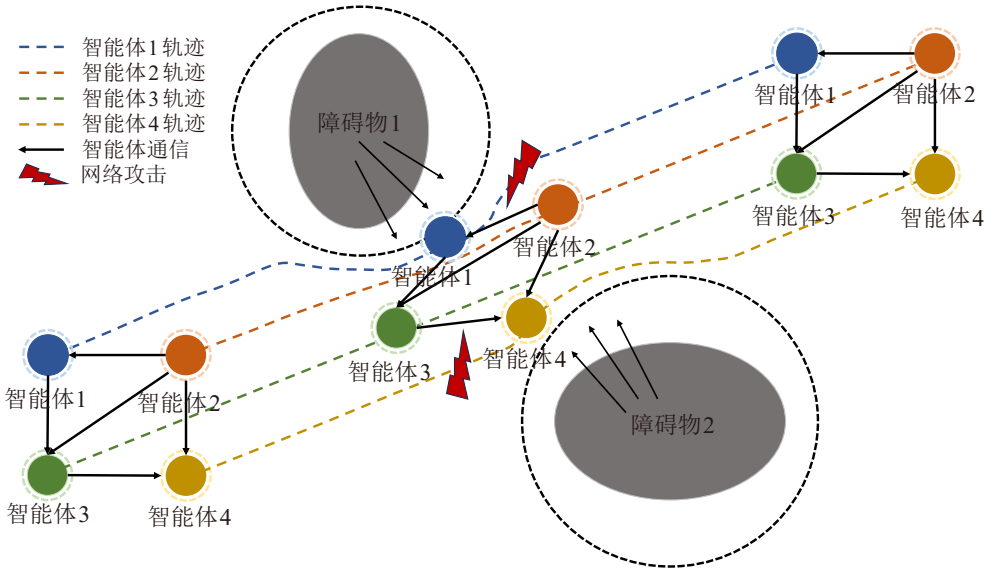

图3 攻击干扰下的避碰避障

击监测器, 设计了一种自适应巡航控制控制器, 最大限度地减小隐秘式 FDI 攻击的影响. 文献 [75] 提出了一种基于新型观测器的辅助信号方法, 确保 CPS 抵御 FDI 攻击, 并根据观测器数据提出一种基于指数控制障碍函数的新型框架, 根据此方法设计出的控制器能够保证系统在 FDI 攻击下实现弹性跟踪和避碰避障. 在文献 [76] 中, 机器人周围障碍物 (包括其他机器人) 的状态需要通过通信来获取, 在此过程中有可能会受到欺骗攻击, 基于此提出一种鲁棒运动规划算法, 将去噪自编码器与深度强化学习模型相结合, 以减轻在不同数量障碍物环境中欺骗攻击的影响. 结果表明, 无论是否受到攻击, 该方法都能学习一个编码器和解码器来近似障碍物的准确位置.

以上调研结果表明, 目前, 在网络攻击下实现避碰避障的方式可以分为两类: 一类是采用基于数据的方法, 如 DRL, 根据历史数据来近似估算出障碍物的位置; 另一类是依靠智能体所装配的传感器测量智能体与障碍物之间的距离, 以此实现智能体的避碰避障. 无论何种方式, 目前关于攻击干扰和避碰避障的理论分析都具有很强的独立性, 关于二者关系 (比如在攻击满足何种条件时能够确保智能体避碰避障的有效性) 的分析较少, 这是未来值得探究的地方.

4 攻击检测

4.1 基于模型的方法

随着 CPS 的蓬勃发展, 其安全问题也越来越受到重视, 同时面临的安全挑战也越来越严峻. 一般而言, 恶意攻击者通常在信息层面展开攻击. 一个经典的 CPS 如图 4 所示, 控制中心接收系统发来的状态信息并进行处理, 之后向控制器发送对应的控制信号, 控制器再对系统进行控制, 以此构成循环. 攻击者可在状态传输信道和控制信号传输信道等信息层面展开攻击, 攻击检测器也常常被安置在这些位置. 常见的攻击手段有 FDI 攻击、DoS 攻击、重放攻击、时间戳攻击等[77-78]. 一旦攻击者得手, 可能在物理层面引发灾难性的后果: 电网瘫痪、工业过程失控或自动驾驶车辆碰撞. 真实案例有伊朗核设施遭受黑客攻击导致失控[79], 乌克兰电网过载导致大面积瘫痪[80] 等. 随着系统互联程度加深, FDI、DoS 等攻击可绕过传统网络安全机制, 直接篡改传感器数据或控制指令, 导致状态估计失真和决策失效. 当前主流检测方法围绕异常感知与攻击辨识两大维度展开: 基于模型的方法利用系统动力学特性, 通过卡尔曼滤波器或观测器生成残差信号, 分析其统计偏差识别异常 (如卡方检测器), 或借助未知输入观测器主动解耦攻击信号. 故基于模型的攻击检测方法可大致分为两类, 即被动检测与主动检测. 下面详细介绍这两类方法.

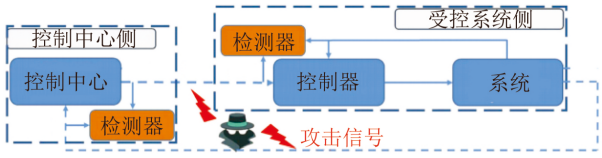

图4 针对 CPS 的攻击检测

4.1.1 被动检测

被动检测是利用提前设置好的检测器接收系统的实时信息, 通过其统计特性进行检测. 由于系统的输出值可能过大, 为了减少计算负担, 检测器通常接收的是系统的残差信号 (即系统输出值与预测值之

间的差), 这些信号通常由卡尔曼滤波器、状态观测器等产生. 为了更好地检测残差的统计特性, 检测器通常是卡方检测器或者 K-L 散度检测器. 卡方检测器的一般形式如下:

$$g_k = \sum_{i=k-J+1}^{k} z_i^{\mathrm{T}} \mathcal{S}^{-1} z_i \underset{H_1}{\overset{H_0}{\gtrless}} \gamma.$$

其中: g_k为待检验参数, γ为检测阈值, J为检测时间窗口大小, $z_i^{\mathrm{T}}\mathcal{S}^{-1}z_i$遵循卡方分布. 卡方检测器的运作流程如下: 首先, 利用卡尔曼滤波器或状态观测器生成残差序列; 其次, 计算残差的协方差矩阵并求其逆矩阵, 将残差归一化为标准统计量; 最后, 构造假设检验 —— 零假设 (H_0) 代表无攻击, 此时归一化残差的平方和 (即检验统计量) 服从卡方分布. 若统计量超过预设阈值 (根据显著性水平确定), 则拒绝 H_0并判定为攻击. 该检测器依赖线性高斯系统假设, 对于带有非高斯噪声的系统可靠性较差, 且要求残差协方差矩阵可逆. 但卡方检测器对突发的强攻击灵敏度高, 被广泛应用于工业控制系统[23]. K-L 散度检测器基于信息论, 通过度量实际数据分布与正常行为模板的差异识别攻击[54], 其工作原理包含两个阶段: 离线阶段建立正常工况下系统信号 (如传感器数据、通信流量) 的概率分布模型 (如高斯混合模型); 在线阶段实时计算当前数据窗口分布与参考模型的 K-L 散度 (即相对熵), 该值反映了两分布间的信息损失量. 若散度值持续超过动态阈值, 则判定存在攻击. K-L 散度的优势在于能捕捉细微的统计特性偏移 (如隐蔽攻击导致的分布缓慢漂移), 且适用于非线性、非高斯系统, 典型应用包括检测电网中的 FDI 攻击和工业物联网的异常通信模式, 但对建模精度要求较高, 且计算开销较大.

4.1.2　主动检测

以上被动检测方法均存在一定的缺陷, 为突破被动检测的瓶颈, 基于未知输入观测器的方法被提出. 该方法通过设计特殊观测器结构, 将攻击信号建模为“未知输入”并实现解耦估计. 这类方法不仅生成残差, 更能主动重构攻击信号本身, 为检测提供新维度. 文献 [40] 进一步发展了这一思路: 设计攻击信号无偏估计器, 通过并行于状态估计的观测器实时提取攻击信号, 进而创新性地提出平均恶意扰动功率检测机制, 利用攻击信号的功率统计特征构建假设检验 —— 无攻击时信号功率稳定在理论噪声水平, 而攻击注入会显著抬升功率谱密度. 该方法无需依赖残差协方差矩阵可逆等强假设, 可同时识别隐蔽攻击与非稳态攻击 (如 DoS 攻击), 实现了从被动响应到主动感知的范式转变, 为集体可观测系统的安全防护提供了新路径. 然而, 应当指出的是, 主动检测方法也存在局限性, 如计算复杂度高, 对系统模型精确性依赖强, 且易受模型失配和噪声干扰影响, 可能导致误报或实现困难.

4.2　基于数据驱动的方法

相较于依赖精确系统动力学模型的检测方法, 基于数据的方法直接从系统运行产生的海量数据 (如网络流量包、系统日志、传感器读数等) 中学习正常行为模式或攻击特征, 无需对系统内部机制有先验的精确建模. 这种方法在处理高维、非线性、动态演变的网络攻击, 特别是面对 CPS 中复杂的异构数据融合时, 展现出强大的适应性和潜力. 根据学习范式和对攻击知识的利用方式, 数据驱动的检测方法主要可分为误用检测、异常检测、混合检测、对抗生成式检测和其他检测方法[81].

4.2.1　误用检测

误用检测是网络入侵检测中应用最广泛的范式之一, 其核心思想是将观测到的网络流量或系统行为与已知攻击的预定义模式 (“签名”) 数据库进行匹配. 在深度学习语境下, 这意味着使用有标签的数据集 (包含标记的正常流量和各类已知攻击样本) 以监督学习方式训练模型 (如深度神经网络 (deep neural network, DNN)、卷积神经网络 (convolutional neural network, CNN)、LSTM 等)[82-85]. 模型通过学习这些已知攻击的特征模式, 实现对同类或高度相似攻击的高精度识别.

文献 [86] 强调前馈 DNN 模型在识别网络入侵中的有效性, 但与传统的机器学习模型相比具有较高的计算成本, 会降低训练速度, 并可能导致次优解决方案. 文献 [82] 采用 DNN 和 LSTM 等深度学习算法来检测已知与未知的 DoS/分布式 DoS 攻击, 此外, 跨数据集测试表明, 模型初始无法识别未知攻击, 但通过定期重新训练, 其检测性能获得了显著提升. 文献 [85] 提出了内核辅助主成分分析和混沌蜜獾优化算法用于特征提取和选择, 并采用门控注意力双长短期记忆模型对各种类型的攻击进行分类.

误用检测范式的最大优势在于对已知攻击的高检测率和低误报率. 然而, 其核心局限性在于高度依赖先验的已知攻击签名库. 对于训练数据中未出现过的未知攻击、特征分布与已知攻击显著不同的非典型攻击以及能够持续动态改变自身特征 (签名) 以逃避检测的多态攻击, 传统误用检测模型的识别能力通常较差. 提升其对新型攻击的适应性往往需要

持续收集新攻击样本并频繁进行模型再训练, 带来较高的维护成本和延迟. 因此, 该范式更适合防御已知的、模式相对固定的攻击.

4.2.2 异常检测

异常检测范式基于一个基本假设: 攻击行为会显著偏离系统或网络的正常行为模式. 与误用检测不同, 它不依赖于具体的攻击签名知识, 在深度学习中, 这通常通过无监督或半监督学习实现. 模型 (如深度自编码器 (deep autoencoder, DAE)、变分自编码器 (variational autoencoder, VAE)、LSTM-自编码器等) 仅使用大量的正常行为数据进行训练, 学习并构建“正常”的基准模型[87-90]. 在检测阶段, 任何导致模型重构误差显著升高或不符合学习到的正常分布的数据实例, 即被判定为异常.

文献 [87] 提出一种基于 LSTM-自编码器和单类支持向量机的混合攻击检测方法. LSTM-自编码器学习数据中的潜在特征表示, 然后将该潜在信息送到单类支持向量机以用于进一步分类. 实验结果表明, 所提出的混合模型能够有效识别网络流量中的异常. 文献 [88] 提出一种混合两阶段学习技术, 第 1 阶段使用条件 VAE 减少已知异常的错误分类, 第 2 阶段采用极值理论减少推断未知异常的错误分类风险. 与其他方法相比, 虽然该方法在 NSL-KDD 和 CICIDS 2017 数据集上具有更好的性能, 但对于未知攻击的真正率仍然较低. 文献 [90] 使用 DAE 进行攻击检测并且具有较低的误报率.

异常检测的核心优势在于其无需先验攻击知识, 理论上能够检测任何偏离正常行为的未知攻击、非典型攻击和多态攻击, 这使得它在应对新型和演化攻击方面具有天然优势. 然而, 其面临的主要挑战是高误报率. 此外, 处理高度不平衡的数据 (正常样本远多于攻击样本) 和模型训练中的近似优化损失也是需要关注的问题.

4.2.3 对抗生成式检测

对抗生成式检测利用生成对抗网络 (generative adversarial networks, GAN)、对抗性自动编码器等生成式模型, 生成对抗样本增强模型鲁棒性, 或者生成非典型和多态网络攻击, 并评估其脆弱性[91].

文献 [92] 提出了一种结合去噪自编码器和瓦尔瑟曼生成对抗网络的新架构, 以解决高维度、大规模且不平衡的网络流量数据问题, 并增强了基于异常的攻击检测. 文献 [93] 采用混合深度卷积生成对抗网络来构建智能攻击检测系统, 该系统可以识别各种攻击, 包括物联网网络的对抗攻击. 文献 [91] 采用 GAN 模型来合成可以绕过攻击检测的对抗性多态分布式 DoS 攻击. 文献 [94] 将自编码器与一个通过 GAN 训练的跨导 LSTM 网络相结合, 将其得到的潜在表示传递给多层感知器以执行检测任务.

对抗生成式检测为应对日益复杂的规避性攻击 (如多态攻击) 提供了强大工具, 特别是在增强模型鲁棒性 (通过对抗训练) 和解决训练数据稀缺/不平衡问题方面潜力巨大. 然而, 该范式也面临显著挑战: 训练过程复杂且不稳定 (如 GAN 的模式崩溃问题), 生成样本的真实性和功能性保障困难, 计算开销大, 选择合适的特定应用生成模型并非易事. 此外, 大多数研究假设攻击者拥有目标模型的完整知识 (白盒), 而实际攻击往往是黑盒的, 这之间的差距也需要进一步研究.

4.2.4 混合检测

混合检测范式旨在融合多种检测方法的优势克服各自的不足, 该方法通常利用监督学习组件高精度识别已知攻击, 同时利用无监督/半监督学习组件来捕捉未知或异常行为, 这种结合可以是串行的 (先误用后异常, 或反之) 或并行的 (两者同时运行并综合结果).

文献 [95] 提出了开放集分类网络用于识别已知和未知攻击, 开放集分类网络包括一个基于 CNN 的分类器和一个语义嵌入聚类方法, 实验结果表明了所提出方法在发现和学习未知攻击方面的可行性. 文献 [96] 介绍了一种入侵检测系统, 它可以使用由双向 LSTM、高斯混合模型和增量学习方法组成的混合框架来检测未知的分布式 DoS 攻击. 文献 [97] 提出了一种基于开放集识别的入侵检测方案, 用于分类已知和未知的网络攻击. 该混合模型由 CNN 模型和 Transformer 编码器模型相结合的预训练模块组成. 文献 [98] 提出了 RFG-HELAD 模型用于精细级别的攻击检测, 该模型由一个基于 DNN 的 K 分类模型和一个结合 GAN 与 K 近邻算法的 $K+1$ 分类模型组成.

混合范式理论上具有高检测率 (对已知攻击) 和检测未知/多变攻击的能力, 并有望降低纯异常检测的高误报率. 然而, 设计高效的混合系统面临架构复杂性和集成挑战. 如何最优地选择和组合不同范式的组件, 设计有效的信息融合与决策机制, 平衡检测率与误报率都是关键问题. 此外, 增加的组件通常带来更高的计算开销和训练复杂度, 在资源受限的 CPS 环境中需要特别考虑.

4.2.5 其他基于数据的检测方法

除上述核心检测方法外, 迁移学习 (transfer

learning, TL) 和 DRL 作为新兴技术, 在基于数据的攻击检测, 特别是在 CPS 动态环境中, 展现出独特价值.

TL 旨在将从一个任务 (源域, 通常数据丰富) 学到的知识迁移到相关但不同的任务 (目标域, 可能数据稀缺, 如特定 CPS 场景或新型攻击) 上, 以提升目标任务的性能. 文献 [99] 通过利用 CNN、自适应架构和 TL 技术, 提出了一种创新的分布式 DoS 攻击检测方法, 实验结果表明, 所提出的自适应 TL 方法能够有效识别攻击类别. 文献 [100] 提出一种基于深度 TL 的攻击检测系统, 该框架采用一种 3 层结构的方法, 协同集成了 CNN、遗传算法和引导聚合集成技术, 并表现出优异的性能.

DRL 将深度学习的感知能力与强化学习的决策能力相结合. 在攻击检测中, DRL 智能体通过与环境交互, 根据采取的检测动作 (如分类、告警) 获得的奖励或惩罚 (如检测成功、误报、漏报) 来学习最优的检测策略. 文献 [101] 提出一种新的多智能体强化学习体系结构, 实现了自动、高效、鲁棒的攻击检测, 实验结果表明, 该方法能够有效处理类不平衡问题, 并以极低的误报率提供细粒度的攻击分类. 文献 [102] 提出了一种结合联邦学习和 DRL 的攻击检测框架, 该框架允许分散的智能体使用其数据源训练局部模型, 并在非实时无线网络智能控制器上将模型聚合为全局模型以指导决策.

TL 和 DRL 为解决数据稀缺、环境适应性和持续学习等问题提供了新思路. 然而, TL 面临负迁移 (源域知识对目标域有害) 和过拟合风险, DRL 则需要大量的交互数据进行训练, 计算复杂度非常高, 且通常依赖于模拟环境 (可能与真实环境存在差距), 构建能充分体现 CPS 复杂性和攻击多样性的高保真模拟环境本身也是一大挑战.

总体而言, 基于模型的方法在系统模型精确、噪声特性已知的理想情况下, 具备坚实的理论保障和可预测的性能, 其计算效率高, 特别适合资源受限、对确定性要求高的场景 (如底层设备控制回路). 然而, 其性能高度依赖于模型的精确性, 对于非线性、强耦合或模型不确定的复杂系统, 其检测效果会显著下降.

相比之下, 基于数据驱动的方法在检测复杂、隐蔽和未知攻击方面表现出更强的适应性和更高的准确率, 其擅长从海量、高维的异构数据 (如网络流量、系统日志) 中学习异常模式, 但代价是需要大量的训练数据、较高的计算资源, 且决策过程如同“黑箱”, 可解释性差. 此外, 数据驱动方法常受困于高误报率, 且对训练数据未覆盖的新型攻击变种可能失效.

以深度学习为代表的数据驱动攻击检测方法已成为应对 CPS 复杂安全威胁的重要手段. 误用检测在已知攻击上高效精准, 但对新型攻击适应性差; 异常检测能发现未知偏差, 却饱受高误报率困扰; 混合检测试图融合优势, 但设计复杂; 对抗生成式方法在提升鲁棒性和数据增强上潜力巨大, 但训练复杂且样本真实性难保障; TL 和 DRL 则为数据稀缺、环境适应和持续学习提供了新途径, 但各有其应用挑战 (负迁移、高计算成本). 选择何种范式需综合考虑 CPS 的具体应用场景、可用数据、资源限制以及对检测已知/未知攻击、误报率容忍度的不同需求.

5 总结与展望

本文立足于防护者视角, 对近年来网络攻击下 CPS 安全防护的研究进展进行了系统梳理, 重点围绕安全状态估计、安全控制和攻击检测 3 大核心防御技术展开深入分析. 尽管上述技术在理论与应用层面均取得了显著进展, 为构建安全、可靠的 CPS 奠定了坚实基础, 但纵观现有研究体系, 仍存在诸多共性瓶颈与待突破的难题. 同时, 随着新一代信息技术的融合与应用场景的不断拓展, CPS 安全防护也面临着新的挑战与需求.

5.1 存在的问题与挑战

值得强调的是, 随着 5G/6G、数字孪生、群体智能等技术的深度融合, CPS 安全防护面临复杂的问题与挑战:

1) 强模型依赖与场景敏感: 模型方法依赖精确动力学与噪声假设, 难以覆盖复杂工况与隐蔽攻击; 数据驱动方法虽对未知攻击更具适应性, 但对训练分布与传感条件敏感.

2) 数据与仿真环境不足: 高质量攻击数据稀缺, 高保真数字孪生/仿真环境构建成本高且与真实系统存在分布差异, 制约了检测与鲁棒控制策略的可迁移性评估.

3) 工程实用性约束: 在资源受限与毫秒级响应需求下, 如何在可计算与可验证前提下维持检测精度、控制性能与稳定冗余, 仍缺乏系统化方案.

5.2 未来展望

未来研究需在深化现有防护技术的同时, 重点关注以下方向:

1) 分布式协同防御架构: 针对大规模分布式 CPS (如无人机集群、智能电网), 设计轻量级协同估计与控制框架, 解决恶意节点隔离、通信延迟与资源

约束下的全局一致性难题.

2) 跨场景泛化能力: 提升数据驱动方法在多种攻击模式、动态环境及跨系统迁移中的适应性, 突破数据依赖与迁移瓶颈.

3) 安全-效率-实时性权衡: 优化高维状态空间下的算法复杂度 (如凸松弛精度), 满足工业控制系统的毫秒级响应需求.

4) 跨层一体化防护: 建立信息层攻击检测-物理层安全控制-状态估计的闭环联动机制, 阻断攻击在信息-物理层的传播 (如 FDI 攻击诱发物理碰撞).

参考文献 (References)

[1] 郭楠, 贾超. 《信息物理系统白皮书 (2017)》解读 (下) [J]. 信息技术与标准化, 2017(5): 42-47.
(Guo N, Jia C. Interpretation of “cyber-physical systems white paper (2017)”(part two)[J]. Information Technology & Standardization, 2017(5): 42-47.)

[2] 李洪阳, 魏慕恒, 黄洁, 等. 信息物理系统技术综述[J]. 自动化学报, 2019, 45(1): 37-50.
(Li H Y, Wei M H, Huang J, et al. Survey on cyber-physical systems[J]. Acta Automatica Sinica, 2019, 45(1): 37-50.)

[3] Kim S, Park K J, Lu C Y. A survey on network security for cyber-physical systems: From threats to resilient design[J]. IEEE Communications Surveys & Tutorials, 2022, 24(3): 1534-1573.

[4] Chen T M. Stuxnet, the real start of cyber warfare?[J]. IEEE Network, 2010, 24(6): 2-3.

[5] 汤奕, 陈倩, 李梦雅, 等. 电力信息物理融合系统环境中的网络攻击研究综述[J]. 电力系统自动化, 2016, 40(17): 59-69.
(Tang Y, Chen Q, Li M Y, et al. Overview on cyber-attacks against cyber physical power system[J]. Automation of Electric Power Systems, 2016, 40(17): 59-69.)

[6] Zhang D, Wang Q G, Feng G, et al. A survey on attack detection, estimation and control of industrial cyber-physical systems[J]. ISA Transactions, 2021, 116: 1-16.

[7] 杨光红, 芦安洋, 安立伟. 网络攻击下的信息物理系统安全状态估计研究综述[J]. 控制与决策, 2023, 38(8): 2093-2105.
(Yang G H, Lu A Y, An L W. A survey on secure state estimation of cyber-physical systems under cyber attacks[J]. Control and Decision, 2023, 38(8): 2093-2105.)

[8] Ding D R, Han Q L, Ge X H, et al. Secure state estimation and control of cyber-physical systems: A survey[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(1): 176-190.

[9] Lu A Y, Yang G H. Detection and identification of sparse sensor attacks in cyber-physical systems with side information[J]. IEEE Transactions on Automatic Control, 2023, 68(9): 5349-5364.

[10] He W L, Xu W Y, Ge X H, et al. Secure control of multiagent systems against malicious attacks: A brief survey[J]. IEEE Transactions on Industrial Informatics, 2022, 18(6): 3595-3608.

[11] Zhang J N, Ma Y C. Complex dynamic networks for multiple attacks: A jump-like event-triggered controller based on neural network model[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(5): 4470-4480.

[12] Lu A Y, Yang G H. Input-to-state stabilizing control for cyber-physical systems with multiple transmission channels under denial of service[J]. IEEE Transactions on Automatic Control, 2018, 63(6): 1813-1820.

[13] Wang Q J, Qian Y N, Lu A Y. Data and model-based switching observer for cyber-physical systems against sparse sensor attacks[J]. International Journal of Systems Science, 2025: 1-13.

[14] 符莎, 李平, 赵民新. 一类混合噪声系统的重放攻击检测方法[J]. 控制与决策, 2025, 40(10): 3065-3072.
(Fu S, Li P, Zhao M X. Replay attack detection method for a class of mixed noise systems[J]. Control and Decision, 2025, 40(10): 3065-3072.)

[15] Su L, Fang S N, Liu Z J, et al. Secure control for discrete-time hidden Markov jump systems subject to replay attacks via output feedback[J]. Journal of Control and Decision, 2023, 10(4): 584-595.

[16] Zhao D, Shi Y, Ding S X, et al. Replay attack detection based on parity space method for cyber-physical systems[J]. IEEE Transactions on Automatic Control, 2025, 70(4): 2390-2405.

[17] Sun Q D, Yang G H. Secure state estimation for continuous-time cyber-physical systems under stochastic attacks and faults[J]. IEEE Transactions on Automatic Control, 2025, 70(9): 6119-6126.

[18] Zhang J N, Ma Y C, Xu Y N, et al. Parallel adaptive event-triggered asynchronous control for two-time-scale fuzzy semi-Markov jump systems under deception attacks[J]. Journal of the Franklin Institute, 2023, 360(17): 12941-12968.

[19] Song H Y, Yao H Y, Shi P, et al. Distributed secure state estimation of multi-sensor systems subject to two-channel hybrid attacks[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2022, 8: 1049-1058.

[20] Lu A Y, Yang G H. Distributed secure state estimation for linear systems against malicious agents through sorting and filtering[J]. Automatica, 2023, 151: 110927.

[21] Gao R, Yang G H, Wasly S. Resilient distributed state estimation with multi-hop communication[J]. Automatica, 2024, 168: 111823.

[22] Shoukry Y, Nuzzo P, Puggelli A, et al. Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach[J]. IEEE Transactions on Automatic Control, 2017, 62(10): 4917-4932.

[23] Lu A Y, Yang G H. Secure state estimation for multiagent systems with faulty and malicious agents[J]. IEEE Transactions on Automatic Control, 2020, 65(8): 3471-3485.

[24] An L W, Yang G H. State estimation under sparse sensor attacks: A constrained set partitioning approach[J]. IEEE Transactions on Automatic Control, 2019, 64(9): 3861-3868.

[25] Mishra S, Shoukry Y, Karamchandani N, et al. Secure state estimation against sensor attacks in the presence of noise[J]. IEEE Transactions on Control of Network Systems, 2017, 4(1): 49-59.

[26] An L W, Yang G H. Fast state estimation under sensor attacks: A sensor categorization approach[J]. Automatica, 2022, 142: 110395.

[27] Li J H, Yang G H. Disturbance decoupled secure state estimation: An orthogonal projection-based method[J]. Automatica, 2023, 147: 110740.

[28] An L W, Yang G H. Secure state estimation against sparse sensor attacks with adaptive switching mechanism[J]. IEEE Transactions on Automatic Control, 2018, 63(8): 2596-2603.

[29] Lu A Y, Yang G H. Switched projected gradient descent algorithms for secure state estimation under sparse sensor attacks[J]. Automatica, 2019, 103: 503-514.

[30] Wang H M, Hou Q. Secure state estimation for affine T-S fuzzy systems under sparse sensor attacks[J]. Journal of the Franklin Institute, 2024, 361(8): 106793.

[31] Mao Y W, Mitra A, Sundaram S, et al. On the computational complexity of the secure state-reconstruction problem[J]. Automatica, 2022, 136: 110083.

[32] Lu A Y, Yang G H. Secure state estimation under sparse sensor attacks via saturating adaptive technique[J]. IEEE Transactions on Control of Network Systems, 2023, 10(4): 1890-1898.

[33] Li Z S, Mo Y L. Secure distributed dynamic state estimation against sparse integrity attack via distributed convex optimization[J]. IEEE Transactions on Automatic Control, 2024, 69(9): 6089-6104.

[34] Niu M F, Wen G H, Lv Y Z, et al. Innovation-based stealthy attack against distributed state estimation over sensor networks[J]. Automatica, 2023, 152: 110962.

[35] Zhao X D, Liu L, Basin M V, et al. Event-triggered reverse attacks on remote state estimation[J]. IEEE Transactions on Automatic Control, 2024, 69(2): 998-1005.

[36] Han D, Mo Y L, Xie L H. Convex optimization based state estimation against sparse integrity attacks[J]. IEEE Transactions on Automatic Control, 2019, 64(6): 2383-2395.

[37] Chen Y, Kar S, Moura J M F. Resilient distributed parameter estimation with heterogeneous data[J]. IEEE Transactions on Signal Processing, 2019, 67(19): 4918-4933.

[38] Lu A Y, Yang G H. A polynomial-time algorithm for the secure state estimation problem under sparse sensor attacks via state decomposition technique[J]. IEEE Transactions on Automatic Control, 2023, 68(12): 7451-7465.

[39] Li Z S, Mo Y L. Efficient secure state estimation against sparse integrity attack for regular linear system[J]. International Journal of Robust and Nonlinear Control, 2023, 33(1): 209-236.

[40] Lu A Y, Yang G H. Attack detection from the perspectives of control center and controlled systems[J]. IEEE Transactions on Automatic Control, 2024, 69(12): 8662-8673.

[41] Jin Z W, Zhang S T, Hu Y Y, et al. Security state estimation for cyber-physical systems against DoS attacks via reinforcement learning and game theory[J]. Actuators, 2022, 11(7): 192.

[42] Naz R, K S, Shrivastava N A. Cyber attack detection against state estimation in CPPS: A data-driven approach[C]. The 23rd National Power Systems Conference. Indore, 2025: 1-6.

[43] Nagaraj K, Aljohani N, Zou S, et al. State estimator and machine learning analysis of residual differences to detect and identify FDI and parameter errors in smart grids[C]. The 52nd North American Power Symposium. Tempe, 2021: 1-6.

[44] Lu K D, Zhou L, Wu Z G. Evolutionary fractional-order extended Kalman filter of cyber-physical power systems[J]. IEEE Transactions on Cybernetics, 2025, 55(3): 1395-1408.

[45] Lu K D, Zhou L, Wu Z G. Representation-learning-based CNN for intelligent attack localization and recovery of cyber-physical power systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6145-6155.

[46] Raghuvamsi Y, Teeparthi K. Detection and reconstruction of measurements against false data injection and DoS attacks in distribution system state estimation: A deep learning approach[J]. Measurement, 2023, 210: 112565.

[47] Wang H Z, Ruan J Q, Wang G B, et al. Deep learning-based interval state estimation of AC smart grids against sparse cyber attacks[J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 4766-4778.

[48] VakilKandi S J, Bayat F, Jalilvand A, et al. Cyber-physical systems under hybrid cyber-attacks: Resilient event-triggered H_∞[J]. ISA Transactions, 2025, 163: 51-64.

[49] Miao Z H, Li M, Chen Y, et al. Event-triggered security defense control for remote motor under DoS attack[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(7): 4485-4493.

[50] Ma Y J, Li Z J. Neural network-based secure event-triggered control of uncertain industrial cyber-physical systems against deception attacks[J]. Information Sciences, 2023, 633: 504-516.

[51] Wu C W, Li X L, Pan W, et al. Zero-sum game-based optimal secure control under actuator attacks[J]. IEEE Transactions on Automatic Control, 2021, 66(8): 3773-3780.

[52] Fei C, Shen J, Qiu H L, et al. Data driven secure control for cyber-physical systems under hybrid attacks: A Stackelberg game approach[J]. Journal of the Franklin Institute, 2024, 361(6): 106715.

[53] Lu A Y, Yang G H. Event-triggered secure observer-based control for cyber-physical systems under adversarial attacks[J]. Information Sciences, 2017, 420: 96-109.

[54] Zhang C L, Yang G H, Lu A Y. Resilient observer-based control for cyber-physical systems under denial-of-service attacks[J]. Information Sciences, 2021, 545: 102-117.

[55] Yin X H, Zhang L, Zong G D. Composite sliding mode control for cyber-physical systems with multi-source disturbances and false data injection attack[J]. International Journal of Robust and Nonlinear Control, 2024, 34(15): 10649-10665.

[56] Jin X, Haddad W M, Yucelen T. An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems[J]. IEEE Transactions on Automatic Control, 2017, 62(11): 6058-6064.

[57] Chen Z B, Yu Z X, Li S G. Output feedback adaptive fuzzy inverse optimal security control against sensor and actuator attacks for nonlinear cyber-physical systems[J]. IEEE Transactions on Fuzzy Systems, 2024, 32(5): 2554-2566.

[58] An L W, Yang G H, Deng C, et al. Event-triggered reference governors for collisions-free leader-following coordination under unreliable communication topologies[J]. IEEE Transactions on Automatic Control, 2024, 69(4): 2116-2130.

[59] Liang S, Xu B, Sun S S, et al. Dynamic-command-limiting-based AOA constraint control of hypersonic flight vehicle[J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(1): 1163-1174.

[60] Liu J L, Yin T T, Cao J, et al. Security control for T-S fuzzy systems with adaptive event-triggered mechanism and multiple cyber-attacks[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(10): 6544-6554.

[61] An L W, Yang G H. Collisions-free distributed cooperative output regulation of nonlinear multiagent systems[J]. IEEE Transactions on Automatic Control, 2024, 69(11): 8072-8079.

[62] An L W, Yang G H, Wasly S. Obstacle avoidance in distributed optimal coordination of multirobot systems: A trajectory planning and tracking strategy[J]. IEEE Transactions on Control of Network Systems, 2024, 11(3): 1335-1344.

[63] Han H R, Cheng J, Lv M L, et al. Enhancing collision-free formation control in multiagent systems: An approach based on time-derivative of artificial potential functions[J]. IEEE Transactions on Cybernetics, 2025, 55(7): 3445-3456.

[64] Yang H J, He Y Q, Xu Y, et al. Collision avoidance for autonomous vehicles based on MPC with adaptive APF[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 1559-1570.

[65] Orozco-Rosas U, Picos K, Pantrigo J J, et al. Mobile robot path planning using a QAPF learning algorithm for known and unknown environments[J]. IEEE Access, 2022, 10: 84648-84663.

[66] Zhang L L, Yang G H. Secure adaptive trajectory tracking control for nonlinear robot systems under multiple dynamic obstacles: Safety barrier certificates[J]. IEEE Transactions on Industrial Electronics, 2022, 69(11): 11549-11559.

[67] Gou F D, Du H K, Zhao C Y, et al. A policy-guided reinforcement learning method for encirclement control in multiobstacle environment[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(9): 17034-17046.

[68] Yuan Z K, Yao C H, Liu X X, et al. Multiagent formation control and dynamic obstacle avoidance based on deep reinforcement learning[J]. IEEE Transactions on Industrial Informatics, 2025, 21(6): 4672-4682.

[69] Zhang J, Ning J, Tong S C. Adaptive fuzzy secure collision-free formation control for nonlinear MASs with DoS attacks[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2025, 55(8): 5705-5716.

[70] Yin J Y, Yang G H, Wang H M. Resilient collision-avoidance distributed optimal coordination strategy for Euler-Lagrangian systems subjected to asynchronous DoS attacks[J]. IEEE Transactions on Industrial Informatics, 2025, 21(9): 7142-7152.

[71] Wu X J, Zong G D, Wang H Q, et al. Collision-free distributed adaptive resilient formation control for underactuated USVs subject to intermittent actuator faults and denial-of-service attacks[J]. IEEE Transactions on Vehicular Technology, 2025, 74(9): 13606-13617.

[72] Gong J, Murguia C, Bayuwindra A, et al. Resilient controller synthesis against DoS attacks for vehicular platooning in spatial domain[J]. IEEE Transactions on Vehicular Technology, 2025, 74(5): 8251-8266.

[73] Yang T C, Murguia C, Lv C. Risk assessment for connected vehicles under stealthy attacks on vehicle-to-vehicle networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 13627-13638.

[74] Yang T C, Murguia C, Nešić D, et al. Toward crash-free autonomous driving: Anomaly detection and control for resilience to stealthy sensor attacks[J]. IEEE Internet of Things Journal, 2025, 12(1): 276-287.

[75] Zhang D T, Shi P, Lim C P, et al. Resilient tracking control of cyber-physical systems against false data injection attacks and obstacle avoidance[J]. IEEE

Transactions on Automation Science and Engineering, 2025, 22: 12596-12605.

[76] Tang W B, Zhou Y, Liu Y, et al. Robust motion planning for multi-robot systems against position deception attacks[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2157-2170.

[77] Guo H B, Sun J, Pang Z H. Analysis of replay attacks with countermeasure for state estimation of cyber-physical systems[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2024, 71(1): 206-210.

[78] Qu F Y, Yang N C, Liu H, et al. Time-stamp attacks on remote state estimation in cyber-physical systems[J]. IEEE Transactions on Control of Network Systems, 2024, 11(1): 450-461.

[79] Farwell J P, Rohozinski R. Stuxnet and the future of cyber war[J]. Survival, 2011, 53(1): 23-40.

[80] Sui T J, Mo Y L, Marelli D, et al. The vulnerability of cyber-physical system under stealthy attacks[J]. IEEE Transactions on Automatic Control, 2021, 66(2): 637-650.

[81] Sabeel U, Heydari S S, El-Khatib K, et al. Unknown, atypical and polymorphic network intrusion detection: A systematic survey[J]. IEEE Transactions on Network and Service Management, 2024, 21(1): 1190-1212.

[82] Sabeel U, Heydari S S, Mohanka H, et al. Evaluation of deep learning in detecting unknown network attacks[C]. 2019 International Conference on Smart Applications, Communications and Networking. Sharm El Sheikh, 2020: 1-6.

[83] Ho S, Al Jufout S, Dajani K, et al. A novel intrusion detection model for detecting known and innovative cyberattacks using convolutional neural network[J]. IEEE Open Journal of the Computer Society, 2021, 2: 14-25.

[84] Apruzzese G, Colajanni M, Ferretti L, et al. On the effectiveness of machine and deep learning for cyber security[C]. 2018 10th International Conference on Cyber Conflict. Tallinn, 2018: 371-390.

[85] Devendiran R, Turukmane A V. Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy[J]. Expert Systems with Applications, 2024, 245: 123027.

[86] Bangyal W H, Ahmad J, Rauf H T, et al. Evolving artificial neural networks using opposition based particle swarm optimization neural network for data classification[C]. International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies. Sakhier, 2019: 1-6.

[87] Said Elsayed M, Le-Khac N A, Dev S, et al. Network anomaly detection using LSTM based autoencoder[C]. Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks. New York: ACM, 2020: 37-45.

[88] Yang J, Chen X, Chen S W, et al. Conditional variational auto-encoder and extreme value theory aided two-stage learning approach for intelligent fine-grained known/unknown intrusion detection[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3538-3553.

[89] Ma J, Su W. Collaborative DDoS defense for SDN-based AIoT with autoencoder-enhanced federated learning[J]. Information Fusion, 2025, 117: 102820.

[90] Kumar V, Kumar K, Singh M, et al. NIDS-DA: Detecting functionally preserved adversarial examples for network intrusion detection system using deep autoencoders[J]. Expert Systems with Applications, 2025, 270: 126513.

[91] Chauhan R, Shah Heydari S. Polymorphic adversarial DDoS attack on IDS using GAN[C]. International Symposium on Networks, Computers and Communications. Montreal, 2020: 1-6.

[92] Arafah M, Phillips I, Adnane A, et al. Anomaly-based network intrusion detection using denoising autoencoder and wasserstein GAN synthetic attacks[J]. Applied Soft Computing, 2025, 168: 112455.

[93] Wu Y X, Nie L S, Wang S P, et al. Intelligent intrusion detection for Internet of things security: A deep convolutional generative adversarial network-enabled approach[J]. IEEE Internet of Things Journal, 2023, 10(4): 3094-3106.

[94] Yang J, Wu Y G, Yuan Y P, et al. LLM-AE-MP: Web attack detection using a large language model with autoencoder and multilayer perceptron[J]. Expert Systems with Applications, 2025, 274: 126982.

[95] Zhang Z, Zhang Y, Guo D, et al. A scalable network intrusion detection system towards detecting, discovering, and learning unknown attacks[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(6): 1649-1665.

[96] Shieh C S, Lin W W, Nguyen T T, et al. Detection of unknown DDoS attacks with deep learning and Gaussian mixture model[J]. Applied Sciences, 2021, 11(11): 5213.

[97] Hu X Y, Gu C X, Chen Y H, et al. OpenCBD: A network-encrypted unknown traffic identification scheme based on open-set recognition[J]. Wireless Communications and Mobile Computing, 2022, 2022(1): 1746373.

[98] Zhong Y, Wang Z L, Shi X G, et al. RFG-HELAD: A robust fine-grained network traffic anomaly detection model based on heterogeneous ensemble learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 5895-5910.

[99] Anley M B, Genovese A, Agostinello D, et al. Robust DDoS attack detection with adaptive transfer learning[J]. Computers & Security, 2024, 144: 103962.

[100] Latif S, Boulila W, Koubaa A, et al. DTL-IDS: An optimized intrusion detection framework using deep transfer learning and genetic algorithm[J]. Journal of Network and Computer Applications, 2024, 221: 103784.

[101] Tellache A, Mokhtari A, Korba A A, et al. Multi-agent

reinforcement learning-based network intrusion detection system[C]. 2024 IEEE Network Operations and Management Symposium. Seoul, 2024: 1-9.

[102] Abou El Houda Z, Moudoud H, Brik B. Federated deep reinforcement learning for efficient jamming attack mitigation in O-RAN[J]. IEEE Transactions on Vehicular Technology, 2024, 73(7): 9334-9343.

作者简介

芦安洋 (1991−), 男, 教授, 博士, 博士生导师, 主要研究方向为信息物理系统安全性、切换系统、自适应控制, E-mail: luanyang@ise.neu.edu.cn;

张佳楠 (1999−), 女, 博士生, 主要研究方向为信息物理系统安全性、模糊控制, E-mail: zhangjianan199@163.com;

王庆杰 (1995−), 男, 博士生, 主要研究方向为信息物理系统安全性、复杂系统, E-mail: qj_paper@163.com;

纪寒康 (1998−), 男, 博士生, 主要研究方向为信息物理系统安全性、复杂网络建模与分析, E-mail: jihankang001@126.com;

尹利榜 (2000−), 男, 硕士生, 主要研究方向为多智能体跟踪控制、避碰避障控制, E-mail: yin_libang@163.com;

朱立秋 (2001−), 男, 硕士生, 主要研究方向为多智能体避碰避障、信息物理系统安全性, E-mail: 2301061@stu.neu.edu.cn;

孙秉旭 (2002−), 男, 硕士生, 主要研究方向为信息物理系统安全性, E-mail: 15668466769@163.com.

科研团队简介

芦安洋教授科研团队隶属于东北大学“辽宁省自主无人系统安全运行技术重点实验室”, 依托东北大学控制科学与工程一流学科优势, 长期专注于自主无人系统安全运行与信息物理系统安全性等方向的研究. 目前, 实验室以杨光红教授、叶丹教授和董久祥教授等为学术骨干, 汇聚控制理论、人工智能、信息物理系统安全等多学科交叉背景的青年科研力量, 围绕无人系统智能认知、状态监测与故障诊断以及容侵控制等重大科学问题, 构建从基础理论、关键技术到工程示范的完整创新链条.

芦安洋教授主持国家自然青年基金项目 B 类 (原优青)、C 类和面上项目, 以及辽宁省优秀青年基金等多项科研课题, 并承担多项面向天然气管道、钢铁等行业的工程项目. 入选 2020 年国家博士后创新人才支持计划, 获得 2023 年中国自动化学会自然科学奖一等奖、2021 年中国自动化学会优秀博士学位论文奖、2020 年博士后创新人才支持计划优秀创新成果奖等. 发表 SCI 论文 37 篇, 其中一作/通信论文 24 篇, SCI 他引 1600 余次. 一作论文包括控制领域两大顶刊 TAC 及 Automatica 论文 14 篇 (含长文 7 篇), 单篇引用最高 322 次.

目前课题组正承担国家自然科学基金面上项目“非可靠通信环境下的多智能体系统多目标可靠跟踪方法研究”、青年基金项目“通信和计算资源受限下信息物理系统的安全状态估计研究”、辽宁省自然科学基金优秀青年基金“非可靠通信环境下车路协同系统信息物理两侧协同防护技术研究”和国家管网集团横向项目“压缩机顺控模拟系统搭建项目”等多个项目.