军事智能数据安全问题：对抗攻击威胁*

陆正之[1]，黄希宸[2]，彭　勃[1]
（1. 国防科技大学　试验训练基地，陕西　西安　710106；
2. 国防科技大学　电子科学学院，湖南　长沙　410073）

摘　要：人工智能技术已深入军事作战的各个领域，对现代战争形态进行了全面革新。数据作为军事智能模型的核心驱动力，为模型的有效运转提供了保障。然而，由于深度学习的不可解释性，对抗攻击技术的存在给当前军事智能模型带来了严峻的数据安全问题。这种威胁在智能系统的训练和推理过程中均可能产生，形式多样，难以防范。同时，受到对抗样本干扰的军事数据类型多样，敌方采取的欺骗手段也日趋复杂。因此，分析军事智能数据安全风险样态，并进一步给出军事智能数据风险的防范措施，希望能够为增强军事智能数据的安全性提供有益的参考和借鉴。

关键词：军事人工智能；数据安全；对抗攻击；物理对抗攻击
中图分类号：TP391　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2024.11.005
引用格式：陆正之，黄希宸，彭勃. 军事智能数据安全问题：对抗攻击威胁［J］. 网络安全与数据治理，2024，43（11）：23-28.

The data security of military intelligence: adversarial attacks

Lu Zhengzhi[1], Huang Xichen[2], Peng Bo[1]
（1. Test Center, National University of Defense Technology, Xi′an 710106, China;
2. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China）

Abstract: Artificial intelligence technology has now been deeply applied in various fields of military operations, comprehensively changing the shape of modern warfare. Data is the core driving force of military intelligence models, providing a guarantee for the effective operation of the models. However, due to the non-interpretability of deep learning, the existence of adversarial attack techniques has brought serious data security problems to current military intelligence models. On the one hand, such security threats come in various forms and can be affected during the full life cycle of training and reasoning of intelligent systems. On the other hand, the types of military data interfered by adversarial samples are complicated, and the means of implementing deception show a diversified trend. Therefore, this paper will analyse the security risk pattern of military intelligent data, and further give specific measures on how to prevent the risk of military intelligent data in the hope that it can provide certain references and lessons for improving the security of military intelligent data.

Key words: artificial intelligence for military; data security; adversarial attacks; physical adversarial attacks

0　引言

深度学习技术的迅猛进步已使全球达成共识：人工智能技术（Artificial Intelligence，AI）将与核技术、生物技术和航空航天技术并驾齐驱，成为影响国家安全的关键因素[1]。2017 年，美国率先提出“算法战”的构想，此后，全球主要大国纷纷将 AI 技术融入陆、海、空、天、网及电等多元化军事领域，以执行探测识别、威胁评估、情报分析及指挥决策等核心任务。2022 年，美国退役上将约翰·艾伦预言，未来战争将迈入“极速战”的新纪元。所谓“极速战”，指的是一种高度依赖 AI 主导，人类指挥官极少介入的超快速战争模式。这预示着人工智能成为“重塑战争法则”的革命性技术。

目前，在国际军事舞台上，已有诸如美国“捕食者”和“死神”无人机、以色列的“铁穹”防御系统、英国的“塔洛斯”无人步战车等多个实例展示了人工智能技术的实战应用。此外，在国际联合军事演习中频繁使用

*基金项目：长沙市杰出创新青年培养计划（kq2107002）

人工智能辅助决策系统，能够整合来自不同来源的数据，为指挥官提供快速、准确的战场信息，从而加快决策过程。

对抗性攻击与人工智能技术的安全鲁棒应用息息相关。由于 AI 模型内部工作机理存在着不可解释性，一旦所使用的数据遭受攻击或篡改，将对模型的输出结果造成严重影响。这种通过对数据施加某种恶意干扰使得人工智能发生错误的技术就被称为对抗性攻击（Adversarial Attacks）技术[2]。目前，已经有大量研究表明对抗攻击所带来的数据安全问题可以极大程度地影响军事智能模型。2021 年 8 月，兰德公司发布了《对抗性攻击如何影响美国军事人工智能系统》[3]，通过实例深入剖析了对抗攻击对军事智能系统和作战行动的影响，并建议应切实加强对模型与数据集的安全防护。同年，韩国提出使用对抗性迷彩贴图来在物理场景下对数据施加扰动，实现军事目标的智能对抗[4]。2022 年，英国研究了对抗攻击和数据的不确定性在混合战争（Hybrid Warfare）中的威胁。同年，Chen Yuwei 评估了对抗攻击在搜寻、锁定、追踪、瞄准、交战和评估各个阶段可能造成的数据安全威胁，并使用仿真作战软件进行了验证[5]。

关注对抗攻击所带来数据安全问题对军事智能模型的可靠稳定部署有着重要意义。本文旨在分析军事智能模型所面临的数据安全风险及其具体形态，阐述对抗攻击技术的应用场景与部署方式，以期为智能化建设的稳健推进提供坚实的安全保障。本文的核心内容如图 1 所示。首先对军事智能数据可能遭遇的四种风险形态进行了详尽阐述；然后列举了六种易受到对抗攻击威胁的军事领域常用数据类型，并详细描述了相关对抗攻击的具体部署方法；最后总结了为确保军事数据安全性所需采取的具体措施。

1　军事智能数据安全风险样态

数据主要作用于模型的训练和推理两个阶段。在这两个阶段中，数据信息面临着多种对抗攻击技术的威胁，不仅可能干扰军事智能模型的判断结果，更有可能导致数据泄露，对指挥决策产生深远影响。

1.1　训练数据安全风险

1.1.1　数据投毒攻击

数据投毒（Data Poisoning）攻击[6]主要作用于数据集的构建阶段，是通过向训练集中投放被污染数据导致数据分布异常，影响模型的决策边界，从而降低模型的可用性和有效性。现阶段主要数据投毒攻击方式有两种：模型偏斜（Model Skewing）和反馈误导（Feedback Weaponization）。模型偏斜就是通过污染训练数据导致模型的决策边界发生偏斜，影响模型性能。反馈误导则是利用

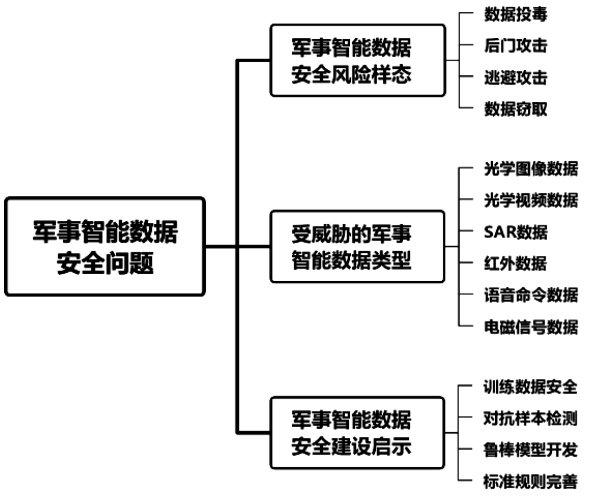

图 1　本文核心内容示意图

模型的自我训练和交互反馈机制，将错误或伪装的数据反馈给模型，滥用反馈机制来操纵系统朝着错误的方向发展。

1.1.2　后门植入攻击

后门植入（Backdoor）攻击[7]是近些年兴起的针对智能模型的新型攻击样式。攻击者在模型的训练过程中通过污染数据的方式对模型植入后门。当后门未被触发时，被攻击的模型与正常模型无异；而当隐藏的后门被推理数据中触发器（Trigger）激活时，模型将输出预设的错误标签。后门攻击与数据投毒的主要区别在于其高度的隐蔽性，因为后门的存在不会干扰模型在常规情况下的输出结果，使得其难以被察觉。

后门攻击可能发生在多种非完全受控的训练场景中，如利用第三方数据集，在第三方平台上训练或直接使用第三方模型等。由于许多军事智能模型会先调用公开数据集进行预训练（Pre-trained）或者直接调用预训练好的模型，然后对模型进行微调对齐（Fine-tune），有可能在浑然不知的情况下被攻击者植入后门，造成极大的安全隐患。

1.2　推理数据安全风险

1.2.1　逃避对抗攻击

逃避攻击，又称对抗样本攻击[8]。在模型推理阶段，通过给输入数据施加某种人类难以察觉的特定扰动，可导致 AI 模型输出错误的结果。逃避攻击是研究最为广泛的对抗攻击方法，也是军事智能系统面临的最直接的数据安全威胁。目前，针对军事智能系统的逃避攻击朝着实用性逐渐增强、物理可行性逐渐提升和应用任务更加多样的方向发展，将作为新型智能欺骗干扰方式影响未来战场。

根据不同的分类标准，对抗样本可以被分为多种类

型。首先，根据攻击者对目标模型的了解程度不同可以划分为白盒攻击与黑盒攻击。其中，白盒攻击指攻击者完全掌握目标模型所有信息和知识。而黑盒攻击是指攻击者在对目标系统的内部结构和参数一无所知情况下进行攻击，相较于白盒是一种更为实际但难度更大的攻击方式。其次，根据攻击目标的不同可以划分为定向攻击与非定向攻击。定向攻击是希望通过输入对抗数据将模型误导至某个特定的错误输出，例如将某型号坦克识别成民用汽车。而非定向攻击的目标则是扰乱模型的输出，确保输出不正确即可。最后，根据攻击的应用场景不同可以分为数字攻击与物理攻击。数字攻击是指直接在数字域输入图像上添加扰动的攻击方式。而物理攻击则更为复杂，涉及使用如对抗贴图、补丁等物理手段，使生成的扰动能够在现实世界中部署。

1.2.2 数据窃取攻击

数据窃取攻击[9]主要利用模型参数和预测结果等关键信息来非法获取训练数据，或创建功能近似的替代模型。目前，主要的数据窃取攻击方式分为逆向攻击和萃取攻击两大类。其中逆向攻击可细分为成员推理攻击（Member Inference Attack）和模型反演攻击（Model Inversion Attack）。成员推理攻击通过模型结果来推断训练数据集中是否包含特定样本，从而部分获取训练数据集的信息，进而实现窃取敏感数据的目的。而模型反演攻击则依据模型输出来逆向推测输入数据的某些特征。需明确的是，模型反演并不能直接生成训练数据，而只能描绘出某类别数据的显著性特征，生成类似“人群画像”的共性特征描述，以便为其他攻击方式如逃避攻击或数据投毒等提供辅助。萃取攻击则是通过访问模型的应用程序接口，推测模型的内部结构、参数、超参数等关键信息，或构建一个与目标模型功能相似的新模型。

由于无人设备的大规模普及，数据窃取攻击已成为一种切实的安全隐患。一旦敌方利用网络通信手段成功入侵或捕获我方无人装备，并获取其内部存储的智能模型的访问权限，就可以通过成员推理或模型反演攻击来获得数据集的相关特征，同时构建出替代模型。一方面，敌方可以在试验训练中模拟我方装备的实际响应结果，采取对应措施，提升实战效果；另一方面，将会显著提升敌方逃避攻击的威胁程度，降低我方装备效能。

2 对抗攻击威胁的军事数据类型

对抗攻击一开始只是在光学图像数据识别模型中发现。但随着研究的逐步深入，对抗攻击技术已逐步拓展至各类军事应用数据类型，包括但不限于合成孔径雷达[10]（Synthetic Aperture Radar，SAR）、一维电磁信号以及指控命令等。由于获取敌方信息和在数字域中篡改数据难以实现，这对对抗攻击技术的实用性和物理可实现性提出了更高的要求。在军事领域可用的对抗攻击技术需要在敌方信息较为匮乏的情况下依然可靠，并且具备物理实现手段，可以在现实世界中部署。因此，本文主要关注由物理攻击[11]可能带来的智能数据安全问题。

2.1 光学图像数据

光学图像数据，作为军事领域中的核心数据类型，广泛应用于无人装备自动驾驶、关键目标识别与精确定位以及场景细分等多个方面。然而，这些图像数据在执行各项任务时，却面临多样化的对抗性攻击威胁，如对抗性车辆迷彩涂装、欺骗性迷彩服以及对抗性激光束等。

对抗性车辆迷彩涂装[12]通过施加特定设计的迷彩图案于车辆表面，能有效误导智能目标检测器，实现隐匿或者误导目的，如图2（a）所示。2022年联合研发出一种具有鲁棒性的迷彩涂装攻击方法[13]。这种方法能在多种拍摄角度和环境条件下干扰深度检测模型，导致检测失败或错误分类，进而提高了对抗迷彩涂装在实际应用中的实用性。

另一方面，对抗性迷彩服[14]通过在服装上印制特殊纹理，这些纹理能够有效地欺骗智能检测器，从而实现类似“隐形衣”的效果，如图2（b）所示。与传统迷彩服相比，对抗性迷彩服在颜色选择上更为科学，避免了以往颜色过于鲜艳、容易被人类视觉察觉的缺陷[15-16]，有助于规避敌方的智能监控系统，提高人员的隐蔽性和生存能力。

此外，对抗性激光束[17]作为一种新型智能攻击手段，通过向目标物体投射激光束，能够干扰无人驾驶装备或深度识别器的正常运行，如图2（c）所示。这种攻击方式具有远程部署、低成本、隐蔽性强、瞬时性高、弱光环境适用以及多环境兼容等特点。恶意攻击者仅需借助一支大功率便携式激光笔，将激光投射到交通信号牌或者障碍物上，即可对无人装备实施干扰，甚至迫使其停止工作或自我销毁。这种攻击手段不仅对无人步战车、无人机等无人装备构成威胁，还可能对侦察设备中的深度识别系统产生影响。

2.2 光学视频数据

视频数据常被用在目标追踪、动作识别与预测以及场景理解等任务中，并在导引头制导、战术手势识别等领域发挥重要作用。与静态图像数据相比，视频数据的关键特性在于其动态时序变化。由于背景环境和拍摄视角的频繁变动，实施对抗性攻击的难度增加，对扰动生成的实时性提出了更高要求。目前，针对视频数据的物理对抗攻击主要包括对抗性广告板和对抗性补丁等形式。

(a) 对抗性迷彩　(b) 对抗性服装　(c) 对抗性激光

图 2　针对光学图像数据的对抗攻击方式

对抗性广告板通过打印广告板或大屏幕展示特殊设计的对抗性图案来改变背景信息，进而威胁视频数据安全，如图 3（a）所示。事实上，这种背景板攻击已能在真实环境中导致无人驾驶装备无故转向避障[18]或追踪目标失效[19]。而对抗性补丁[20]则通过将补丁附着在动态目标上，具有更好的鲁棒性，如图 3（b）所示。近期，国内研究团队提出了一种针对无人机目标追踪的对抗性补丁方法[21]，证实了智能对抗技术对无人机采集的视频数据安全构成潜在威胁。

(a) 对抗性广告版

(b) 针对目标追踪的对抗性补丁

图 3　针对光学视频数据的对抗攻击方式

2.3　合成孔径雷达数据

合成孔径雷达（SAR）是一种主动传感器，它通过微波感知技术实现对目标的持续观测，且不受天气和光照等环境条件的限制。与光学图像不同，SAR 通过主动发射电磁波与目标地物相互作用产生调制效应，接收其后向散射形成的回波信号，并通过成像处理算法生成 SAR 图像。尽管 SAR 与光学图像在成像机理上存在显著差异，但最近的研究表明，SAR 数据同样面临着对抗性攻击带来的安全威胁。

与光学数据不同，SAR 数据具有明确的电磁物理含义，即目标反射回波的相干能量累积。因此，光学对抗样本的物理实现方法，如对抗补丁和对抗涂装等，并不适用于 SAR 数据。因此，国内研究者们[22-23]提出运用属性散射中心模型（Attributed Scattering Center Model，ASCM）来构建 SAR 对抗样本（如图 4 所示），通过结合对抗样本与 ASCM，可以模拟真实场景中二面角、三面角等电磁散射中心的分布特征，进而实现对雷达回波信号的有效调控。

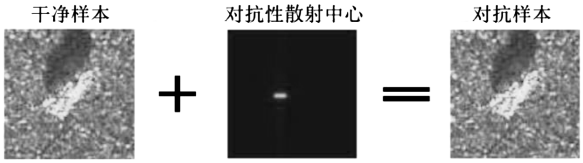

图 4　针对 SAR 数据的对抗散射中心攻击

2.4　红外数据

目前，红外数据对抗攻击技术的研究主要聚焦于红外行人识别任务。通过运用对抗性白炽灯[24]（图 5（a））和对抗性红外迷彩服[25]（图 5（b））等手段，实现在红外行人检测器中的“隐身”效果。2023 年，国内研究团队提出了一种基于气凝胶材料的可控温对抗性补丁红外攻击方式[26]，如图 5（c）所示，不仅使攻击行为更难以被人眼察觉，而且将攻击范围扩展至红外车辆检测与识别领域，进一步证实了智能对抗技术对红外数据安全构成的严重威胁。

(a) 对抗性灯泡　(b) 对抗红外服　(c) 对抗热补丁

图 5　针对红外数据的对抗攻击方式

2.5　语音命令数据

一维语音数据在军事领域具有不可忽视的应用价值，其在军事通信、保密工作、情报收集以及指挥自动化等多个方面均发挥了至关重要的作用。智能语音控制系统的使用，显著减轻了各型装备驾驶员的工作负担。

然而，值得注意的是，智能语音控制系统面临着潜在的数据安全风险，可能会受到小型信号发射器的隐蔽对抗性攻击，以极小的代价造成重大的安全威胁。香港城市大学的研究者们提出了一种具备实时性的对抗性语音发射机 Metamorph[27]，如图 6 所示。这种物理可行的语音攻击方式在 6 米的攻击范围内，对 DeepSpeech 智能语音控制模型的攻击成功率高达 90%。Metamorph 能够使智能语音控制系统错误地将诸如“and you know it”的正常语句解读为“open the camera”或“restart”等指令，进而引发错误操作或干扰，甚至可能对装备正常运行构成严重威胁。

2.6　电磁信号数据

现代信息化战场中，“制电磁权”已经成为敌我双方

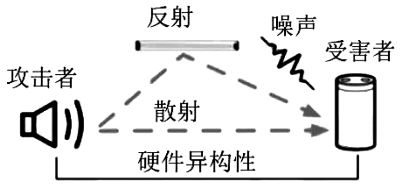

图 6　针对语音的对抗性语音发射器

竞争的新制高点。当前电子战系统对 AI 的高度依赖使得电磁信号数据安全也受到对抗性攻击技术的威胁。目前，针对电磁信号数据的对抗攻击主要有两种实现方式，即基于基带调制的攻击与基于干扰发射机的攻击。基于基带调制的攻击方式由英国帝国理工大学的 Muhammad 等人率先提出[28]，通过将电磁信号对抗攻击看作是一种特殊的基带调制方式，将对抗扰动在信号发射之前通过调制的方式添加到原信号中去，从而实现对敌方信号数据的安全威胁。而基于干扰发射机的方式[29-30]则是在信号传输过程中，通过独立的对抗扰动发射机来添加对抗扰动，利用信道反演的方法对电磁信号实施干扰，具有更强的物理可实现性。这些方法的成功实施意味着对抗样本可以有效愚弄智能电磁信号处理系统。电磁信号的智能数据安全问题将成为未来电子对抗作战的重点之一。

3　对抗攻击带来的智能数据安全建设启示

目前，军事智能系统面临着多种多样的数据安全问题。随着对抗攻击技术朝着实用性、多领域适用性和物理可实现性进一步发展，这种威胁将进一步加剧。因此，智能化时代的军事数据安全建设需要采取一系列措施来应对这一挑战。

（1）重视对军事训练数据集的保护。加强对训练数据的安全防护技术是至关重要的。针对不同类型的军事数据，需要研发相应的安全防护技术，包括数据加密、数据隔离、入侵检测等。这些技术能够有效地防止投毒样本和后门触发器的注入和攻击，保证数据安全性。

（2）加强对所采集数据的对抗性检测。在智能系统的使用过程中，需要对所采集的数据进行对抗性检测[31]，以发现和过滤潜在的对抗样本。常用的对抗性检测算法包括基于分布统计的方法、基于特征学习的方法以及基于对抗检测网络的方法等。这些方法能够有效地识别出对抗样本，从而防止它们对智能系统造成损害。

（3）提高智能系统的鲁棒性和安全性。为了提高智能系统的鲁棒性和安全性，需要采用更加先进的防御技术。这包括使用对抗训练[32]、防御蒸馏[33]等防御手段，以增强智能系统对对抗样本的抵抗能力。

（4）加强数据安全标准建设。为了规范智能系统的研发和使用，需要加强数据安全标准建设。这包括制定相关标准和规范，明确数据安全的要求和责任，加大数据安全监管和追责力度。

4　结论

随着智能化作战的快速发展，数据资源在军事探测通信和指挥控制中所扮演的角色愈发关键。然而，深度学习技术的不可预测性与黑盒特性使得对抗攻击技术对军事智能数据安全构成了严重挑战。本文详尽阐述了智能数据在训练和推理阶段所面临的四种主要数据安全风险形态，并对六种不同类别的数据可能遭遇的对抗攻击策略进行了深入探讨。为了有效应对这些安全威胁，本文还从数据保护、对抗检测、模型鲁棒性增强以及标准规范四个方面提出了切实可行的建议。期望通过相关内容的探讨，能够为提升智能数据安全水平贡献一份力量。

参考文献

[1] 人工智能颠覆未来战争 [J]. 军事文摘，2023（7）：80.

[2] AKHTAR N, MIAN A, KARDAN N, et al. Advances in adversarial attacks and defenses in computer vision: a survey [J/OL]. IEEE Access, 2021（9）: 155161-155196.

[3] ZHANG L, HARTNETT G S, AGUIRRE J, et al. Operational feasibility of adversarial attacks against artificial intelligence [EB/OL]. (2022-12-XX) https://www.rand.org/content/dam/rand/pubs/research_reports.

[4] KIM J, LEE K, LEE H, et al. Camouflaged adversarial attack on object detector [C]//2021 21st International Conference on Control, Automation and Systems (ICCAS), 2021: 613-617.

[5] CHEN Y. The risk and opportunity of adversarial example in military field [C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022: 100-107.

[6] FAN J, YAN Q, LI M, et al. A survey on data poisoning attacks and defenses [C]//2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), 2022: 48-55.

[7] LI Y, JIANG Y, LI Z, et al. Backdoor learning: a survey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35（1）: 5-22.

[8] 李明慧，江沛佩，王骞，等. 针对深度学习模型的对抗性攻击与防御 [J]. 计算机研究与发展，2021，58（5）：909-926.

[9] HU H, SALCIC Z, SUN L, et al. Membership inference attacks on machine learning: a survey [J]. ACM Computing Surveys, 2022, 54（11s）: 235: 1-235: 37.

[10] 冯博迪，杨海涛，李高源，等. 神经网络在 SAR 图像目标识别中的研究综述 [J]. 兵器装备工程学报，2021，42（10）：15-22.

[11] WEI X, PU B, LU J, et al. Visually adversarial attacks and defen-

ses in the physical world: a survey [J]. arXiv: 2211.01671, 2022.

[12] SURYANTO N, KIM Y, KANG H, et al. DTA: physical camouflage attacks using differentiable transformation network [C/OL]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 15284 - 15293.

[13] WANG D, JIANG T, SUN J, et al. FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial Attack [J]. arXiv: 2109.07193, 2021.

[14] HU Z, HUANG S, ZHU X, et al. Adversarial texture for fooling person detectors in the physical world [J]. arXiv: 2203.03373, 2022.

[15] HU Y C T, CHEN J C, KUNG B H, et al. Naturalistic physical adversarial patch for object detectors [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 7828 - 7837.

[16] HU Z, CHU W, ZHU X, et al. Physically realizable natural-looking clothing textures evade person detectors via 3D modeling [C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 16975 - 16984.

[17] DUAN R, MAO X, QIN A K, et al. Adversarial laser beam: effective physical-world attack to DNNs in a blink [J]. arXiv: 2103.06504, 2021.

[18] KONG Z, GUO J, LI A, et al. PhysGAN: generating physical-world-resilient adversarial examples for autonomous driving [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 14242 - 14251.

[19] WIYATNO R, XU A. Physical adversarial textures that fool visual object tracking [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, 2019: 4821 - 4830.

[20] HOORY S, SHAPIRA T, SHABTAI A, et al. Dynamic adversarial patch for evading object detection models [J]. arXiv: 2010.13070, 2020.

[21] RASOL J, XU Y, ZHANG Z, et al. Bilateral adversarial patch generating network for the object tracking algorithm [J]. Remote Sensing, 2023, 15 (14): 3670.

[22] PENG B, PENG B, ZHOU J, et al. Scattering model guided adversarial examples for SAR target recognition: attack and defense [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1 - 17.

[23] QIN W, WANG F. SCMA: a scattering center model attack on CNN-SAR target recognition [J]. IEEE Geoscience And Remote Sensing Letters, 2023 (20). DOI: 10.1109/LGRS.2023.3253189.

[24] Zhu Xiaopei, Li Xiao, Li Jianmin, et al. Fooling thermal infrared pedestrian detectors in real world using small bulbs [J]. arXiv: 2101.08154, 2021.

[25] ZHU X, HU Z, HUANG S, et al. Infrared invisible clothing: hiding from infrared detectors at multiple angles in real world [J]. arXiv: 2205.05909, 2022.

[26] WEI X, YU J, HUANG Y. Physically adversarial infrared patches with learnable shapes and locations [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12334 - 12342.

[27] CHEN T, SHANGGUAN L, LI Z, et al. Metamorph: injecting inaudible commands into over-the-air voice controlled systems [C]// Proceedings 2020 Network and Distributed System Security Symposium, Internet Society, 2020.

[28] HAMEED M Z, GYÖRGY A, GÜNDÜZ D. The best defense is a good offense: adversarial attacks to avoid modulation detection [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1074 - 1087.

[29] KIM B, SAGDUYU Y E, ERPEK T, et al. Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond [J]. arXiv: 2013.13989, 2021.

[30] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers [J]. IEEE Transactions on Wireless Communications, 2022, 21 (6): 3868 - 3880.

[31] LIANG B, LI H, SU M, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18 (1): 72 - 85.

[32] BAI T, LUO J, ZHAO J, et al. Recent advances in adversarial training for adversarial robustness [J]. arXiv: 2102.01356, 2021.

[33] SOLL M, HINZ T, MAGG S, et al. Evaluating defensive distillation for defending text processing neural networks against adversarial examples [C]// Artificial Neural Networks and Machine Learning-ICANN 2019: Image Processing. Cham: Springer International Publishing, 2019: 685 - 696.

（收稿日期：2024 - 09 - 03）

作者简介：

陆正之（1995 - ），男，博士，助理研究员，主要研究方向：对抗样本攻击、深度学习和仿真技术。

黄希宸（2000 - ），男，硕士生，主要研究方向：SAR 对抗攻击。

彭勃（1986 - ），通信作者，男，博士，副研究员，主要研究方向：对抗攻击、智能感知。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部