自动驾驶系统相关的对抗攻击与防御技术研究综述

王嘉宁 PB22051022

我们证明本报告中所有非我们自己工作的材料都已注明出处

2025 年 6 月 5 日

自动驾驶系统相关的对抗攻击与防御技术研究综述

PB22051022 王嘉宁

2025 年 6 月 5 日

摘要

随着深度学习技术在自动驾驶视觉感知系统中的广泛应用，对抗攻击及其防御策略已成为保障智能交通安全的核心研究课题。本综述系统梳理了近五年内 35 篇代表性研究成果，全面分析了自动驾驶场景下图像分类模型面临的对抗威胁及其防护机制。研究表明，对抗样本通过微小扰动即可导致模型误判，而物理世界的补丁攻击对交通标志识别构成严峻威胁。在防御技术方面，融合对抗训练的主动防御、基于输入重构的被动防御以及针对物理攻击的专用防护构成了多层次防御体系。本文深入探讨了攻击与防御技术在动态环境中的对抗效应，总结了实验评估指标与性能对比，并指出物理攻击防御、轻量化防护架构、多模态融合防御、对抗样本可解释性及标准化测试框架五大未来研究方向。该综述为提升自动驾驶系统的对抗鲁棒性提供了理论支持和技术参考。

1 引言

本学期我选修了人工智能安全这门课程，学习到了很多以前从未接触过的知识，其中对抗攻击与防御技术本门课程的重点，在考核一中也作为重点考核对象，因此我想依次为方向进行进一步的调研。我在中国知网上查阅了近五年对抗攻击与防御方向的多篇相关论文，并进学习与整理，现进行对这些文献的综述汇报。

在进行正式文献调研前，我首先研究了和对抗攻击与防御有关的现实应用，其中重点关注了自动驾驶系统中对抗攻击与防御性能的研究 [1]。由于城市中的交

通系统具有复杂的环境变化因素，因此在传统列车控制系统上加入自主感知系统可以增强系统对环境探测的敏感性和驾驶的安全性。基于深度学习的视觉感知系统作为自动驾驶的“眼睛”，承担着环境感知、目标检测和决策支持等关键任务，其中至关重要的一步就是对感知到的图像在处理过程中进行图像分类。

正如本学期课程所学到的，深度学习模型在面对各种对抗攻击时表现出不同程度的脆弱，这种对抗攻击生成的样本称为对抗样本，即通过对原始输入数据添加精心构造的、通常人眼/耳难以察觉的微小扰动而生成的恶意输入。在自动驾驶场景中，对抗攻击可能导致交通标志误识别、障碍物漏检等严重后果，威胁行车安全。通过利用不同的对抗攻击方式对交通标志添加人眼难以察觉的扰动，攻击成功率最高可达 97.56%（RFGSM，keps=0.093）[1]，从而使自动驾驶系统的识别准确率骤降。

那么我们应当如何解决这个问题呢？接下来我将进行一系列对抗攻击与防御相关问题的学习，最终根据学到的知识解决以上提出的问题。

2 对抗攻击技术研究进展

知己知彼百战不殆，要想防御对抗攻击，首先要研究对抗攻击。本小节我们将对当下对抗攻击技术的研究进展进行统计整理，主要包括两个部分：攻击分类与特征，以及各种类型攻击的攻击方法。

2.1 攻击分类与特征

为了更有条理性的研究各种不同特性的对抗攻击，首先我们对对抗攻击进行分类处理，可根据攻击环境、攻击目标和攻击知识三个维度进行分类。[2]

根据攻击环境，可分为数字域攻击和物理域攻击。数字域攻击直接在图像像素层面添加扰动，而物理域攻击则通过在真实物体表面添加“物理对抗补丁”，即扰动（如贴纸、涂鸦）实现攻击 [3]。

根据攻击目标，可分为定向攻击和非定向攻击，前者诱使模型输出特定错误类别，后者仅需导致模型错误输出即可 [4]。

根据攻击知识，可分为白盒攻击（攻击者完全了解模型结构、参数甚至防御

表 1: 对抗攻击分类及典型特征

攻击类型	典型特征	自动驾驶场景案例
数字域攻击	像素级扰动，易于实施	篡改车载摄像头输入的交通标志图像
物理域攻击	需考虑环境因素，鲁棒性强	在真实交通标志上粘贴对抗补丁
定向攻击	诱导特定错误输出	将“停车”标志识别为“限速 60”
非定向攻击	仅需导致错误识别	导致任何交通标志误分类
白盒攻击	攻击效率高，威胁来自内部	针对已知模型架构的精确攻击
黑盒攻击	依赖迁移性，更具现实威胁	对未知商用自动驾驶系统的攻击

机制)、黑盒攻击（攻击者对目标模型一无所知）和灰盒攻击（介于两者之间，攻击者可能了解部分模型信息，但对防御机制并不了解）[5]。

下面我将列表 1来总结各种攻击的典型特征，以及在自动驾驶场景中的案例体现，使这些概念更加通俗易懂。

对抗攻击在自动驾驶场景中表现出三大核心特征：隐蔽性、迁移性和物理可实现性 [4]。

隐蔽性指扰动在人类视觉难以察觉的范围内；

可转移性指针对一个模型生成的对抗样本对其他模型同样有效；

物理可实现性则强调扰动在真实物理环境中依然有效。

2.2 数字域攻击方法

数字域攻击即我们在课程中学到的图像攻击，是通过在数字图像中添加微小扰动实现攻击。接下来我将列举各种图像攻击 [6][7]，并分析它们的效果 [8]（因优势已在设计理念中体现，故只讨论攻击成功率与劣势），如表 2所示。

L-BFGS 攻击。最早被设计攻击 DNN 魔性的对抗样本攻击方法, 最终目标位在输入的约束空间中找到一个不可察觉的最小输出扰动，并使模型分类出错。

FGSM 攻击。是一种利用模型梯度信息来构造对抗样本的方法，通过在原始输入数据的梯度上升方向加入微小扰动实现误分类。

JSMA 攻击。对抗显性图，即找到对分类器特定输出影响程度最大的输入，

表 2: 部分攻击方法优劣

方法名称	攻击成功率	劣势
FGSM	较高	鲁棒性相对较低，容易被防御机制识别
PGD	较高	需要更多的计算资源和时间
C&W 攻击	高	计算成本较高，需要专业知识来调整参数
DeepFool	中	可能受到模型非线性特性影响
I-FGSM	中等至高	增加计算复杂性，时间消耗大，可额能过度拟合
L-RFGS	中等	计算资源需求高，收敛速度慢，对参数选择敏感

对原始图像添加扰动。

DEEPFOOL 攻击。基于超几何平面分类的 DeepFool 攻击方法，从简单的线性二分类问题出发，通过添加扰动使得位于分类面一侧的样本被分类面分为另一侧的类别。

C&W 攻击。对抗样本用优化参数表示，优化参数须达到两个目标：一是对抗样本和对应的干净样本差距越小越好，二是对抗样本应该使得模型分类出错

One Pixel attack 攻击。仅改动一个像素点就能产生对抗样本的方法。

Zoo 攻击。将 C&W 中的损失函数进行对数处理，进而使该损失函数在目标函数中占比更大，优先满足该损失函数变小的条件，并使用对称差分商对梯度和 Hessian 矩阵对梯度进行估计。

GAN 攻击。生成对抗网络，可以在复杂数据（如图像、音频或视频）上进行学习和训练，生成模型与判别模型相互进行对抗性游戏，以所部方式进行训练，直至最终生成模型输出的对抗样本。(是的，不只是静态图像，动态视频中也有对抗攻击和防御的应用。目前已经可以实现 DNQ 智能体模型，并在此基础上进行对抗攻击与防御的训练。[9])

MI-FGSM 攻击。引入栋梁概念，考虑前一次迭代梯度值，以累加权重的方式进行优化，有助于更准确地估计梯度方向，使优化过程更加平滑。

PGD 攻击。迭代共计，在每次迭代中计算损失函数相对于输入的梯度，病句次更新输入样本，以最大化模型的预测损失。

BIM 攻击。通过迭代将非线性模型在局部区域近似为线性模型，从而更准确

的捕捉梯度变化方向，生成更有效的对抗样本。

2.3 物理域攻击方法

物理域攻击需要克服现实环境中的光线变化、视角偏移、天气干扰等挑战，技术难度更高。对抗补丁攻击是物理域攻击的主流形式是指在图像的局部区域生成的连续像素块，打印后可粘贴在物理世界的目标对象上，进而能够实现误判 [3]。

物理补丁攻击需考虑三个方面：物理对抗补丁的形式及部署方式、自然性和鲁棒性之间的平衡 [3]。

提出一种通用防御物理空间补丁对抗攻击方法, 遮罩防御方法，首先将待检测图片分割成若干像素块，在利用快速傅里叶变换和二值化处理求这些像素块中的高频信息含量，依次对含有较多高频信息的像素块使用遮罩，最后用目标检测器验证。此方法能够通用预防与物理空间的对抗补丁攻击 [10]。

物理对抗补丁攻击在现实中有很多实例，除了今天要讨论的主要问题涉及到的自动驾驶攻击，还有人脸识别共计，交通标识牌识别攻击等，攻击形式更是千奇百怪，贴纸，涂鸦，阴影，眼镜，二维码等生活中随处可见的东西都可能形成无力对抗补丁攻击，而当下研究防御物理域攻击的文章较少，可以作为未来的一个重点发展目标。

2.4 白盒攻击、黑盒攻击与迁移攻击

白盒攻击的其中一种是给予直接优化的攻击方法，基本原理为通过优化算法对目标函数进行优化并产生较小的扰动，该方法可能导致时间成本过大，此时间成本主要用于搜索模型的超参数。还有一种白盒攻击时基于梯度优化的百合攻击方法，沿着梯度变化方向对输入样本添加一定扰动，使得分类误判。操作相对简单，切成功率高，是不前流行的对抗攻击技术。典型有 FGSM。另还有基于决策边界分析的攻击方法、基于 GAN 进行对抗样本生成的攻击方法等。[11][12]

迁移攻击是黑盒场景下的主要攻击手段，利用对抗样本的跨模型迁移性实现攻击。其核心思想是打破对抗样本对特定模型的过拟合，增强泛化能力。

对抗训练基本流程：遍历 SNCOD 训练集，以一定概率利用 BTIA 方法生成五中不同对抗样本，并加入训练集中。加载原始目标检测模型权重，小批量训练，

先冻结骨干网络训练，再回复骨干网络训练，获得最终更鲁棒目标检测防御模型[13]。

2.5 后门攻击

后门攻击是一种针对机器学习模型（尤其是深度学习模型）的隐蔽攻击方式。其核心思想是：攻击者通过某种方式在模型训练过程中植入一个隐藏的“后门”，使得该模型在绝大多数正常输入上都表现良好（与未植入后门的干净模型几乎无异），但当输入包含特定的、攻击者预先设定的触发模式时，模型的输出就会被恶意操控，产生攻击者期望的错误结果。

根据《基于 JSMA 对抗攻击的去除深度神经网络后门防御方案》[14] 一文来研究后门防御方案。文中方案使用 BadNets 后门攻击植入后门触发器。首先采用 JSMA 生成对抗样本，通过特定扰动模拟出发后门条件，获得扰动信息并还原后门触发器的图案。接着计算后门触发器的权重分布，可视化关键神经元激活状态，并利用热力图对神经元权重进行计算。最后试用脊回归函数将关键神经元权重置零，去除后门触发器。

3 对抗防御技术体系

3.1 主动防御机制

主动防御旨在通过增强模型自身鲁棒性，从根本上抵御对抗攻击。

对抗训练是最广泛应用的主动防御策略，其核心思想是将对抗样本注入训练过程，使模型学习正确识别此类样本。

知识蒸馏是一种神经网络压缩方法，核心思想是让一个轻量级的学生网络学习比他性能更强的教师网络的知识。知识蒸馏通过教师网络输出的概率向量训练学生网络，从而实现教师网络和学生网络之间的知识迁移。防御性蒸馏机制通过知识蒸馏来提高深度神经网络的鲁棒性。剪枝与鲁棒蒸馏融合方法通过知识蒸馏将鲁棒性从大型模型迁移至轻量模型。实验表明，该方法在 CIFAR-10 数据集上将轻量模型的鲁棒性显著提升，同时保持实时性要求。[15]

表 3: 主动防御技术对比

防御方法	优势	局限性
对抗训练	直接提升模型鲁棒性	训练复杂度高，可能降低标准准确率
鲁棒蒸馏	实现轻量化防御，满足实时性	依赖教师模型质量
陷阱网络	主动防御，高灵活性	需精确设计陷阱策略

表 4: 续表

防御方法	核心技术	适用场景
对抗训练	对抗样本注入训练集	高安全性要求的端到端系统
鲁棒蒸馏	知识从鲁棒教师模型迁移	车载边缘计算设备
陷阱网络	预设可攻击空间诱导攻击者	多模型协同决策系统

陷阱式集成对抗防御网络创新性地利用“可攻击空间假设”[16]。(可攻击空间：在 DNN 的特征空间中，对不隶属于训练数据的数据流形之外的 DNN 特征空间定义为广义上的对抗样本的可攻击空间。)(邻近可攻击空间：位于目标数据流形所处的特征敏感空间之中。)(背景可攻击空间：位于整体目标数据流形之外的广袤特征空间之中。)

三者的特征如表 3

该网络包含多个子模型，每个子模型设置不同的决策陷阱，诱使攻击者进入预设的无效攻击方向。当检测到对抗样本时，系统激活陷阱机制，将其导向安全输出。实验结果显示 Trap-Net 探测后每种攻击方法的防御效力都接近 100%，效果非常好。

基于高斯增强和迭代攻击的对抗训练防御方法 GILLC，在训练过程中引入高斯噪声增强和迭代攻击样本生成，显著提升了模型在复杂环境下的鲁棒性。实验采用五种白盒攻击，分别为 FGSM、BIM、PGD、DeepFool 和 CW，该方法在 CIFAR10 数据集上的测试表明，模型对主流攻击的防御成功率都很高。此方法在保持对干净样本良好分类效果的同时，有效提高了模型对单不攻击和迭代攻击的防御能力 [17]。

基于攻击图与随机森林算法的网络电子对抗攻击主动防御方法。首先手机网络电子对抗攻击的相关数据并做预处理工作，然后将提取网络特征得到状态信息矩阵中的数据集划分为训练稽核测试机，用训练集数据构建随机森林模型，实现

对网络电子对抗攻击的最有效防御，最大限度保障网络环境在攻击作用下安全性。实验结果显示该方法在面对不同攻击时胜率均高于 90%，明显高于其他三个对比方法，是一种有效的对抗攻击主动防御策略。[18]

3.2 被动防御机制

被动防御不改变模型本身，而是通过预处理输入数据或检测异常样本实现防护。

基于插值法的防御算法 (IDA) 是被动防御机制的典型代表 [19]。

基于插值法的对抗攻击防御算法：

首先输入样本图像、未加 IDA 的神经网络模型和样本的真实类别标签。

接着输出识别模型对于样本的识别率。

如果神经网络识别准确率大于 90%，说明未收到对抗攻击，返回神经网络模型识别原始图片的预测值；

低于 90%，基于像素区域插值算法对样本进行插值缩放，生成 img。立方卷积插值算法将 img 还原到原尺寸，返回神经网络经过处理后的图片的预测值。

实验证明，这种方法对 FGSM 攻击生成的对抗样本识别率从 4.82% 提升至最高 82.28%（0.8IDA），且不影响人类视觉识别。

自适应像素去噪方法也是一种被动防御机制，通过分析图像局部统计特征，区分正常像素与对抗扰动像素 [20]。

对实验结果进行分析，相比于蒸馏防御，PGD 对抗训练与 DAE 防御，APDD 的综合房与效果更好，并且不需要对目标分类模型及其训练过程进行修改，是一种非常好的防御机制。

3.3 物理攻击专用防御

物理攻击防御需考虑环境动态性与传感器噪声。针对交通路标防御提出一种多阶段对抗防御方法，通过融合多个摄像头数据，降低单点攻击风险 [21]。该方案的核心思想是：单个摄像头可能被对抗补丁欺骗，但多个视角同时被欺骗的概

率极低。系统通过比较不同视角的识别结果，检测并排除异常输出。

该防御方法分为三个阶段——焦点损失对抗训练阶段（基于焦点损失函数消除政府样本不平衡类别的对抗训练）、知识蒸馏阶段（对第一阶段的教师模型群体进行知识蒸馏，迁移到学生模型群体）、模型投票决策阶段（对第二阶段学生模型群体预测结果进行加权平均）。

实验结果显示该方法训练后学生模型群体取得 85% 分类准确率，具有较好防御效果。该方法能有效抵抗对抗样本的攻击，提高了模型鲁棒性。

采用宽度学习系统 (BLS) 替代传统深度学习模型，构建轻量级防御框架 [22]。

BLS 无需深度结构，通过增量学习动态更新，在资源受限环境下仍保持高效。

实验表明，该系统在物理对抗样本检测中达到 82% 的准确率，比传统深度学习模型提高 32 个百分点。

针对物理攻击的有效防御需结合数字防御与物理防护双重策略。数字防御指模型层面的鲁棒性提升，而物理防护则包括在交通标志表面添加防护涂层或特殊材料，使对抗贴纸难以附着，或安装物理屏障阻止攻击者接近关键交通设施。

对抗防御的研究方向主要沿着三个方向发展：[23]

修改目标模型以增强鲁棒性。如通过交替训练样本修改模型权重的对抗训练，使用量化模型来抵抗及预提读的攻击等。

修改输入以消除扰动。基本思想是通过对输入图像进行修改，使其不受攻击者添加的扰动的影响。可分为两类，一类是通过输入图像进行滤波处理，以消除攻击者添加的扰动。另一类是通过重建输入图像，以消除攻击者添加的扰动。

向模型添加外部模块。分为两类，一类通过将多个模型融合在一起，提高计算机视觉系统鲁棒性，另一类通过在计算机视觉系统中添加外部模块，以提高计算机视觉系统鲁棒性。

三者比较如表 5

表 5: 对抗防御方法比较

防御方法	优点	缺点
修改目标模型	防御效果好	计算开销大，泛化能力弱
修改输入	计算开销较低	防御效果一般
添加外部模块	使用便利，可以多个模块叠加使用	单一模块防御力一般

4 自动驾驶应用场景的特殊挑战

4.1 场景特性与威胁分析

自动驾驶系统面临的环境具有高度动态性、不可预测性和安全关键性三大特征，使其对抗防御面临独特挑战。自动驾驶场景的对抗攻击可能导致多级连锁反应：就像之前的特斯拉追尾事故，很有可能就是存在不法分子恶意布置“补丁”[24]，而使智能车辆发生误判。当单个车辆模型受到攻击做出错误判断时，这些错误可能在整个交通系统中传播，引发一系列后续错误。攻击单个车辆模型可能引发错误的交通决策，进而导致交通堵塞甚至事故。

自动驾驶视觉系统面临的三重威胁模型：传感器层面攻击（干扰摄像头输入）、模型层面攻击（操纵深度学习模型）和系统层面攻击（破坏多传感器融合机制）。这些攻击可能单独实施，也可能协同进行，形成复杂攻击链。[25]

在环境动态性方面，自动驾驶系统需应对光照变化、天气条件、运动模糊等挑战。郭敏等人的研究表明，雨雾天气会显著降低现有防御机制的效果，因为自然噪声与对抗扰动的叠加增加了辨识难度 [26]。他们提出结合天气自适应预处理与对抗训练的方法，在恶劣天气下仍保持较高防御性能。

4.2 防御方案设计

针对自动驾驶场景的特殊需求，研究者提出了多种专用防御方案。考虑云边端协同防御架构，通过分布式计算实现高效防护 [27]。在该架构中，云端负责复杂模型训练和全局更新，边缘节点（路侧单元）负责区域威胁情报共享，车载端执行实时轻量级防御。这种分层架构既满足实时性要求，又能应对新型攻击。

研究对交通路标对抗攻击的防御方案，提出一种结合空间变换与区域分割的

方法 [21]。该方法首先检测图像中的交通标志区域，然后应用随机旋转和缩放等空间变换削弱扰动效果，最后使用分割网络提取标志特征进行分类。这种方案在真实道路测试中成功防御了多种物理补丁攻击。

针对联邦学习场景提出选择性防御策略，解决分布式训练中的对抗攻击问题 [28]。该方法通过分析参数更新贡献度识别恶意节点，动态调整聚合权重，保护车联网中的协同学习系统。

研究未知攻击的泛化性对抗防御技术，提出了统一评估框架 [29]。该框架包含 20 多种攻击类型和 15 种防御方法，通过自动化测试管道评估防御方案的泛化能力。结果表明，目前没有单一防御方法能应对所有攻击类型，但集成防御策略（如检测 + 重构 + 对抗训练的组合）在未知攻击面前表现最佳。

针对时序数据的对抗防御研究为动态环境下的防御 [30]，设计了一种基于时空特征一致性的检测机制，通过分析连续帧间的逻辑一致性识别对抗攻击。该方法在高速公路场景测试中，对动态对抗攻击的检测准确率大大提高。

5 未来研究方向

基于对现有研究的系统分析，我们提出以下研究方向：

物理攻击防御的泛化能力提升是亟待突破的难点。现有物理防御方法多针对特定攻击模式设计，缺乏通用性。邓欢等人指出，开发物理扰动统一表征框架是重要方向，该框架应能建模不同材质、光照和距离下的扰动影响。同时，结合计算机图形学与材料科学，设计可抵抗对抗贴纸的智能交通基础设施，如自清洁涂层或光学防伪标志。

轻量化防御架构对资源受限的车载环境至关重要。王滨等人融合剪枝与蒸馏的尝试显示，紧凑模型可实现与大型模型相当的鲁棒性。未来研究可探索神经架构搜索 (NAS) 技术自动设计高效防御架构，平衡实时性、准确性与鲁棒性。同时，开发专用硬件加速器支持车载对抗样本检测，满足毫秒级响应要求。

多模态融合防御是提升系统级安全的关键。单一视觉模态易受攻击，而融合激光雷达、毫米波雷达等多源数据可构建互补的感知通道。李前等人的云边端协同框架提供了基础，但需进一步研究跨模态对抗攻击及防护，如攻击者可能同时干扰摄像头与 LiDAR 数据。开发多模态一致性验证机制，通过交叉验证检测异常

输入，是值得探索的方向。

对抗样本的可解释性研究有助于深入理解防御机制。目前防御方法多基于经验设计，缺乏理论解释。周大为等人指出，通过可视化决策过程分析防御机制的工作机理，可指导更有效的防御设计。结合因果推断理论，建模扰动与误分类的因果关系，可能为防御提供新范式。

标准化测试框架的建立是推动领域发展的基础设施。现有研究使用不同数据集和评估指标，导致结果难以直接比较。肖子勤等人倡议建立自动驾驶对抗攻防基准平台，包含统一数据集、攻击库和评估协议。该平台应模拟多样化驾驶环境（城市、高速、乡村）及气象条件，全面评估防御方案的实用价值。

6 结论

本综述系统分析了面向自动驾驶图像分类模型的对抗攻击与防御技术研究进展。研究表明，对抗攻击尤其是物理补丁攻击，已成为自动驾驶安全的重大威胁；而融合多层次防御策略（主动加固、被动检测、物理防护）的综合方案是提升系统鲁棒性的有效途径。

在攻击技术方面，数字域攻击已发展出多种高效生成方法，物理域攻击特别是对抗补丁技术日益成熟，对交通标志识别构成现实威胁。防御技术呈现多元化发展趋势：主动防御如对抗训练从本质上提升模型鲁棒性；被动防御如输入重构和异常检测不改变模型即可过滤对抗样本；专用物理防御则结合算法与基础设施增强综合防护能力。

自动驾驶场景的动态环境特性和安全关键要求使其对抗防御面临独特挑战。云边端协同防御、多视角融合验证等创新方案针对性地解决了部分问题，但复杂环境下的防御可靠性仍需提升。实验评估表明，现有先进防御方案可将对抗攻击成功率大幅降低，但计算开销和泛化能力仍是瓶颈。

未来研究需在物理防御泛化性、轻量化架构、多模态融合、可解释性和标准化评估五大方向寻求突破。随着自动驾驶技术走向大规模应用，对抗攻防研究将从单纯的算法竞争转向系统级安全工程，需要跨学科协作构建更可靠的防护体系。只有确保视觉感知系统在对抗环境下的稳定性，自动驾驶技术才能真正实现安全落地应用。

参考文献

[1] 唐军, 黄文静, 李爽, 等. 面向自动驾驶系统图像分类模型的对抗攻击及防御性能研究 [J]. 机车电传动,2025,(01):25-34.DOI:10.13890/j.issn.1000-128X.2025.01.100.

[2] 王文萱, 汪成磊, 齐慧慧, 等. 面向深度模型的对抗攻击与对抗防御技术综述 [J]. 信号处理,2025,41(02):198-223.

[3] 邓欢, 黄敏桓, 李虎, 等. 物理对抗补丁攻击与防御技术研究综述 [J]. 信息安全学报,2025,10(01):75-90.DOI:10.19363/J.cnki.cn10-1380/tn.2025.01.06.

[4] 汪欣欣, 陈晶, 何琨, 等. 面向目标检测的对抗攻击与防御综述 [J]. 通信学报,2023,44(11):260-277.

[5] 刘文钊, 郭凯威. 面向深度神经网络视觉模型对抗鲁棒性的攻击与防御方法研究综述 [J]. 网络安全技术与应用,2025,(01):42-48.

[6] 秦书晨, 王娟, 朱倪宏, 等. 图像对抗样本检测与防御方法研究进展 [J]. 智能安全,2024,3(04):81-95.

[7] 陈国凯, 冯辉. 深度学习中对抗样本攻击与防御方法研究 [J]. 唐山师范学院学报,2024,46(03):59-66+77.

[8] 肖子勤, 史涯晴. 针对图像分类模型的对抗攻击及防御技术综述 [J]. 西安邮电大学学报,2024,29(06):86-97.DOI:10.13682/j.issn.2095-6533.2024.06.012.

[9] 熊水彬. 基于深度强化学习的对抗攻击和防御在动态视频中的应用 [J]. 通信技术,2023,56(09):1115-1120.

[10] 翔云, 韩瑞鑫, 陈作辉, 等. 一种通用防御物理空间补丁对抗攻击方法 [J]. 信息安全学报,2023,8(02):138-148.DOI:10.19363/J.cnki.cn10-1380/tn.2023.03.11.

[11] 韩家宝, 王成, 钟炜. 面向 AI 系统的攻击与防御方法研究 [J]. 信息对抗技术,2025,4(01):1-21.

[12] 王兴宾, 侯锐, 孟丹. 深度神经网络的对抗样本攻击与防御综述 [J]. 广州大学学报 (自然科学版),2020,19(04):1-10.

[13] 周栋, 孙光辉, 吴立刚. 面向空间视觉目标检测的对抗攻击与防御算法 [J].

控制与决策,2024,39(07):2161-2168.DOI:10.13195/j.kzyjc.2022.1669.

[14] 张光华, 刘亦纯, 王鹤, 等. 基于 JSMA 对抗攻击的去除深度神经网络后门防御方案 [J]. 信息网络安全,2024,24(04):545-554.

[15] 王滨, 李思敏, 钱亚冠, 等. 基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法 [J]. 网络与信息安全学报,2022,8(06):102-109.

[16] 孙家泽, 温苏雷, 郑炜, 等. 基于可攻击空间假设的陷阱式集成对抗防御网络 [J]. 软件学报,2024,35(04):1861-1884.DOI:10.13328/j.cnki.jos.006829.

[17] 王丹妮, 陈伟, 羊洋, 等. 基于高斯增强和迭代攻击的对抗训练防御方法 [J]. 计算机科学,2021,48(S1):509-513+537.

[18] 刘洋. 基于攻击图与随机森林算法的网络电子对抗攻击主动防御方法 [J]. 长江信息通信,2023,36(12):54-56.

[19] 范宇豪, 张铭凯, 夏仕冰. 基于插值法的对抗攻击防御算法 [J]. 网络空间安全,2020,11(04):74-78.

[20] 张帅, 张晓琳, 刘立新, 等. 基于自适应像素去噪的对抗攻击防御方法 [J]. 计算机工程与设
计,2023,44(05):1336-1344.DOI:10.16208/j.issn1000-7024.2023.05.008.

[21] 孙安临, 钱亚冠, 顾钊铨, 等. 自动驾驶场景下对交通路标对抗攻击的防御 [J]. 浙江科技学院学报,2022,34(01):52-60.

[22] 冼卓滢, 陈国明, 罗家梁, 等. 基于宽度学习防御对抗攻击的图像分类 [J]. 现代计算机,2023,29(17):49-56.

[23] 葛佳伟, 王娟, 石磊, 等. 计算机视觉对抗攻击与防御方法分析 [J]. 智能安全,2023,2(02):48-56.

[24] 杨弋🄵, 邵文泽, 王力谦, 等. 面向智能驾驶视觉感知的对抗样本攻击与防御方法综述 [J]. 南京信息工程大学学报 (自然科学
版),2019,11(06):651-659.DOI:10.13878/j.cnki.jnuist.2019.06.003.

[25] 刘佳玮, 张文辉, 寇晓丽, 等. 增强型深度对抗样本攻击防御算法 [J]. 西安电子科技大学学报,2021,48(06):23-31.DOI:10.19665/j.issn1001-2400.2021.06.004.

[26] 郭敏, 曾颖明, 于然, 等. 基于对抗训练和 VAE 样本修复的对抗攻击防御

技术研究 [J]. 信息网络安全,2019,(09):66-70.

[27] 李前, 蔺琛皓, 杨雨龙, 等. 云边端全场景下深度学习模型对抗攻击和防御 [J]. 计算机研究与发展,2022,59(10):2109-2129.

[28] 陈卓, 江辉, 周杨. 一种面向联邦学习对抗攻击的选择性防御策略 [J]. 电子与信息学报,2024,46(03):1119-1127.

[29] 周大为, 徐一搏, 王楠楠, 等. 针对未知攻击的泛化性对抗防御技术综述 [J]. 中国图象图形学报,2024,29(07):1787-1813.

[30] 刘坤, 曾恩, 刘博涵, 等. 基于多变量时序数据的对抗攻击与防御方法 [J]. 北京工业大学学报,2023,49(04):415-423.