国防科技大学学报
Journal of National University of Defense Technology
ISSN 1001-2486,CN 43-1067/T

《国防科技大学学报》网络首发论文

题目：梯度对齐下的目标检测模型对抗鲁棒性提升
作者：薛伟，李婕，杜明洋
收稿日期：2025-03-21
网络首发日期：2025-10-16
引用格式：薛伟，李婕，杜明洋．梯度对齐下的目标检测模型对抗鲁棒性提升[J/OL]．国防科技大学学报. https://link.cnki.net/urlid/43.1067.t.20251015.1406.002

doi:10.11887/j.issn.1001-2486.25030021

梯度对齐下的目标检测模型对抗鲁棒性提升

薛伟[1*]，李婕[1]，杜明洋[2]

（1. 安徽工业大学 计算机科学与技术学院，安徽 马鞍山 243032；
2. 国防科技大学 电子对抗学院，安徽 合肥 230037）

摘　要：针对目标检测模型在受到对抗攻击时对抗鲁棒性不足的问题，提出了基于梯度对齐的目标检测模型对抗鲁棒性提升方法。该方法在模型对抗训练阶段，构建基于对抗损失和对齐损失的复合损失函数，并通过引入梯度对齐策略，约束对抗样本与干净样本之间的梯度差异。结合知识蒸馏的监督信号和自监督学习的表征能力，最大化对抗样本与干净样本的特征相似度。在 PASCAL VOC 和 MS COCO 数据集上的实验结果表明，所提方法可以有效提升模型在对抗样本上的对抗鲁棒精度。

关键词：目标检测；对抗攻击；对抗鲁棒性；对抗训练；梯度对齐

中图分类号：TP391　　**文献标志码：**A

Enhancing adversarial robustness in object detection: a gradient alignment approach

XUE Wei[1], LI Jie[1], DU Mingyang[2]*

(1. School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China;
2. College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China)

Abstract: To better address the problem of insufficient adversarial robustness of object detection models under adversarial attacks, an adversarial robustness enhancement method for object detection models based on gradient alignment was proposed. During the adversarial training phase, a composite loss function was constructed based on adversarial loss and alignment loss. By introducing a gradient alignment strategy, the gradient differences between adversarial examples and clean examples were constrained. Furthermore, with the supervisory signals of knowledge-distilled and the representational capability of self-supervised learning being incorporated, the feature similarity between adversarial and clean examples was maximized. Experimental results on the PASCAL VOC and MS COCO datasets demonstrate that the proposed method effectively improves the model's adversarial robustness accuracy against adversarial examples.

Keywords: object detection; adversarial attack; adversarial robustness; adversarial training; gradient alignment

在当前国家信息安全战略背景下，基于深度学习的智能检测技术已成为关键基础设施安全监测的重要保障。作为典型代表，基于卷积神经网络（convolutional neural network，CNN）的目标检测模型近年来在检测精度、推理效率及泛化能力等方面取得了显著进展，在交通标志检测[1]、桥梁裂缝检测[2]、注油孔检测[3]、合成孔径雷达图像检测[4]等领域得到广泛应用。在目标检测任务中，模型需同时识别图像中目标的类别信息及其空间位置，并输出相应的边界框。然而，随着研究的深入，目标检测模型在面对对抗攻击时所表现出的脆弱性日益凸显，这给目标检测任务的安全性与可靠性带来了严峻挑战。对抗攻击是一种通过在原始输入样本中添加人眼难以察觉的微小扰动，从而生成对抗样本以诱导模型产生错误输出的攻击方式[5]。现有目标检测模型在遭遇对抗样本时，其性能整体下降[6]。因此，探索能够增强模型对抗鲁棒性的方法，提升模型在遭受恶意攻击

收稿日期：2025-03-21

基金项目：安徽省自然科学基金资助项目（2208085MF168）；安徽省高校科研计划资助项目（2023AH040149）

*__第一作者：__薛伟（1986—），男，江苏南通人，副教授，博士，硕士生导师，E-mail：xuewei@ahut.edu.cn

引用格式：薛伟，李婕，杜明洋. 梯度对齐下的目标检测模型对抗鲁棒性提升[J]. 国防科技大学学报

Citation: XUE W, LI J, DU M Y, et al. Enhancing adversarial robustness in object detection: a gradient alignment approach[J]. Journal of National University of Defense Technology

情况下的稳定性和可靠性，对于构建安全可信的智能系统具有重要的意义。

近年来，对抗训练作为一种提升目标检测模型对抗鲁棒性的方法，在一定程度上提升了模型的对抗鲁棒性，但在实际应用中仍面临一些挑战。在不牺牲干净样本的检测性能的前提下，提升模型在面对对抗攻击时的检测精度，仍是需要深入研究的问题。注意到，梯度不仅可以量化模型输出对输入扰动的敏感性，也能够决定参数的更新方向。对抗样本生成是对抗攻击的核心环节，其基本思想是通过在干净样本中添加精心设计的微小扰动（通常基于模型梯度信息生成），使模型产生错误的输出。而干净样本的梯度信息反映模型在标准条件下的学习方向和优化目标，与对抗样本梯度呈现显著差异。因此，从梯度层面分析干净样本与对抗样本差异，成为提高模型对抗鲁棒性的一条潜在路径。Tong 等在文献[7]中提出通过平衡干净样本损失函数与对抗鲁棒损失函数的梯度方向，提升了模型的泛化性与鲁棒性。Dong 等在文献[8]中分析了干净样本与对抗样本梯度之间的干扰，并提出鲁棒模型 RobustDet（robust detector），在干净样本和对抗样本上进行有效学习。Li 等在文献[9]中提出 GradAlign（gradient alignment）方法量化初始化期间每个样本梯度内冲突的程度，避免在对抗训练时的出现过拟合现象。Ganz 等在文献[10]提出感知对齐梯度（perceptually aligned gradients，PAG）方法，揭示了梯度对齐与鲁棒性之间的双向联系。然而，现有研究多集中于图像分类任务，关于利用两种样本梯度信息对目标检测模型对抗鲁棒性提升的探索相对缺乏。

本文从对抗样本梯度与干净样本梯度之间的差异出发，提出一种基于梯度对齐的目标检测模型对抗鲁棒性提升方法。该方法在传统对抗训练框架的上引入梯度对齐策略，以减小对抗样本与干净样本在梯度差异，并结合知识蒸馏与自监督学习的特征对齐策略，最大化两种样本在特征空间中的相似性，从而提升目标检测模型在面对对抗攻击时的对抗鲁棒性。

1　对抗样本与对抗训练

1.1　对抗样本

常用的对抗样本生成方法有快速梯度符号法[5]（fast gradient sign method，FGSM）、基本迭代法[11]（basic iterative method，BIM）、投影梯度下降法[12]（projected gradient descent，PGD）等。其中，PGD 是一种基于梯度的迭代对抗攻击方法，其核心思想是通过给定模型参数θ，多步迭代在输入空间寻找对抗扰动δ，使得添加扰动后的对抗样本在满足约束条件$\|\delta\|_\infty \le \epsilon$的前提下，最大化模型的损失函数$L(f_\theta(x), y)$，从而诱导模型产生错误预测。PGD 生成对抗样本$x^{adv}$的方式如下所示：

$$x^{adv} = \Pi_\Delta\left(x + \eta \cdot sign\left(\nabla_x L(f_\theta(x), y)\right)\right) \quad (1)$$

其中，x 表示干净样本；y 表示 x 对应的真实标签；$f_\theta(x)$ 表示模型的输出；L 表示损失函数；$\nabla_x L(f_\theta(x), y)$ 表示损失的梯度；$sign(\cdot)$ 为符号函数，用于构造扰动方向；Π_Δ 表示投影操作，其作用是将更新后的样本限制在Δ内；η 为步长，通常令$\eta = \varepsilon / k$，ε 表示允许的最大扰动幅度；k 为攻击步数。

1.2　对抗训练

对抗训练[5]是一种在模型训练过程中同时引入对抗样本与原始干净样本的方法，旨在增强模型对输入扰动的鲁棒性。在对抗训练中，模型参数通过联合优化干净样本与对抗样本上的损失函数进行更新，具体表示如下：

$$\min_\theta \sum_{i=1}^{N}[(1-\alpha)\max_{\delta\in\Delta} L(f_\theta(x_i+\delta), y_i) + \alpha L(f_\theta(x_i), y_i)] \quad (2)$$

其中，θ表示模型参数；N 表示样本数量；x_i 和 y_i 分别表示第 i 个样本和该样本的真实标签；$L(f_\theta(x_i+\delta), y_i)$ 表示模型的损失函数；$\alpha \in [0,1]$表示平衡对抗样本损失与干净样本损失的权重系数。模型通过在输入空间中学习更具泛化性和鲁棒性的特征表示，已增强其在对抗攻击场景下的稳定性。

2　基于梯度对齐的对抗训练方法

2.1　YOLOv3

YOLOv3[13]采用 K-means 聚类方法计算数据集中目标框的聚类中心作为先验框，更有效地匹配数据集中目标的尺度分布特征。此外，通过引入多尺度检测机制，YOLOv3 在不同层级的特征图上进行目标预测，增强对不同尺度目标的识别能力和模型的鲁棒性。YOLOv3 的损失函数主要

由边界框回归损失、置信度损失和分类损失构成。

1） 边界框回归损失（L_{box}）

边界框回归损失通过平方差损失计算预测边界框与真实边界框之间的误差，用于预测边界框的位置。对于中心坐标(x,y)，直接计算其平方误差；对于宽高(w,h)，采用先取平方根再计算平方误差的方式，如式（3）所示：

$$L_{box}=\lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}l_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2 +(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] \tag{3}$$

其中，(x_i,y_i)和$(\hat{x}_i,\hat{y}_i)$分别表示预测框和真实框的中心坐标；(w_i,h_i)和$(\hat{w}_i,\hat{h}_i)$分别表示预测框和真实框的宽和高；λ_{coord}表示控制边界框损失的权重参数；S^2表示特征图的网格总数；B表示每个网格单元内预测的边界框的数量；l_{ij}^{obj}表示指示函数，指当前第i个网格的第j个预测边界框负责检测某个真实目标时为1，否则为0。

2） 置信度损失（L_{conf}）

置信度损失采用二分类交叉熵损失函数，在特征空间构建区分目标与背景的判别边界，指导模型更准确地识别包含目标的前景区域与不包含目标的背景区域，具体计算如式（4）所示：

$$L_{conf}=-\sum_{i=0}^{s^2}\sum_{j=0}^{B}l_{ij}^{obj}\log(c_{ij})-\lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B}(1-l_{ij}^{obj})\log(1-c_{ij}) \tag{4}$$

其中，c_{ij}表示模型预测的第i个网格的第j个边界框的置信度；λ_{noobj}表示当预测框没有预测到目标时，置信度误差在损失函数中所占权重。

3） 分类损失（L_{class}）

分类损失采用多类别交叉熵损失函数，对每个边界框的类别预测，分类损失仅针对包含目标的边界框进行计算，如式（5）所示：

$$\mathcal{L}_{\text{class}}=-\sum_{i=0}^{S^2}\sum_{j=0}^{B}l_{ij}^{\text{obj}}\sum_{c\in\text{classes}}[\hat{p}_{ij}^{c}\log(p_{ij}^{c}) +(1-\hat{p}_{ij}^{c})\log(1-p_{ij}^{c})] \tag{5}$$

其中，p_{ij}^{c}和$\hat{p}_{ij}^{c}$分别表示模型预测类别的概率和真实类别的概率。

在YOLOv3的损失函数中，由边界框回归损失、分类损失、置信度损失组成，通过最小化这三部分损失优化模型参数，总损失函数如下：

$$L_{yolo}=L_{box}+L_{conf}+L_{class} \tag{6}$$

2.2 特征对齐

为更有效地学习对抗样本与干净样本之间的特征，在YOLOv3中引入知识蒸馏特征对齐（knowledge-distilled feature alignment，KDFA）和自监督特征对齐（self-supervised feature alignment，SSFA）策略，两种方法均基于特征对齐（feature alignment，FA）[14]进行对抗训练，从而增强模型在面对对抗攻击时的鲁棒性。

KDFA方法引入知识蒸馏机制[15]，通过知识蒸馏提供的监督信号，在干净数据集上训练教师网络指导学生模型，使其在面对对抗扰动时能够学习到与教师网络在干净样本上提取的相似特征表示，从而提升学生网络的稳定性。

SSFA方法利用自监督学习[16]的表征能力，通过干净样本与对抗样本上的特征表示保持一致性，使其能够学习到对输入扰动不敏感的判别特征。具体而言，首先通过PGD方法生成对抗样本，使学生网络提取干净样本和和对应的对抗样本的特征，再将两种样本作为同一数据的不同表示，衡量两种样本特征的相似性。

如图1所示，FA方法结合两种特征对齐方法，利用平均池化作为特征投影算子，将高维特征映射到低维流形空间。通过最大化中间层特征相似度来实现特征对齐，从而提升模型在对抗攻击下的对抗鲁棒性。两种特征对齐的具体计算方法如式（7）所示：

$$\begin{cases}L_{kd}=\beta\cdot\dfrac{1}{N}\sum_{i=1}^{N}\left(1-sim(f_i^{mid},ft_i^{mid})\right)\\ L_{ss}=\gamma\cdot\dfrac{1}{N}\sum_{i=1}^{N}\left(1-sim(f_i^{mid},fss_i^{mid})\right)\end{cases} \tag{7}$$

其中，N表示样本的总数；f_i^{mid}表示每个样本的特征向量；ft_i^{mid}表示教师网络的特征向量；fss_i^{mid}表示自监督特征对齐中的特征向量；sim表示余弦相似度计算操作；β和γ表示控制损失权重的两个参数。另外，设置f^{mid}为YOLOv3结构内Darknet-53的最后一层。Darknet-53由53层卷积层组成，其最后一层用于提取高级特征，并将特征映射到YOLOv3检测头，输出特征图用于后续模型进行边界框预测、类别分类和置信度预测。

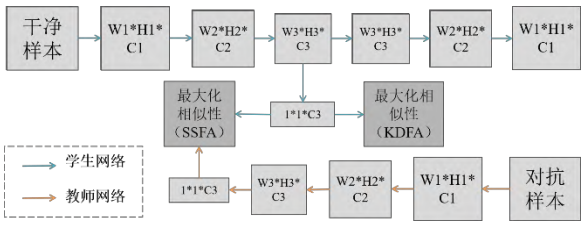


图 1　FA 方法概念流程图

Fig.1　Conceptual flowchart of the FA methodology

2.3　梯度对齐

由于干净样本与对抗样本在梯度更新方向上的不一致性，模型在学习过程中可能难以同时兼顾对两类样本的有效表征，进而影响其在对抗环境下的检测性能。针对这一问题，在 FA 的基础上提出基于梯度对齐的对抗训练方法（gradient alignment，GA）。梯度对齐方法流程如图 2 所示，该方法主要通过 PGD 生成对抗样本，再与干净样本一同输入到 YOLOv3 模型进行对抗训练，并通过约束干净样本和对抗样本的梯度差异，提取两种样本相似的特征表示，从而有效提升模型在面对对抗攻击时的对抗鲁棒性。具体而言，该方法首先计算干净样本的损失函数，并通过反向传播得到模型参数的梯度，如式（8）所示：

$$L_{clean} = L_{yolo}(\hat{y}, y) \tag{8}$$

其中，$\hat{y}$ 是模型对于干净样本的预测；y 是实际标签。通过 L_{clean} 反向传播计算梯度：$\nabla_\theta L_{clean}$。

通过 PGD 生成对抗样本，计算对抗样本的损失函数和梯度，如式（9）所示：

$$L_{adv} = L_{yolo}(\hat{y}_{adv}, y) \tag{9}$$

其中，$\hat{y}_{adv}$ 是模型对于对抗样本的预测。通过 L_{adv} 反向传播计算梯度：$\nabla_\theta L_{adv}$。

通过计算干净样本与对抗样本梯度的 ℓ_2 范数，衡量两种样本在参数更新方向上的差异，并引入相应的优化目标以最小化两者之间的梯度分布差异，从而提升模型在面对对抗扰动时的鲁棒性，如式（10）所示：

$$L_{grad} = \frac{1}{N}\sum_{i=1}^{N} \| \nabla_\theta L_{clean}^{(i)} - \nabla_\theta L_{adv}^{(i)} \|_2^2 \tag{10}$$

其中，N 表示模型训练时的样本数量，$\nabla_\theta L_{clean}^{(i)}$ 和 $\nabla_\theta L_{adv}^{(i)}$ 分别表示第 i 个样本的梯度。

结合对抗损失、知识蒸馏特征对齐损失和自我监督特征对齐损失以及梯度对齐损失，模型总损失函数如下所示：

$$L_{total} = L_{adv} + \lambda L_{grad} + L_{kd} + L_{ss} \tag{11}$$

其中，λ 表示在模型对抗训练时控制梯度对齐损失的权重系数。

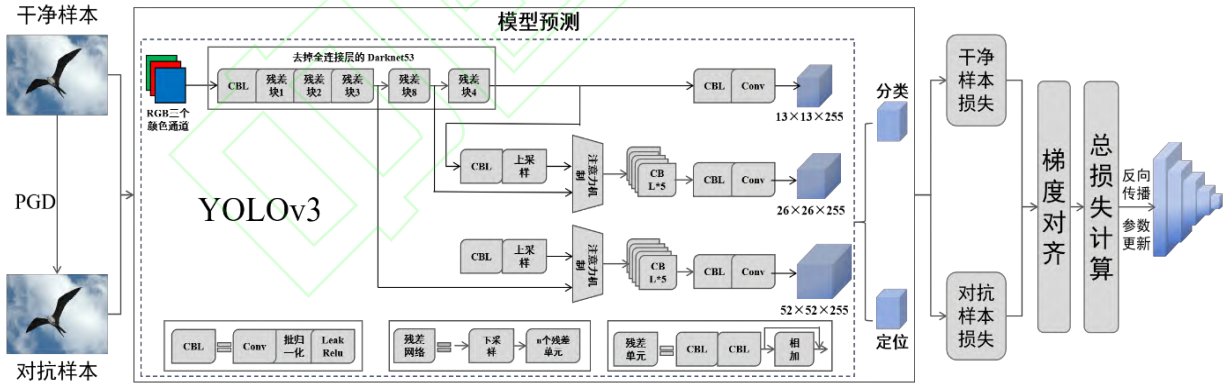


图 2　梯度对齐方法流程图

Fig.2　Flowchart of the gradient alignment method

算法整体流程如算法 1 如示，GA 方法在对抗训练中使用 PGD 生成对抗样本，通过减少干净样本和对抗样本之间的梯度差异，并引入特征对齐策略，最大化干净样本和对抗样本的特征相似性，有效降低对抗攻击对模型的影响，提高模型在对抗样本上的对抗鲁棒性。

算法 1　基于梯度对齐的对抗训练方法

Alg.1　Adversarial training method based on gradient alignment

输入：训练数据集 D，损失函数 L，模型 f，对抗样本 x^{adv}，教师模型 ft，批次大小 B，总轮数 E，系数 β 、 γ 、 λ

输出：训练好的模型 f

初始化模型参数 θ

for $e = 1$ **to** E **do**

for $b = 1$ **to** $ceil(|D|/B)$ **do**

$x^{adv} = PGD(x, y)$

$L_{clean} = L(f(x), y)$

$L_{adv} = L(f(x^{adv}), y)$

$L_{kd} = \beta \cdot (1 - sim(f^{mid}, ft^{mid}))$

$L_{ss} = \gamma \cdot (1 - sim(f^{mid}, fss^{mid}))$

$g_{clean} = \nabla_\theta L_{clean}$

$g_{adv} = \nabla_\theta L_{adv}$

$L_{grad} = \| \nabla_\theta L_{clean}^{(i)} - \nabla_\theta L_{adv}^{(i)} \|_2^2$

$L_{total} = L_{adv} + \lambda \cdot L_{grad} + L_{kd} + L_{ss}$

end for

end for

return 训练好的模型 f

3 实验结果及分析

3.1 实验设置

3.1.1 数据集设置

实验采用 PASCAL VOC[17]和 MS-COCO[18]的两种数据集对 GA 进行评估。PASCAL VOC 包含 20 个目标类别，如人、车、动物等，实验采用 2007 年和 2012 年训练的组合，测试集使用 2007 年的测试。MS-COCO 是一个大规模目标检测的数据集，涵盖 80 个类别，涉及日常生活中的多种常见物体，实验采用 train+valminusminival 2014，测试集使用 minival 2014。

3.1.2 评价指标

实验使用的衡量模型检测性能和对抗鲁棒性表现的评价指标主要包括召回率（recall，R）、平均精度均值（mean average precision，mAP）、交并比（intersection over union，IoU）等。

其中，IoU 用于衡量预测框与真实框的重叠比例，具体计算方法如式（12）所示：

$$IoU = \frac{\text{预测框与真实框的交集面积}}{\text{预测框与真实框的并集面积}} \tag{12}$$

IoU 取值范围为$[0,1]$，数值越大，表明预测框与真实框的重叠区域越多。

R 衡量所有正样本中被正确检测的比例，具体计算方法如式（13）所示：

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \tag{13}$$

其中，N_{tp} 表示模型被正确检测的目标数量；N_{fn} 表示模型被漏检的目标数量。

平均精度（average precision，AP）衡量模型对单类别的检测性能。其中，精度（precision，P）的具体计算方法如式（14）所示：

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \tag{14}$$

其中，N_{fp} 表示模型被误检的目标数量。通过计算精确率-召回率曲线下的面积计算 AP，如式（15）所示：

$$AP = \int_0^1 P(R)dR \tag{15}$$

AP 值越高，表明检测结果的准确性越高。

mAP 通过计算所有类别 AP 的平均值评估多类别目标。mAP 越高，表明模型的综合性能越强，如式（16）所示：

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{16}$$

其中，N 表示类别总数；AP_i 表示第 i 类的 AP。实验所使用的两种数据集中 mAP 的计算方式如下所示：

1）VOC 测评方式：使用 mAP@0.5 测试模型，表示交并比阈值固定为 0.5；

2）COCO 测评方式：使用 mAP@0.5:0.95 测试模型，表示在 0.5 到 0.95 的 IoU 阈值下计算 AP 并取均值。

3.1.3 实验参数设置

在测试检测模型的对抗鲁棒性时，参数设置如下：$\varepsilon = 0.01$、$\alpha = 0.5$、$\beta = 10$、$\gamma = 10$、$\lambda = 0.1$。考虑到不同 PGD-k 攻击下的不同性能，实验在对抗训练中使用不同攻击步数、训练相应模型，固定步长大小为 0.01，评估目标检测模型的对抗鲁棒精度。实验使用对抗样本精度（adv_mAP）和干净样本精度（cle_mAP）评价目标检测模型的性能，并与 KDFA 与 SSFA 两种对抗训练方法进行对比消融实验。

模型对抗训练时，实验使用分段学习率衰减策略，提高训练的稳定性并加速收敛，并以随机梯度下降作为优化器。设定初始学习率为lr_0，在训练总轮数的 2/3 处进行学习率衰减，缩小为原来的 1/10，以提升模型在收敛阶段的稳定性与精度。具体计算方法如式（17）所示：

$$lr_t = \begin{cases} lr_0, t < \dfrac{2}{3}T \\ 0.1 \times lr_0, t \geq \dfrac{2}{3}T \end{cases} \tag{17}$$

其中，lr_t 表示第 t 轮学习率；T 表示总训练轮数。

3.2 结果分析

3.2.1 对比消融分析

在 PASCAL VOC 数据集上，实验采用对比消融分析模型对抗鲁棒性，并且对于不同的对抗训练方法的设置具有一致性。其中，YOLOv3 表示未经过对抗训练的模型，以下简称 v3；KD 表示 KDFA 方法；KDSS 表示训练时使用 KDFA 与 SSFA 两种方法；KD_GA 表示训练时使用 KDFA 与 GA 两种方法；KDSS_GA 表示训练时使用 KDFA、SSFA 和 GA 三种方法。

如表 1 实验结果表明，各模块对模型性能的影响呈现差异。v3+KD 独立作用时，对抗样本精度有所提升，但干净样本精度从 0.78 降至 0.523。v3+SS 独立作用时，模型的对抗鲁棒性有所提升，但在干净样本上的精度却大幅降低。

表 1 VOC 数据集下各种方法的模型 mAP 对比

Tab.1 Comparison of model mAP of various methods under the VOC dataset

方法	cle_mAP	adv_mAP	精度下降
YOLOv3	0.780	0.257	0.523
v3+SS	0.502	0.261	0.241
v3+KD	0.523	0.200	0.323
v3+KDSS	0.723	0.301	0.422
v3+SS_GA	0.590	0.377	0.113
v3+KD_GA	0.586	0.422	0.164
v3+KDSS_GA	0.696	0.487	0.209

在不引入 GA 时，模型表现较为局限。虽然 v3+KDSS 方法能使干净样本精度保持至 0.723，较 KD 或 SS 方法单独使用时更具稳定性。但 KDSS 方法训练后的模型对抗样本精度提升过低，模型整体性能较低。相比之下，GA 方法的引入使模型表现出更好的对抗鲁棒性。v3+SS_GA 相比 v3+SS 将精度下降幅度从 0.241 降低至 0.113，进一步提升了模型的对抗鲁棒精度。相比 v3+KD，在 v3+KD_GA 方法训练后模型的对抗样本精度将精度从 0.2 提升至 0.422，同时将对比干净样本的精度下降幅度控制在 0.164，有效提升模型的对抗鲁棒精度。

结合三个模块的 KDSS_GA 方法训练后的模型展现出较好的检测性能。在干净样本上的 mAP 达到 0.696，在对抗样本上的 mAP 提升至 0.487，相比 YOLOv3 在对抗样本下的 mAP 有明显提升，有效提升了模型的对抗鲁棒性。

实验另外使用 COCO 数据集训练模型并测试其性能。COCO 数据集下各种方法的模型 mAP 对比如表 2 所示，由于 COCO 数据集中的小目标复杂场景的特性，YOLOv3 的精度下降幅度远超 VOC 数据集。而 v3+KDSS_GA 对比干净样本，精度下降仅 0.162，在对抗样本的检测精度上也达到更好的对抗鲁棒性。

表 2 COCO 数据集下各种方法的模型 mAP 对比

Tab.2 Comparison of model mAP of various methods under the COCO dataset

方法	cle_mAP	adv_mAP	精度下降
YOLOv3	0.545	0.065	0.480
v3+KD	0.387	0.115	0.272
v3+KDSS	0.489	0.187	0.302
v3+KD_GA	0.436	0.251	0.185
v3+KDSS_GA	0.444	0.282	0.162

综上所述，v3+KDSS_GA 方法在两个数据集上表示出较好的对抗鲁棒性。相较于其他方法，KDSS_GA 方法在干净样本和对抗样本上的表现均较为均衡，模型在对抗样本上的精度有所提升，在干净样本上的精度下降幅度进一步降低，体现该方法在对抗鲁棒性方面的优越性。如图 3 所示，展示 KDSS_GA 方法在 VOC 数据集上训练所得模型对干净样本的部分可视化检测结果。

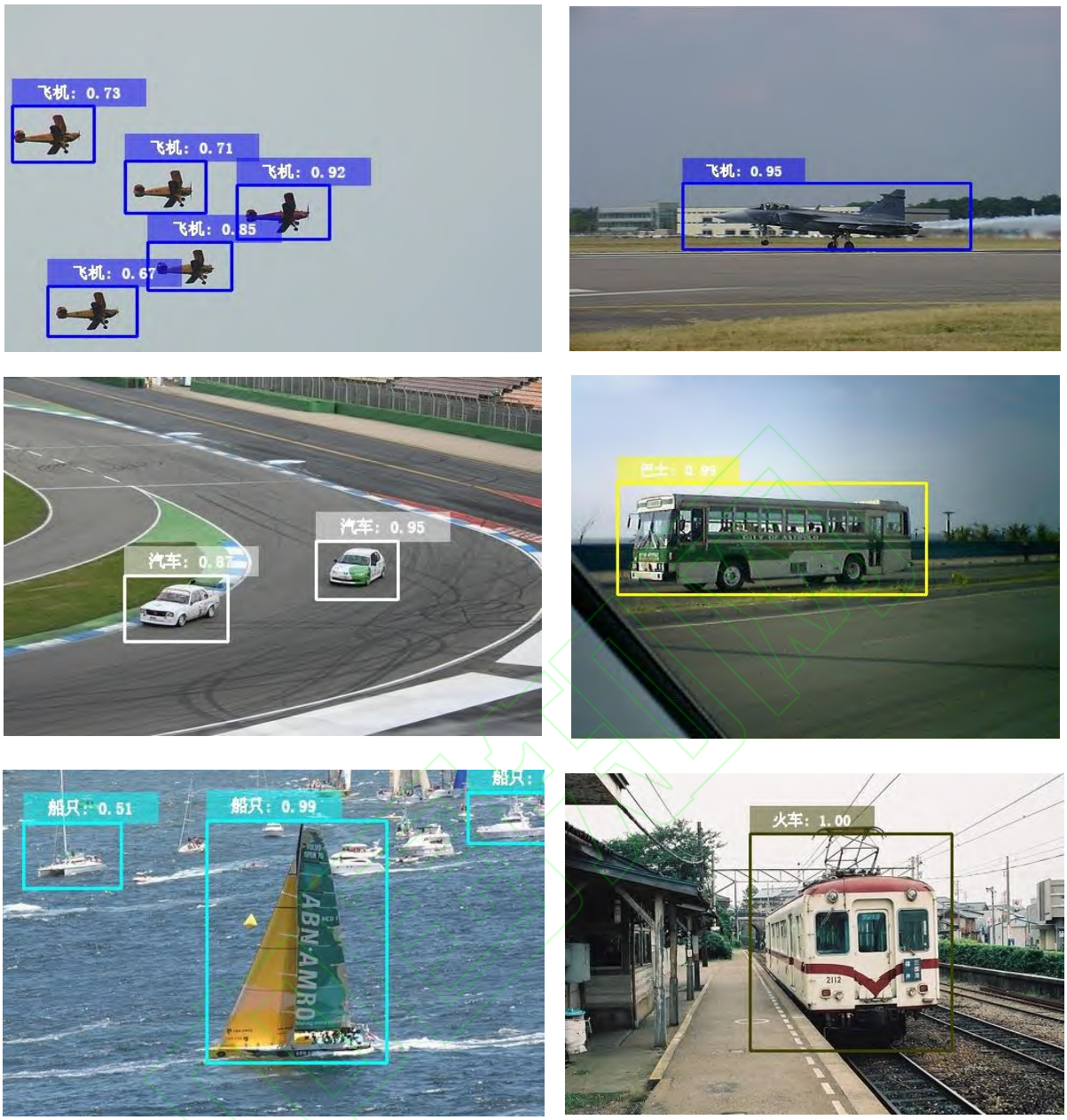


图 3　KDSS_GA 方法在 VOC 数据集的干净样本上的部分可视化结果

Fig.3　Partial visualisation results of the KDSS_GA method on clean examples of the VOC dataset

3.2.2　参数实验分析

实验使用不同攻击步数和步长大小对模型进行对抗训练，对不同方法训练后的模型在 VOC 数据集上进行性能分析。

不同攻击步数下各种方法的模型 mAP 对比如表 3 所示，当使用 PGD-1 攻击步数训练时，v3+KDSS_GA 训练后的模型在干净样本和对抗样本上的表现最佳，表示出较好的对抗鲁棒性。随着攻击步数增加，模型在两种样本上的检测精度均有所下降。较小的对抗训练攻击步数能够在保持较高干净样本性能的同时，有效提升 YOLOv3 模型在对抗样本上的鲁棒性。

表 3　不同攻击步数下各种方法的模型 mAP 对比

Tab.3　Comparison of model mAP for various methods at different attack steps

VOC	方法	cle_ mAP	adv_ mAP	精度下降
PGD-1	YOLOv3	0.780	0.257	0.523
	v3+KD	0.523	0.200	0.323
	v3+KDSS	0.723	0.301	0.422

	v3+KD_GA	0.586	**0.422**	**0.164**
	v3+KDSS_GA	0.696	**0.487**	**0.209**
PGD-3	YOLOv3	0.780	0.051	0.729
	v3+KD	0.498	0.129	0.369
	v3+KDSS	0.663	0.187	0.476
	v3+KD_GA	0.514	0.165	0.349
	v3+KDSS_GA	0.574	**0.356**	**0.218**
PGD-5	YOLOv3	0.780	0.016	0.764
	v3+KD	0.468	0.133	0.335
	v3+KDSS	0.617	0.180	0.437
	v3+KD_GA	0.551	0.214	**0.337**
	v3+KDSS_GA	0.570	**0.229**	**0.341**
PGD-10	YOLOv3	0.780	0.004	0.776
	v3+KD	0.410	0.145	0.265
	v3+KDSS	0.600	0.209	0.391
	v3+KD_GA	0.530	0.201	**0.329**
	v3+KDSS_GA	0.552	**0.216**	0.336

使用不同攻击步数进行对抗训练时，对各种方法在 VOC 数据集上的性能也有一定影响。KDSS、KD_GA 和 KDSS_GA 方法在对抗样本上的表现更优于 YOLOv3，其中 v3+KDSS_GA 方法在对抗样本上的精度更优。随着攻击步长的增加，所有方法的性能均逐渐下降。在 PGD-3 攻击步数下，YOLOv3 在对抗样本上的检测精度下降过多，v3+KDSS、v3+KD_GA 和 v3+KDSS_GA 方法的对抗样本性能都有所下降，但 v3+KDSS_GA 方法仍保持相对较高的对抗鲁棒性。模型在 PGD-5 和 PGD-10 攻击步数的测试下，v3+KDSS、v3+KD_GA 和 v3+KDSS_GA 方法在面对对抗样本时的检测精度有所下降，但下降幅度相对较小。

总体而言，v3+KDSS_GA 方法在使用不同的攻击步数训练后的模型均表现出较好的对抗鲁棒性，在使用较少攻击步数的对抗训练下，模型的在准确率与鲁棒性上表现出更优的综合性能。

实验使用不同攻击步数训练模型，并对模型分别测试不同对抗攻击强度和不同步长大小生成的对抗样本的精度。将表 4 中的三种步长大小的对抗样本精度与表 3 中 v3+KDSS_GA 方法下干净样本的检测精度进行对比，模型在对抗样本上的精度略有下降，但在三种对抗样本上整体精度下降幅度较小。其中，在 PGD-5 训练下对模型进行 PGD-3 的测试，其 mAP 达到 0.361，比步长大小为 0.02 时提升 60%。PGD-5 训练后的模型在步长为 0.005 时对其进行 PGD-1 测试，其 mAP 为 0.460，优于 PGD-1 和 PGD-10，说明过强或过弱的训练攻击都会降低模型性能。较小的步长大小能够在保持较高干净样本精度的同时，提升模型在对抗样本上的检测精度。

在未受到对抗攻击时，模型能够较好地检测目标，KDSS_GA 方法训练的模型在干净样本上的召回率如图 4 所示，召回率普遍接近或高于 0.7，模型能够识别大部分真实目标。尽管对抗攻击会导致召回率下降，模型仍能在较小的攻击步数生成的对抗样本中保持一定的性能，表示出一定的抗干扰能力。KDSS_GA 方法训练的模型在对抗样本上的召回率如图 5 所示，在步长大小较小的情况下，模型的对抗鲁棒性相对较强，能够在轻度扰动下保持较好的性能。

表 4　KDSS_GA 模型不同攻击步数和步长大小下的 mAP 对比

Tab.4　Comparison of mAP for different attack steps and step sizes for the KDSS_GA model

训练	测试	步长大小 0.02	步长大小 0.01	步长大小 0.005
PGD-1	PGD-1	0.294	**0.400**	**0.487**
	PGD-3	0.156	0.182	0.315
	PGD-5	0.121	0.144	0.205
	PGD-10	0.097	0.109	0.136
PGD-3	PGD-1	0.325	**0.411**	**0.461**
	PGD-3	0.220	0.237	0.356
	PGD-5	0.200	0.214	0.265
	PGD-10	0.188	0.194	0.210
PGD-5	PGD-1	0.326	**0.412**	**0.460**

	PGD-3	0.225	0.240	0.361
	PGD-5	0.205	0.219	0.270
	PGD-10	0.196	0.201	0.215
PGD-10	PGD-1	0.318	**0.398**	**0.445**
	PGD-3	0.225	0.238	0.349
	PGD-5	0.209	0.220	0.267
	PGD-10	0.202	0.206	0.216

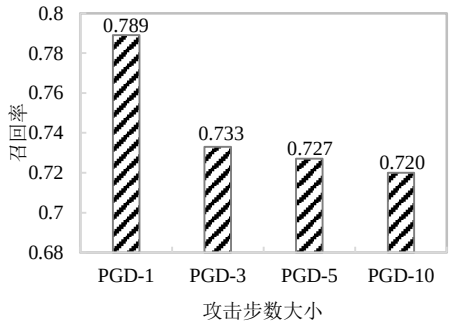
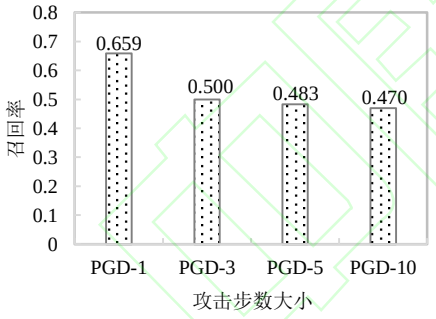


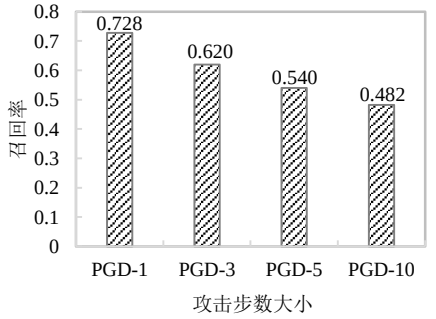
图 4　KDSS_GA 方法训练的模型在干净样本上的召回率

Fig.4　Recall of the model trained by the KDSS_GA method on clean examples



（a）步长大小为 0.01 时的召回率

（a）Recall at a step size of 0.01



（b）步长大小为 0.005 时的召回率

（b）Recall at a step size of 0.005

图 5　KDSS_GA 方法训练的模型在对抗样本上的召回率

Fig.5　Recall of the model trained by the KDSS_GA method on adversarial examples

4　结论

本文提出了一种基于梯度对齐的对抗鲁棒性提升方法，通过最小化干净样本与对抗样本之间的梯度差异，提升目标检测模型在面对对抗攻击时的对抗鲁棒性。所提方法可为基于卷积神经网络的目标检测模型在实际应用中的安全性提供一定的技术支撑。未来工作将进一步扩展该方法在多种对抗攻击下的检测验证。

参考文献（References）

[1] 郭君斌，于琳，于传强. 改进 YOLOv5s 算法在交通标志检测识别中的应用[J]. 国防科技大学学报, 2024, 46(6): 123-130.
GUO J B, YU L, YU C Q. Application of improved YOLOv5s algorithm in traffic sign detection and recognition[J]. Journal of National University of Defense Technology, 2024, 46(6): 123-130. (in Chinese)

[2] 蒋仕新，邹小雪，杨建喜，等. 复杂背景下基于改进 YOLO v8s 的混凝土桥梁裂缝检测方法[J]. 交通运输工程学报, 2024, 24(6): 135-147.
JIANG S X, ZOU X X, YANG J X, et al. Concrete bridge crack detection method based on improved YOLO v8s in complex backgrounds[J]. Journal of Traffic and Transportation Engineering, 2024, 24(6): 135-147. (in Chinese)

[3] 孟瑞，卢宗远，丛英浩. 基于轻量级卷积神经网络的注油孔检测算法[J]. 安徽工业大学学报(自然科学版), 2023, 40(2): 166-172, 190.
MENG R, LU Z Y, CONG Y H. Oil injection hole detection algorithm based on lightweight convolution neural network[J]. Journal of Anhui University of Technology (Natural Science), 2023, 40(2): 166-172, 190. (in Chinese)

[4] 韩子硕，王春平，付强，等. 联合生成对抗网络和检测网络的 SAR 图像目标检测[J]. 国防科技大学学报, 2022, 44(3): 164-175.
HAN Z S, WANG C P, FU Q, et al. Target detection in SAR images based on joint generative adversarial network and detection network[J]. Journal of National University of Defense Technology, 2022, 44(3): 164-175. (in Chinese)

[5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2015-03-20)[2025-03-01].

https://arxiv.org/abs/1412.6572.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19)[2025-03-01]. https://arxiv.org/abs/1312.6199.

[7] TONG H Y, ZHANG X Y, JIN Y L, et al. Balancing generalization and robustness in adversarial training *via* steering through clean and adversarial gradient directions[C]//Proceedings of the 32nd ACM International Conference on Multimedia, 2024: 1014-1023.

[8] DONG Z Y, WEI P X, LIN L. Adversarially-aware robust object detector[C]// Proceedings of European Conference on Computer Vision, 2022: 297-313.

[9] LI Y X, GUO Y H. GradAlign for training-free model performance inference[EB/OL]. (2024-11-29)[2025-03-01]. https://arxiv.org/abs/2411.19819v1.

[10] GANZ R, KAWAR B, ELAD M, et al. Do perceptually aligned gradients imply robustness? [C]//Proceedings of the 40th International Conference on Machine Learning, 2023: 10628-10648.

[11] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]. Artificial intelligence safety and security. Boca Raton: Chapman and Hall/CRC, 2018: 99-112.

[12] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2019-09-04)[2025-03-01]. https://arxiv.org/abs/1706.06083.

[13] REDMON J, FARHADI A. Yolov3: an incremental improvement [EB/OL]. (2018-04-08)[2025-03-01]. https://arxiv.org/abs/1804.02767.

[14] XU W P, HUANG H C, PAN S Y. Using feature alignment can improve clean average precision and adversarial robustness in object detection[C]//Proceedings of the IEEE International Conference on Image Processing (ICIP), 2021: 2184-2188.

[15] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2018-03-09)[2025-03-01]. https://arxiv.org/abs/1503.02531v1.

[16] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]// Proceedings of the 37th International Conference on Machine Learning, 2020: 1597-1607.

[17] EVERINGHAM M, ALI ESLAMI S M, VAN GOOL L, et al. The pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[18] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of European Conference on Computer Vision, 2014: 740-755.