面向自动驾驶中强大而实用的 2D 目标检测补丁攻击

Yuxin Cao[1,2] Yedi Zhang[2] Wentao He[3] Yifan Liao[2] Yan Xiao[4]
Chang Li[1] Zhiyong Huang[2] Jin Song Dong[2]

[1]*State Key Laboratory of Intelligent Vehicle Safety Technology, Changan Automobile, Chongqing, China*
[2]*National University of Singapore, Singapore*
[3]*Ningbo University, Ningbo, China*
[4]*Sun Yat-Sen University, Guangzhou, China*

Fig. 1. mAP 基础攻击指标（mAP 和 ASR）中固有的高估问题的例子。(a) 多个检测框。(b) 低于阈值的 IoU 匹配。(c) 攻击失败和成功。尽管在 (a) 和 (b) 中 mAP 在所有图像上几乎降到了零，但没有人是完全未被检测到的。然而，我们提出的 PASR 总是能够反映这个人是否被隐藏。补丁来自 T-SEA。

***Abstract*—尽管取得了进展，基于学习的自动驾驶系统仍然对对抗性补丁严重脆弱，这在其真实环境部署中带来了严重的安全和保障风险。特别令人担忧的是黑盒攻击，这种攻击无需模型知识便能成功，并且它的可迁移性被广泛研究以减少与查询攻击方法相比的计算成本。以前基于可迁移性的黑盒攻击通常采用平均精度（mAP）作为评估指标，并据此设计训练损失。然而，由于存在多个检测到的边界框和相对宽松的交并比（IoU）阈值，这些方法的攻击效果常常被高估，导致在实际攻击场景中的成功率降低。此外，在低分辨率数据上训练的补丁往往无法在高分辨率图像上保持有效性，限制了其在高分辨率自动驾驶数据集上的迁移能力。为了填补这一空白，我们提出了 P^{3}A，一个面向自动驾驶 2D 目标检测的强大且实用的补丁攻击框架，专门针对高分辨率数据集进行优化。首先，基于 IoU，我们引入了一种新的评价指标，实用攻击成功率（PASR），以更准确地量化对抗性补丁攻击的效果，并与自动驾驶中的行人安全更相关。其次，我们提出了一种专门的损失函数，定位-置信抑制损失（LCSL），以在 PASR 下提高攻击的迁移能力。最后，为了保持在高分辨率数据集上的迁移能力，我们进一步将概率性尺度保持填充（PSPP）融入补丁攻击流程中，作为数据预处理步骤。大量实验证明，无论是在提出的基于 IoU 的实用评价指标下还是之前的 mAP 评价指标下，P^{3}A 在未见过的模型和未见过的高分辨率数据集上都超越了最新的攻击技术。**

I. 引言

学习驱动的二维目标检测在自动驾驶等应用中已达到人类水平表现 [?], [?] ，但仍易受对抗性扰动的影响 [?] ，这威胁到现实世界部署的可靠性和安全性 [?], [?] 。在各种攻击方法中，对抗性补丁攻击由于在物理环境中的实用性而引起了特别关注 [?] 。最初针对交通标志识别，这些攻击已发展到危及行人检测 [?], [?] ，对自动驾驶车辆构成了严重的风险。与基于查询的扰动不同 [?] ，补丁通常在多个代理模型的集成上进行训练，以增强在模型、数据集和环境之间的可迁移性 [?] ，将负担从测试时的查询转移到离线训练。最近的工作如 T-SEA [?] 通过采用自集成策略进一步提高了可迁移性，展示了仅使用一个代理模型的有效攻击。

现有的大多数关于可转移补丁攻击的研究主要依赖于平均精度均值（mAP）作为主要评估指标，并专门设计相应的损失函数和优化策略以最小化 mAP。然而，正如图 1 所示，补丁攻击后 mAP 值通常会下降，这主要由两个因素导致：i）对于单个行人预测多个重叠检测框，ii）预测框与真实值（GT）边界框之间的交并比（IoU）值低于 IoU 阈值。具体来说，多个重叠框的存在会导致假阳性（FP）的增加，从而降低 mAP 的精确性组件。此外，如果所有检测框的 IoU 均低于阈值，相应的 GT 无法匹配到任何预测，这导致真正例（TP）的损失，进一步降低精确性和召回率，最终导致 mAP 的下降。由于在自动驾驶中这样的情况不会阻止车辆在行人面前停下，因此它们并不构成真正的安全隐患 [?] 。因此，迫切需要一种更具实用性的指标——能够有效反映检测器是否直接忽略行人。此外，现有的补丁是在低分辨率（LR）数据（≈ 600 像素）上训练的，并且在行人占据较小区域且补丁难以主导检测的高分辨率（HR）自动驾驶数据（≥ 1200 像素）上表现不佳。

在这项工作中，我们引入了一个强大且实用的补丁攻击框架，P^{3}A，用于提高对 HR 自动驾驶数据集的对抗性补丁的实际攻击迁移性。为此，我们首先提出了“实用攻击成功率（Practical Attack Success Rate，PASR）”，这是一种新评估指标，通过仅在没有检测框与 GT 行人重叠时认为攻击成功来更准确地反映真实世界的风险，因为这种情况会在现实场景中导致与行人的直接碰撞。受新指标和模型训练阶段使用的 IoU 损失启发，我们提出了一种“定位-置信度压制损失（Localization-Confidence Suppression Loss，LCSL）”，它同时利用 IoU 和置信度得分来促进攻击并提高迁移性。注意到几乎所有现有工作仅依赖置信度分数（对象性 & 分类分数），而忽视了 IoU 在增强攻击效果中的关键作用。最后，我们提出了“概率尺度保持填充（Probabilistic Scale-Preserving Padding，PSPP）”，以保持 HR 数据中人的绝对大小，这可以增强已训练补丁对自动驾驶数据集的迁移能力。通过全面的实验，我们展示了 P^{3}A 在 11 个主流目标检测器和 9 个广泛使用的数据集上，就 PASR 以及基于 mAP 的指标而言，优于最新的攻击方法。具体来说，P^{3}A 可以实现高达 52 % 的模型迁移性和 54 % 的数据迁移性。

我们的贡献可以总结如下。

- 我们总结了现有基于 mAP 的指标的高估问题，并提出了一种新的指标 PASR，该指标准确反映了对抗性

补丁的实际攻击性能。

- 受目标检测器训练中使用的 IoU 损失的启发，我们提出了一种定位-置信度抑制损失（LCSL），该方法充分利用了这两种信息，以增强训练补丁的有效性和可转移性。
- 我们提出了一种新颖的黑箱补丁攻击框架，P^{3}A，旨在解决现有补丁对 HR 自动驾驶数据的有限可转移性问题。
- 我们展示了 P^{3}A 在 11 个检测器和 9 个数据集（包括 7 个 HR 数据集）上优于 SOTA 攻击，显示出更强的迁移性。

II. 相关工作

黑箱对抗攻击。由于其在实际场景中的实用性，尤其是在针对商业模型时，黑箱对抗攻击近年来受到越来越多的关注。这些攻击通常根据对目标模型的访问程度分为基于查询的攻击 [?], [?] 和基于迁移的攻击 [?], [?] 。基于查询的攻击需要广泛的模型访问，这通常导致高成本并且容易引发警报。相比之下，基于迁移的攻击通过在本地替代模型上训练生成对抗样本，并且不依赖重复查询。由于这些攻击可以实时执行，因此在受限或高风险环境中特别实用且具有可扩展性。

基于迁移的目标检测补丁攻击。对抗性补丁已被广泛用于攻击目标检测器，特别是在人员检测任务中。由于需要实时响应和强大的迁移能力，基于迁移的攻击已经成为主流的攻击方法。现有工作主要集中在提高对抗性补丁的隐蔽性 [?], [?], [?] 或者增强攻击效果 [?], [?] 。为了确保跨模型的迁移能力，可以使用多样化检测器的集成 [?], [?] 在本地训练补丁，从而提高对未见模型的泛化能力。T-SEA [?] 提出了一个自集成框架，实现了最先进的跨模型迁移能力，同时只需一个替代模型进行补丁训练。另一方面，现有的基于迁移的攻击通常仅在低分辨率数据集（INRIA [?] ）上训练对抗性补丁，当应用于高分辨率自动驾驶场景时，通常表现出较差的攻击性能。此外，虽然常用的指标 mAP 对于一般目标检测有效，但在涉及多个重叠检测和低于阈值的 IoU 匹配的场景中则不那么可靠。本文中，我们为社区提出了一种新的实用指标，以更准确和可靠地评估实用补丁攻击，特别是在自动驾驶情境中。为了在此指标下提高攻击迁移能力，我们在攻击过程中引入了基于 IoU 的损失，并专注于解决高分辨率数据集的挑战。

III. 前言

威胁模型。P^{3}A 在两个现实的约束下操作。首先，与 T-SEA 一致，攻击者在训练过程中只能访问一个替代检测器，这减少了计算成本，并符合真实世界中识别和训练检测器资源密集的限制。其次，训练仅限于 LR 数据（INRIA），因为 HR 自动驾驶数据集（$\geq$ 1K）收集成本高且很少公开。此外，HR 数据集需要更高的训练成本，并且由于领域特异性可能缺乏泛化能力。在这些约束下，我们的关键研究问题是：如何仅使用一个替代和一个 LR 数据集，来提高补丁对黑盒检测器和未见过的 HR 自动驾驶数据的可迁移性?

问题表述。给定输入图像，物体检测器产生 N 个候选预测 $\bar{\boldsymbol{O}} = \{(\mathbf{b}_i, \bar{o}_i, \mathbf{s}_i)\}_{i=1}^{N}$ ，其中 $\mathbf{b}_i$ 表示边界框坐标，$\bar{o}_i$ 表示目标度分数，$\mathbf{s}_i \in \mathcal{R}^C$ 表示在 C 类上的分类分数。置信度分数低于置信度阈值的预测首先被丢弃，然后对剩余的框应用非极大值抑制（NMS），抑制那些 IoU 超过给定 IoU 阈值的框以消除重复。最终预测集是 $\boldsymbol{O} = \{(\boldsymbol{b}_j, p_j, c_j)\}_{j=1}^{M}$ ，其中 $\boldsymbol{b}_j$ 表示保留的边界框，c_j 表示预测的类别，p_j 是最终置信度分数（目标度分数 $\bar{o}_j$ 和类别 c_j 的分类分数 $\bar{s}_j^{c_j}$ 的乘积）。

TABLE I
用于可转移补丁攻击的度量和损失。OBJ：目标性分数，CLS：分类分数。

Research	Metric	Adversarial Loss
AdvPatch	mAP	obj/cls/obj $\times$ cls
NAP	mAP	obj $\times$ cls
LAP	mAP	obj
T-SEA	mAP	obj
AdvART	mAP	obj $\times$ cls
AdvTexture	mAP, ASR	obj
MVPatch	mAP, ASR	cls
DePatch	mAP, ASR	obj, affiliated by IoU
FDA	ASR	cls+IoU
P^{3}A	PASR, mAP, ASR	obj $\times$ cls $\times$ IoU

在目标检测攻击中，干净图像 $\boldsymbol{x}$ 的预测集 $\boldsymbol{O}(\boldsymbol{x})$ 被视为真实值 (GT) [?], [?] 。攻击者然后生成一个补丁 $\boldsymbol{\xi}$ ，通过掩码 $\boldsymbol{M}$ 应用在 $\boldsymbol{b}_j$ 内，从而产生 $\boldsymbol{x}_{adv} = \boldsymbol{x} \odot (1 - \boldsymbol{M}) + \boldsymbol{\xi} \odot \boldsymbol{M}$ 。因此，攻击者的目标是找到一个 ξ 来导致 $\boldsymbol{O}(\boldsymbol{x}_{adv}) \neq \boldsymbol{O}(\boldsymbol{x})$ 。由于目标检测涉及到回归和分类，传统的度量标准 mAP 认为如果

$$\mathrm{IoU}(\mathbf{b}_j(\boldsymbol{x}_{adv}), \mathbf{b}_j(\boldsymbol{x})) < 0.5 \quad \text{or} \quad c_j(\boldsymbol{x}_{adv}) \neq c_j(\boldsymbol{x}), \quad (1)$$

，其中在目标检测中常用的 IoU 阈值是 0.5，则攻击是成功的。然而，我们将展示这种成功攻击可能不会阻止自动驾驶汽车停止。

IV. 现有技术的分析

A. 攻击指标

表 I 总结了现有生成可转移对抗补丁的方法的主流手段。几乎所有方法都使用 mAP 作为数字攻击的评估指标，并且有几种方法还另外使用攻击成功率 (ASR) 来评估物理攻击的有效性。然而，我们注意到 mAP 和 ASR 都存在由于多重检测框和低于阈值的 IoU 匹配导致的高估问题。因此，这两个指标都无法反映出真实的攻击影响。

平均精度的均值。mAP 的定义是:

$$\begin{aligned} &\mathrm{P} = \tfrac{\mathrm{TP}}{\mathrm{TP+FP}}, \mathrm{R} = \tfrac{\mathrm{TP}}{\mathrm{TP+FN}}, \\ &\mathrm{mAP} = \tfrac{1}{n-1} \sum_{i=1}^{n-1} (\mathrm{R}_{i+1} - \mathrm{R}_i)\, \mathrm{P}_{i+1}, \end{aligned} \quad (2)$$

其中 TP、FP、TN 和 FN 分别表示真正例、假正例、真负例和假负例样本，P 表示精度，R 表示召回率。mAP 是通过计算 P-R 曲线下的面积来获得的。虽然被广泛使用，但 IoU 面临以下两个高估问题。

1) 多个检测框: 如图 1 (a) 所示，当单个目标被多个重叠的检测框检测到时，检测器无法将它们中的任何一个与 GT 匹配，因为任何预测框与 GT 框之间的 IoU 都不大于 0.5。似乎由于对抗性补丁，原始预测框被分解为多个检测框，导致多个 FP 且没有 TP。FP 的增加导致精度下降，而 TP 的损失则降低了精度和召回率，两者共同导致 mAP 下降。因此，这些情况夸大了攻击的有效性，无法准确反映它们对检测性能的真实影响。

2) 阈值以下的 IoU 匹配：如图 1 (b) 所示，当检测器仅为一个对象预测一个边框，但其与 GT 框的 IoU 低于阈值（通常为 0.5）时，该预测被视为 FP 而不是 TP，这同样导致 mAP 的下降。然而，不可忽视的是，这些较小的框仍然捕捉到了人物的重要语义信息（如身体的上半部分）。从自动驾驶车辆的角度来看，这样的检测仍将触发适当的响应，比如减速或停车以避免“较小的”行人。因此，尽管 mAP 下降了，但物体检测器的功能性未受影响，攻击并未真正危害安全。

总而言之，这两个高估问题说明了 mAP 的一个关键限制：它通过惩罚非关键边界框（减少 TP 并增加 FP）夸大了补丁攻击的影响。因此，仅依赖 mAP 进行数字和物理场景下的安全评估可能会导致对攻击性能的误导性结论。

攻击成功率。在有限数量的研究中，ASR 用于评估物理攻击，特别是在真实场景中。ASR 的定义如下。

$$\mathrm{ASR} = 1 - \frac{\mathrm{TP}}{\mathrm{TP}'}, \tag{3}$$

其中 TP' 表示 GT 样本的数量。ASR 衡量的是检测器未能产生有效检测的对象的比例，方程 1 。尽管 ASR 仅关注 TP，并不惩罚冗余或不准确的边界框，但它仍采用与 mAP 相同的 TP 定义。因此，ASR 可以被视为基于 mAP 的度量。如图 1 所示，ASR 也受到上述两个高估问题的影响：当一个人被一个或多个边界框检测到时，所有这些边界框与 GT 框的 IoU 都低于阈值，ASR 错误地将该情况视为成功攻击，因为预测的框中没有一个被计为 TP。然而，这样的检测仍可能保留足够的语义提示，并不能阻止自动驾驶车辆减速或停止，因此不构成潜在的安全威胁。我们认为这类情况不应被视为成功攻击。

B. 对抗性损失

基于得分的损失。回忆一下，检测器输出两种得分：物体性得分 $\bar{o}$ 和分类得分 $\mathbf{s}$ 。降低任一得分都会增加在 NMS 阶段被丢弃的可能性，从而可能导致原始框被移除。因此，为了有效误导检测器，AdvPatch [**?**] 首次提出三种基于得分的对抗损失，其中最小化最大物体性得分的表现最好。此策略自那时起已被广泛采用，如表 I 所示。

基于 IoU 的损失。近年来，IoU 在一些研究中开始受到有限的关注，尽管尚未被详细探讨，却已被融入到了对抗损失的设计中。Depatch [**?**] 使用目标性得分和 IoU 的加权和来选择最佳边界框，但仅优化对应于该框的目标性得分，IoU 未被作为直接优化目标使用。为了确保补丁在不同距离下的有效性，FDA [**?**] 优化了分类得分和 IoU 结合的平均值。尽管平均化可能与其使用的期望变换相匹配，但与优化最大值相比，平均化带来了显著更高的计算开销。在光学遥感领域，PA-ORSI [**?**] 设计了基于目标性得分和最大 IoU 之和的对抗性补丁，以避免梯度泛滥。然而，两个项是独立处理的，未能捕捉到它们的联合效应。例如，具有最高 IoU 的边界框可能仍然具有低目标性得分，削弱了攻击效果。总体来看，现有方法未能充分利用 IoU 和置信度分数的联合影响，而这对于提高攻击可转移性至关重要。

根据上述分析，我们得出结论：依赖公式 1 来衡量攻击性能无法真实反映现实场景中的实际风险。为了解决这一限制，我们提出了实际攻击成功率（PASR），该新指标能更有效地捕捉对抗性补丁攻击的实际影响。

正式地，给定一个图像 $\mathcal{G} = \left\{\mathbf{b}_1^{\mathrm{GT}}, \mathbf{b}_2^{\mathrm{GT}}, ...\right\}$ 的 GT 框和检测器 $\mathcal{D} = \left\{\mathbf{b}_1^{d}, \mathbf{b}_2^{d}, ...\right\}$ 预测的边界框，我们首先定义如果预测框中没有任何一个与其相交，则一个 GT 框 $\mathbf{b}_i^{\mathrm{GT}}$ 被认为是成功被攻击。对象级攻击成功指示函数可以表示为：

$$A\left(\mathbf{b}_i^{\mathrm{GT}}\right) = \begin{cases} 1, \text{if } \forall \mathbf{b}_j^{d} \in \mathcal{D}, \mathrm{IoU}\left(\mathbf{b}_i^{\mathrm{GT}}, \mathbf{b}_j^{d}\right) = 0, \\ 0, \text{otherwise.} \end{cases} \tag{4}$$

这种严格的定义确保了预测框与 GT 框之间即使是微小的重叠，也会使攻击不被视为成功，这与直觉相符：如果检测器仍然能感知到物体的任何部分，那么从安全性角度来说，攻击就没有成功。然后我们将评估从物体层面（mAP, ASR）提升到图像层面，因为在自动驾驶中，即使存在一个未检测到的行人也足以引发安全问题。我们定义，如果图像中的任何 GT 框被成功攻击，那么这幅图像就被认为成功遭到攻击：

$$\tilde{A}\left(\boldsymbol{x}\right) = \begin{cases} 1, \text{if } \exists \mathbf{b}_i^{\mathrm{GT}} \in \mathcal{G}, A\left(\mathbf{b}_i^{\mathrm{GT}}\right) = 1, \\ 0, \text{otherwise.} \end{cases} \tag{5}$$

最后，给定一个图像数据集，我们首先筛选出检测器无法输出任何人物框的图像。假设这产生了一组剩余的 T 张图像。对每个人物框覆盖对抗性 patch 以生成对抗图像 $\left\{\boldsymbol{x}_{adv}^1, \boldsymbol{x}_{adv}^2, ..., \boldsymbol{x}_{adv}^T\right\}$ 。然后，我们计算 PASR，即这些被成功攻击的对抗图像所占的比例：

$$\mathrm{PASR} = \frac{1}{T}\sum\nolimits_{t=1}^{T} \tilde{A}\left(\boldsymbol{x}_{adv}^t\right). \tag{6}$$

总之，PASR 在图像层面上操作，并通过要求最严格的攻击成功标准来提供以安全为导向的评估。这尤其适用于物理攻击场景，在这些场景中，攻击性能通常通过视频序列中攻击有效的帧数来评估，而不是每次检测的对象级别准确性。有关更多理由和监管背景，请参阅附录。

单独使用现有的基于分数的策略只能降低检测框的置信度，使得检测器不太可能为具有大 IoU 的框输出高置信度分数。结果是，具有较低 IoU 和中等置信度的框可能在 NMS 后成为最终输出。这导致 mAP 的某种下降，但无法减少 PASR。

我们进行了一项试验，仅使用 IoU 作为对抗损失，结果表明即使在单个图像上也无法成功攻击。这是因为检测结果是由边界框的位置及其相关的置信得分共同决定的。尽管仅优化 IoU 可以将其降低到大约 0.6，但这并不能充分降低置信得分，因此目标仍然可检测到。相反，仅优化置信得分可以在一定程度上取得攻击成功，因为当高 IoU 框的置信得分下降到置信阈值以下时，检测框仍然存在但不会被输出。实际上，现有攻击利用了目标检测中分类和定位组件之间的不平衡，从而在基于 mAP 的指标下产生有效的攻击。

LCSL 设计。基于这些分析，我们直观地认为将 IoU 引入基于得分的对抗损失中能够同时抑制 IoU 和置信得分，从而提高训练补丁在 PASR 方面的效果，并增强其泛化能力。具体而言，我们首先选择前 T_k 个具有最高目标性和分类得分乘积的预测结果，为了简化起见，这里省略了对抗图像 $\boldsymbol{x}_{adv}^i$ ，$\bar{s}_j^0$ 表示类别 0（人）的分类得分。对于这些框中的每一个，我们计算其与 $\boldsymbol{b}_{\max}^{\mathrm{GT}}$ 的 IoU，即所有最终 GT 框中面积最大者。这是由于在自动驾驶场景中，更大的人框通常意味着离车辆更近，因此漏检这样的目标带来更高的安全风险。此外，由于我们的目标是让检测器输出中至少有一个人消失，攻击最大的 GT 框既足够且计算