Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1433-1443
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146134

Hans汉斯

解释性陷阱：目标检测模型中性能与解释可靠性研究

叶　阳

西华大学汽车测控与安全四川省重点实验室，四川　成都

收稿日期：2025年9月30日；录用日期：2025年11月6日；发布日期：2025年11月17日

摘　要

目标检测模型在自动驾驶等安全关键领域广泛应用，其决策过程的可解释性对系统可靠性至关重要。当前研究重点主要集中在如何提高模型性能上，对于模型的可解释性以及解释质量与决策可靠性之间的内在关联关注较少。鉴于这一问题，本研究系统探究了目标检测模型性能与解释可靠性之间的关系。基于KITTI数据集，采用三种Faster R-CNN变体在简单、中等和困难三类场景下进行实验，通过Grad-CAM和SmoothGrad-IG两种解释方法，结合新提出的Energy-based Pointing Game和Performance-Explanation Correlation指标进行量化评估。结果表明：性能最佳的ResNet50_Epoch20在简单场景中PEC值为−0.189，揭示了“解释性陷阱”现象——高置信度预测反而伴随低质量解释；Grad-CAM生成的热力图分散且模型间差异显著，而SmoothGrad-IG产生的热力图高度收敛且模型间一致性高；简单场景中的解释可靠性问题最为严重，与直觉相反。

关键词

Faster R-CNN，可解释性，解释性陷阱，Grad-CAM，SmoothGrad-IG

Explainability Traps: A Study on the Relationship between Object Detection Model Performance and Explanation Reliability

Yang Ye

Vehicle Measurement, Control and Safety Key Laboratory of Sichuan Province, Xihua University, Chengdu Sichuan

Received: September 30, 2025; accepted: November 6, 2025; published: November 17, 2025

文章引用: 叶阳. 解释性陷阱: 目标检测模型中性能与解释可靠性研究[J]. 人工智能与机器人研究, 2025, 14(6): 1433-1443. DOI: 10.12677/airr.2025.146134

Abstract

Target detection model is widely used in key safety fields such as automatic driving, and the interpretability of its decision-making process is very important for system reliability. The current research focuses on how to improve the performance of the model, and pays less attention to the interpretability of the model and the internal relationship between the interpretation quality and decision reliability. In view of this problem, this study systematically explores the relationship between the performance of target detection model and interpretation reliability. Based on the Kitti data set, three fast R-CNN variants were used to carry out experiments in simple, medium and difficult scenarios. The two interpretation methods of grad cam and smooth grad Ig were combined with the newly proposed energy based pointing game and performance explanation correlation indicators for quantitative evaluation. The results show that the PEC value of resnet50_poch20 with the best performance is −0.189 in simple scenarios, which reveals the phenomenon of "interpretative trap"—high confidence prediction is accompanied by low quality interpretation; The thermal maps generated by grad cam are scattered and have significant differences among models, while the thermal maps generated by smooth grad IG are highly convergent and have high consistency among models; The problem of interpretation reliability is the most serious in simple scenarios, which is contrary to intuition.

Keywords

Faster R-CNN, Interpretability, Interpretative Traps, Grad-CAM, SmoothGrad-IG

1. 引言

随着深度学习技术的飞速发展，目标检测模型已在自动驾驶、智能交通监控等安全关键领域得到广泛应用[1]。在这些应用中，模型不仅需要具备高精度的检测能力，更需要提供可解释、可信赖的决策过程，以建立用户信任并确保系统安全[2]。

目标检测作为计算机视觉的核心任务，其模型架构主要可分为单阶段(one-stage)和双阶段(two-stage)两大类[3]。双阶段模型首先生成区域建议(Region Proposals)，再对建议区域进行分类和边界框精修，代表了早期目标检测的主流范式。R-CNN [4]系列是该类模型的典型代表，从 2014 年的 R-CNN 到 2015 年的 Faster R-CNN [5]，通过引入区域建议网络(RPN)实现了端到端训练，大幅提升了检测效率。后续发展包括 Mask R-CNN [6] (2017)增加了实例分割分支，Cascade R-CNN [7] (2018)采用级联检测头提高精度，以及 Libra R-CNN [8] (2019)通过均衡采样和特征金字塔改进检测质量。相比之下，单阶段模型直接预测目标类别和边界框，无需区域建议步骤，具有更高的推理速度。代表性工作包括 YOLO [9]系列(2016~2023)、SSD [10] (2016)、RetinaNet [11] (2017)引入 Focal Loss 解决样本不平衡问题，以及 FCOS [12] (2019)采用无锚框设计。

与此同时，可解释人工智能(XAI)方法也取得较大发展，可解释人工智能(XAI)方法根据解释时机可分为内在解释(Intrinsic Explanations)和事后解释(Post-hoc Explanations)两大类[13]。内在解释方法在模型构建阶段就融入可解释性设计，使模型本身具备解释能力，包括决策树和基于规则模型等本身就具有可解释性的模型。相比之下，事后解释方法在已训练好的黑盒模型上应用，不改变原始模型结构，而是通过外部手段提供解释，主要包括梯度类方法(如 Grad-CAM [14]、Score-CAM [15])、扰动类方法(LIME [16]、

SHAP [17])以及集成梯度方法(如 SmoothGrad-IG [18] [19])等。

当前的目标检测模型评估主要关注性能指标(如 mAP) [20]。对于解释方法现有研究主要聚焦于解释方法本身的改进，如 Score-CAM、Layer-CAM [21]等梯度类方法因其计算效率高、可视化效果直观而备受青睐。然而，这些研究大多停留在“如何解释”的层面，缺乏对“解释是否可靠”的系统评估。尽管已有研究表明解释方法本身存在局限性(如梯度类方法对输入扰动敏感)，但解释质量与模型实际决策可靠性之间的关系尚未得到充分探究。特别是在安全关键应用中，如果模型在高置信度预测时提供的解释质量低下，将导致严重的安全隐患，这一问题目前在目标检测领域尚未引起足够重视[22]。

针对上述问题，本研究探究了目标检测模型性能与解释可靠性之间的关系，提出“解释性陷阱”的概念——即模型在高置信度预测时反而提供低质量解释的现象。为此，本研究设计了 Energy-based Pointing Game (EBPG)和 Performance-Explanation Correlation (PEC)两个量化指标，前者评估解释与目标对象的空间对应性，后者衡量检测分数与解释质量之间的相关性。

本研究采用 Grad-CAM 和 SmoothGrad-IG 两种解释方法，对三种不同网络深度及训练周期的 Faster R-CNN 变体(ResNet18_Epoch5、ResNet50_Epoch5 和 ResNet50_Epoch20)进行对比分析。基于 KITTI 官方难度标准，本研究将验证集图像划分为简单、中等和困难三类场景，通过 EBPG 和 PEC 指标系统评估模型在不同场景下的解释质量。特别地，本研究深入分析了 Grad-CAM 与 SmoothGrad-IG 在实例区分能力上的差异，揭示了 Grad-CAM 在多目标场景中注意力分散的局限性，以及 SmoothGrad-IG 在揭示“解释性陷阱”现象中的优越性。

本研究的贡献在于：(1) 在目标检测领域发现并量化了“解释性陷阱”现象；(2) 提出 PEC 指标作为评估解释可靠性的新标准；(3) 揭示了场景难度、模型复杂度与解释可靠性之间的复杂关系。

2. 实验设置

2.1. 场景划分

为系统探究目标检测模型在不同复杂度场景下的解释可靠性差异，本研究基于 KITTI 官方难度标准对验证集进行科学划分。选择场景划分策略的理论依据在于：目标检测任务中，场景复杂度直接影响模型的决策机制和注意力分配模式，而现有研究表明，模型在不同难度场景中可能表现出截然不同的解释质量特性。特别地，针对“解释性陷阱”现象的研究需要区分简单与复杂场景，因为模型在简单场景中可能过度依赖数据集偏差而非目标真实特征，这一假设需要通过场景特异性分析予以验证。

本研究采用 KITTI 2D 目标检测数据集进行实验验证，该数据集是自动驾驶领域最具代表性的基准数据集之一，包含 7481 张训练图像和 7518 张测试图像，聚焦于车辆、行人和骑行者等目标的检测任务。为确保实验结果的可靠性和可比性，本研究严格遵循 KITTI 官方难度标准，基于目标物理特性(高度、遮挡、截断)将 1497 张验证集图像划分为三个难度级别：

① 简单场景(simple)：包含 351 张图像(23.4%)，主要特征为目标满足“高度 ＞40 像素、遮挡程度为 0、截断 ＜0.15”条件，且困难目标比例 ≤15%。此类场景中目标平均数量为 1.6 个，困难目标比例为 0%。

② 中等场景(moderate)：包含 674 张图像(45.0%)，特征为包含满足“高度 ＞25 像素、遮挡程度 ≤1、截断 ＜0.30”条件的目标，且困难目标比例 ≤15%。此类场景中目标平均数量为 4.7 个，困难目标比例为 2%，中等目标比例为 54%。

③ 困难场景(hard)：包含 472 张图像(31.5%)，特征为困难目标(满足“高度 ＞25 像素、遮挡程度 ≤2、截断 ＜0.50”条件)比例 ＞15%。此类场景中目标平均数量为 7.8 个，困难目标比例高达 34%。

这种划分方法具有以下优势：(1) 符合 KITTI 官方评估标准，确保结果可比性；(2) 从目标可见性角度客观衡量场景难度，避免主观判断；(3) 不同难度级别间存在明确的物理界限，便于结果分析。

2.2. 模型构建

为系统探究目标检测模型性能与解释可靠性之间的关系，本研究选取 Faster R-CNN 作为基础框架，并设计三种变体进行对比分析。选择 Faster R-CNN 的原因在于：(1) 作为两阶段目标检测算法的代表，其在 KITTI 等自动驾驶数据集上具有广泛应用和良好性能；(2) 其模块化设计便于集成不同骨干网络，适合进行可解释性研究；(3) 作为工业界和学术界广泛采用的标准模型，其研究结果具有较高的实用价值。

本研究设计的模型变体旨在系统考察模型复杂度与训练充分性对解释可靠性的影响：

① ResNet18_Epoch5：采用轻量级 ResNet18 作为骨干网络，训练 5 个周期。

② ResNet50_Epoch5：采用标准 ResNet50 作为骨干网络，训练 5 个周期。

③ ResNet50_Epoch20：采用 ResNet50 作为骨干网络，但延长训练至 20 个周期。

所有模型均采用完全相同的训练配置以确保对比的公平性：初始学习率 0.001，批量大小 2，优化器为 SGD (动量 0.9，权重衰减 0.0001)。区域建议网络(RPN)与检测头的损失权重比设为 1:1，锚框比例与 KITTI 官方推荐设置保持一致。训练过程中采用早停策略，验证集性能连续 5 个周期未提升则终止训练。

2.3. 解释方法

2.3.1. Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping)作为梯度类解释方法的代表，通过计算目标类别对卷积特征图的梯度来生成类别判别性热力图。其核心思想是识别对特定类别预测有贡献的图像区域，计算公式如下：

$$L_{Grad-CAM}^{c}\left(x,y\right)=ReLU\left(\sum_{k}\alpha_{k}^{c}A^{k}\left(x,y\right)\right) \tag{1}$$

$$\alpha_{k}^{c}=\frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A^{k}\left(i,j\right)} \tag{2}$$

式中，A^{k} 表示第 k 个卷积特征图，α 是类别 c 对应的权重(通过对特征图梯度进行全局平均池化获得)，Z 为归一化因子。由于高层特征层捕获了更丰富的语义信息，能够反映模型对目标对象的高级理解，同时其较大的感受野有助于捕获目标对象的全局上下文，因此本研究选择 Faster R-CNN 模型 backbone 的 layer4 [−1]作为目标层。

2.3.2. SmoothGrad-IG

SmoothGrad-IG 是集成梯度(Integrated Gradients, IG)与 SmoothGrad 的结合方法，通过路径积分和噪声平均机制提供更稳定、更精确的解释。其理论基础源于以下两个关键公式：

$$IG_{i}\left(x\right)=\left(x_{i}-x_{i}'\right)\times\int_{\alpha=0}^{1}\frac{\partial F\left(x'+\alpha\left(x-x'\right)\right)}{\partial x_{i}}d \tag{3}$$

$$SGIG_{i}\left(x\right)=E_{\nu\sim\mathbb{N}\left(0,\sigma^{2}\right)}\left[IG_{i}\left(x+\nu\right)\right]\approx\frac{1}{N}\sum_{j=1}^{N}IG_{i}\left(x+\nu_{j}\right) \tag{4}$$

式中，x 为输入图像，x' 为基线输入(本研究采用全黑图像)，ν 为服从正态分布 N (0, σ^{2})的噪声，σ 为标准差(本研究设置为 0.10)，N 为噪声样本数(本研究取 10)。为确保路径积分的精确性，积分步数设为 25。

2.4. 量化指标

2.4.1. EBPG 指标设计

为定量评估目标检测模型的可解释性质量，本文提出 Energy-based Pointing Game (EBPG)评估指标，

其计算公式如下：

$$\mathrm{EBPG}=\frac{\sum_{(x,y)\in \mathrm{bbox}} \mathrm{heatmap}(x,y)}{\sum_{(x,y)\in \mathrm{image}} \mathrm{heatmap}(x,y)} \tag{5}$$

式中，heatmap (x,y)表示位置(x,y)处的显著图能量值，bbox 为目标边界框区域，image 表示全图区域。该指标衡量了模型注意力机制与目标对象的空间对应性，值域范围为[0, 1]，值越高表示模型关注区域与目标对象的一致性越好，即模型关注区域与目标对象高度重合，解释质量越高；值越接近 0，表示模型注意力主要集中在背景区域，解释质量低下。该指标通过计算目标边界框内热力值占全图热力值的比例，量化模型注意力与目标对象的匹配程度。

EBPG 作为 Pointing Game [23]的量化扩展，旨在客观评估目标检测模型的注意力机制与目标对象的空间一致性。传统 Pointing Game 仅通过二元判断(是否指向目标)评估解释质量，存在粒度粗糙的局限。

与传统 Pointing Game 相比，EBPG 具有以下三点优势：(1) 提供连续量化评估，避免二元判断的粗糙性；(2) 适用于多目标场景，可针对每个检测实例单独计算；(3) 与视觉解释直观对应，便于结果解读。此外，通过[min (EBPG), max (EBPG)]评估模型解释的稳定性，窄范围表示解释质量稳定，宽范围则表明解释可靠性波动大。

2.4.2. PEC 指标设计

为系统评估目标检测模型的决策可靠性，本研究提出 Performance-Explanation Correlation (PEC)指标，其核心思想是通过非参数相关性分析量化模型预测置信度与解释质量之间的内在关联。

考虑到 Spearman 秩相关系数对非线性关系敏感，且不依赖于数据分布假设，更适合评估解释质量与置信度的复杂关联。本研究中 PEC 采用 Spearman 秩相关系数计算检测分数与 EBPG 值之间的相关性。

$$PEC=\rho=spearmanr\left(detection\_scores, ebpg\_scores\right) \tag{6}$$

$$\rho=1-\frac{6\sum d_i^2}{n\left(n^2-1\right)} \tag{7}$$

其中，detection_scores 表示模型对检测目标的置信度分数，ebpg_scores 为对应预测的 EBPG 值，di 表示第 i 个检测实例的秩次差，即检测分数 detection_scores 与 ebpg_scores 的排序位置之差；n 为参与分析的有效检测实例总数。该公式通过比较两个变量排序的一致性来衡量模型置信度与解释质量之间的单调相关关系。

PEC 值的解释实践意义：正值表示模型在高置信度预测时提供了高质量的解释，决策过程可靠；零值表明检测分数与解释质量无显著相关性；而负值则揭示了“解释性陷阱”现象——模型越自信，其决策依据反而越不可靠。当 PEC 为负值时，即使模型性能指标(mAP)优异，其决策过程也可能依赖于数据集偏差或背景线索。

3. 实验结果

3.1. 性能对比

表 1 展示了三种模型在 KITTI 数据集三个难度级别上的 mAP@0.5 结果。总体而言，ResNet50_Epoch20 在所有场景中均表现最佳，ResNet50 次之，ResNet18 性能最低。这符合预期，因为更深的网络和更长的训练周期通常能带来更高的检测精度。

由表可知，所有模型在简单场景中的性能均显著高于困难场景，这与 KITTI 数据集的特性一致——简单场景中的目标更大、更清晰，更容易被检测到。然而，性能优势并不一定意味着决策过程更可靠，这需要通过解释质量评估进一步验证。

Table 1. Model performance comparison
表 1. 模型性能对比

场景	ResNet18_Epoch5	ResNet50_Epoch5	ResNet50_Epoch20
简单	0.707	0.713	0.731
中等	0.615	0.643	0.670
困难	0.594	0.631	0.655

3.2. Grad-CAM

表 2 展示了采用 Grad-CAM 方法对三种模型在不同场景下的解释质量评估结果。分析表明：

① 解释质量与场景难度的关系：所有模型在中等和困难场景中均表现出较高的解释质量(EBPG > 0.35)，其中 ResNet50_Epoch20 在困难场景中达到最高值 0.497。这一结果与表 1 中性能表现一致，表明在更具挑战性的场景中，模型倾向于关注目标对象的真实特征。

② 性能与解释可靠性的初步脱节现象：值得注意的是，ResNet50_Epoch20 在困难场景中虽然 EBPG 值最高(0.497)，但其 PEC 值接近于零(−0.008)，表明该模型的高置信度预测与解释质量之间几乎没有相关性。这一发现提示了“高性能不一定意味着高解释可靠性”的现象。

③ 简单场景中的潜在风险：在简单场景中，尽管所有模型的 EBPG 值均较高(>0.38)，但 ResNet50_Epoch20 的 PEC 值(0.285)显著低于其他模型。这表明，即使是性能最佳的模型，在简单场景中也可能存在解释可靠性不足的问题，暗示了模型可能依赖于背景线索而非目标特征进行决策。

Table 2. Grad-CAM evaluation results
表 2. Grad-CAM 评估结果

模型	场景	平均 EBPG	PEC	样本数
ResNet18_Epoch5	简单	0.380	0.320	100
ResNet18_Epoch5	中等	0.354	0.597	100
ResNet18_Epoch5	困难	0.363	0.662	100
ResNet50_Epoch5	简单	0.400	0.377	100
ResNet50_Epoch5	中等	0.426	0.333	100
ResNet50_Epoch5	困难	0.422	0.584	100
ResNet50_Epoch20	简单	0.422	0.285	100
ResNet50_Epoch20	中等	0.470	0.424	100
ResNet50_Epoch20	困难	0.497	-0.008	100

图 1 展示了使用 Grad-CAM 方法对三种模型(ResNet18_Epoch5、ResNet50_Epoch5 和 ResNet50_Epoch20)在 car、pedestrian 和 cyclist 三类目标上的解释结果。观察发现，Grad-CAM 生成的热力图呈现出明显的分散特性：热力区域不仅覆盖目标对象本身，还广泛分布于同类目标的周边区域。例如，在行人检测案例中，当模型识别出特定行人时，热力图同时高亮了图像中其他行人区域，表明模型注意力被分散至同类对象的多个实例。值得注意的是，当模型正确检测并分类特定目标时(如图中 cyclist 类别所示)，Grad-CAM 生成的热力图却显著聚焦于图像中其他同类目标，而非当前检测实例。这一现象在三种 Faster R-CNN 变体(ResNet18_Epoch5、ResNet50_Epoch5 和 ResNet50_Epoch20)上均一致出现，对于以上现象，本研究认为 Grad-CAM 其设计初衷为图像分类任务，导致其关注机制本质上是“类别级”而非“实例级”。在多目标场景中，Grad-CAM 倾向于将注意力分散至图像中所有同类对象，无法有效区分同一类别中的不同实例。

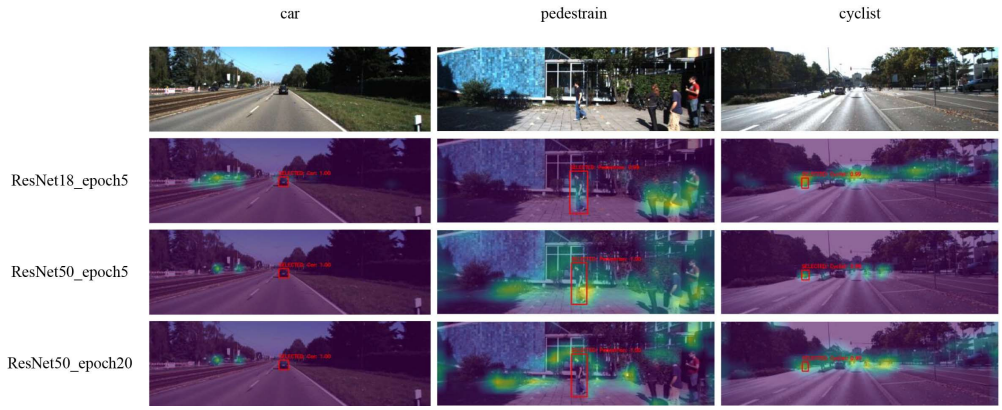

Figure 1. Grad-CAM method explains the results of three models on three types of targets
图 1. Grad-CAM 方法对三种模型在三类目标上的解释结果展示

此外，不同模型间的 Grad-CAM 热力图表现出显著差异。ResNet18_Epoch5 的热力图相对集中于目标区域，而 ResNet50_Epoch20 则显示出更广泛的热力分布，尤其是在简单场景中。这种模型间差异与表 2 中 Grad-CAM 评估结果一致，特别是与 ResNet50_Epoch20 在简单场景中 PEC 值相对较低(0.285)的现象相呼应，暗示了该模型在高置信度预测时可能存在解释可靠性问题。

3.3. SmoothGrad-IG

Table 3. SmoothGrad-IG evaluation results
表 3. SmoothGrad-IG 评估结果

模型	场景	平均 EBPG	PEC	样本数	Map@0.5	EBPG 范围
ResNet18_Epoch5	简单	0.310	0.007	100	0.707	0.000~0.682
ResNet18_Epoch5	中等	0.236	0.129	100	0.615	0.000~0.828
ResNet18_Epoch5	困难	0.253	0.148	100	0.594	0.000~0.783
ResNet50_Epoch5	简单	0.254	−0.090	100	0.713	0.000~0.672
ResNet50_Epoch5	中等	0.102	−0.145	100	0.643	0.000~0.759
ResNet50_Epoch5	困难	0.112	−0.210	100	0.631	0.000~0.637
ResNet50_Epoch20	简单	0.350	−0.189	100	0.731	0.000~0.782
ResNet50_Epoch20	中等	0.195	0.327	100	0.670	0.000~0.722
ResNet50_Epoch20	困难	0.233	0.146	100	0.655	0.000~0.738

表 3 采用更为稳健的 SmoothGrad-IG 方法进行评估，揭示了更为深刻的发现：

① “解释性陷阱”的明确证据：所有模型在简单场景中的 PEC 值均为负值或接近零，特别是性能最佳的 ResNet50_Epoch20，其 PEC 值达到−0.189，表明在简单场景中，模型的高置信度预测与解释质量呈显著负相关。这一现象可被称为“解释性陷阱”——模型越自信，其决策依据反而越不可靠。对此，本研究认为高性能模型由于深层结构和充分训练，可能发展出多种决策路径：有时基于目标对象的真实特征进行决策(高 EBPG 值)，有时则依赖背景线索或数据集偏差进行决策(低 EBPG 值)。特别值得注意的是，ResNet50_Epoch20 在简单场景中 EBPG 范围最宽(0.000~0.782)，与其负 PEC 值形成鲜明对比，表明该模型在某些情况下能提供高质量解释，而在其他情况下几乎不关注目标对象。与直觉相悖的是，简单场景中的“解释性陷阱”最为严重，这可能是因为在目标显著、数量少的场景中，模型更容易依赖背景线索做出高置信度预测。相比之下，性能较低的 ResNet18_Epoch5 在所有场景中均表现出稳定的 PEC 值(接近零或正值)，表明浅层网络虽然性能较低，但提供更一致的决策依据。

② 解释方法对评估结果的影响：ResNet50_Epoch20 在困难场景中的 PEC 值在两种方法下呈现显著差异：Grad-CAM 评估结果为−0.008，而 SmoothGrad-IG 评估结果为 0.146。这一差异表明，Grad-CAM 可能低估了困难场景中的解释质量问题，而 SmoothGrad-IG 通过噪声平均提供了更为敏感和可靠的评估结果。

③ 模型结构与训练周期的影响：ResNet18 在所有场景中均表现出最稳定的 PEC 值，特别是在简单场景中 PEC 接近于零(0.007)，远优于其他两种模型。这表明较浅的网络结构可能提供更一致的决策可靠性，尽管其绝对性能较低。同时，增加训练周期(ResNet50_Epoch20 相比 ResNet50_Epoch5)并未改善甚至可能损害解释可靠性。

④ 解释质量的波动性分析：EBPG 范围分析提供了关于解释稳定性的关键信息。例如，ResNet50_Epoch20 在简单场景中的 EBPG 范围为 0.000~0.782，表明其解释质量波动极大，可靠性较低。相比之下，ResNet18 在相同场景中的 EBPG 范围为 0.000~0.682，波动相对较小，解释更为稳定。性能最佳的 ResNet50_Epoch20 模型在简单场景中表现出最宽的 EBPG 范围(0.000~0.782)，与其负 PEC 值(−0.189)形成鲜明对比。对此，本研究认为，该模型在简单场景况下能提供高质量解释(EBPG = 0.782)，而在其他情况下几乎不关注目标对象(EBPG = 0.000)，且这种波动与置信度呈负相关——模型越自信，其决策依据反而越不可靠。相比之下，性能较低的 ResNet18_Epoch5 在相同场景中 EBPG 范围较窄(0.000~0.682)，PEC 值接近于零(0.007)，表明其解释质量虽不高但更为稳定。同时，简单场景中 EBPG 范围普遍宽于困难场景，这与直觉完全相悖。简单场景中目标更大、遮挡更少、数量较少，本应更容易提供稳定解释，却表现出最宽的 EBPG 范围。这一现象同样指向在简单场景中，模型更容易依赖背景线索或数据集偏差做出决策，导致解释质量波动极大。而在困难场景中，由于目标不明显、遮挡严重，模型被迫关注目标的真实特征，反而表现出相对一致的决策依据。

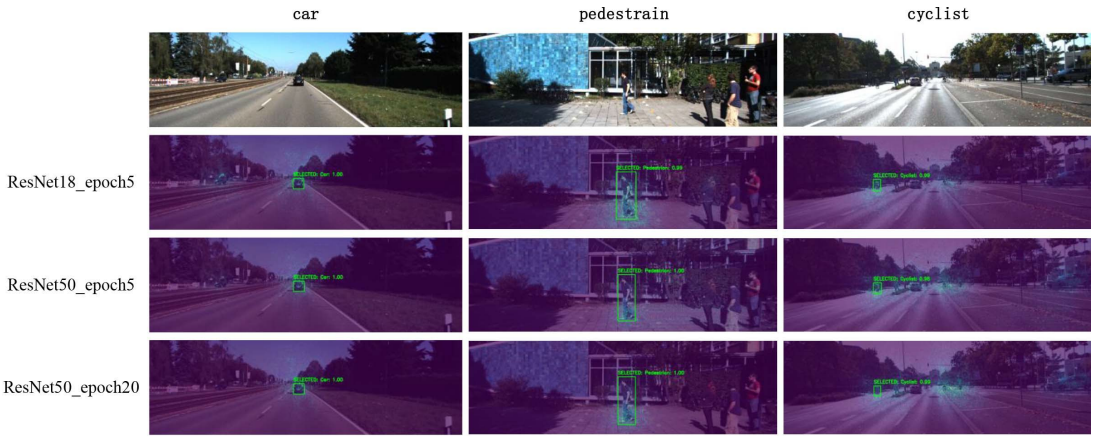

Figure 2. The SmoothGrad-IG method shows the explanation results of three models on three types of targets
图 2. SmoothGrad-IG 方法对三种模型在三类目标上的解释结果展示

图 2 展示了采用 SmoothGrad-IG 方法生成的热力图结果。相比于 Grad-CAM 方法，SmoothGrad-IG 通过路径积分与噪声平均机制，能够更精确地定位对特定检测实例决策真正重要的局部区域，其热力图高度集中于当前检测目标周围，显著提高了实例级解释的准确性。SmoothGrad-IG 产生的热力图表现出高度的收敛性：高热力值区域主要集中于目标对象的精确边界内，且对同类目标的其他实例干扰较小。在行人检测案例中，热力图精准聚焦于被识别行人的身体区域，而对附近其他行人的关注度显著降低。

此外，三种模型(ResNet18_Epoch5、ResNet50_Epoch5 和 ResNet50_Epoch20)使用 SmoothGrad-IG 生

成的热力图在空间分布上表现出高度一致性，即使在模型架构和训练周期存在差异的情况下。这种一致性与表 3 中 SmoothGrad-IG 的 EBPG 范围分析结果相吻合，特别是 ResNet50_Epoch20 在简单场景中 EBPG 范围为 0.000~0.782，表明其解释质量波动较大，但热力图的空间分布模式仍保持相对稳定。

3.4. 极端案例分析

为验证对解释性陷阱形成机制的猜想，本研究对 ResNet50_Epoch20 的 EBPG 极端值案例进行了可视化分析。如图 3 所示，低 EBPG 值(0~0.1)的案例多为简单目标(大且无遮挡)，此时模型注意力集中在目标边界框之外，但预测结果仍然正确；而高 EBPG 值(≥0.6)的案例多为困难目标(小或遮挡严重)，如图 4 所示，此时模型注意力集中在目标边界框内。

从数据拟合的角度看，深度神经网络作为高维空间中的多解函数拟合器，其优化过程会优先选择计算上更简单的解。在 KITTI 数据集中，存在明显的场景结构化特征(如“目标车辆旁总是存在其他车辆”)，这些统计规律构成了模型可以利用的“捷径”。对于简单目标，这些背景线索足以支持正确预测，模型因此发展出“捷径路径”——依赖背景线索进行决策，导致低 EBPG 值(0~0.1)但预测正确。而对于困难目标(小或遮挡严重)，背景线索不足以支持正确预测，模型被迫学习“正确路径”——关注目标真实特征，导致高 EBPG 值(≥0.6)且预测正确。

这一结果解释了 PEC 负值的形成：在简单场景中，模型对简单目标的高置信度预测往往基于“捷径”路径(解释质量低)，而低置信度预测可能偶然使用了“正确”路径(解释质量高)，导致置信度与解释质量呈负相关。相比之下，在困难场景中，模型必须关注目标特征才能正确预测，因此高置信度预测通常伴随着高质量解释，形成正相关。表 3 数据显示，ResNet50_Epoch20 在简单场景中 PEC 值为−0.189，而在困难场景中 PEC 值为 0.146，这一数据对比证实了上述论述。

Figure 3. Case display of extremely low score of SmoothGrad IG method
图 3. SmoothGrad-IG 方法极端低分案例展示

高性能模型(如 ResNet50_Epoch20)由于其更大的模型容量，能够同时学习并存储多种决策路径。表 3 数据显示，ResNet50_Epoch20 在简单场景中 EBPG 范围最宽(0.000~0.782)，表明其在不同简单目标上采用了不同的决策策略：有时使用“捷径路径”，有时使用“正确路径”。更长的训练周期(20 个周期 vs 5 个周期)使模型更充分地学习了数据集中的统计规律，强化了“捷径学习”行为，导致简单场景中 PEC 值从 ResNet50_Epoch5 的−0.090 降至 ResNet50_Epoch20 的−0.189。

Figure 4. Case display of extremely high score of SmoothGrad IG method
图 4. SmoothGrad-IG 方法极端高分案例展示

这一数据拟合机制解释了为何简单场景中的“解释性陷阱”最为严重：模型对“捷径路径”的预测往往更自信(高置信度)，但解释质量低；而对“正确路径”的预测可能置信度较低，但解释质量高。相比之下，浅层网络(ResNet18)由于容量有限，难以充分学习复杂的“捷径”，因此在所有场景中 PEC 值稳定且接近零(简单场景：0.007)，解释质量虽不高但更为一致。

4. 结论

本研究通过在 KITTI 数据集上对三种 Faster R-CNN 变体的对比分析，揭示了“解释性陷阱”现象的存在。研究采用 Grad-CAM 和 SmoothGrad-IG 两种解释方法，结合新提出的 EBPG 和 PEC 量化指标，对模型在不同难度场景下的解释质量进行了全面评估。主要有以下结论。

评估方法的选择对揭示模型解释可靠性具有决定性影响。Grad-CAM 由于其设计原理(基于全局平均池化梯度和上采样操作)，倾向于生成分散的注意力图，反映了模型对图像中同类目标的整体关注；而 SmoothGrad-IG 通过积分路径和噪声平均机制，能够更精确地定位对特定目标预测真正重要的局部区域。这种差异导致 SmoothGrad-IG 比 Grad-CAM 更能揭示模型解释中的潜在问题，特别是在识别“解释性陷阱”方面更为敏感。

场景难度与解释可靠性之间存在反直觉的非线性关系：简单场景中的解释可靠性问题最为严重。高置信度预测与解释质量呈显著负相关。这一现象可能源于模型在简单场景中过度依赖数据集偏差或背景线索，而在困难场景中被迫关注目标的真实特征。此外，较浅的网络结构(ResNet18)在所有场景中均表现出最稳定的 PEC 值，表明模型复杂度增加可能损害解释可靠性。

模型性能(mAP)与解释可靠性(PEC)之间存在明显的脱节现象。ResNet50_Epoch20 在所有场景中性能最佳(简单场景 mAP = 0.731)，但在简单场景中解释可靠性最差(PEC = −0.189)，这一发现对当前仅关注性能指标的模型评估范式提出了重要挑战。

参考文献

[1] Peng, Y. (2023) Deep Learning for 3D Object Detection and Tracking in Autonomous Driving: A Brief Survey.
[2] Mirzaie, M. and Rosenhahn, B. (2025) Interpretable Decision-Making for End-to-End Autonomous Driving.
[3] 董文轩, 梁宏涛, 刘国柱, 等. 深度卷积应用于目标检测算法综述[J]. 计算机科学与探索, 2022, 16(5): 1025-1042.

[4] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/cvpr.2014.81

[5] Ren, S., He, K., Girshick, R., *et al*. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149.

[6] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/iccv.2017.322

[7] Cai, Z. and Vasconcelos, N. (2018) Cascade R-CNN: Delving into High Quality Object Detection. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6154-6162. https://doi.org/10.1109/cvpr.2018.00644

[8] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W. and Lin, D. (2019) Libra R-CNN: Towards Balanced Learning for Object Detection. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 821-830. https://doi.org/10.1109/cvpr.2019.00091

[9] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/cvpr.2016.91

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., *et al*. (2016) SSD: Single Shot Multibox Detector. In: *Lecture Notes in Computer Science*, Springer, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[11] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/iccv.2017.324

[12] Tian, Z., Shen, C., Chen, H. and He, T. (2019) FCOS: Fully Convolutional One-Stage Object Detection. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 9626-9635. https://doi.org/10.1109/iccv.2019.00972

[13] Atakishiyev, S., Salameh, M. and Goebel, R. (2025) Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, 1-20. https://doi.org/10.1109/tits.2025.3574738

[14] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/iccv.2017.74

[15] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., *et al*. (2020) Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 14-19 June 2020, 111-119. https://doi.org/10.1109/cvprw50498.2020.00020

[16] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?". *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[17] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4768-4777.

[18] Sundararajan, M., Taly, A. and Yan, Q. (2016) Gradients of Counterfactuals. https://arxiv.org/abs/1611.02639

[19] Smilkov, D., Thorat, N., Kim, B., *et al*. (2017) Smooth Grad: Removing Noise by Adding Noise. https://arxiv.org/abs/1706.03825

[20] Badithela, A., Wongpiromsarn, T. and Murray, R.M. (2022) Evaluation Metrics for Object Detection for Autonomous Systems. https://arxiv.org/abs/2210.10298

[21] Jiang, P.T., Zhang, C.B., Hou, Q., *et al*. (2021) Layercam: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing*, **30**, 5875-5888.

[22] Li, X., Chen, Z., Zhang, J.M., Sarro, F., Zhang, Y. and Liu, X. (2025) Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems. *ACM Transactions on Software Engineering and Methodology*, **34**, 1-24. https://doi.org/10.1145/3702989

[23] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X. and Sclaroff, S. (2017) Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, **126**, 1084-1102. https://doi.org/10.1007/s11263-017-1059-x