3D 目标检测稳定性分析

王家宝 [1]　孟强 [2]　刘国超 [2]　颜柳江 [2]　王珂 [2]

程明明 [1,3]　侯淇彬 [1,3]

[1] 南开大学，计算机学院　[2] 卡尔动力

[3] 南开大学，深圳福田研究院

February 18, 2025

Abstract

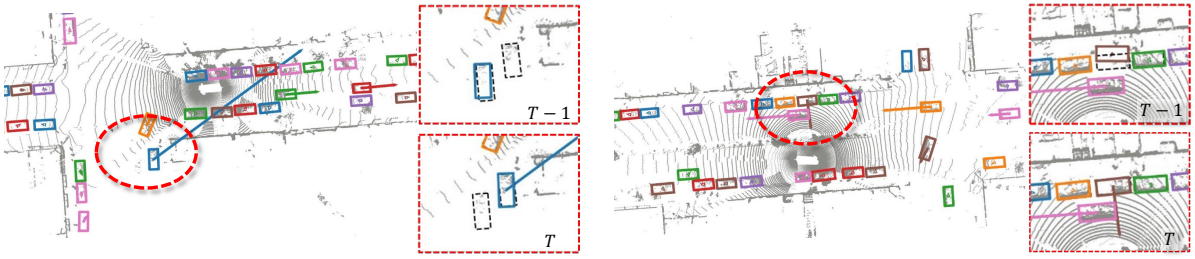
在自动驾驶中，三维物体检测的时间稳定性对驾驶安全性具有重要影响。然而，现有的指标如 mAP 和 MOTA 无法评估检测的稳定性，因此在学术界中对此问题的研究较少。为弥补这一空白，本文提出了 Stability Index (SI)，一种新的度量标准，可以全面评估三维检测器在预测置信度、框定位、尺寸和朝向方面的稳定性。通过在 Waymo 开放数据集上对最先进的物体检测器进行基准测试，SI 揭示了物体稳定性的一些有趣特性，这些特性是其他指标未曾发现的。为了帮助模型提高稳定性，我们进一步提出了一种通用且有效的训练策略，称为预测一致性学习（PCL）。PCL 鼓励在不同时间戳和增强条件下对同一物体的一致性预测，从而提高检测稳定性。此外，我们通过 CenterPoint 模型验证了 PCL 的有效性，并在车辆类别上获得了 86.00 的 SI，比基础模型高出 5.48。我们希望我们的工作能为三维物体检测领域提供一个可靠的基准，并引起研究者对稳定性这一关键问题的关注。

关键词: 三维物体检测, 时间稳定性

1　引言

三维物体检测旨在感知周围环境中的目标物体，利用来自多种数据源的信息，如点云数据 [49, 42, 45, 18, 12, 31]、相机图像 [21, 39]、多传感器数据 [26, 22, 8] 等。作为自动驾驶中的基础性任务，三维物体检测受到了学术界和工业界的广泛关注。近年来，许多高性能的检测器被提出 [16, 3, 20, 48, 43, 38, 39, 50]，显著推动了三维物体检测的发展。

然而，与直觉不同，即使是高性能的检测器也常常出现检测结果不稳定性的现象。传感器噪声、模型的敏感性、微小的场景变化和非确定性操作等因素都可能导致检测结果不稳定。现有的最先进的检测器依然主要集中于提高单次检测的精度，而常常忽略了检测时的时间稳定性。



(a) confidence variations *(b) geometric variations*

Figure 1: 检测不稳定可能引发的安全威胁。左图中，置信度波动导致框闪烁，从而引发物体关联不准确并导致异常的速度估计。右图中，尽管车辆实际静止，但由于框抖动，错误预测出车辆试图并入车流。其中，虚线框表示真实值。检测结果由 [42] 预测，物体跟踪使用 SimpleTrack [25] 进行。

检测稳定性不仅仅是指鲁棒性，它还涉及到确保自动驾驶中人的安全。如 fig. 1 所示，置信度和边界框的不稳定检测可能导致通过跟踪估算的异常速度。这些错误的估算可能会导致对周围目标行为的错误判断，从而误导自车做出不当甚至危险的决策。然而，系统性地弥补较差的检测稳定性需要额外的模块（例如，卡尔曼滤波器 [5, 41, 4]，这些模块通常需要人工调优的精细的参数）。这不仅增加了系统的复杂性和延迟，还需要繁琐的工程工作。因此，提升检测稳定性是实现安全、可靠的自动驾驶的重要步骤。

根据调查，目前还没有工作专门研究三维物体检测器的检测稳定性。主要原因之一是缺乏适当的指标来量化这种稳定性。目前用于评估检测精度的指标，如 mAP [14]，通常忽略了时间信息，但时间信息是稳定性评估的核心。另一方面，专门为时序物体跟踪设计的指标（如 MOTA 和 MOTP [2]）旨在评估物体在时间上的跟踪质量，但跟踪质量与检测稳定性是两个正交的概念。设计的跟踪器对检测噪声具有鲁棒性。一个好的跟踪算法自然会掩盖上游检测器的不稳定性。如 fig. 1(b) 所示，尽管检测器在两帧之间产生了不一致的偏航角和位置，跟踪器仍能将两个框关联在一起并融合稳定的框信息。

我们认为需要一个新的指标来评估检测稳定性。为了这一目标，我们对任务进行了全面分析，我们认为一个有效的指标应具备四个核心特性：*1) 完整性*：该指标必须考虑所有检测属性。*2) 同质性*：所有属性应均匀地集成到指标中。*3) 对称性*：指标应该保持一致性，无论输入顺序如何。*4) 边际单峰性*：当任何元素的稳定性恶化时，指标值永远不会增加。基于我们的分析，我们提出了一种新的度量标准 Stability Index (SI)，通过量化置信度、框定位、尺寸和朝向的时间一致性来评估稳定性。通过我们精心设计的方案，所提出的 SI 完全符合上述所有要求，严格的理论证明验证了这一点。

在大规模的 Waymo 开放数据集（WOD） [36] 上，我们全面地测试了各种先进的三维物体检测器，并观察到现有的指标（如 mAP 和 MOTA）与我们提出的稳定性指标 SI 之间没有明显的相关性。此外，我们的实验揭示了物体检测中的一些技巧（如使用更多的数据增强和多帧策略）在稳定性方面并未带来显著的改进。

我们还引入了一个名为预测一致性学习（PCL）的框架，其本质上是对同一对象在

不同时间戳和增强处理下的预测误差差异进行一致性约束。值得注意的是，我们的 PCL 是一个适用于所有检测器的通用框架，并且在推理过程中不会引入额外成本。简而言之，PCL 将 CenterPoint [45] 车辆类别的 SI 从 80.52 显著提升至 86.00，超越了所有最先进的检测器。我们的贡献总结如下：

1. 首次全面分析了检测稳定性，并提出了 Stability Index (SI) 指标，该指标均匀评估并积极指示所有检测元素的稳定性。同时，提供了严格的理论证明，验证了 SI 的有效性。
2. 提出了一个通用框架——预测一致性学习（PCL），以提升检测稳定性。在 Waymo 开放数据集上的大量实验揭示了物体稳定性的若干有趣见解，并展示了 PCL 的有效性。

2 相关工作

2.1 三维物体检测

三维物体检测是自动驾驶中的一个基础性任务，旨在准确定位三维空间中的物体。该领域的先前研究工作可以根据输入模式大致分为基于激光雷达、基于图像的方法和多模态方法。

大多数现有的基于雷达的方法 [49, 42, 45, 18, 12, 31, 15, 13, 19, 50] 将非均匀点云转换为规则的 2D 柱体或 3D 体素，并在后续阶段使用卷积进行高效处理。除了基于体素的方法之外，该任务还可以通过其他点云表示形式来完成，包括距离视图 [16, 3, 24, 7, 35]、基于点的表示 [9, 29, 20, 44, 46, 28, 48, 43, 34, 27, 30]，以及它们的组合 [35, 32]。一些研究 [36, 23, 38] 引入了最近流行的 Transformer 架构，并在检测精度方面取得了显著的提升。

Transformer 架构在将相机图像转换为鸟瞰图特征方面也取得了巨大成功。这种算法上的突破为自动驾驶汽车的基于视觉的 3D 检测 [21, 39] 和融合 [26, 22, 8] 铺平了道路。值得注意的是，我们提出的度量标准和方法对输入模式不敏感，因此适用于所有 3D 物体检测方法。

2.2 相关指标

对于任何机器学习任务而言，准确衡量性能都至关重要，对于 3D 物体检测来说更是如此。KITTI 数据集 [17] 在评估自动驾驶任务方面发挥了先驱作用，采用了成熟的平均精度（AP）作为指标。Waymo 开放数据集 [36] 进一步将该指标扩展为 APH，考虑了方向误差。相比之下，nuScenes 数据集 [6] 对基于交并比（IOU）的指标是否适用于仅基于视觉的方法提出了质疑，因为这类方法通常存在较大的定位误差。因此，提出了一种新的指标 NDS，通过利用阈值化的 2D 中心距离来评估容易出错的预测。

多目标跟踪（MOT）作为目标检测的下游任务，是自动驾驶的另一个关键组成部分。Bernardin 和 Stiefelhagen [2] 提出了 MOTA 和 MOTP 这两个指标，其中 MOTA 综合考虑了包括假阴性、假阳性和身份切换在内的错误，而 MOTP 则侧重于检测序列与真实值的重叠程度。Weng 等人 [40] 指出这两个指标都没有考虑分数，并通过在不同召回率水平上平均分数将其扩展为 AMOTA 和 AMOTP。总体而言，检测指标忽略了检测框之间的时间关系，而跟踪指标主要关注对象在不同帧之间是否正确关联。先前的方法在捕捉跨帧检测稳定性方面存在不足，这是本研究的关键动机。

3 方法

在本节中，我们首先全面分析了 3D 目标检测中的稳定性问题。基于我们的分析，我们引入了一种新的度量标准，称为 Stability Index(SI)，并证明了其关键性质。最后，我们介绍了预测一致性学习（PCL）以增强检测的稳定性。

3.1 符号说明

3D 目标检测器的有效预测 P 包括一个置信度分数 c 和一个 3D 边界框，定义为 $B = (x, y, z, l, w, h, \theta)$。其中，$(x, y, z)$ 是边界框中心的坐标，(l, w, h) 表示边界框的尺寸，θ 表示偏航角。

给定两个边界框 B_1 和 B_2，我们定义了一个变换函数 $T_{B_1 \to B_2}(\cdot)$，表示从 B_1 到 B_2 的映射。因此，我们可以将此自定义变换应用于任意边界框 B，得到 $\hat{B} = T_{B_1 \to B_2}(B)$，其中

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} \cos(\theta_2 - \theta_1) & \sin(\theta_2 - \theta_1) & 0 \\ -\sin(\theta_2 - \theta_1) & \cos(\theta_2 - \theta_1) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x - x_1 \\ y - y_1 \\ z - z_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix},$$
$$\begin{pmatrix} \hat{l} \\ \hat{w} \\ \hat{h} \end{pmatrix} = \begin{pmatrix} l_2/l_1 \times l \\ w_2/w_1 \times w \\ h_2/h_1 \times h \end{pmatrix}, \qquad \hat{\theta} = \theta + (\theta_2 - \theta_1).$$

本质上，此操作将 B_1 和 B 之间的差异转换为 B_2 和 $\hat{B}$ 之间的差异。

3.2 检测稳定性分析

在自动驾驶的背景下，检测器所预测的任何属性出现变化都可能导致危险情况。例如，边界框位置和方向的波动可能导致速度估计不准确，从而可能导致不安全的交互决策。不稳定的置信度分数可能导致预测闪烁，阻碍自动驾驶系统准确跟踪对象。此外，附近车辆尺寸的不稳定预测可能促使自车采取不当的避让操作。总之，必须全面考虑所有检测元素的稳定性，以确保自动驾驶的安全性。

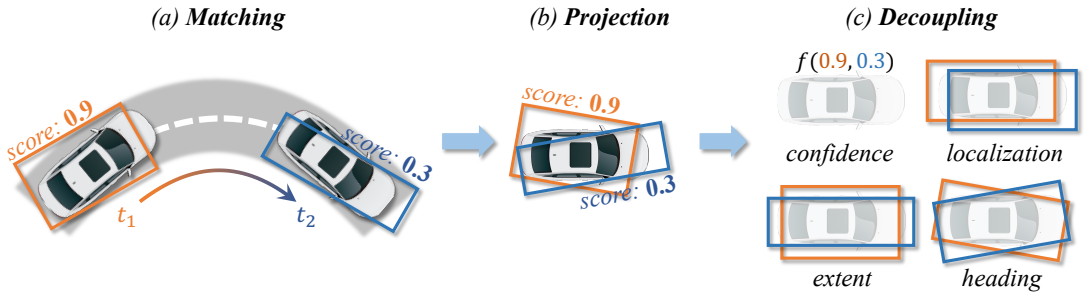


Figure 2: 计算 Stability Index 的过程。橙色 和 蓝色 框表示通过匈牙利算法搜索到的预测与真实值之间的最佳匹配。这些框随后通过其对象 ID 标签在帧之间进行关联。在将预测投影到预构建的基准框后，SI 将其解耦为逐元素计算，然后聚合以进行最终的检测稳定性评估。

评估稳定性的一种简单方法是汇总所有这些元素的变化，这本质上是对 2D 视频检测 [47] 的扩展。然而，这些变化不应直接相加，因为这些检测属性表示对象的不同物理属性。此外，元素变化与对象属性无关，因此无法捕捉不稳定预测引起的危险程度。例如，偏航角的大幅抖动可能导致大体积对象（如车辆）的行为快速变化。相比之下，行人更容易受到中心偏移和边界框尺寸不稳定性的影响。因此，如何将这些元素标准化为单一且一致的单位仍然是一个具有挑战性的问题。

一种可能的统一边界框相关元素物理单位的方法是采用 mAP 度量中使用的交并比（IoU）。为此，我们首先在最小单元中评估检测稳定性，涉及两个时间戳 t_1, t_2 的单个对象，如 fig. 2 所示。将真实值框表示为 B_1^g, B_2^g，预测表示为 $P_i = \{c_i, B_i\}, i \in \{1, 2\}$。由于对象移动，两个预测的 3D 框之间的 IoU 无法直接计算。相比之下，我们预定义的操作通过将框投影到其中一个真实值上来实现测量。例如，我们可以将 B_1 投影到第二个真实值上，得到 $\hat{B}_1 = T_{B_1^g \to B_2^g}(B_1)$，并计算 $\text{IoU}(\hat{B}_1, B_2)$。此 IoU 可以在一定程度上反映检测稳定性。然而，此测量存在两个显著缺陷（补充材料中的性质 1 和性质 2 证明了这一点）：(1) IoU 随帧顺序变化，即 $\text{IoU}(\hat{B}_1, B_2) \neq \text{IoU}(B_1, \hat{B}_2)$。(2) IoU 不是边际单峰的。换句话说，增强元素的稳定性有时会导致 IoU 值变差。这两个缺陷使得 IoU 无法作为检测稳定性的有效评估指标。

通过对稳定性的详细分析和潜在解决方案的探索，我们确定了有效度量标准应满足的四个关键性质：

- **全面性**：度量标准应全面反映所有相关检测元素的影响。
- **同质性**：所有元素的影响应被处理为统一的物理单位。
- **对称性**：当应用于正向和反向输入时，度量值应一致。
- **边际单峰性**：对于每个元素，当其他元素固定时，度量标准应在其稳定性方面是单峰的。

3.3 Stability Index

虽然 IoU 是一个有希望的出发点，但要满足这四个性质，需要精心设计以有效整合置信度分数并解决不对称性和非单峰性设计缺陷。为此，我们引入了基准框投影、元素解耦和稳定性聚合方案，如 fig. 2 所示。最终，我们评估连续帧中对象对的稳定性，并将度量标准表示为 Stability Index (SI)。

基准框投影。由于投影到任一真实值框可能会引入不对称性问题，因此我们提出将预测投影到中间基准框 $B^p = (0, 0, 0, l^p, w^p, h^p, 0)$ 上。这里，我们利用几何平均值 $l^p = \sqrt{l_1^g l_2^g}$、$w^p = \sqrt{w_1^g w_2^g}$、$h^p = \sqrt{h_1^g h_2^g}$ 来确保基准框的尺寸与真实值框 B_1^g, B_2^g 的尺寸紧密匹配。这对于准确的稳定性测量至关重要，因为不同尺寸的对象对波动的敏感程度不同。最终，我们得到 $\hat{B}_1 = T_{B_1^g \to B^p}(B_1), \hat{B}_2 = T_{B_2^g \to B^p}(B_2)$，如 fig. 2(b) 所示。

元素解耦。为了实现边际单峰性，度量标准在除了一个任意元素外其他所有元素都固定时必须表现出以下两个特征：(1) 当且仅当元素稳定时，度量标准达到峰值。(2) 当元素的稳定性在任何连续方向上恶化时，度量值单调非增。我们认识到，由于元素之间的相互干扰，IoU 无法满足这些特征，因此我们提出将其解耦为四个独立部分，如 fig. 2(c) 所示。例如，为了测量定位稳定性，我们使 $\hat{B}_1, \hat{B}_2$ 中除边界框中心外的元素相同。具体来说，我们用基准框中的元素替换它们，得到 $\hat{B}_i^{loc} = (\hat{x}_i, \hat{y}_i, \hat{z}_i, l^p, w^p, h^p, 0), i \in \{1, 2\}$。类似地，我们可以得到 $\hat{B}_i^{ext}, \hat{B}_i^{hdg}, i \in \{1, 2\}$ 用于边界框尺寸和方向。然后，我们通过以下两个方程评估边界框定位和尺寸的稳定性：

$$SI_l = \mathrm{IoU}(\hat{B}_1^{loc}, \hat{B}_2^{loc}), \quad SI_e = \mathrm{IoU}(\hat{B}_1^{ext}, \hat{B}_2^{ext}). \tag{1}$$

直接使用 $\mathrm{IoU}(\hat{B}_1^{hdg}, \hat{B}_2^{hdg})$ 会在 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 之间的角度差超过 $\pi/4$ 时违反单峰性（补充材料中的引理 3 证明了这一点）。因此，我们将此情况视为失败，并明确将度量值设置为 0。边界框方向的稳定性最终为：

$$SI_h = \begin{cases} 0, & \text{如果} |\hat{\theta}_1 - \hat{\theta}_2| \geq \pi/4, \\ \mathrm{IoU}(\hat{B}_1^{hdg}, \hat{B}_2^{hdg}), & \text{否则。} \end{cases} \tag{2}$$

置信度的稳定性可以通过分数 c_1, c_2 之间的差异来捕捉，即使用 $1 - |c_1 - c_2|$。还有一个问题是，此函数容易受到目标检测器固有置信度尺度的影响。例如，如果所有分数都除以一个缩放因子，检测性能和稳定性应保持不变。然而，$1 - |c_1 - c_2|$ 的值会增加，导致稳定性测量不准确。为了解决这个问题，我们计算所有置信度的 99% 和 1% 百分位数，分别为 $c^{0.99}$ 和 $c^{0.01}$。然后，通过以下公式校准置信度稳定性：

$$SI_c = \max\left(0, 1 - |c_1 - c_2| / (c^{0.99} - c^{0.01})\right). \tag{3}$$

稳定性聚合。在最后一步中，我们使用以下公式聚合所有组件的稳定性：

$$SI = SI_c \times (SI_l + SI_e + SI_h)/3. \tag{4}$$

这里，$SI_c \in [0, 1]$ 被视为边界框稳定性的权重。SI_l, SI_e, SI_h 由于具有相同的 IoU 单位，可以取平均值。最终，SI 成功满足了有效稳定性评估器的四个性质，如 lemmas 1 and 2 所示。详细分析和理论证明可在我们的补充材料中找到。

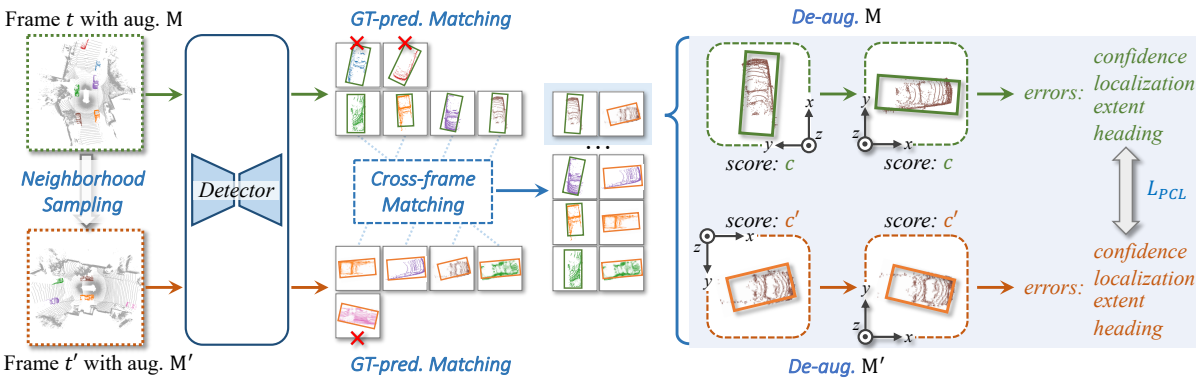


Figure 3: 提出的预测一致性学习（PCL）的流程。在每次迭代中，PCL 在相邻时间戳 t 和 t' 处采样一对帧，并对配对样本应用增强 $\mathbf{M}$ 和 $\mathbf{M}'$。然后，真实值-预测匹配和跨帧匹配协作关联两帧之间来自同一对象的检测器预测。在去增强过程后，PCL 计算置信度、定位、尺寸和方向的预测误差，这些误差定义在对象自身坐标系中。最后，PCL 惩罚所有预测对之间的误差差异，以强制执行时间一致性。在图中，pred. 和 aug. 分别表示预测和增强。

Lemma 1. *SI 是一个对称度量标准，统一评估所有元素对检测稳定性的影响。*

Lemma 2. *SI 在所有元素方面是边际单峰的。当且仅当检测在帧之间完全稳定时，度量值达到最大值 1。*

我们之前的讨论集中在由两个连续时间戳的单个对象组成的最小集合上。为了评估大规模基准测试的 SI，我们首先使用匈牙利算法将每个真实值与预测配对。通过标记的对象 ID，我们将评估分割为计算多个最小集合的 SI。最终结果是所有值的平均值。更多细节（如处理极端情况）在补充材料中。

3.4 预测一致性学习

除了度量标准的设计外，我们还尝试提高 3D 目标检测器的检测稳定性。为此，我们引入了一种通用且有效的训练策略，称为预测一致性学习（PCL），如 fig. 3 所示。我们的 PCL 建立在鼓励在不同增强和时间戳下跨帧预测一致性的核心思想之上。它由四个关键阶段组成：邻域采样、预测配对、去增强和预测一致性损失。

邻域采样。对于每个时间戳为 t 的帧 F，我们首先从范围 $[-n, n]$ 中均匀采样一个整数 Δt，其中 n 是预定义的参数。随后，我们获取时间戳为 $t+\Delta t$ 的帧 F'，并将 F, F' 捆绑为网络的成对输入。帧进一步通过随机翻转、旋转和缩放分别进行增强。我们将增强记录到矩阵 $\mathbf{M}$ 和 $\mathbf{M}'$ 中，其中 $\mathbf{M}$ 可以描述如下：

$$\mathbf{M} = \begin{pmatrix} i_x & 0 & 0 \\ 0 & i_y & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot s. \tag{5}$$

这里，i_x 和 i_y 表示帧是否在 x 和 y 方向上翻转，-1 表示发生相应的翻转，1 表示未发生。α 是随机旋转应用的角度，s 表示随机缩放的因子。

预测配对。在检测器从配对样本生成预测后，下一步是收集相应的预测以进行比较。我们首先执行真实值-预测匹配，将每个真实值框与最佳匹配的预测匹配，这可以通过匈牙利算法或任何其他合理的方法完成。随后，跨帧匹配通过相应的对象 ID 关联两帧之间的预测，并创建预测对以供后续比较。

去增强。训练期间使用的数据增强可能会大大改变检测误差的模式，阻碍每对预测的公平比较。例如，随机缩放可能会放大边界框位置和尺寸的误差，而随机翻转可能会改变误差方向。因此，我们对每个预测应用去增强步骤，以消除增强的影响。对于预测 $P=\{c,x,y,z,l,w,h,\theta\}$，我们使用相应的 $\mathbf{M}$ 将其恢复为 $\bar{P}=\{\bar{c},\bar{x},\bar{y},\bar{z},\bar{l},\bar{w},\bar{h},\bar{\theta}\}$：

$$\begin{cases}\bar{c}=c,\\ (\bar{x},\bar{y},\bar{z})^T=\mathbf{M}^{-1}(x,y,z)^T,\\ (\bar{l},\bar{w},\bar{h})^T=(l,w,h)^T/s,\\ \bar{\theta}=i_x\cdot i_y\cdot(\theta-\alpha).\end{cases}\tag{6}$$

预测一致性损失。在引入一致性损失之前，我们首先计算去增强预测 $\bar{P}$ 相对于真实值框 B^g 的预测误差。我们将置信度的误差定义为 $e_c=1-\bar{c}$。边界框定位、尺寸和方向的预测误差在对象的自身坐标系中计算。具体来说，边界框中心的误差计算如下：

$$\mathbf{e}_l=\begin{pmatrix}\cos\theta^g & \sin\theta^g & 0\\ -\sin\theta^g & \cos\theta^g & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\bar{x}-x^g\\ \bar{y}-y^g\\ \bar{z}-z^g\end{pmatrix}.\tag{7}$$

边界框尺寸的预测误差公式为：

$$\mathbf{e}_e=\left(\bar{l}/l^g,\bar{w}/w^g,\bar{h}/h^g\right)^T.\tag{8}$$

最后，边界框方向的误差 $\mathbf{e}_h$ 编码为三角向量：

$$\mathbf{e}_h=\left(\sin(\bar{\theta}-\theta^g),\cos(\bar{\theta}-\theta^g)\right)^T.\tag{9}$$

我们的最后一步是鼓励每对预测在预测误差方面表现出相似的模式。为此，我们收集成对误差 $\{e_{c,i},e'_{c,i}\}$、$\{\mathbf{e}_{l,i},\mathbf{e}'_{l,i}\}$、$\{\mathbf{e}_{e,i},\mathbf{e}'_{e,i}\}$ 和 $\{\mathbf{e}_{h,i},\mathbf{e}'_{h,i}\}$，其中 $i\in\{1,2,\cdots,N\}$，N 是帧 F 和 F' 之间成功关联的对象数量。最终，我们的预测一致性损失为：

$$\begin{aligned}L_{PCL}=&\frac{1}{N}\sum_{i=1}^{N}\big(w_1\cdot\mathrm{MSE}(e_{c,i},e'_{c,i})+w_2\cdot L_1(\mathbf{e}_{l,i},\mathbf{e}'_{l,i})\\ &+w_3\cdot L_1(\mathbf{e}_{e,i},\mathbf{e}'_{e,i})+w_4\cdot L_1(\mathbf{e}_{h,i},\mathbf{e}'_{h,i})\big).\end{aligned}\tag{10}$$

这里，w_1、w_2、w_3 和 w_4 是用于平衡损失中不同部分的权重，如果未指定，则默认为 1。MSE 和 L_1 分别是均方误差和 L_1 距离的损失函数。原始检测损失和我们的预测一致性损失都被用于训练目标检测器。

Table 1: Waymo 开放数据集上的基准测试结果。模型根据车辆类别的 mAPH 排序。我们使用两种不同强度的颜色突出每列中的最高和第二高结果。"†" 表示该模型不仅基于雷达。CenterPoint*[45] 表示 CenterPoint 的 pillar 版本。

方法	车辆 (%)						行人 (%)						骑行者 (%)					
	mAPH	SI	SI_c	SI_l	SI_e	SI_h	mAPH	SI	SI_c	SI_l	SI_e	SI_h	mAPH	SI	SI_c	SI_l	SI_e	SI_h
Second [42]	72.60	81.37	90.2	84.2	92.0	92.2	59.81	63.07	83.9	69.6	87.8	67.6	61.95	67.21	81.1	76.1	88.3	83.8
CenterPoint*[45]	72.82	80.61	89.0	85.4	91.0	92.8	65.28	64.57	83.2	74.4	87.4	68.9	65.87	68.06	80.8	77.7	87.0	85.9
Pointpillar [18]	72.84	80.84	89.6	84.4	92.3	91.6	54.64	62.03	84.7	72.1	88.8	57.9	59.51	66.14	82.2	74.9	88.0	77.4
CenterPoint [45]	73.73	80.52	89.0	85.3	90.7	92.9	69.50	68.40	85.7	73.3	88.6	75.0	71.04	68.40	80.3	78.5	87.4	89.8
PartA2Net [31]	75.02	82.86	91.4	85.4	91.7	91.7	66.16	65.08	84.6	73.6	86.7	67.0	67.90	72.73	85.9	79.3	87.0	84.3
PV R-CNN [32]	75.92	83.73	91.9	86.4	92.3	91.7	66.28	66.17	86.0	73.5	87.4	66.6	68.38	73.53	86.8	78.9	88.4	83.2
Voxel R-CNN [12]	77.19	84.26	92.0	86.7	92.1	93.3	74.21	69.50	86.9	75.3	88.1	73.6	71.68	73.23	84.4	80.1	87.7	89.3
VoxelNext [10]	77.84	84.82	92.9	86.3	91.6	94.2	76.24	74.74	92.7	75.7	88.0	75.8	75.59	76.48	90.0	79.2	84.9	87.8
PV R-CNN+[33]	77.88	84.49	92.1	87.2	92.4	93.2	73.99	69.27	86.8	75.3	88.1	73.2	71.84	73.05	84.2	80.3	87.7	89.2
DSVT [38]	78.82	84.90	92.5	86.9	91.5	94.8	76.81	74.58	91.9	76.5	88.7	75.9	75.44	76.20	88.2	80.5	86.1	89.9
TransFusion† [1]	79.00	82.32	89.3	86.8	92.7	95.7	76.52	69.11	84.5	75.4	89.9	78.8	70.11	70.35	80.6	79.5	90.6	91.1

4 实验

4.1 Waymo 开放数据集上的基准测试

实现细节。我们在 OpenPCDet [37] 和 MMDetection3D [11] 上复现了常用的基于雷达和融合的 3D 检测器。所有检测器均在 Waymo 开放数据集（WOD）[36] 上使用默认配置进行训练。我们的训练使用完整版本的训练集，包含 798 个序列和 158,361 个样本。我们在验证集上评估这些模型，验证集包含 202 个序列和 40,077 个样本，使用 LEVEL 1 mAP（加权 Heading 准确率，mAPH）和提出的 SI 进行评估。除了 SI，我们还展示了其在置信度 (SI_c)、定位 (SI_l)、尺寸 (SI_e) 和方向 (SI_h) 上的子指标。

SI 与 mAPH 的关系。 table 1 展示了模型在车辆、行人和骑行者类别上的结果。模型根据车辆类别的 mAPH 排序，我们突出显示了每列中表现最好的两个模型。从结果中我们发现，检测精度和模型稳定性之间没有明显的相关性。例如，TransFusion 在车辆类别上具有最高的 mAPH，但其 SI 远低于具有相似检测指标的基于雷达的模型。这可能是因为融合模型通过相机图像的额外信息提高了检测精度。然而，视觉信息在推断精确的 3D 位置时是间接使用的，从而增加了检测的不确定性。另一方面，CenterPoint 在车辆检测上达到了 73.73 mAPH，高于 Second 和 PointPillar。但其 SI 为 80.52，是所有检测器中的最低值。这些结果证明了这两个指标之间的没有明确正相关关系。

对象属性的影响。 fig. 4 展示了各种对象属性如何影响检测稳定性。我们根据指定属性对对象进行分组，并使用 CenterPoint 进行检测。fig. 4(a) 展示了检测稳定性与对象距离之间的负相关关系，距离越远，对象通常越难学习。对于所有类别，SI 随着对象点数的增加而增加，并在点数达到 5^3 时趋于饱和，如 fig. 4(b) 所示。fig. 4(c) 和 (d) 进一步探讨了车辆体积和长宽比的影响。我们发现小型车辆往往具有更稳定的检测结果。

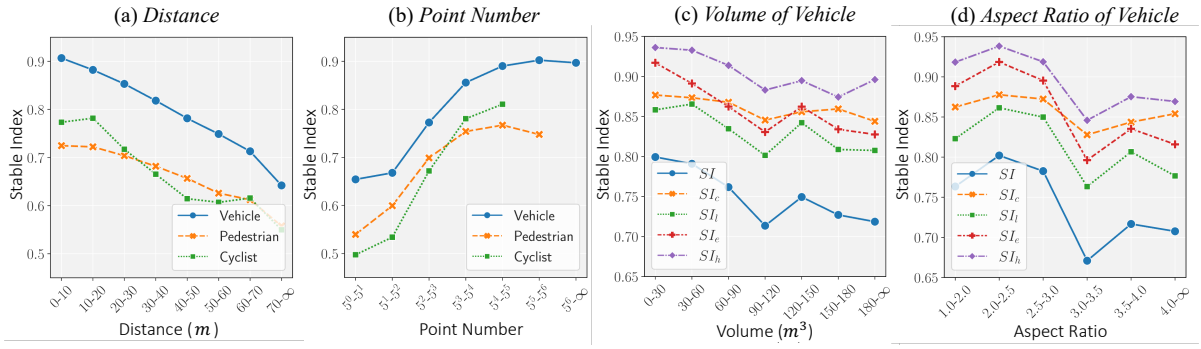


Figure 4: 对象属性与检测稳定性之间的关系。

Table 2: 多帧策略对检测稳定性的影响。

方法	帧数	车辆 (%)		行人 (%)		骑行者 (%)	
		mAPH	SI	mAPH	SI	mAPH	SI
CenterPoint	1	73.73	80.52	69.50	68.40	71.04	68.40
	2	75.04	80.86	75.17	70.40	71.23	69.39
	4	75.85	81.74	75.38	71.69	71.68	69.93
PV R-CNN	1	78.33	85.17	75.75	70.15	72.47	73.31
	2	79.62	86.39	80.37	73.79	73.66	76.78
	4	80.51	87.50	81.12	75.32	74.77	76.34

长宽比在 2 到 3 之间的车辆表现出相对较高的 SI 值。这可能是因为这类车辆在现实场景中更为常见。长宽比较大的车辆（如卡车/电车/公交车）在数据集中相对较少，且需要更大的感受野，导致其检测结果更不稳定。

多帧策略的效果。将多个连续点云合并为一个输入是解决 LiDAR 数据稀疏性问题的常用策略。table 2 显示，该策略不仅提高了模型精度，还提升了检测稳定性。以车辆为例，使用四帧后，CenterPoint 和 PV R-CNN 的检测精度分别达到 75.85 和 80.51 mAPH，比基线提高了 2.12 和 2.18 mAPH。同时，CenterPoint 和 PV R-CNN 的 SI 值分别提高了 1.22 和 2.33。这一趋势在所有类别中一致，表明该策略在提升检测精度和稳定性方面具有普遍有效性。

总结。我们的实验验证了所提出的 SI 是检测精度的补充指标。该指标值在不同模型类型之间差异较大，并展示了与对象属性相关的几种模式。我们还检验了数据增强（见补充材料）和多帧策略两种常用方案。增加数据增强的程度对检测稳定性的影响较小。虽然使用多帧被证明是有效的，但在将数据编码为体素特征时会带来较大的计算开销。相比之下，我们提出的 PCL 在推理过程中不引入额外计算，同时显著提高了检测稳定性，后续实验将进一步说明这一点。

Table 3: 所提出的 PCL 的效果。“-” 表示基础模型，“w/o PCL” 表示未使用预测一致性损失微调的模型。

方法	车辆 (%)		行人 (%)		骑行者 (%)	
	mAPH	SI	mAPH	SI	mAPH	SI
-	73.73	80.52	69.50	68.40	71.04	68.40
w/o PCL	73.70	80.93	69.55	68.35	71.27	68.20
PCL ($n = 0$)	75.57	85.42	70.18	71.87	70.86	68.80
PCL ($n = 4$)	75.26	85.83	69.56	72.76	70.65	69.22
PCL ($n = 8$)	75.04	85.94	68.82	72.87	70.31	69.32
PCL ($n = 12$)	74.64	85.93	68.50	72.95	70.85	69.33
PCL ($n = 16$)	74.54	86.00	67.82	73.14	70.25	69.16

4.2 PCL 实验

实现细节。我们采用广泛使用的 CenterPoint 作为基础模型，使用 OpenPCDet 中的默认设置进行训练。具体来说，我们使用 Adam 优化器训练模型 36 个周期。初始学习率为 0.003，采用单周期策略，前 40% 周期学习率逐渐增加到 0.03，然后在剩余训练中逐渐降低。

我们没有选择端到端训练，而是在 PCL 的基础上对基础模型进行微调，训练几个周期。训练配置与端到端训练相同，只是周期数减少到 5，学习率降低到原来的 1/10。值得注意的是，该方案不仅大大降低了训练成本，还展示了 PCL 的有效性。

PCL 的有效性。我们比较了使用和不使用 PCL 微调的模型性能，如 table 3 所示。可以观察到，直接微调模型对模型精度和稳定性的影响较小。相比之下，当使用 PCL 且不涉及跨帧信息时（即 $n = 0$），在车辆、行人和骑行者上的表现已经分别达到了 84.54、70.95 和 68.80 的 SI 值。这些结果显示出显著的提升，与基础模型相比分别提高了 +4.49、+3.52 和 +0.60。对于 mAP，我们发现了一个有趣的现象：三个类别的 mAP 分别变化了 +1.87、+0.63 和 -0.39。这得出了两个有价值的结论：(1) 我们的 PCL 不仅增强了稳定性，还提高了整体检测精度，尤其是车辆类别。(2) 无论 mAP 如何变化，SI 都一致提高，进一步验证了这两个指标评估的是模型的不同属性。

帧对间隔的影响。 PCL 的一个关键超参数是帧对之间的最大间隔 n，如邻域采样中所述。n 越大，PCL 对比的两帧之间的时间跨度越长。table 3 中的结果显示，随着 n 的变化，检测精度和稳定性呈现出相反的趋势。以车辆类别为例。当 $n = 0$ 时，我们在所有 PCL 模型中获得了最高的 mAP（75.57）和最低的 SI（85.42），这是所有 PCL 模型中的最佳表现。随着 n 的增大，mAP 最终下降到 74.54。相反，模型稳定性随着 n 达到 16 时逐渐上升到 86.00 SI。这可能是因为当帧间隔 n 增大时，物体形态可能发生显著变化。强行对齐可能会损害模型的准确性，但这种对齐有助于为相同物体提供一致的预测，进而导致稳定的检测。

PCL 中损失组件的影响. 我们引入的连贯性损失由置信度、定位、范围和方向四

Table 4: 在车辆类别上，PCL 中不同组件的实验结果（%）。“C”、“L”、“E” 和 “H” 分别表示应用与置信度、定位、范围和方向相关的损失部分。

组件				mAPH	SI	SI_c	SI_l	SI_e	SI_h
C	L	E	H						
				73.70	80.93	88.90	85.90	91.50	93.64
✓				75.64	84.04	92.28	85.80	91.44	93.65
	✓			74.15	81.80	89.34	87.16	91.79	93.73
		✓		73.86	81.73	89.14	86.13	93.37	93.67
			✓	73.44	80.89	88.88	85.64	91.47	93.85
✓	✓	✓	✓	75.57	85.42	92.81	86.78	93.31	93.90

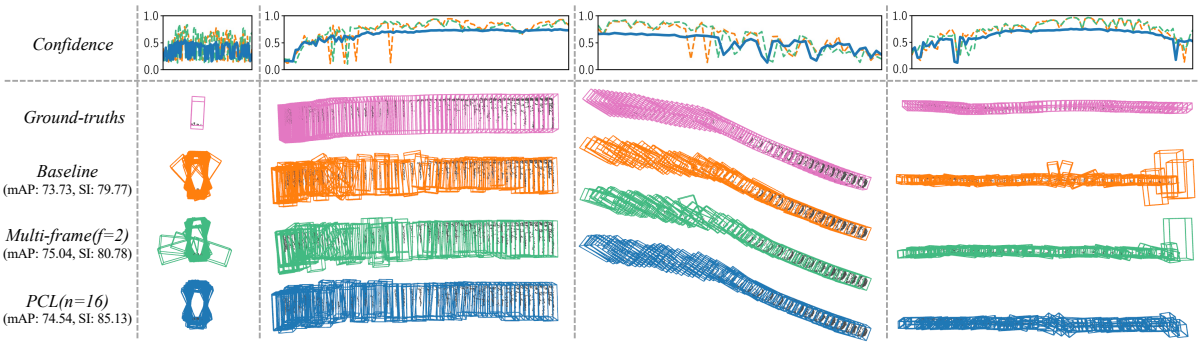


Figure 5: Ground-truth（真值，粉色）和使用不同训练策略（基准、双帧输入、PCL 策略）的 CenterPoint 模型预测结果（分别用橙色，绿色，蓝色表示）可视化。上排展示了预测的置信度变化，下排展示了 3D 框的预测。

个部分组成。我们研究了这些损失组件的不同组合对模型性能的影响，并在 table 4 中报告了结果。这些结果表明，PCL 中的每个损失部分都能从相应的方面提升模型的稳定性，验证了每个组件的有效性。当所有损失部分都被应用时，检测精度和稳定性达到了最高。

与置信度得分相关的损失部分在最终的 SI 上产生了最大的提升。这可能是因为用于训练检测器的分类损失主要关注是否正确分类了一个物体而留出了足够的空间来提升一致性。相反，框参数本身已经具有一致预测的潜力，因为它们使用真实标签作为目标。强行在这些参数上施加一致性并没有对置信度得分产生同样大的影响。此外，我们观察到与方向组件相关的损失带来的提升最小，表明保持方向一致性是一个具有挑战性的任务。

可视化. 在 fig. 5 中，我们展示了少量真值数据和来自三种不同模型的检测结果：基础的 CenterPoint 模型，输入两帧的 CenterPoint 模型，以及使用 $n = 16$ 的 PCL 模型。在第一排，我们绘制了置信度得分随时间变化的趋势，发现 PCL 模型在抑制置信度波动方面优于其他两种模型。对于预测的 3D 框，PCL 模型的结果也比其他模型更加稳定。值得注意的是，尽管我们的 PCL 模型相比多帧版本的 CenterPoint 模型，mAP

较低，但在 SI 上的表现明显优于它。这进一步验证了检测精度和稳定性捕获的是模型表现的独立方面。这些现象充分展示了 PCL 在增强检测稳定性方面的有效性。

5 结论与局限性

在本研究中，我们全面探讨了目标检测中一个关键但被忽视的问题——检测稳定性。为了评估这种稳定性，我们精心设计了一种经过验证的度量指标，称为 Stability Index（SI）。我们进一步提出了预测一致性学习框架，旨在提升模型稳定性。大量实验验证了 SI 的合理性以及所提框架的有效性。我们希望我们的工作能够作为一个可靠的基准，并引起学术界对三维目标检测中这一关键问题的关注。

为了激发未来的研究工作，我们根据当前的理解总结了几点局限性：(1) 提出的 SI 仅关注于默认的检测元素，目的是为了通用性。然而，一些检测器还会输出额外的预测结果，如速度和属性。将这些额外元素的稳定性纳入度量，同时保持 SI 的特性，是一个具有实际价值的方向。(2) 在追求通用基准方法的过程中，我们将 PCL 的设计限制为与现有目标检测器兼容，避免在推理过程中引入额外的计算，从而确保其广泛适用性。未来的工作可能会突破这些限制，探索提升性能的更多可能性。

References

[1] Xuyang Bai et al. “Transfusion: Robust lidar-camera fusion for 3d object detection with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1090–1099.

[2] Keni Bernardin and Rainer Stiefelhagen. “Evaluating multiple object tracking performance: the clear mot metrics”. In: *EURASIP Journal on Image and Video Processing* 2008 (2008), pp. 1–10.

[3] Alex Bewley et al. “Range conditioned dilated convolutions for scale invariant 3d object detection”. In: *arXiv preprint arXiv:2005.09927* (2020).

[4] Alex Bewley et al. “Simple online and realtime tracking”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468.

[5] Gary Bishop, Greg Welch, et al. “An introduction to the kalman filter”. In: *Proc of SIGGRAPH, Course* 8.27599-23175 (2001), p. 41.

[6] Holger Caesar et al. “nuscenes: A multimodal dataset for autonomous driving”. In: *CVPR*. 2020, pp. 11621–11631.

[7] Yuning Chai et al. “To the point: Efficient 3d object detection in the range image with graph convolution kernels”. In: *CVPR*. 2021, pp. 16000–16009.

[8] Xiaozhi Chen et al. “Multi-view 3d object detection network for autonomous driving”. In: *CVPR*. 2017, pp. 1907–1915.

[9] Yilun Chen et al. “Fast point r-cnn”. In: *ICCV*. 2019, pp. 9775–9784.

[10] Yukang Chen et al. “Voxelnext: Fully sparse voxelnet for 3d object detection and tracking”. In: *CVPR*. 2023, pp. 21674–21683.

[11] MMDetection3D Contributors. *MMDetection3D: OpenMMLab next-generation platform for general 3D object detection*. `https://github.com/open-mmlab/mmdetection3d`. 2020.

[12] Jiajun Deng et al. “Voxel r-cnn: Towards high performance voxel-based 3d object detection”. In: *AAAI*. Vol. 35. 2021, pp. 1201–1209.

[13] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[14] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2005.

[15] Lue Fan et al. “Embracing single stride 3d object detector with sparse transformer”. In: *CVPR*. 2022, pp. 8458–8468.

[16] Lue Fan et al. “Rangedet: In defense of range view for lidar-based 3d object detection”. In: *ICCV*. 2021, pp. 2918–2927.

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *CVPR*. 2012.

[18] Alex H Lang et al. “Pointpillars: Fast encoders for object detection from point clouds”. In: *CVPR*. 2019, pp. 12697–12705.

[19] Jinyu Li, Chenxu Luo, and Xiaodong Yang. “PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds”. In: *CVPR*. 2023, pp. 17567–17576.

[20] Zhichao Li, Feng Wang, and Naiyan Wang. “Lidar r-cnn: An efficient and universal 3d object detector”. In: *CVPR*. 2021, pp. 7546–7555.

[21] Zhiqi Li et al. “Bevformer: Learning bird’ s-eye-view representation from multi-camera images via spatiotemporal transformers”. In: *ECCV*. Springer. 2022, pp. 1–18.

[22] Zhijian Liu et al. “Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation”. In: *IEEE International Conference on Robotics and Automation*. IEEE. 2023, pp. 2774–2781.

[23] Zhijian Liu et al. “FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer”. In: *CVPR*. 2023, pp. 1200–1211.

[24] Gregory P Meyer et al. “Lasernet: An efficient probabilistic 3d object detector for autonomous driving”. In: *CVPR*. 2019, pp. 12677–12686.

[25] Ziqi Pang, Zhichao Li, and Naiyan Wang. “SimpleTrack: Understanding and Rethinking 3D Multi-object Tracking”. In: *arXiv preprint arXiv:2111.09621* (2021).

[26] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. “Multi-modal fusion transformer for end-to-end autonomous driving”. In: *CVPR*. 2021, pp. 7077–7087.

[27] Charles R Qi et al. “Deep hough voting for 3d object detection in point clouds”. In: *ICCV*. 2019, pp. 9277–9286.

[28] Charles R Qi et al. “Frustum pointnets for 3d object detection from rgb-d data”. In: *CVPR*. 2018, pp. 918–927.

[29] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. “Pointrcnn: 3d object proposal generation and detection from point cloud”. In: *CVPR*. 2019, pp. 770–779.

[30] Shaoshuai Shi et al. “From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network”. In: *IEEE TPAMI* 43.8 (2020), pp. 2647–2664.

[31] Shaoshuai Shi et al. “Part-aˆ 2 net: 3d part-aware and aggregation neural network for object detection from point cloud”. In: *arXiv preprint arXiv:1907.03670* 2.3 (2019).

[32] Shaoshuai Shi et al. “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection”. In: *CVPR*. 2020, pp. 10529–10538.

[33] Shaoshuai Shi et al. “PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection”. In: *arXiv preprint arXiv:2102.00463* (2021).

[34] Weijing Shi and Raj Rajkumar. “Point-gnn: Graph neural network for 3d object detection in a point cloud”. In: *CVPR*. 2020, pp. 1711–1719.

[35] Pei Sun et al. “Rsn: Range sparse net for efficient, accurate lidar 3d object detection”. In: *CVPR*. 2021, pp. 5725–5734.

[36] Pei Sun et al. “Scalability in perception for autonomous driving: Waymo open dataset”. In: *CVPR*. 2020, pp. 2446–2454.

[37] OpenPCDet Development Team. *OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds*. `https://github.com/open-mmlab/OpenPCDet`. 2020.

[38] Haiyang Wang et al. “Dsvt: Dynamic sparse voxel transformer with rotated sets”. In: *CVPR*. 2023, pp. 13520–13529.

[39] Yue Wang et al. “Detr3d: 3d object detection from multi-view images via 3d-to-2d queries”. In: *Conference on Robot Learning*. PMLR. 2022, pp. 180–191.

[40] Xinshuo Weng et al. “3d multi-object tracking: A baseline and new evaluation metrics”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2020, pp. 10359–10366.

[41] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple online and realtime tracking with a deep association metric”. In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

[42] Yan Yan, Yuxing Mao, and Bo Li. “Second: Sparsely embedded convolutional detection”. In: *Sensors* 18.10 (2018), p. 3337.

[43] Zetong Yang et al. “3dssd: Point-based 3d single stage object detector”. In: *CVPR*. 2020, pp. 11040–11048.

[44] Zetong Yang et al. “Ipod: Intensive point-based object detector for point cloud”. In: *arXiv preprint arXiv:1812.05276* (2018).

[45] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. “Center-based 3d object detection and tracking”. In: *CVPR*. 2021, pp. 11784–11793.

[46] Haoyang Zhang et al. “Varifocalnet: An iou-aware dense object detector”. In: *CVPR*. 2021, pp. 8514–8523.

[47] Hong Zhang and Naiyan Wang. “On the stability of video detection and tracking”. In: *arXiv preprint arXiv:1611.06467* (2016).

[48] Yifan Zhang et al. “Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds”. In: *CVPR*. 2022, pp. 18953–18962.

[49] Yin Zhou and Oncel Tuzel. “Voxelnet: End-to-end learning for point cloud based 3d object detection”. In: *CVPR*. 2018, pp. 4490–4499.

[50] Zixiang Zhou et al. “Centerformer: Center-based transformer for 3d object detection”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 496–513.