arXiv:2302.02858v3 [cs.CV] 5 Dec 2023

TR3D：迈向实时室内三维目标检测

DanilaRukhovich Anna Vorontsova Anton Konushin

三星研究院

摘要

近年来，稀疏三维卷积改变了三维目标检测领域。三维卷积神经网络的表现可与基于投票的方法相媲美，同时内存效率更高，也更适用于大型场景。不过，这一领域仍有改进空间。我们以务实、注重实践的方式分析此类方法的性能，并找出其弱点。针对发现的问题逐一进行改进后，我们提出了 TR3D：一种快速的全卷积三维目标检测模型，可通过端到端方式训练，并在标准基准数据集 ScanNet v2、SUN RGB-D 和 S3DIS 上取得最先进的结果。此外，为了兼顾点云和 RGB 输入，我们引入了二维与三维特征的早期融合。我们采用融合模块将传统三维目标检测方法扩展为多模态方法，并展示了显著的性能提升。我们将采用早期特征融合的模型称为 TR3D+FF，该模型在 SUN RGB-D 数据集上优于现有的三维目标检测方法。总的来说，TR3D 和 TR3D+FF 模型不仅准确，而且轻量、内存高效、速度快，标志着实时三维目标检测迈出了又一个里程碑。代码可在 https://github.com/SamsungLabs/tr3d 获取。

索引词— 三维目标检测、室内场景理解、点云

1. 引言

近期，自动驾驶、增强现实/虚拟现实应用、三维建模和家用机器人技术的兴起，使三维目标检测这一核心场景理解技术受到关注。

现代三维目标检测方法可分为基于投票的方法、基于 Transformer 的方法和三维卷积方法。基于投票的方法通过特征提取网络处理点，并利用中心投票生成目标提议，再累积各组中的点特征。许多基于投票的方法可扩展性较差，限制了其应用。

基于 Transformer 的方法不再依赖领域特定启发式方法和超参数，而是采用端到端学习，并在推理时执行前向传播。尽管泛化能力更强，但在处理大型场景时仍存在问题。

三维卷积方法将点云表示为体素，从而能够高效处理稀疏三维数据。稠密体积特征需要大量内存，因此转而采用稀疏表示和稀疏三维卷积。与其他方法相比，三维稀疏卷积方法内存效率更高，并且能够很好地扩展到大型场景，同时不会牺牲点密度。直到最近，这类方法的准确率仍然不足，但随着该领域近期的进展，研究人员开发出了快速、可扩展且准确的方法 [1]。

总体而言，现代三维检测方法仅使用几何输入就取得了令人瞩目的结果；与此同时，研究人员也探索了利用其他模态的数据进行三维目标检测的可能性。点云难以获取且需要额外设备，而 RGB 相机则容易得多。RGB 相机通常集成在采集设备中，能够提供廉价、易用且信息极其丰富的数据。现有方法在后期阶段加入 RGB数据；这些方法要么采用复杂且耗费内存的架构 [2], ，依赖限制其应用的定制流程 [3], ，要么运行缓慢的迭代方案 [4]。与之相反，我们提出了一种早期融合策略，该策略也可以集成到其他基于点云的模型中。结合所提出的简单而高效的全卷积流水线，该策略能够在基于点云和 RGB数据的三维目标检测中取得最先进的结果。

2. 相关工作

点云中的三维目标检测。 基于投票的方法开创了这一领域，VoteNet [5]是首个将点投票引入三维目标检测的方法。VoteNet 从三维点中提取特征，根据各点投票得到的中心将点分配到每个目标候选项，并从每个点组中计算目标特征。BRNet [6]利用投票中心中的代表点来优化投票结果，从而更好地捕获精细的局部结构特征。H3DNet [7] 通过预测一组混合几何基元，改进了点组生成流程。RBGNet [8]提出了一种基于射线的特征分组模块，该模块从簇中心均匀发射射线，以聚合目标表面上的逐点特征。

基于 Transformer 的方法。在基于 Transformer 的方法中，分组不会显式地由一组超参数引导，因此这类方法的领域特异性较低。GroupFree

[9]采用 Transformer 模块迭代更新目标查询位置并累积中间结果。3DETR [10]率先使用 Transformer 模型解决三维目标检测任务。

基于体素的方法。 基于体素的三维目标检测方法 [1, 11, 12] 将点转换为体素，并使用三维卷积网络进行处理。然而，稠密的体积特征仍会消耗大量内存。GSDN [12] 通过稀疏三维卷积缓解了这一问题，但准确率明显低于现代基于投票的方法。与 GSDN 不同， [12], FCAF3D [1] 不使用锚框，而是以数据驱动方式处理三维目标检测，成为首个性能可与最先进的基于投票的方法相媲美的基于体素的方法。在我们的工作中，我们使用 FCAF3D 作为强基线。

点云与 RGB 融合。 图像中包含的语义数据可能为三维目标检测提供额外线索。近期工作 [13, 14, 15]以 RGB 特征为引导生成初始区域提议，然后对其进行细化。I mVoteNet [13] 利用二维检测结果进行投票。EPNet ++ [15] 将图像特征与特征提取模型的中间输出进行融合。MMTC [3]采用级联结构，其中在三维目标检测网络的第一阶段和第二阶段之间加入了二维分割网络。TokenFusion [4] 使用 Transformer 模型融合点云与 RGB 信息，并学习用投影和聚合后的跨模态特征替换无信息量的令牌。

3. 提出的方法

TR3D 基于 FCAF3D [1]，继承了其简洁的全卷积设计（图 1），同时进行了一些更新，以提升其效能和准确率。此外，我们针对 RGB 和点云的三维目标检测提出了早期特征融合，并在 TR3D 中引入融合模块，构建多模态 TR3D+FF 模型。

3.1. TR3D：三维目标检测方法

以 FCAF3D [1] 作为基线，我们进行了若干修改。TR3D 是联合应用所有这些修改后的结果（图 1）。为确保这些修改以预期方式促进最终性能，我们逐步、一次一个地应用这些修改，并评估其在 S3DIS 上的检测准确率、内存占用和推理速度（以下简称 FPS；实际上指每秒处理的场景数量）。我们在表 1 中报告了这一系列实验的结果。

效果。 首先，我们旨在将基线模型改造成快速模型且更加轻量。性能检查显示，检测头中第一层级的单个生成式转置卷积层消耗了总内存的三分之一。因此，我们移除了第一级的检测头：这不仅显著降低了内存占用——

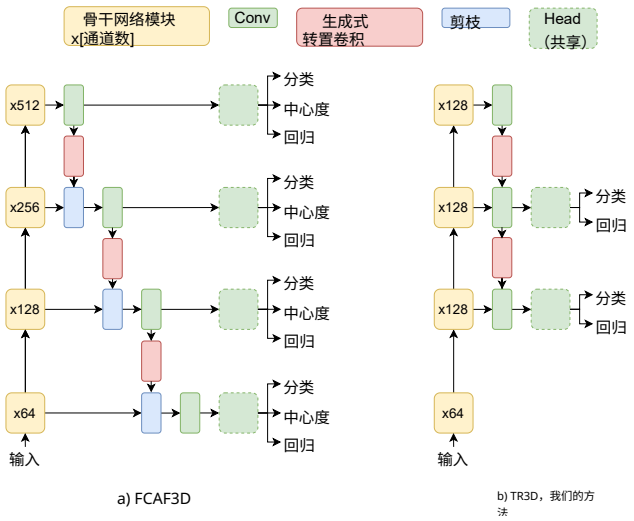

图 1：TR3D 与 FCAF3D 基线的方案对比。

从 661 Mb 降至 415 Mb，减少了 1.5 倍，还增加了 +6 FPS。移除该检测头后，剪枝层似乎变得多余，因此我们也将其省略。接着，我们移除了第四级的检测头，因为它负责处理大型物体，而这类物体在室内场景中很少见。此外，如果限制骨干网络模块中的输出通道数量，参数数量便会从 68.3 M 大幅降至 14.7 M，内存消耗也减半。

修改	mAP	FPS	#参数量、M	内存、Mb
基线 [1]	66.7 (64.9)	10.9	70.5	661
- 第1层检测头	62.9(61.3)	16.	70.1	415
- 剪枝	62.9 (61.3)	17.3	70.1	415
- 第4层检测头	61.5(59.6)	17.3	68.3	408
- >128通道	61.2（58.8）	20.8	14.7	207
- 中心度	61.5（59.8）	21.0	14.7	207
+ TR3D分配器	72	.0	14.7	207
+ DIoU 损失	**74.5（72.1）**	**21.0**	**14.7**	**207**

表1：TR3D 各组件在 S3DIS 上的研究结果。蓝色单元格表示最显著的提升。我们的方法内存占用减少了 3 倍，参数数量减少了 4.5 倍，推理速度几乎快了 2 倍。

准确率。 在第二轮改进中，我们主要关注检测准确率。我们的实验表明，中心度对预测质量没有贡献，因此可以直接省略，不会产生不良影响。FCAF3D 分配器仅考虑 3D 边界框内的点。因此，这可能导致遗漏细长或较小的物体（例如白板），因为这些物体可能位于各个位置之间，因而无法与真实框匹配。因此，我们引入了 TR3D 分配器，考虑的是最近的点，而非框内的点——

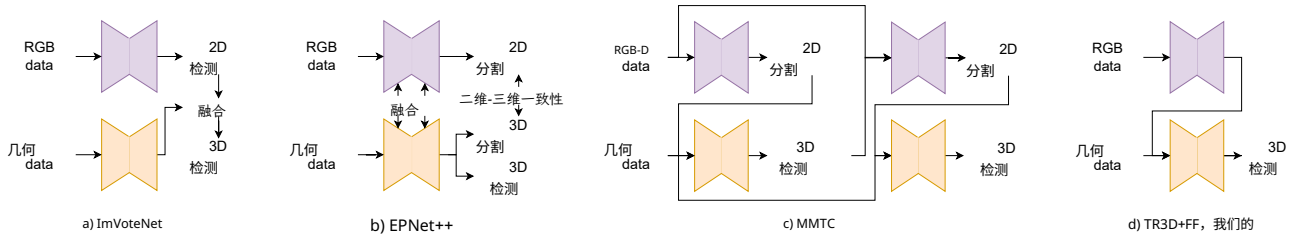

图 2：我们早期的特征融合策略与现有的 RGB 和点云融合方法：ImVoteNet [13], EPNet++[15]、MMTC[3]。凭借简单直接的设计，它在检测准确率方面似乎更有优势。

这些点也可能位于边界框之外。此外，我们为每个物体类别预先指定检测头层级：通常，大型物体（例如床 或沙发）在第三层处理，而较小的物体（例如椅子 或床头柜）则在第二层处理。改用新的多层分配器后，性能从 61.5 mAP 提升至 72.9 mAP。采用新型分配器后，分配的点可能位于真实值边界框之外，因此交并比可能为零。为确保训练过程正常进行，我们用 DIoU 损失替代 IoU损失，成功解决了这类情况。最终这项改进使模型以 21 FPS 达到 74.5 mAP；模型轻量，包含 14.7M 个参数，峰值内存消耗为 207 Mb。总体而言，TR3D 的内存消耗比基线低 3 倍，参数少 4.5 倍，处理时间缩短至一半。

3.2. TR3D+FF：RGB 与点云融合

TR3D采用早期二维-三维特征融合。首先，使用经过预训练以完成二维目标检测任务的冻结 ResNet50+FPN 网络处理 RGB 图像。然后，与 [2] 中相同，将提取的二维特征投影到三维空间。最后，将投影后的二维特征与三维特征逐元素相加，实现融合。与现有方法相比，这种策略明显更加简单，也更节省时间和内存（图 2）；令人意外的是，它还能取得更好的结果。

我们将融合模块集成到 VoteNet [5] 和 TR3D 中，并在基于点云和多模态的设置下评估其性能。此外，我们将 VoteNet+FF 与 ImVoteNet [13]进行比较，后者在 VoteNet [5]的基础上实现了另一种特征融合策略。根据表 2 中给出的指标，我们的早期特征融合使 VoteNet 的检测准确率提升了 +6.8 mAP@0.25，而 ImVoteNet 的提升较小，为 5.7 mAP@0.25。同时，通过特征融合，TR3D 的检测准确率在 mAP@0.25 和 mAP@0.5 指标上分别提升了 +2.3 和 +3.0。这一细微但稳定的提升表明，几何数据是主要信息来源，而视觉数据则起到补充作用，可引导调整预测结果。

方法	输入	mAP@0.25	mAP@0.5
VoteNet[5]	PC	57.7	-
ImVoteNet[13]	点云+RGB	63.4	-
VoteNet+FF，我们的方法	点云+RGB	64.5 (63.7)	39.2 (38.1)
TR3D，我们的方法	PC	67.1 (66.3)	50.4 (49.6)
TR3D+FF，我们的方法	点云+RGB	**69.4** (68.7)	**53.4** (52.4)

表2：在 SUN RGB-D 上，基于点云的三维目标检测方法与采用我们早期融合的多模态方法的结果。

4. 实验

4.1. 实验设置

数据集。 实验在 SUN RGB- D [16], ScanNet v2 [17], 和 S3DIS [18]上进行。ScanNet v2 [17] 包含 1513 个重建扫描，并带有丰富标注：其中 1201 个构成训练子集，剩余 312 个用于验证。按照文献[13]提出的方法，我们根据逐点语义标注计算 18 个物体类别的轴对齐边界框。SUN RGB-D [16] 是一个包含 10 355 张 RGB-D图像的单目数据集。训练集和验证集分别包含 5285 个和 5050 个点云，并标注了方向边界框。按照文献 [5]提出的方法，我们选取最常见的 10 个物体类别进行评估。S3DIS[18]包含 6 个大型区域中的 271 个场景。按照标准评估协议，我们使用区域 5中的扫描数据评估检测准确率，涵盖 5 个语义类别。

指标。对于所有数据集，我们使用交并比阈值为 0.25 和 0.5 时的平均精度均值（mAP）作为评估指标。为消除随机性导致的离群值并获得具有统计显著性的结果，我们将模型训练五次，并对每个训练后的模型分别评估五次。为了与采用相同流程的现有方法进行公平比较，我们报告每项指标在 25 次试验中的最佳值和平均值（括号内）。

实现细节。我们的模型基于 mmdetection3d 框架 [20]实现，并在单张 NVidia 4090 GPU 上训练和测试。我们遵循 FCAF3D [1], 的训练流程，使用相同的损失函数、优化器、学习

方法	发表于	ScanNet			SUN RGB-D			S3DIS		
		mAP@0.25	mAP@0.5	FPS	mAP@0.25	mAP@0.5	FPS	mAP@0.25	mAP@0.5	FPS
VoteNet[5]	2019年国际计算机视觉大会	58.6	33.5	14.1	57.7	-	24.9	-	-	-
GSDN[12]	ECCV 2020	62.8	34.8	-	-	-	-	47.8	25.1	-
BRNet[6]	CVPR 2021	66.1	50.9	<14	61.1	43.7	<24	-	-	-
3DETR[10]	ICCV 2021	65.0	47.0	<7	59.1	32.7	<12	-	-	-
H3DNet[7]	ECCV 2020	67.2	48.1	7.9	60.1	39.0	<24	-	-	-
GroupFree [9]	ICCV 2021	69.1 (68.6)	52.8(51.8)	8.3	63.0 (62.6)	45.2 (44.4)	<24	-	-	-
RBGNet [8]	CVPR 2022	70.6 (69.9)	55.2(54.7)	<14	64.1 (63.6)	47.2 (46.3)	<24	-	-	-
HyperDet3D [19]	CVPR 2022	70.9	57.2	<14	63.5	47.3	<24	-	-	-
FCAF3D [1]	ECCV'22	71.5 (70.7)	57.3(56.0)	15.7	64.2 (63.8)	48.9 (48.2)	17.9	66.7(64.9)	45.9 (43.8)	10.9
TR3D, ours	-	**72.9** (72.0)	**59.3**(57.4)	**23.7**	**67.1** (66.3)	**50.4** (49.6)	**27.5**	**74.5** (72.1)	**51.7** (47.6)	**21.0**

表 3：仅基于几何输入的 TR3D 与现有三维目标检测方法的结果。TR3D 的表现优于强基线 FCAF3D 在所有基准数据集上。

率调度方案和数据增强方法。

4.2. 与先前工作比较

基于点云的三维目标检测。 我们在表 3 中报告了 TR3D 在 ScanNet v2、SUN RGB-D 和 S3DIS 上的定量结果。可以看出，TR3D 在所有基准数据集上、就各项指标而言均展现出显著优势。最明显的准确率提升出现在 S3DIS 上，差异达到 +7.8 mAP@0.25 和 +5.8 mAP@0.5。总的来说，我们认为，仅使用几何输入的方法使我们在室内三维目标检测方面达到了新的最先进水平。

多模态三维目标检测。 表 3 列出了包括 TR3D+FF 在内的多模态方法在 SUN RGB-D 上的结果。TR3D+FF 超越了先前最先进水平的 MMTC [3]，在 mAP@0.25 和 mAP@0.5 上分别提升了 4.1 和 4.8。

方法	发表于	mAP@0.25	mAP@0.5	FPS
ImVoteNet [13]	CVPR'20	63.4	-	14.8
EPNet [14]	ECCV 2020	64.6	-	-
TokenFusion [4]	CVPR 2022	64.9（64.4）	48.3（47.7）	-
EPNet++ [15]	TPAMI'22	65.3	-	-
MMTC [3]	BMVC'21	65.3(64.7)	48.6 (48.2)	<7
TR3D+FF，我们的方法。		**69.4** (68.7)	**53.4**(52.4)	**17.5**

表 4：在 SUN RGB-D 上，使用点云和 RGB 的 TR3D+FF 与现有 3D 物体检测方法的结果。TR3D+FF 是最先进的多模态 3D 物体检测方法。

4.3. 定性结果

图3展示了输入点云、真实标注和预测边界框。

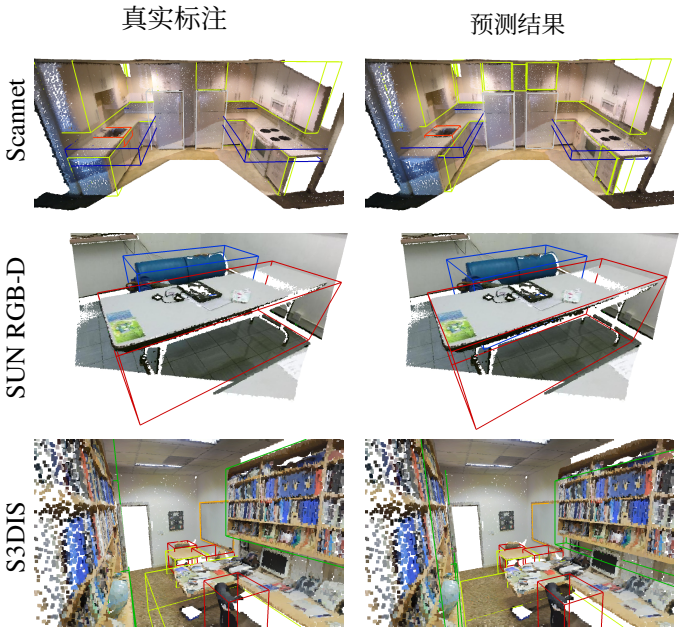

图2：ScanNet、SUN RGB-D 和 S3DIS 点云中的真实标注物体以及 TR3D 检测到的物体。

5. 结论

本文介绍了一种新颖的三维目标检测方法 TR3D。此外，我们开发了一种早期融合策略，将视觉特征融入三维处理流程，并提出了 TR3D 的改进版本 TR3D+FF，该方法同时利用点云和 RGB 输入。我们在标准基准数据集 ScanNet v2、SUN RGB-D 和 S3DIS 上评估了所提出的方法。实验结果表明，TR3D 在准确率和速度方面均优于现有方法，从而在基于点云的三维目标检测领域达到新的最先进水平；而 TR3D+FF 在使用 RGB 和点云的方法中取得了最佳结果。

6. 参考文献

[1]Danila Rukhovich、Anna Vorontsova 和 Anton Konushin，《Fcaf3d：全卷积无锚框三维目标检测》，载于 *IEEE/CVF* 欧洲计算机视觉会议（*ECCV*）。Springer，2022 年，第 477–493 页。

[2]Danila Rukhovich、Anna Vorontsova 和 Anton Konushin，《Imvoxelnet：面向单目和多视图通用三维目标检测的图像到体素投影》，*arXiv*，2021。

[3] Jinhyung Park、Xinshuo Weng、Yunze Man 和 Kris Kitani，《用于三维目标检测的多模态任务级联》，发表于英国机器视觉会议（*BMVC*），2021。

[4] Yikai Wang、Xinghao Chen、Lele Cao、WenbingHuang、Fuchun Sun 和 Yunhe Wang，《面向视觉变换器的多模态令牌融合》，发表于*IEEE* 计算机视觉与模式识别会议（*CVPR*），2022，第12186–12195页。

[5] Charles R Qi、Or Litany、Kaiming He 和 Leonidas J Guibas，《用于点云三维目标检测的深度霍夫投票》，发表于 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2019，第9277–9286页。

[6] Bowen Cheng、Lu Sheng、Shaoshuai Shi、Ming Yang 和 Dong Xu，《面向点云中基于投票的三维目标检测的代表点回溯》，载于*IEEE* 计算机视觉与模式识别会议（*CVPR*），2021 年，第 8963–8972 页。

[7] Zaiwei Zhang、BoSun、Haitao Yang 和 Qixing Huang，《H3DNet：使用混合几何基元进行三维目标检测》，载于 *IEEE/CVF* 欧洲计算机视觉会议（*ECCV*），2020 年，第 311–329 页。

[8] Haiyang Wang、Shaoshuai Shi、Ze Yang、Rongyao Fang、Qi Qian、Hongsheng Li、Bernt Schiele 和 Liwei Wang，《RBGNet：用于三维目标检测的基于射线的分组》，载于*IEEE* 计算机视觉与模式识别会议（*CVPR*），2022 年，第 1110–1119 页。

[9]Ze Liu、Zheng Zhang、Yue Cao、Han Hu 和 Xin Tong，“通过变换器进行无分组式三维目标检测”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*），2021年，第2949–2958页。

[10] Ishan Misra、Rohit Girdhar 和 Armand Joulin，“用于三维目标检测的端到端Transformer模型”，载于*IEEE/CVF*计算机视觉国际会议（*ICCV*），2021年，第2906–2917页。

[11] Ji Hou、Angela Dai 和 Matthias Nießner，《3D-SIS：RGB-D 扫描的 3D 语义实例分割》，载于 *IEEE Conf*会议 *C*计算机 *Visi*在……上 *d P*模式 *R*识别 *i -tion*（计算机视觉与模式识别会议），2019，第 4421–4430 页。[12]JunYoung Gwak、Christopher Choy 和 Silvio Savarese，《用于三维单次目标检测的生成式稀疏检测网络》，载于 *IEEE/CVF* 欧洲计算机视觉会议（*ECCV*）。Springer，2020，第 297–313 页。

[13] Charles R Qi、Xinlei Chen、Or Litany 和 Leonidas J Guibas，《ImVoteNet：利用图像投票提升点云中的三维目标检测》，载于*IEEE* 计算机视觉与模式识别会议（*CVPR*），2020，第 4404–4413 页。

[14] Tengteng Huang、Zhe Liu、Xiwu Chen 和 Xiang Bai，《EPNet：利用图像语义增强点特征以进行三维目标检测》，载于 *IEEE/CVF* 欧洲计算机视觉会议（*ECCV*），Andrea Vedaldi、Horst Bischof、Thomas Brox 和 Jan-Michael Frahm 编，2020，第 35–52 页。

[15]Zhe Liu、Tengteng Huang、Binglin Li、Xiwu Chen、Xi Wang 和 Xiang Bai，“EPNet++：用于多模态三维目标检测的级联双向融合”，*arXiv*，第 abs/2112.11088 卷，2021 年。

[16]Shuran Song、Samuel P Lichtenberg 和 Jianxiong Xiao，“SUN RGB-D：RGB-D 场景理解基准套件”，载于 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2015 年，第 567–576 页。

[17]Angela Dai、Angel X. Chang、Manolis Savva、Maciej Halber、Thomas Funkhouser 和 Matthias Nießner，“ScanNet：带有丰富标注的室内场景三维重建”，载于 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2017 年。

[18] Iro Armeni、Ozan Sener、Amir R. Zamir、Helen Jiang、Ioannis Brilakis、Martin Fischer 和 Silvio Savarese，“大规模室内空间的三维语义解析”，载于 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2016 年，第 1534–1543 页。

[19] Yu Zheng、Yueqi Duan、Jiwen Lu、Jie Zhou 和 Qi Tian，“Hyperdet3D：学习一种场景条件化三维目标检测器”，载于 *IEEE* 计算机视觉与模式识别会议（*CVPR*），2022 年，第 5575–5584 页。

[20]MMDetection3D贡献者，“MMDetection3D：OpenMMLab 新一代通用三维目标检测平台”，https://github.com/ open-mmlab/mmdetection3d，2020 年。