清华大学自动化系

**专业实践技术报告**

题目：SparseInteraction++: 用于激光雷达、雷达和摄像头3D目标检测的稀疏语义引导

班 级： 自11

姓 名： 李昭阳

学 号： 2021013445

实践地点： 北京市海淀区西二旗中路33号院

实践单位： 北京小米科技软件有限责任公司

实践部门： 汽车部-自动驾驶部

指导教师： 徐少清

2024年 9 月 5 日

目 录

[一、课题背景及需求分析 2](#_Toc176457040)

[1.1 概述 2](#_Toc176457041)

[1.2 背景与分析 2](#_Toc176457042)

[1.3 工作贡献 3](#_Toc176457043)

[二、算法实现 3](#_Toc176457044)

[2.1 模态特征提取器 4](#_Toc176457045)

[2.2 噪声融合滤波器 4](#_Toc176457046)

[2.3 检测头与损失函数： 6](#_Toc176457047)

[三、实验结果分析 6](#_Toc176457048)

[3.1 实验设置： 6](#_Toc176457049)

[3.2 实验结果： 6](#_Toc176457050)

[四、总结与展望 7](#_Toc176457051)

[参考文献 7](#_Toc176457052)

一、课题背景及需求分析

**1.1 概述**

多模态融合技术（例如结合主动测距传感器与图像数据）在环境感知中提供了一种互补且具有成本效益的方法，无论在何种光照与天气条件下都能有效运行。然而，当前针对环视图像与雷达的融合方法常常受到雷达数据固有噪声与位置模糊性的影响，导致性能显著下降。为克服这些挑战，我们提出了一种鲁棒的端到端融合框架，称为SparseInteraction++。我们首先引入了Noisy Fused Filter（NFF）模块，该模块利用从图像中查询到的语义特征来过滤噪声数据，从而提取准确的前景特征。此外，我们实现了Sparse Cross-Attention Encoder（SCAE），以在稀疏层面上有效融合雷达、激光雷达与图像特征，解决位置模糊问题。为进一步增强模型的收敛性与性能，我们引入了包含融合前景位置信息的前景先验查询，并将其与预定义查询连接后输入到后续基于Transformer的解码器中。实验结果表明，这些融合策略显著提升了检测性能，在nuScenes基准测试上建立了新的性能标杆。

**1.2 背景与分析**

感知3D障碍物是计算机视觉和机器人领域的基础任务，通常需要使用多种传感器。LiDAR与摄像头技术的结合在性能上表现出高精度[1][2][3][4]。然而，摄像头在能见度低的条件下（如大雾或大雨）表现不佳，而仅依赖LiDAR又成本高昂，这为感知任务增加了额外的挑战。

雷达传感器因其成本低廉、在各种天气条件下的稳健性以及在长距离上准确估计速度的能力而脱颖而出[5][6][7]。然而，直接使用主动测距传感器的数据面临噪声和位置不确定性等挑战。此外，单个雷达的有限范围使得捕捉全面的全局特征变得困难。

近期研究通过将图像特征投影到鸟瞰图（BEV）并将其与雷达数据融合，显示出了有前景的结果。特别是，CRN[8]通过基于Transformer的架构，将图像的BEV特征与雷达数据融合。这种融合有助于解决雷达固有的位置模糊性，从而实现更准确、可靠的感知。然而，在融合前存在的雷达数据噪声导致了次优的特征融合效果。

为了滤除雷达信息中的噪声，CramNet[9]（如图1-(a)所示）采用了一种直接方法，通过雷达特征本身来消除噪声雷达数据。另一种方法，如图1-(b)所示的HVDetFusion[10]，则利用图像的检测边界框来过滤噪声。而如图1-(c)所示的MVFusion[11]，通过将3D雷达点投影到2D图像像素上，利用图像的前景分类来过滤无关数据。然而，这些方法的共同局限性在于它们在过滤前过于依赖单一模态的特征，往往忽视了多模态融合的潜在优势，导致结果不够理想。

**1.3 工作贡献**

在本文中，我们提出了一种基于Transformer的鲁棒端到端融合框架，称为SparseInteraction++。为了解决雷达数据噪声对特征融合的负面影响，我们设计了Noisy Fused Filter模块，该模块包括语义优化传感器编码器（Semantic-Optimized Sensor Encoder，SSE）和前景检测掩码（Foreground Detection Mask，FDM），利用图像语义数据引导来保留有用的雷达信息。随后，我们利用稀疏前景雷达特征及其对应的图像特征区域，通过稀疏交叉注意力编码器（Sparse Cross-Attention Encoder，SCAE）增强多模态特征交互，有效缓解特征错位问题。此融合方法不仅解决了主动测距传感器的位置模糊性问题，还能在传感器失效时提供稳健的结果。此外，我们从前景融合特征和增强的鸟瞰图（BEV）融合特征中高效提取前景目标查询。最终，将包含前景雷达位置信息的前景先验查询与预定义查询结合后输入到基于Transformer的解码器中，这一过程有助于模型收敛和性能提升。总体而言，通过融合来自雷达、激光雷达和摄像头的特征，我们的框架在3D目标检测中取得了令人印象深刻且稳健的结果。总结来说，本文的贡献如下：

* 我们引入了一种新颖的Noisy Fused Filter模块，该模块通过使用图像语义和雷达特征作为引导，高效解决了误报问题，过滤掉了无关信息。
* 为有效应对雷达位置模糊性问题，我们提出了一种基于稀疏查询的模块——稀疏交叉注意力编码器（Sparse Cross-Attention Encoder），通过将高质量的3D目标先验融入自适应3D查询，促进了多模态表示的交互。
* 我们的SparseInteraction++框架在nuScenes数据集上采用Transformer技术实现了最先进的性能，所提出的前景融合过滤方法预计将激发该领域的未来研究。

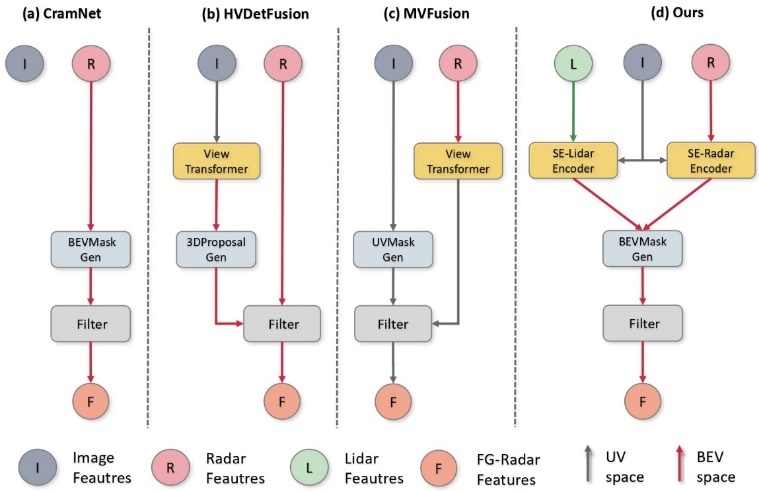


Figure 1不同设计流程的过滤噪声数据比较

二、算法实现

首先，我们从雷达点云和多视角图像中提取特征。为处理噪声数据，我们引入了Noisy Fused Filter模块，该模块将雷达特征与来自图像的高层语义特征相结合，以过滤掉无关信息。随后，增强后的雷达和图像特征被输入到稀疏交叉注意力融合编码器中，该编码器采用一种可变形注意力机制[14]。最后，前景融合特征中的先验信息被编码为前景先验查询，这些查询与预定义查询连接后，作为目标查询输入到Transformer解码器层中。图2展示了我们提出的SparseInteraction++框架的概述。

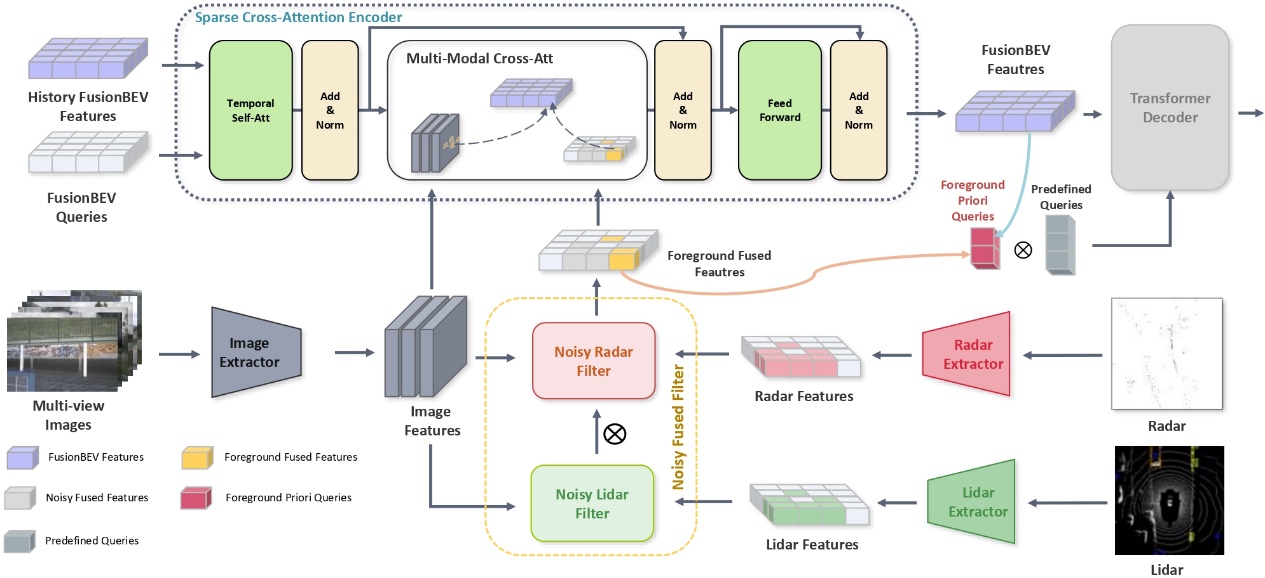
**2.1 模态特征提取器**

1. **图像特征提取器：**参考BEVFormer[12]，我们采用ResNet101[15]作为骨干网络，从多个摄像头视角中提取图像特征。在此骨干网络的基础上，我们应用标准的特征金字塔网络（FPN）[16]，记为，其中，i表示每个摄像头视角。
2. **雷达特征提取器：**由于雷达在提供可靠的高程数据方面存在局限性，我们对雷达点云使用柱状表示方法。参考FUTR3D[13]中的方法，我们采用PillarNet[17]结合多层感知机（MLP）来提取鸟瞰图（BEV）空间中的雷达特征。这些特征表示为，其中C表示雷达特征的维度，(H,W)对应鸟瞰图的分辨率。
3. **激光雷达特征提取器：**激光雷达特征提取过程与雷达特征提取类似。我们对激光雷达点云采用类似的柱状表示方法，利用PillarNet架构结合MLP来提取鸟瞰图（BEV）空间中的特征。这些激光雷达特征表示为，其中C表示激光雷达特征的维度，(H,W)对应鸟瞰图的分辨率。

**2.2 噪声融合滤波器**

主动测距传感器，如雷达和激光雷达，在捕捉场景中物体的速度和位置方面非常有效。然而，它们往往会无意中记录非目标元素的数据，导致误报问题。为了解决这一挑战，我们引入了一种新颖且轻量级的模块，称为噪声融合滤波器（Noisy Fused Filter，NFF）。该模块专门设计用于减轻无关或噪声数据对检测准确性的影响。

NFF模块由两个主要组件组成：噪声雷达滤波器和噪声激光雷达滤波器。每个组件又进一步分为两个关键子模块：**语义优化传感器编码器**和**前景检测掩码**。语义优化传感器编码器负责通过将图像中提取的高层语义信息与原始雷达和激光雷达数据进行整合，从而丰富这些数据。这一过程有助于更好地区分目标物体和非目标元素。同时，前景检测掩码在隔离相关特征方面起着关键作用，通过识别和过滤掉那些被认为无关或可能导致误报的特征来实现这一目的。

****与CramNet[9]和HVDetFusion [10]等以往的方法不同，我们的NFF模块采用了一种独特的方法来过滤无关的融合特征。通过利用雷达、激光雷达和图像语义的结合优势，NFF模块能够有效抑制可能导致错误检测的噪声数据，从而得到更加干净和准确的特征集，这些特征将用于后续的检测任务。

1. **语义优化传感器编码器**：为了减轻噪声数据对检测过程的影响，我们采用了一种创新的方法，利用中的非空特征作为查询，聚合邻近特征。这种聚合由从图像中提取的语义信息引导，有助于优化特征并抑制无关数据。我们面临的一个重大挑战是雷达和激光雷达点的固有稀疏性。这种稀疏性导致编码器输出中存在大量空洞或空白特征，这可能会降低融合数据的质量。为了解决这个问题，我们采用了一个3x3的无偏置卷积层，专门用于将前景融合特征分布到这些空白区域。通过这样做，我们显著增强了雷达和激光雷达数据的整体密度和丰富性。这一操作有效填补了稀疏点云所造成的空白，确保在后续阶段使用的特征更加稳健和信息丰富。随后，丰富的融合特征与对应的图像特征对齐，生成语义优化特征，记为。这一对齐过程至关重要，因为它结合了传感器数据和图像语义的优势，从而生成更准确且具有上下文感知的特征表示。鉴于主动测距传感器通常提供的高度信息存在不确定性，我们采用了一种策略，即将每个在位置(x, y)处的有效雷达或激光雷达特征像素提升到具有不同高度的个三维点。这些三维点随后被投影到第i个图像的对应像素位置。这一投影步骤使我们能够在空间上将雷达或激光雷达特征与图像特征对齐，确保最终的语义优化特征准确反映场景的三维结构。最后，通过应用可变形注意力[14]进一步优化有效的雷达或激光雷达特征。这种高级的注意力机制与图像特征协同工作，使模型能够动态地聚焦于输入数据中最相关的部分。可变形注意力机制适应场景中物体的不同形状和大小，从而提高特征优化过程的精度。整个优化过程，包括特征的对齐和优化，可以用数学表达如下：

Figure 2 SparseInteraction++的整体架构

其中， ​是位于 (x,y) 处的语义优化特征像素。表示三维点 在第 i 个摄像头图像坐标系中的投影点

1. **前景检测掩码**：鉴于点云往往在物体周围聚集，因此必须解决雷达和激光雷达传感器在定位方面固有的模糊性。这种模糊性通常导致在3D边界框内的点分布稀疏，仅捕捉到有限数量的点，这些点可能无法完全代表物体的边界。然而，我们的观察表明，这些点，尤其是位于边界框边缘附近的点，仍然可以为检测过程提供有价值的信息。尽管这些边界点数量较少，但它们携带的关键空间线索可以提升整体目标检测的准确性。基于此，我们提出了一种新颖的方法，旨在最大化这些边界点的贡献。具体而言，我们将从真实标签中得出的3D边界框扩大1.5倍。此扩展使我们能够包括可能位于原始边界之外的附加点，确保不会忽略潜在的有用数据。然后，将扩展后的3D边界框投影到BEV中，作为二值掩码地面真值，记为。这个二值掩码作为区分BEV中前景和背景区域的参考，这对于精确的目标检测至关重要。

接下来，需要对前一阶段生成的语义优化特征进行分类。为此，我们采用了1x1卷积层，这在将高维特征转化为更易于管理的形式进行分类方面特别有效。该层与Sigmoid激活函数结合使用，生成每个特征的概率分数，表示其属于前景的可能性。选择Sigmoid函数是因为它能够将输出映射到0和1之间的范围，这使其非常适合二元分类任务。为了提高模型的召回率，我们使用了一个较低的前景阈值γ。这一较低的阈值至关重要，因为它允许模型包括那些可能置信度较低但仍可能属于前景的特征。通过这样做，我们减少了遗漏可能对准确检测至关重要的相关目标或特征的风险。因此，前景分数超过阈值γ的雷达特征被指定为前景雷达特征。这些特征被视为关键，并被传递到模型的后续处理阶段。相反，分数低于阈值γ的雷达特征被分类为背景雷达特征。这些背景特征通常对检测任务的重要性较低，可能会在模型中被过滤掉或以不同的方式使用。通过这种方法，我们能够有效利用雷达和激光雷达传感器提供的空间信息，特别是那些位于物体边界附近的点，从而提高检测的准确性。通过扩大3D边界框并仔细分类特征，我们确保模型能够捕捉到场景的更全面表示，从而在挑战性检测场景中获得更好的性能。

**2.3 检测头与损失函数：**

我们使用了类似于 Deformable DETR[14] 的检测头，该检测头直接输出物体类别概率和 3D 检测框，无需进行 NMS 后处理。整体损失函数由两个部分组成：掩码损失 ​ 和检测损失 ​。掩码损失 旨在优化前景掩码模块中的二元分类，而检测损失则专注于优化 3D 检测头。

考虑到前景融合特征的有限性，我们在训练前景掩码模块时应用了焦点损失（focal loss）。为了与 BEVFormer 保持一致，我们同样在 3D 检测头中使用焦点损失进行分类，并使用 L1 损失进行边界框回归。

三、实验结果分析

**3.1 实验设置：**

我们使用nuScenes数据集[21]来评估我们的方法，这是一个全面的自动驾驶研究资源，包括700个场景用于训练，150个用于验证，以及150个用于测试。该数据集提供了360度全景视角，并结合了各种传感器的数据：六个摄像头、一个激光雷达和五个雷达。对于nuScenes检测任务，性能通过两个主要指标进行评估：平均精度均值（mAP）和nuScenes检测分数（NDS）。mAP指标是通过测量预测对象和地面真实对象在鸟瞰图（BEV）中的中心距离来计算的，平均在0.5、1、2和4米的阈值距离上。NDS指标将mAP与其他真实正指标结合起来，包括平均平移误差（mATE）、平均尺度误差（mASE）、平均方向误差（mAOE）、平均速度误差（mAVE）和平均属性误差（mAAE），通过加权平均计算。

**3.2 实验结果：**

为了评估我们的SparseInteraction++方法的性能，我们在nuScenes[37]验证集上进行了评估，结果详细列在表I中。我们的SparseInteraction++方法展示了明显优于先前融合技术的优越性，在验证集上达到了51.1%的mAP和59.5%的NDS。与我们的基线模型BEVFormer[12]相比，SparseInteraction++在NDS和mAP上分别取得了7.8%和9.5%的显著性能提升。此外，我们将我们的方法应用于VoD[22]数据集，结果呈现在表II中，进一步证明了我们方法的有效性。

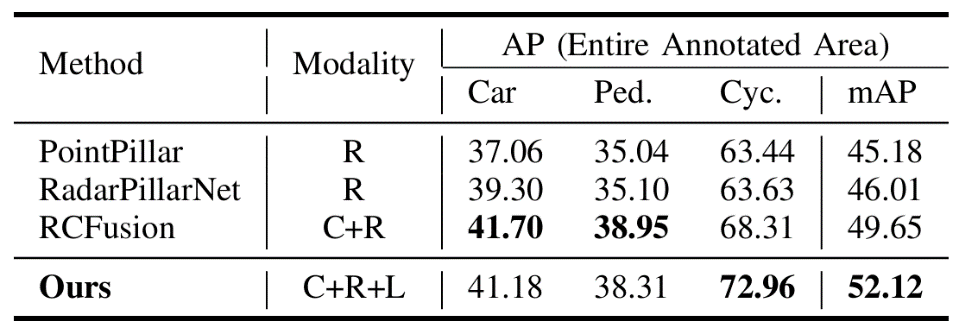


Table 2 在VoD验证数据集上的比较结果

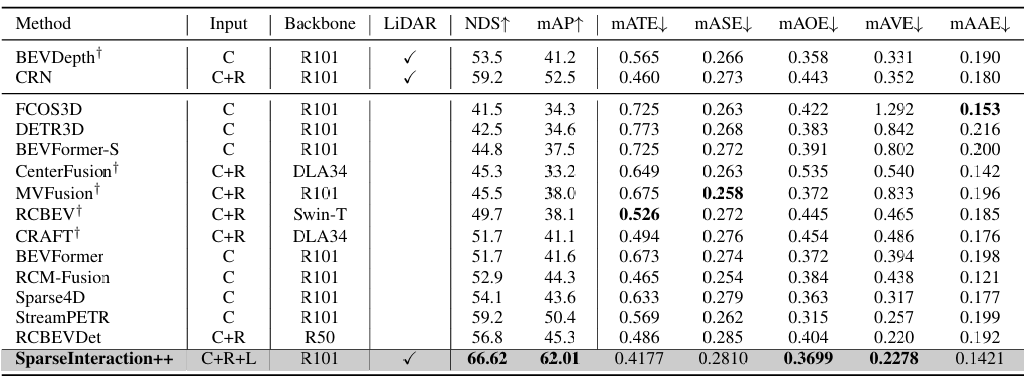


Table 1 在nuScenes验证集上的性能比较。

"C" 表示摄像头，"C+R" 表示摄像头和雷达，"C+R+L" 表示摄像头、雷达和激光雷达。""LIDAR"：使用额外的激光雷达数据源作为深度监督。

四、总结与展望

在这项研究中，我基于SparseInteraction[18]，融合激光雷达信号，设计了一种新颖的基于transformer的前景激光雷达-雷达-摄像头融合框架，该框架具有一个轻量级模块，该模块在图像语义信息的指导下提取前景融合特征。我们的方法有效地缓解了与噪声和位置歧义相关的挑战，从而实现了最先进的性能。对于未来的工作，我们打算将速度预测集成到时间模块中，以进一步提高检测精度**。**

参考文献

1. S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, “Fusion painting: Multimodal fusion with adaptive attention for 3d object detection,” in JOURNAL OF LATEX CLASS FILES, VOL. 18, NO. 9, SEPTEMBER 2020 10 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 3047–3054, IEEE, 2021.
2. S. Xu, F. Li, Z. Song, J. Fang, S. Wang, and Z.-X. Yang, “Multisem fusion: Multimodal semantic fusion for 3d object detection,” IEEE Transactions on Geoscience and Remote Sensing, pp. 1–1, 2024.
3. Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, “Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation,” in 2023 IEEE international conference on robotics and automation (ICRA), pp. 2774–2781, IEEE, 2023.
4. Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, and W. Zhan, “Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection,” arXiv preprint arXiv:2304.14340, 2023.
5. C. Cui, Y. Ma, J. Lu, and Z. Wang, “Redformer: Radar enlightens the darkness of camera perception with transformers,” IEEE Transactions on Intelligent Vehicles, 2023.
6. L. Wang, X. Zhang, W. Zeng, W. Liu, L. Yang, J. Li, and H. Liu, “Global perception-based robust parking space detection using a low cost camera,” IEEE Transactions on Intelligent Vehicles, vol. 8, no. 2, pp. 1439–1448, 2022.
7. Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, “Full-velocity radar returns by radar-camera fusion,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16198–16207, 2021.
8. Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, “Crn: Camera radar net for accurate, robust, efficient 3d perception,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17615–17626, 2023.
9. J.-J. Hwang, H. Kretzschmar, J. Manela, S. Rafferty, N. ArmstrongCrews, T. Chen, and D. Anguelov, “Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection,” in European Conference on Computer Vision, pp. 388–405, Springer, 2022.
10. K. Lei, Z. Chen, S. Jia, and X. Zhang, “Hvdetfusion: A simple and robust camera-radar fusion framework,” arXiv preprint arXiv:2307.11323, 2023.
11. Z. Wu, G. Chen, Y. Gan, L. Wang, and J. Pu, “Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion,” arXiv preprint arXiv:2302.10511, 2023.
12. Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, “Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” in European conference on computer vision, pp. 1–18, Springer, 2022.
13. X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, “Futr3d: A unified sensor fusion framework for 3d detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 172–181, 2023.
14. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.
15. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
16. T.-Y. Lin, P. Doll´ ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21172125, 2017.
17. G. Shi, R. Li, and C. Ma, “Pillarnet: Real-time and high-performance pillar-based 3d object detection,” in European Conference on Computer Vision, pp. 35–52, Springer, 2022.
18. S. Jiang, S. Xu, li fang, L. Liu, Z. Song, Y. Bo, and Z.-X. Yang,“Sparseinteraction: Sparse semantic guidance for radar and camera 3d object detection,” in ACM Multimedia 2024, 2024
19. V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3d packing for self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2485–2494, 2020.
20. Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, and Z. Li, “Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1486–1494, 2023.
21. H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1162111631, 2020.
22. I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv preprint arXiv:1711.05101, 2017.