智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于多模态融合的三维目标检测方法研究

陆军, 赵颢然, 鲁林超

引用本文:
陆军, 赵颢然, 鲁林超. 基于多模态融合的三维目标检测方法研究[J]. 智能系统学报, 2025, 20(5): 1167-1177.
LU Jun, ZHAO Haoran, LU Linchao. Research on 3D object detection based on multi-modal fusion[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1167-1177.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202502015

DOI: 10.11992/tis.202502015
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250814.1400.006

基于多模态融合的三维目标检测方法研究

陆军，赵颢然，鲁林超
（哈尔滨工程大学 智能科学与工程学院，黑龙江 哈尔滨 150001）

摘　要：在自动驾驶场景中，由于多模态的融合，三维目标检测效果易受传感器未充分校准的影响，同时，对于目标密集的复杂场景，检测过程中易对目标造成误检，从而降低模型的召回率和检测精度。针对以上问题，设计了多模态融合网络 SoftFusion-QC(softfusion with query contrast) 用以实现三维目标检测。为了自适应地融合来自激光雷达的点云数据和摄像头捕获的图像信息，提出可变形跨模态特征聚合模块 (deformable cross-modality feature aggregate, DCFA)，实现深层次的特征融合。为了有效应对传感器校准不足问题，引入查询对比机制 (query contrast, QC)，通过基于 Transformer 的查询交互策略和查询框对比学习策略，显著提升了检测的精度和鲁棒性，解决了密集目标检测的误检问题。在 nuScenes 自动驾驶数据集上，取得了 69.8% 的 mAP(mean average precision) 与 72.8% 的 NDS(normalized detection score)。通过定量的性能分析和消融实验验证了算法的有效性。

关键词：三维目标检测；多模态融合；深度学习；深度估计；特征聚合；注意力机制；激光雷达；自动驾驶
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1673−4785(2025)05−1167−11

中文引用格式：陆军，赵颢然，鲁林超. 基于多模态融合的三维目标检测方法研究 [J]. 智能系统学报，2025, 20(5): 1167–1177.
英文引用格式：LU Jun, ZHAO Haoran, LU Linchao. Research on 3D object detection based on multi-modal fusion[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1167–1177.

Research on 3D object detection based on multi-modal fusion

LU Jun, ZHAO Haoran, LU Linchao
(College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: In the context of autonomous driving, the performance of 3D object detection via multimodal fusion is susceptible to insufficient sensor calibration. Additionally, in complex scenes with dense targets, the detection process is prone to false positives, thereby reducing the model's recall and precision. To address these challenges, we have designed a multimodal fusion network, SoftFusion-QC (softFusion with query contrast), for 3D object detection. To adaptively fuse point cloud data from LiDAR with image information from cameras, we propose a Deformable cross-modality feature aggregate (DCFA) module, which facilitates deep-level feature fusion and effectively mitigates the issue of inadequate sensor calibration. To resolve the problem of false positives in dense object detection, we introduce a query contrast (QC) mechanism. By employing a Transformer-based query interaction strategy and a query box contrastive learning strategy, this mechanism significantly enhances detection accuracy and robustness. On the nuScenes autonomous driving dataset, our method achieves 69.8% mAP (mean average precision) and 72.8% NDS (normalized detection score). The effectiveness of our algorithm is validated through quantitative performance analysis and ablation studies.

Keywords: 3D target detection; multimodal fusion; deep learning; depth estimation; feature aggregation; attention mechanism; LiDAR; autonomous driving

收稿日期：2025−02−26.　**网络出版日期：**2025−08−15.
基金项目：黑龙江省自然科学基金项目 (F201123).
通信作者：陆军. E-mail: lujun0260@sina.com.

自动驾驶技术[1] 正推动全球汽车产业进行战略调整，该技术关键包含 3 个部分：环境感知、路

线规划及车辆操控。环境感知系统[2]的根本目标是精确捕捉周边环境信息，以减少碰撞的可能性，并为规划和控制提供必要的环境数据。相较于二维目标检测[3]，三维目标检测[4-5]能够更准确地获取目标的大小、方向等空间信息，大幅提升车辆与真实世界的交互能力。然而，由于增加了维度，其计算复杂度和计算量也相应大幅提升。与二维目标检测仅依赖基础网络提取特征以识别物体类别不同，三维目标检测任务还涉及物体在三维空间中的位置和姿态的精确确定。多模态三维目标检测技术[6]通过整合不同传感器提供的数据特征来实现优势互补，尤其在图像数据与点云数据融合方法上表现明显。这种融合技术将图像数据中的语义信息和点云数据中的深度和几何结构信息相结合，能够更准确地对场景进行感知。多模态三维目标检测方法中，基于投影的技术在特征融合阶段利用投影矩阵实现点云和图像特征的整合。例如 PointPainting[7]网络通过顺序融合策略，将点云与图像数据的输出进行融合。PointAugmenting[8]进一步优化，用更丰富的边缘信息和更大感受野的卷积神经网络 (convolutional neural networks, CNN)[9]特征替换 PointPainting 中的分割分数，从而显著提升性能。FusionPainting[10]则在语义层面上融合 2D 图像与 3D 点云数据，通过不同维度的分割方法提取语义信息，并利用基于语义的融合模块对分割结果进行自适应整合以输出给三维检测器，实现目标检测。然而，投影过程中的不精确性可能限制其性能。非投影式的三维目标检测方法则避免了从相机到激光雷达的投影限制。基于 Query 学习的方法通过注意力机制在特征融合前实现特征对齐，以获得高度鲁棒的多模态特征。例如，TransFusion[11]利用 Transformer[12]解码器分别预测激光雷达 (LiDAR) 点云的初始边界框，并自适应融合图像特征。DeepFusion[13]通过逆向几何相关增强实现 LiDAR 点云与图像像素之间的几何对齐，并利用跨模态注意力机制动态捕捉特征间的相关性。在统一特征空间的方法中，通常在特征融合前通过投影实现异构模态的预融合统一。BEVFusion(bird's-eye view fusion)[14]采用两个独立流程处理点云和图像数据，并在鸟瞰图层面进行融合。EA-BEV(enhanced attention bird's-eye view)[15]结合边缘感知深度融合模块和深度估计模块，解决深度跳跃问题，以更准确地融合两种视图，生成更准确的深度分布。与此不同，CMT(cross-modal Transformer)[16]和 Uni-TR(unified Transformer)[17]采用 Transformer 进行点云和图像的标记化，通过 Transformer 编码构建隐式统一空间。

在多模态融合的三维检测背景下，检测准确性受到传感器质量的影响，校准不足的传感器会导致算法在目标定位和识别上产生误差，从而降低检测系统的整体性能。同时，在目标密集和场景复杂的环境，如城市交通场景，高密度的目标分布提高了模型的误检率，进而降低检测的召回率和精度。提升传感器校准的准确度以及增强模型在复杂场景下的鲁棒性，成为提升多模态融合性能的关键。

本研究聚焦于多模态融合领域中的非投影方案中基于 Query 学习的方法。在面对多模态融合过程中传感器校准不足的问题时，提出了一种基于 Transformer 架构的跨模态特征聚合模块。该模块旨在建立图像与点云特征之间的软关联关系，以此减轻校准矩阵对目标检测性能的影响。此外，针对目标密集型场景，本研究提出了一种查询对比优化策略。该策略旨在提升模型对于相似度较高的查询提案框的区分能力，从而保持模型的高鲁棒性。最终，在 nuScenes 数据集[18]上进行的消融实验验证了所提出算法的有效性。

1　SoftFusion-QC 目标检测网络

1.1　网络整体结构

SoftFusion-QC 三维目标检测算法网络框架如图 1 所示。将 LiDAR 点云与多视图相结合作为输入数据。在点云处理分支中，原始点云数据经由体素特征编码 (voxel feature encoding, VFE) 进行规则化处理，以生成结构化的体素网格，这些体素被送入到 3D 骨干特征网络中以提取深层特征，并转换成鸟瞰图 (bird's eye view, BEV) 特征图。在 Transformer 编码阶段之后，引入前馈神经网络 (feed-forward neural network, FFN)，旨在初始化对象查询，加速网络收敛过程。提出可变形跨模态特征聚合模块 (deformable cross-modality feature aggregate, DCFA)，通过 2D 骨干网络处理多视角图像，以提取多尺度的图像特征图，并借助点云分支中的目标查询来引导图像与点云特征的融合，实现自适应的特征聚合。为了提高模型在密集场景下的检测精度，引入查询对比机制 (query contrast, QC)，通过对生成的预测查询和真实值之间的对比学习，增强模型对任务的理解和性能。

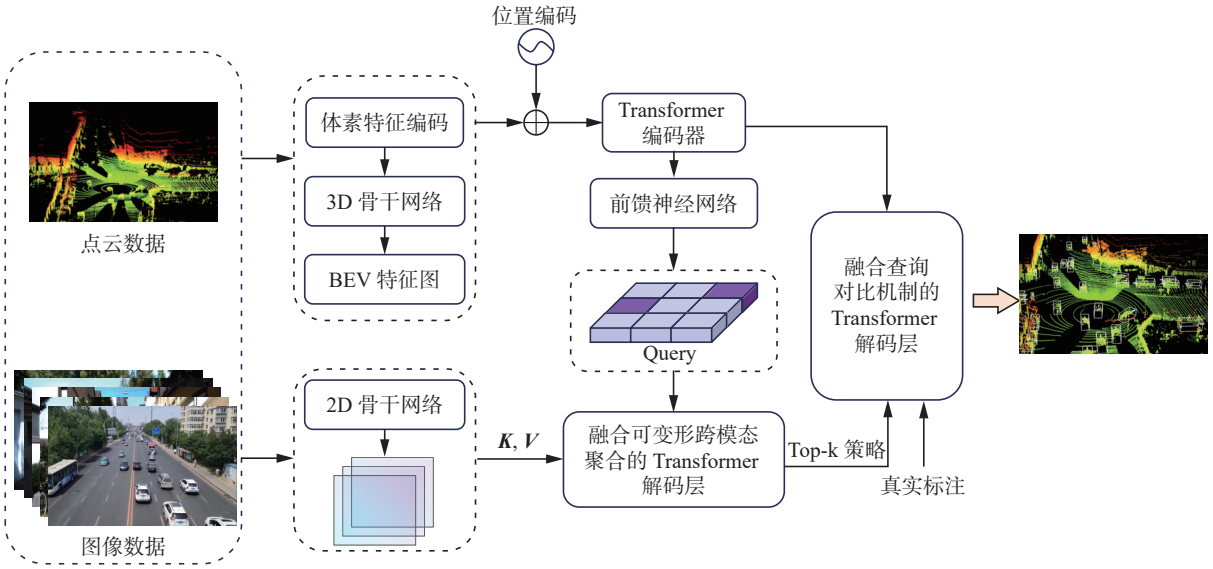

图 1　SoftFusion-QC 网络整体框架

Fig. 1　Overall framework of the SoftFusion-QC network

1.2　基于 Transformer 的特征编码器

Transformer 模型凭借其自注意力机制和编码器−解码器架构, 在计算机视觉领域中处理序列数据方面展现出了显著的灵活性与效率。这种架构赋予了 Transformer 处理各类任务时的卓越性能。其并行化处理能力和全局注意力特征使得模型能够有效捕捉视觉数据中的长距离依赖, 突破了传统卷积神经网络局部感受野的限制。

1.2.1　位置编码

在处理输入点云信息序列时, 采用体素特征编码方法将原始点云数据规则化处理, 通过输入嵌入层, 将每个点的原始低维特征向量映射到一个更高维度的空间, 对序列进行维度提升, 以增强点特征的表征能力, 使其包含更多潜在的几何和上下文信息。利用 3D 骨干网络提取深层特征并生成 BEV 特征图。同时引入位置编码以保留序列中各元素的位置信息。在对 BEV 特征图进行位置编码时, 采用正弦位置编码 (sinusoid position encoding, SPE) 的方法, 将空间信息有效地嵌入到点云特征图中。对于一个给定的 BEV 特征图 $\boldsymbol{F}\in\mathbb{R}^{H\times W\times C}$, 其尺寸由高度 H、宽度 W 和通道数 C 定义, 位置编码的目的是为每个位置(h,w)引入一个位置嵌入向量, 对于高度和宽度维度上的嵌入向量, 计算公式为

$$\mathrm{PE}(p,2k)=\sin(p/10\,000^{2k/d_{\mathrm{model}}})$$

$$\mathrm{PE}(p,2k+1)=\cos(p/10\,000^{2k/d_{\mathrm{model}}})$$

式中: p 为位置索引; k 为位置编码的维度; d_{model} 为位置向量的维度, 用于波长参数的归一化, 从而确定位置编码信息, 通过直接相加元素的方式, 得到包含位置感知信息的 BEV 特征图。

1.2.2　多头注意力机制

利用注意力机制, 通过评估经过位置编码后的点云序列内不同元素的加权贡献来识别全局依赖性。编码后的点云序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$, 首先利用权重矩阵对序列进行线性变换, 得到查询 (Query, $\boldsymbol{Q}$)、键 (Key, $\boldsymbol{K}$)、值 (Value, $\boldsymbol{V}$) 向量, 计算公式为

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{Q}}$$

$$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{K}}$$

$$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{V}}$$

式中: $\boldsymbol{W}^{\boldsymbol{Q}}\in\mathbb{R}^{d\times d_q}$, $\boldsymbol{W}^{\boldsymbol{K}}\in\mathbb{R}^{d\times d_k}$, $\boldsymbol{W}^{\boldsymbol{V}}\in\mathbb{R}^{d\times d_v}$ 分别是可学习的权重矩阵, d 是输入向量的维度, d_q、d_k、d_v 分别是 Query、Key 和 Value 的维度。

注意力机制通过评估计算 Query 和 Key 之间的相似度, 进而确定注意力权重:

$$\alpha_{ij}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_i\cdot\boldsymbol{K}_j^{\mathrm{T}}}{\sqrt{d_k}}\right)$$

式中 α_{ij} 表示第 i 个 Query 和第 j 个 Key 之间的注意力权重, 将这些权重与 Value 相乘求和, 得到加权表示输出结果:

$$\boldsymbol{Z}_i=\sum_{j=1}^{n}\alpha_{ij}\boldsymbol{V}_j$$

最后, 通过一个前馈神经网络对 $\boldsymbol{Z}$ 进行进一步处理, 得到最终的自注意力输出:

$$\boldsymbol{Y}=\mathrm{FFN}(\boldsymbol{Z})$$

式中 $\boldsymbol{Y}$ 是注意力机制的输出序列。

多头注意力机制同时捕获点云序列中不同位置的多样信息, 这些信息分布在不同的表示子空间。通过对输入序列进行分割, 形成多个独立的头, 每个头都有自己的 $\boldsymbol{K}$、$\boldsymbol{Q}$、$\boldsymbol{V}$ 矩阵, 从而允许模型在不同的表示空间中学习到更丰富的信息, 整个过程描述为

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_h)\boldsymbol{W}^{\mathrm{O}}$$

式中：$\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_h$ 分别表示 h 个头的各自输出，$\boldsymbol{W}^{\mathrm{O}}\in\mathbb{R}^{dxd}$ 是一个可学习的权重矩阵。

利用 Transformer 编码器的全局信息捕获功能，采用前馈神经网络来生成点云分支的初始对象查询 $\boldsymbol{Q}_{\mathrm{L}}$。这些初始对象查询将被送入后续的可变形跨模态特征聚合模块解码器流程中，不包含具体的类别标签，而是提供了目标位置的初始信息。这样的初始化策略使解码器能够在初始阶段获得关于目标位置的大致估计，有助于更专注于特定目标的检测和识别。随着解码过程的深入，这些初始查询将经过调整和优化，实现更精确的目标检测结果。

1.3　可变形跨模态特征聚合模块

在图像与点云特征融合的研究领域中，尽管点级融合策略在一定程度上提升了性能，但其效果仍受限于 LiDAR 点云稀疏性。当目标在点云数据中仅由少数 LiDAR 点表示时，相应的图像特征提取能力也将受到限制，只能反映少量的图像信息。这种方法未能充分利用高分辨率图像中丰富的语义信息，造成资源的浪费。此外，点级融合策略依赖于精确的传感器校准，通常需要复杂的校准流程和额外的硬件支持。

为了缓解以上问题，提出了一种可变形跨模态特征聚合模块 (DCFA)，不依赖于激光雷达点与图像像素间的直接硬关联，而是将多视角图像特征完整保留，提取每张图像的高分辨率语义特征图，为每个特征图的每一处空间位置计算并关联其精确的空间位置信息，所有视角、层级的特征图及其对应的位置信息被展平并拼接成特征序列和对应的空间位置信息序列，最终构建为一个结构化、空间可查询的特征池，作为统一查询对象。在 Transformer 解码器的框架内，通过交叉注意力机制灵活地实现特征的整合。这不仅全面挖掘了高分辨率图像中蕴含的语义信息，还实现了对不同空间位置特征的动态加权，降低对精确传感器校准的需求，同时提升模型的检测效能。如图 2 所示，本文设计的可变形跨模态特征聚合模块通过 Transformer 编码器获取的激光雷达对象查询及其对应的初始查询框的物体中心 C_{Q}，利用激光雷达与相机之间的投影矩阵 $\boldsymbol{M}_{\text{cam-lidar}}$，计算查询对象 $\boldsymbol{Q}_{\mathrm{L}}$ 在对应图像上的参考点 P_{ref}，公式为

$$P_{\mathrm{ref}}=\boldsymbol{R}\boldsymbol{K}\cdot\boldsymbol{M}_{\text{cam-lidar}}\cdot C_{\mathrm{Q}}$$

式中：$\boldsymbol{R}$ 为相机坐标系与世界坐标系之间的校准旋转矩阵，$\boldsymbol{K}$ 为相机标定矩阵。

给定图像输入表示 $\boldsymbol{x}$ 作为值特征，首先根据初始对象查询特征 $\boldsymbol{Q}_{\mathrm{L}}$ 学习采样偏移 Δp^{x} 和注意力权重 A_x，具体公式为

$$A_x=\mathrm{Softmax}(\mathrm{Linear}(\boldsymbol{Q}_{\mathrm{L}}))$$

$$\Delta p^{x}=\mathrm{Linear}(\boldsymbol{Q}_{\mathrm{L}})$$

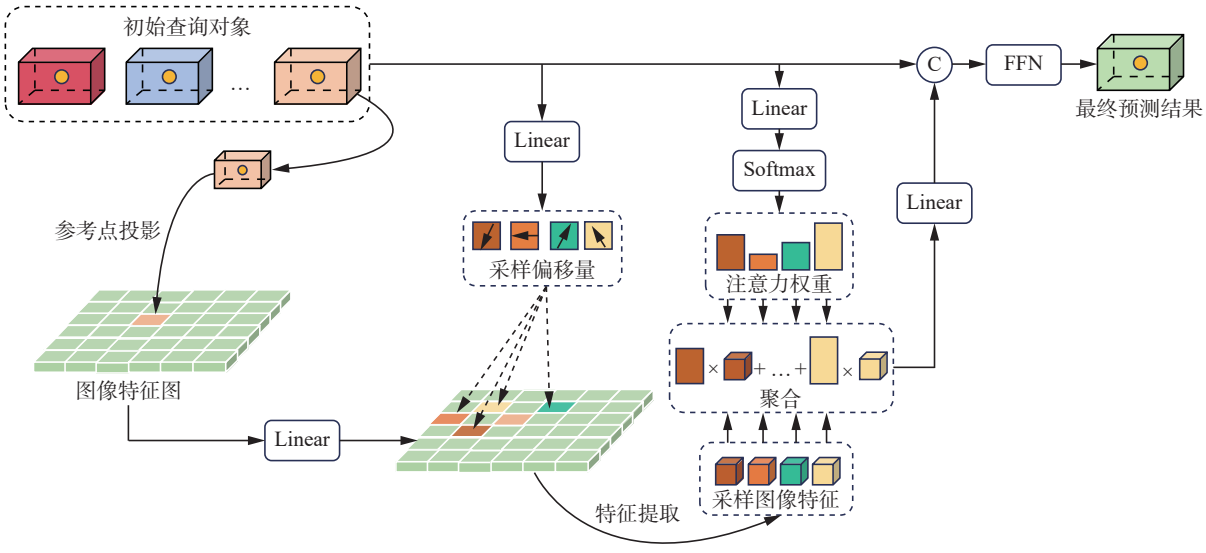

图 2　可变形跨模态特征聚合模块

Fig. 2　Deformable cross-modal feature aggregation module

通过特征图参考点 P_{ref} 和采样偏移量 Δp^{x} 可以确定位置的关键点，进一步在这些位置上提取 Key，通过插值方法对邻近的键值进行处理以获得 Value，确保了注意力机制不仅关注局部区域，还能通过插值来捕获更为精细的特征细节，最后，将计算得到的局部且稀疏的注意力权重应用于这些采样的 Value 上，以生成最终的注意力输出，整个过程可以表示为

$$\mathrm{DAttn}(\boldsymbol{Q}_{\mathrm{L}},P_{\mathrm{ref}},\boldsymbol{x})=\sum_{m=1}^{M}\boldsymbol{W}_m\left[\sum_{k=1}^{K}A_{mk}^{x}\cdot\boldsymbol{W}_m'\boldsymbol{x}\left(p_{\mathrm{ref}}+\Delta p_{mk}^{x}\right)\right]$$

式中：M 为多头注意力的头数，K 为参考点附近采样点的个数，Δp_{mk} 表示第 m 个注意力头中第 k 个采样

点的采样偏移量, $\boldsymbol{W}_m$ 和 $\boldsymbol{W}'_m$ 为可学习的权重矩阵。

针对每个对象查询, 只关注投影的 2D 中心周围的区域, 网络可以更好地学习在何处选择基于输入 LiDAR 特征的图像特征。在可变形跨模态特征聚合模块之后, 通过使用全连接层来利用包含 LiDAR 和图像信息的对象查询生成最终的边界框预测。

1.4　查询对比优化策略

自动驾驶场景中, 由于物体分布的密集性, 如果固定前 k 得分策略 (Top-k) 预测, 会导致在局部密集区域出现过多误报, 影响模型准确性。针对这个问题, 引入一种查询对比机制 (QC), 旨在减少密集区域内的误预测。误报的根本原因在于缺乏有效的监督信号来区分高度相似的查询。本研究提出的机制专注于增强与真实标签匹配度最高的查询, 并抑制那些与真实标签匹配度较低的查询的预测结果。通过这种策略, 可以显著提升目标检测的准确性和鲁棒性, 有效减少在物体密集区域出现的误报现象。

1.4.1　构造正负 GT 查询对

为了增强模型对高度相似查询的区分能力, 基于匈牙利匹配结果构建了真实标注 (ground truth, GT) 与查询之间的配对。将每个真实标注与其最佳匹配查询定义为正样本对, 而将同一真实标注对应的其他非匹配查询定义为负样本对。对于每个真实标注, 通过匈牙利匹配算法为其分配一个最优的查询对象, 形成正样本对。这一过程可以视为解决二分图匹配问题, 即在不相交边的集合中选择尽可能多的边。

在检测器中初步生成了一组固定大小的 N 个边界框, 该数量远超实际场景中存在的目标数量。为实现 GT 与初始化对象查询之间的一一对应关系, 将 GT 扩展成 N 个检测框。此外, 使用了一个额外的特殊类标签 ϕ 表示 GT 中未标记任何对象或视为背景的情况, 使得预测集合与 GT 集合总的二分图匹配数就有 A_N^N 个, 即预测集和真实集均含有 N 个元素。在此背景下优化预测集和真实集元素之间的配对, 以最小化匹配损失, 最优匹配的过程可用公式表示为

$$\hat{\sigma}=\underset{\sigma\in\Sigma_N}{\operatorname{argmin}}\sum_{i}^{N}L_{\text{match}}\left(y_i,\hat{y}_{\sigma(i)}\right)$$

式中: $\hat{\sigma}$ 为最优匹配, y_i 和 $\hat{y}_{\sigma(i)}$ 分别表示 GT 值和预测值, Σ_N 为一个包含 N 个目标的集合, $L_{\text{match}}\left(y_i,\hat{y}_{\sigma(i)}\right)$ 表示真实值 y_i 与预测索引 $\sigma(i)$ 之间的匹配损失。

在对匹配损失 $L_{\text{match}}\left(y_i,\hat{y}_{\sigma(i)}\right)$ 进行计算时, 综合考虑目标分类与目标边界框回归的影响。对于第 i 个真值 $y_i=(c_i,b_i)$, 其中 c_i 为类别标签值, b_i 为边界框参数, 对于预测的索引为 σ_i 的匹配元素, 设 $\hat{p}_{\sigma(i)}(c_i)$ 为划分为类别 c_i 的概率, 其对应的边界框预测为 $\hat{b}_{\sigma(i)}$, 匹配损失计算公式为

$$L_{\text{match}}(y_i,\hat{y}_{\sigma(i)})=-\text{sign}_{\{c_i\neq\varnothing\}}\hat{p}_{\sigma(i)}(c_i)+\text{sign}_{\{c_i\neq\varnothing\}}L_{\text{box}}(b_i,\hat{b}_{\sigma(i)})$$

式中: sign 为符号函数, 其含义是只计算非空集合; L_{box} 为边界框损失, 计算边界框损失时, 考虑了交并比 (intersection over union, IoU) 损失以及 L1 损失, 具体公式为

$$L_{\text{box}}(b_i,\hat{b}_{\sigma(i)})=\lambda_{\text{IoU}}L_{\text{IoU}}(b_i,\hat{b}_{\sigma(i)})+\lambda_{\text{L1}}\left\|b_i,\hat{b}_{\sigma(i)}\right\|_1 \tag{1}$$

式中 λ_{IoU} 和 λ_{L1} 为超参数。

总体而言, 为了提升查询与 GT 之间的匹配精度, 采用一种策略, 即选取与给定 GT 匹配成本最低的查询构成正查询对。相应地, 与同一 GT 匹配的其他查询则被定义为负查询对。通过建立这种 GT 与查询之间的正负配对关系, 在训练过程中更有针对性地强化或抑制相关对象查询, 进而提升整个网络的查询效率与准确性。

1.4.2　对比学习

在对正负 GT 查询对进行监督之前, 首先需对这些匹配对进行定量分析。然而, 传统的集合度量方法并不足以全面捕捉 GT 与查询之间的相似性。因此, 本研究将 GT 和查询映射到一个高维特征空间中, 以实现全面的相似性度量。在这个高维特征空间中, 对象查询被表示为提案框, 涵盖了对象的类别、位置、尺寸和方向等信息。这些信息被编码成每个查询的高维特征向量。Transformer 解码器能够将 GT 和查询直接编码到选定层的特征嵌入中, 从而获得它们的高维特征向量。此外, 在每个解码器层添加一个 FFN 预测头作为输出层, 并使用一个共享的多层感知机 (multi-layer perception, MLP)[19] 进行相似性估计。整体架构如图 3 所示。

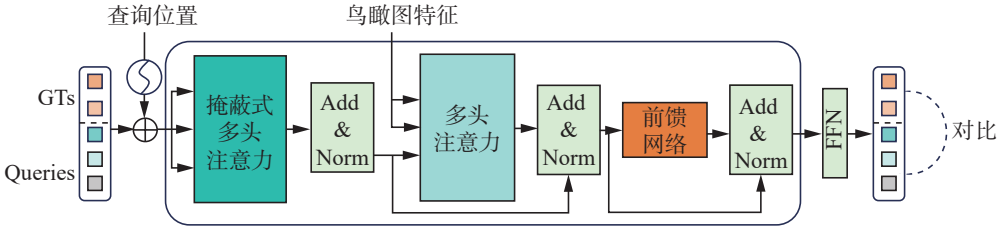

图 3　Transformer 对比学习解码层

Fig. 3　Transformer contrastive learning decoding layer

实际情况中，GT 框和查询框的分布特征可能存在较大差异，GT 框表示的目标之间通常不重叠，并且沿着相对规则的路径 (如车道) 分布，而查询框不仅密集且重叠，还可能随机分布于场景中，显示出无序性。Transformer 解码器通过注意力机制识别真实框与查询提案框之间的关联，但这两种不同分布的特性可能对相似度估计造成影响。为了降低这种分布差异对相似度估计的硬性影响，采用 MLP 对查询框特征进行投影，以与 GT 目标框的高维特征空间相匹配。MLP 的非线性映射能力使得查询嵌入能够被有效地映射到与 GT 嵌入相同的特征空间，从而减少它们之间的分布差异。在对齐 GT 目标框和查询框的嵌入特征之后，利用余弦相似度来评估所有正负 GT-查询对之间的相似度。即对于 n 维空间中的两个向量 $\boldsymbol{A}=(x_{11},x_{12},\cdots,x_{1n})$ 和 $\boldsymbol{B}=(x_{21},x_{22},\cdots,x_{2n})$，两者之间的余弦相似度为

$$\text{Cosine Similarity}(\boldsymbol{A},\boldsymbol{B})=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^2}\sqrt{\sum_{k=1}^{n}x_{2k}^2}}$$

在监督学习过程中，采用信息噪声损失函数 (information noise-contrastive estimation loss, InfoNCE Loss)[20] 来优化查询与其对应真 GT 之间的匹配度。InfoNCE Loss 于信息论的思想，通过比较正样本与负样本之间的相似度来调整模型参数，训练网络区分正负样本。该损失函数的目标是指导模型生成与最匹配查询相对应的真实标签的精确预测，并将所有非匹配查询的输出排斥开，以此增强模型识别不同查询与 GT 之间相关性的能力。具体来说，对于某一场景中的第 i 个目标的 GT，其嵌入表示为 $\boldsymbol{g}_i$，$\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_K\}$表示 K 个查询嵌入，假设经过匈牙利算法进行最佳匹配后，第 i 个 GT 的最佳匹配为查询 $\boldsymbol{q}_j$，那么对于第 i 个 GT 的查询对比损失为

$$L_i^{\text{QC}}=-\log\frac{\exp\left(\dfrac{\cos(\boldsymbol{g}_i,\rho(\boldsymbol{q}_j))}{\tau}\right)}{\sum_{k=1}^{K}\exp\left(\dfrac{\cos(\boldsymbol{g}_i,\rho(\boldsymbol{q}_k))}{\tau}\right)}\tag{2}$$

式中：$\rho(\cdot)$表示 MLP 映射层，以弱化 GT 与查询预测之间的分布差异性；τ表示温度系数，作用是控制模型对负样本的区分度；$\cos(\cdot)$为相似性度量，这里使用的是余弦相似度。

在损失优化过程中，希望最大化正查询对所占的比重，即 GT 与查询预测之间的距离尽可能近，根据式 (2)，在对比损失的最小化过程中，需要使其值尽可能趋近于零，目的是显著增强正 GT 查询对，并抑制负 GT 查询对，从而生成更准确的查询预测框，降低在目标密集场景下的误报率，并进一步提高检测模型的性能。

1.5　损失函数

检测网络首先整合点云与图像分支的特征以形成初始查询，并采用 Top-k 策略筛选出前 k 个查询进行正负查询对匹配，匹配过程被视为二分图的匹配问题[21]，利用匈牙利算法实现查询与真实标注之间的一一对应匹配，从而构造正负查询对以对比学习的方式监督生成高度匹配的查询预测。整个训练过程的损失包括分类损失、边界框几何损失以及对比学习监督损失。

对于分类损失，选用 Focal Loss[22] 作为损失函数，其在传统交叉熵损失的基础上引入了一个调整因子，减少易于分类的样本的损失权重，使模型在训练过程中更注重难以分类的样本，以达到平衡正负样本的效果，公式为

$$L_{\text{cls}}=\begin{cases}-\alpha_t(1-p_t)^{\gamma}\log(p_t), & y=1\\ -(1-\alpha_t)p_t^{\gamma}\log(1-p_t), & \text{其他}\end{cases}$$

式中：p_t是模型对于每个类别的预测概率；α_t是用于平衡正负样本权重的系数；γ是一个调整参数，用于减少简单样本的权重并增强对困难样本的关注；y是真实的类别标签，1 表示正样本，0 表示负样本。

对于边界框损失，与基于锚框的方法不同，如果直接生成边界框，可能会导致损失函数的相对缩放问题。具体而言，常用的损失函数依据预测框与真实框之间的绝对距离来计算损失，这会导致不同尺寸或位置的边界框之间的损失值存在较大差异。为了缓解这一问题，采用 L1 损失和 IoU 损失的线性组合，IoU 损失作为一种尺度不变的损失函数，有助于统一不同边界框的损失尺度。总体来说，边界框损失由式 (1) 给出，改进的检测损失为

$$L_{\text{det}}=\alpha L_{\text{cls}}+\beta L_{\text{L1}}+\delta L_{\text{IoU}}$$

式中α、β、δ为超参数。

综上所述，本文网络的总损失函数为检测损失与对比学习监督损失之和：

$$L=L_{\text{det}}+L_{\text{QC}}$$

2　神经网络训练

2.1　实验环境

本研究的实验操作在 Ubuntu 20.04 操作系统上执行，依托于 Python 3.8 版本、PyTorch 深度学习框架以及 MMDetection3D[23] 目标检测环境，并通过 nuScenes 数据集对算法性能进行评估。

2.2　网络参数设置

点云分支的处理通过 VFE 模块实现, 并且根据主流方案, 将体素网格的尺寸设定为 (0.075 m, 0.075 m, 0.2 m), 点云的覆盖范围在 X 和 Y 方向设置为 [−54 m, 54 m], 在 Z 方向为 [−5 m, 3 m]。3D 骨干网络使用与 ResNet-18 的结构, 但用 3D 稀疏卷积替换了 2D 卷积, 并且不采用预训练权重。经过 3D 骨干网络的特征提取后, 使用特征金字塔网络 (feature pyramid networks, FPN) 结构来获取多尺度的 BEV 特征。为简化流程, 仅选择部分缩减特征作为 Transformer 的输入, 整个 Transformer 架构包含 3 个编码器层和 3 个解码器层, 以提升计算效率。编码器的 FFN 预测头输出的前 300 个高分查询预测被选为对象查询。另一方面, 在图像分支中, 采用经典的 ResNet-50 网络作为二维骨干网络进行特征提取, 可变形跨模态特征聚合模块中每个目标查询的图像采样点数 K 设为 4。在损失函数超参数设置方面, 损失函数 $L=\alpha L_{\text{cls}}+\beta L_{\text{L1}}+\delta L_{\text{IoU}}+L_{\text{QC}}$ 中设置 $\alpha=1,\beta=4,\delta=2$, Focal loss 中调节参数 γ 设置为 2, 平衡因子 α_t 设置为 0.25。式 (2) 对比损失函数中 τ 设置为 0.7。训练过程中, 采用 adamW(adam with weight decay) 优化器[24]和单周期学习率策略[25], 最大学习率设置为 0.001, 衰减率设置为 0.01, batch size 设置为 8, 总共训练 20 个 epoch。

3　实验结果及分析

3.1　目标检测实验结果

在 nuScenes 测试集上对本研究所提的 SoftFusion-QC 算法进行性能评估, 并与当前主流算法的性能进行了对比, 以突显本算法的优势。评估采用的指标依据 nuScenes 数据集官方定义, 包括总体 mAP(mean average precision) 和 NDS(normalized detection score), 以及各类目标下的检测 AP(average precision)。在对比实验中, 对比算法选择基于激光雷达点云的三维目标检测方法 PointPillar[26]、UVTR-L(unified voxel transformer with LiDAR)[27]、CenterPoint[28]、Voxel-NeXt(voxel-based next-generation detection)[29], 以及基于图像与点云融合的多模态目标检测方法 PointPainting、3D-CVF(3D cross-view fusion)[30]、MVP(multi-view point cloud detector)[31]、AutoAlignV2 (automated alignment for multi-modal fusion v2)[32]。

实验数据如表 1 所示, 其中, L 表示 LiDAR 点云输入, L+C 表示 LiDAR 点云和相机照片双模态输入。分别列举对于轿车 (Car)、卡车 (Truck)、工程车 (C.V.)、巴士 (Bus)、拖挂车 (Trail.)、隔离栏 (Bar.)、摩托车 (Mot.)、自行车 (Bike)、行人 (Ped.) 和交通锥桶 (T.C.) 的检测性能, 通过对比 SoftFusion-QC 算法与其他算法在 nuScenes 测试集上的表现, 可以发现本算法利用 Transformer 的注意力机制, 针对点云分支的预测查询框, 能更有效地提取图像中的上下文特征, 实现更优的检测性能。在与仅基于激光雷达点云的检测算法对比中, SoftFusion-QC 在 mAP 上超越 CenterPoint 算法 9.5 百分点, 在 NDS 上提升了 5.5 百分点; 与 VoxelNeXt 相比, mAP 提高了 5.3 百分点, NDS 提升了 2.8 百分点。这些实验结果进一步证实了多模态融合方法相较于仅基于激光雷达点云的方法在性能上的显著提升。此外, 本研究还将 SoftFusion-QC 与近年来提出的基于图像与点云融合的多模态算法进行了对比, 结果显示 SoftFusion-QC 在多数类别的检测性能上均保持领先。与 AutoAlignV2 相比, SoftFusion-QC 在 C.V.和 Bar.类别的检测性能更为突出, 这可能是因为 AutoAlignV2 采用了深度感知的 GT-AUG(ground truth augmentation) 策略, 利用 3D 标注中的深度信息来增强图像, 有助于同步图像与点云特征。对于大型目标, 深度信息的生成更为精确, 因此增强效果更佳, 但在其他类别中, SoftFusion-QC 展现出更大的优势。

表 1　各种算法在 nuScenes 测试集上的结果比较

Table 1　Comparison of various algorithms on the nuScenes test set results

算法	模态	mAP	AP									
			NDS	Car	Truck	C.V.	Bus	Trail.	Bar.	Mot.	Bike	Ped.	T.C.
PointPillar	L	40.1	55.0	76.0	31.0	11.3	32.1	36.6	56.4	34.2	14.0	64.0	45.6
UVTR-L	L	52.8	66.3	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CenterPoint	L	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
VoxelNeXt	L	64.5	70.0	84.6	53.0	28.7	64.7	55.8	74.6	73.2	45.7	85.8	79.0
PointPainting	L+C	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
3D-CVF	L+C	52.7	62.3	83.0	45.0	15.9	48.8	49.6	65.9	51.2	30.4	74.2	62.9
MVP	L+C	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
AutoAlignV2	L+C	68.4	72.4	87.0	59.0	**33.1**	68.6	59.3	**78.0**	76.6	54.9	86.9	81.3

续表 1

算法	模态	mAP	AP										
			NDS	Car	Truck	C.V.	Bus	Trail.	Bar.	Mot.	Bike	Ped.	T.C.
SoftFusion-QC(本文方法)	L+C	**69.8**	**72.8**	**88.2**	**60.3**	30.8	**69.7**	**63.2**	75.6	**77.1**	**55.4**	**89.4**	**87.5**

注：加粗表示本列最优结果。

3.2　消融实验

通过消融实验检验网络中各模块策略的有效性和合理性。在点云分支中，通过一个类未知的 FFN 网络结构生成初始查询，进一步添加一个类别特定的 FFN 网络对查询进行类别预测，以此作为消融实验的基线模型。

为了验证模型对传感器未充分校准问题的鲁棒性，通过随机向从相机到 LiDAR 传感器的变换矩阵添加平移偏移量来模拟传感器校准不足的情况，平移偏移量在 3 个轴向上均限制在 0.8 m 内。此外，为了对比验证本研究所提出的可变形跨模态特征聚合模块的有效性，在基线模型上分别引入类似 EPNet[33](enhanced point-cloud network，以下简称 EP) 和 PointAugmenting(以下简称 PA) 的逐点拼接融合方案进行对比实验，评价指标为 mAP，其中 SoftFusion 指的是本研究所提出的网络去除查询对比学习解码层的部分。

如图 4 所示，与其他融合方法相比，SoftFusion 展现出更优的鲁棒性。当传感器发生 0.8 m 的错位时，SoftFusion 的 mAP 仅下降 1.3 百分点，而 PA 和 EP 的 mAP 分别下降了 2.7 百分点和 3.1 百分点。这是因为 PA 和 EP 的逐点对齐融合策略在很大程度上依赖于投影矩阵，从而在传感器校准不足时性能下降。相比之下，本研究所提方法中，标定矩阵主要用于将查询对象投影到图像上，而融合模块对投影位置的要求不严格，Transformer 注意力机制能够根据上下文信息自适应地寻找相关图像特征，因此对传感器校准不足具有较高的鲁棒性。针对密集场景下的误报率高的问题，通过构建正负 GT-查询对，在 Transformer 解码器层中进行对比学习策略，从而增强模型区分高相似度查询的能力，有效降低了在目标密集场景中的误报率。

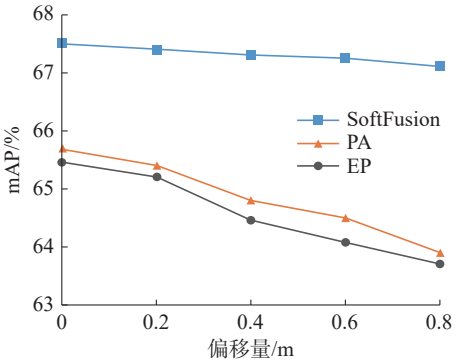

图 4　不同偏移量对不同融合策略模型的性能影响

Fig. 4　Impact of different offsets on performance of fusion strategy models

为了验证查询对比优化策略中相关组件的有效性，进行对比实验。nuScenes 数据集中存在类别数量不均衡的问题 (即长尾效应)，其中部分类别如建筑车辆、公交车、摩托车和自行车等占比很低，而小汽车和行人的占比较高，因此根据场景对应的 Car 类和 Ped.类的真实标注数量，将 nuScenes 验证集划分为 3 个区间：0~20, 21~40 和 41 及以上，其中数字表示相应场景中 Car 类和 Ped.类的真实标注数量之和。基于此，对比本研究算法在有无查询对比机制情况下的实验结果，以 mAP 作为评价指标，实验结果如图 5 所示。

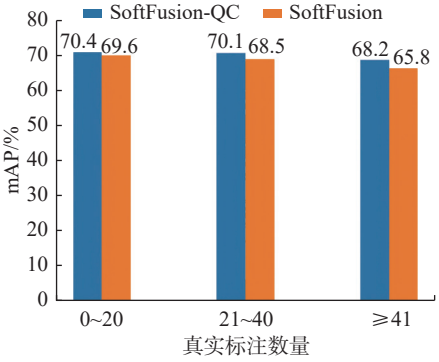

图 5　不同标注数量对模型性能的影响

Fig. 5　Impact of different annotation quantities on model performance

数据显示，真实标注数量在 0~20 之间的场景中，SoftFusion-QC 相较于对比方法，mAP 领先 0.8 百分点；在标注数量为 21~40 的场景中，领先 1.6 百分点；真实标注数量在 41 及以上时，其提升幅度达到 2.4 百分点。这一趋势可以看出，在目标分布更为密集的情境下，采用查询对比机制的检测器性能优势更为显著。

3.3　实验结果可视化

为了定性评估算法的有效性，选择不同的场景进行预测，并通过可视化手段直观展示结果。如图 6~9 所示，选取了 nuScenes 验证集中的两个场景进行预测分析，其中 GT 表示场景的真实标注，而 Pred 表示算法的预测输出。对于每个场景，环绕车辆的 6 张图片被拼接在一起展示，每个场景分为上下两排，每排展示 3 张图片。上排图片分别对应于 nuSceness 数据采集车辆的左前方相机、正前方相机和右前方相机视角，下排图片则对应于左后方相机、正后方相机和右后方相机视角。不同类别的目标通过不同颜色的三维框进行区分。通过比较真实标注和预测标注可以观察

到, 本研究所提出的 SoftFusion-QC 三维目标检测算法在复杂场景下避免了因为遮挡或密集因素造成误检, 与 GT 标注内容保持一致, 表现出较高的检测水平。

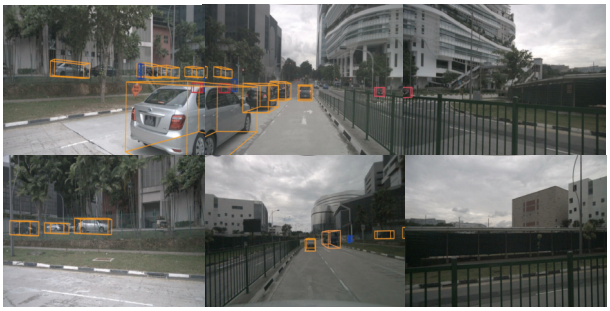

图 6　场景 1 可视化结果 (GT)
Fig. 6　Scene 1 visualization results (GT)

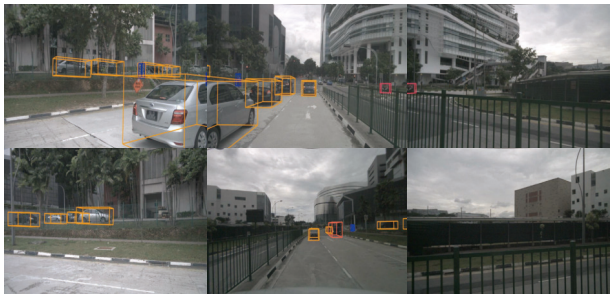

图 7　场景 1 可视化结果 (Pred)
Fig. 7　Scene 1 visualization results (Pred)

图 8　场景 2 可视化结果 (GT)
Fig. 8　Scene 2 visualization results (GT)

图 9　场景 2 可视化结果 (Pred)
Fig. 9　Scene 2 visualization results (Pred)

同时, 作为对有无查询对比机制消融实验的补充, 从定性角度可视化 SoftFusion 与 SoftFusion-QC 的检测性能, 具体可视化结果如图 10 所示。可以看到, 在圆圈标记的密集行人和其他小目标区域, 缺少查询对比机制的 SoftFusion 相较于具备该机制的 SoftFusion-QC 产生了较多的误检。

SoftFusion　　SoftFusion-QC

图 10　有无查询对比机制模型检测结果对比
Fig. 10　Different offsets affect fusion strategy model performance

4　结束语

本文针对激光雷达和相机校准未充分校准以及检测场景目标密集的情况, 分别展开研究。对于传感器校准不足的问题, 设计一种可变形跨模态特征聚合模块, 采用初始化的目标查询框作为 Query, 利用可变形的交叉注意力机制, 实现图像与点云特征的融合, 构建了点云与图像特征之间的软关联, 增强了模型对传感器校准不足情况的鲁棒性。对于目标密集场景的检测问题, 提出一种基于对比学习的查询对比优化策略, 通过引入对比损失, 增强了模型在目标密集场景中区分高相似度查询预测框的能力。实验结果表明, 所提出的算法在 nuScenes 数据集上取得了 69.8% 的 mAP 和 72.8% 的 NDS, 展现出比较优异的性能。未来将探索时序信息建模, 通过多帧动态校准补偿单帧传感器的位姿漂移, 并研究跨传感器泛化框架以适配未标定新设备场景。

参考文献:

[1] 张耀丹. 无人驾驶汽车的现状及发展趋势[J]. 汽车实用技术, 2018, 43(6): 10, 15.
ZHANG Yaodan. The current situation and tendency of driverless cars[J]. Automobile applied technology, 2018, 43(6): 10, 15.

[2] 王世峰, 戴祥, 徐宁, 等. 无人驾驶汽车环境感知技术综述[J]. 长春理工大学学报 (自然科学版), 2017, 40(1): 1−6.
WANG Shifeng, DAI Xiang, XU Ning, et al. Overview on environment perception technology for unmanned ground vehicle[J]. Journal of Changchun University of Science and Technology (natural science edition), 2017, 40(1): 1−6.

[3] JANA P, MOHANTA P P. Recent trends in 2D object detection and applications in video event recognition[EB/OL]. (2022−02−07)[2025−02−26]. https://arxiv.org/abs/2202.03206.

[4] PRAVALLIKA A, HASHMI M F, GUPTA A. Deep learning frontiers in 3D object detection: a comprehens-

ive review for autonomous driving[J]. IEEE access, 2024, 12: 173936−173980.

[5] ZHU Minling, GONG Yadong, TIAN Chunwei, et al. A systematic survey of transformer-based 3D object detection for autonomous driving: methods, challenges and trends[J]. Drones, 2024, 8(8): 412.

[6] TANG Yingjuan, HE Hongwen, WANG Yong, et al. Multi-modality 3D object detection in autonomous driving: a review[J]. Neurocomputing, 2023, 553: 126587.

[7] VORA S, LANG A H, HELOU B, et al. PointPainting: sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4604−4612.

[8] WANG Chunwei, MA Chao, ZHU Ming, et al. PointAugmenting: cross-modal augmentation for 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 11794−11803.

[9] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541−551.

[10] XU Shaoqing, ZHOU Dingfu, FANG Jin, et al. FusionPainting: multimodal fusion with adaptive attention for 3D object detection[C]//2021 IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021: 3047−3054.

[11] BAI Xuyang, HU Zeyu, ZHU Xinge, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1080−1089.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998−6008.

[13] LI Yingwei, YU A W, MENG Tianjian, et al. DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 17161−17170.

[14] LIANG Tingting, XIE Hongwei, YU Kaicheng, et al. BEVFusion: a simple and robust LiDAR-camera fusion framework[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 10421−10434.

[15] HU Haotian, WANG Fanyi, SU Jingwen, et al. EA-BEV: edge-aware bird's-eye-view projector for 3D object detection[EB/OL]. (2023−03−31)[2025−02−26]. https://arxiv.org/abs/2303.17895.

[16] YAN Junjie, LIU Yingfei, SUN Jianjian, et al. Cross modal transformer via coordinates encoding for 3D object dectection[EB/OL]. (2023−01−03)[2025−02−26]. https://arxiv.org/abs/2301.01283.

[17] WANG Haiyang, TANG Hao, SHI Shaoshuai, et al. UniTR: a unified and efficient multi-modal transformer for bird's-eye-view representation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 6792−6802.

[18] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11621−11631.

[19] LEE W, KIM H, AHN J. Defect-free atomic array formation using the Hungarian matching algorithm[J]. Physical review A, 2017, 95(5): 053424.

[20] TOLSTIKHIN I O, HOULSBY N, KOLESNIKOV A, et al. MLP-Mixer: an all-MLP architecture for vision[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 24261−24272.

[21] EGGERT S, KLIEMANN L, SRIVASTAV A. Bipartite graph matchings in the semi-streaming model[C]//Algorithms-ESA 2009. Berlin: Springer Berlin Heidelberg, 2009: 492−503.

[22] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999−3007.

[23] CONTRIBUTORS M. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection [EB/OL]. (2019−06−17)[2025−02−26]. https://arxiv.org/abs/1906.07155.

[24] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//International Conference on Learning Representations. Singapore: OpenReview.net, 2025: 1−18.

[25] SMITH L N, TOPIN N. Super-convergence: very fast training of neural networks using large learning rates[C]//Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. Baltimore: SPIE, 2019: 369−386.

[26] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12689−12697.

[27] LI Yanwei, CHEN Yilun, QI Xiaojuan, et al. Unifying voxel-based representation with transformer for 3D object detection[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems.

New Orleans: Curran Associates Inc., 2022: 18442−18455.

[28] YIN Tianwei, ZHOU Xingyi, KRAHENBUHL P. Center-based 3D object detection and tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 11779−11788.

[29] CHEN Yukang, LIU Jianhui, ZHANG Xiangyu, et al. VoxelNeXt: fully sparse VoxelNet for 3D object detection and tracking[C]//2023 IEEE/CVF conference on computer vision and pattern recognition. Vancouver: IEEE, 2023: 21674−21683.

[30] YOO J H, KIM Y, KIM J, et al. 3D-CVF: generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection[C]//Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 720−736.

[31] YIN Tianwei, ZHOU Xingyi, KRÄHENBÜHL P. Multimodal virtual point 3D detection[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 16494−16507.

[32] CHEN Zehui, LI Zhenyu, ZHANG Shiquan, et al. Deformable feature aggregation for dynamic multi-modal 3D object detection[C]//Computer Vision–ECCV 2022. Cham: Springer Nature Switzerland, 2022: 628−644.

[33] HUANG Tengteng, LIU Zhe, CHEN Xiwu, et al. EPNet: enhancing point features with image semantics for 3D object detection[C]//Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 35−52.

作者简介:

陆军, 教授, 博士生导师, 博士, 主要研究方向为计算机视觉、机器感知和机械臂控制。科技部科技型中小企业创新基金项目评审专家, 国家自然科学基金同行评议专家。发表学术论文 80 余篇, 出版著作 5 部。E-mail: lujun0260@sina.com。

赵颢然, 硕士研究生, 主要研究方向为三维目标检测、计算机视觉。E-mail: 1793961894@qq.com。

鲁林超, 硕士, 主要研究方向为三维目标检测、计算机视觉。E-mail: llczsr@163.com。