飞控与探测
Flight Control & Detection
ISSN 2096-5974,CN 10-1567/TJ

《飞控与探测》网络首发论文

题目：基于 Transformer 的毫米波雷达/激光雷达/相机融合 3D 目标检测方法

作者：陈坤泽，刘晓晨，申冲

网络首发日期：2025-05-19

引用格式：陈坤泽，刘晓晨，申冲．基于 Transformer 的毫米波雷达/激光雷达/相机融合 3D 目标检测方法[J/OL]．飞控与探测.
https://link.cnki.net/urlid/10.1567.tj.20250519.1146.002

网络首发时间：2025-05-19 14:35:27　网络首发地址：https://link.cnki.net/urlid/10.1567.tj.20250519.1146.002

基于 Transformer 的毫米波雷达/激光雷达/相机融合 3D 目标检测方法

陈坤泽[1]，刘晓晨[2]，申　冲[2]

（1.中北大学 信息与通信工程学院・太原・030051；

2.中北大学 仪器与电子学院・太原・030051）

摘　要：针对单一传感器在环境感知任务中的性能局限性，提出了一种基于 Transformer 的毫米波雷达/激光雷达/相机融合 3D 目标检测方法。该方法由三个关键模块组成：1)相机模块，将图像特征与初始 3D 目标预测结合，进行视觉增强；2)雷达模块，针对毫米波雷达点云稀疏问题，提出时序多帧融合算法，对连续 5 帧毫米波雷达数据进行融合，同时提出点云加权融合算法，在毫米波雷达与激光雷达的点云描述能力互补的基础上，构建增强的雷达点云；3)融合模块，采用 Transformer 解码器，充分整合雷达点云与相机特征，以提升 3D 目标检测性能。在自制城市道路数据集上进行了实验，并基于 nuScenes 数据集指标进行评估。实验结果表明，相较于现有方法，本文方法在 mAP 指标上提升 6.38%，在 NDS 指标上提升 5.93%。

关键词：环境感知；多模态融合；目标检测；变换器（transformer）；注意力机制

中图分类号：TN958; TN957.52　　**文献标志码：**A

Transformer-based Radar/Lidar/Camera Fusion Solution for 3D Object Detection

CHEN Kunze[1], LIU Xiaochen[2], SHEN Chong[2]

（1. School of Information and Communication Engineering, North University of China, Taiyuan 030051;
2. School of Instrument and Electronics, North University of China, Taiyuan 030051)

Abstract: In response to the performance limitations of single sensors in environmental perception tasks, this paper proposes a transformer-based multimodal 3D object detection method that integrates millimeter-wave radar, LiDAR, and camera data. The proposed framework consists of three key modules: (1) Camera module, which combines image features with initial 3D object predictions to enhance visual representation; (2) Radar module, which addresses the sparsity issue of millimeter-wave radar point clouds by introducing a temporal multi-frame fusion algorithm that aggregates data from five consecutive frames. Additionally, a point cloud weighted fusion algorithm is proposed to construct an enhanced radar point cloud by leveraging the complementary characteristics of millimeter-wave radar and LiDAR; (3) Fusion module, which employs a Transformer decoder to effectively integrate radar point clouds and camera features, thereby improving 3D object detection performance. Experiments are conducted on a custom urban road dataset, and the proposed method is evaluated based on metrics from the nuScenes dataset. The results demonstrate that, compared to existing methods, the proposed approach achieves a 6.38% improvement in mAP and a 5.93% increase in NDS.

Keywords: environment perception; multi-sensor fusion; object detection; transformer; attention mechanism

0　引　　言

自动驾驶等领域的蓬勃发展，对环境感知系统中 3D 目标检测准确性和可靠性的要求日益提高。在实际的应用场景中，单一传感器所提供的数据往往难以精准地描述周围环境中的目标信息，并存在各自的局限性，如激光雷达[1-4]探测距离近且数据处理复杂、相机[5-8]深度感知能力弱、毫米波雷达[9-10]点云稀疏且分辨率较低等。多传感器融合技术通过整合不同传感器的优势，弥补各自的不足，进而提升整体的目标检测性能。

激光雷达与相机融合的目标检测方法[11-13]利用点云与视觉信息对目标进行检测。但这种方案的检测距离仍然受限。毫米波雷达与相机融合的目标检测方法[14-16]将远距离探测与深度优势结合，但是该方法的点云低分辨率特性与环境光依赖性仍会影响检测结果。针对上述问题，本文利用毫米波雷达、激光雷达和相机三种传感器进行融合目标检测，通过毫米波雷达的长距离探测能力补足了激光雷达和相机在远距离感知方面的不足，从而显著提升了远距离目标检测的精度，扩展了车辆前方环境的感知范围。同时，毫米波雷

基金项目：山西省基础研究计划（202303021211150）；航空科学基金（202400080U0001）；山西省量子传感、精密测量重点实验室基金（201905D121001）；山西省研究生创新实践项目（2024SJ244）
作者简介：陈坤泽（1999—），男，硕士生。
通信作者简介：刘晓晨（1995—），男，博士，副教授，硕士生导师。

达的全天候性能有效克服了激光雷达和相机对天气条件的依赖性。在复杂多变的环境下，这种多传感器融合方法能够保证系统的稳定性和可靠性。因此，通过融合毫米波雷达、激光雷达与相机的信息进行 3D 目标检测，有望大幅提高环境感知的精准度与系统的鲁棒性。

多传感器融合的关键在于将不同类型传感器的数据准确联合。现阶段主要依赖于传感器标定来实现像素级[15]、特征级[11,12,16]或检测级[13-14]的数据联合。然而，这一方法并不适用于毫米波雷达、激光雷达和相机的融合。毫米波雷达的横向探测角度为 120°，激光雷达可以对周围 360°的范围进行扫描，相机的横向视场角为 50°到 90°，毫米波雷达和激光雷达的横向探测范围远大于相机视场，并且雷达波束在实际环境中可以发生多次反射。这导致在传感器探测视场重叠的区域内，三种传感器所能检测到的目标不一致。针对此问题，本文选择基于 Transformer 注意力机制的方法完成多传感器的融合。Transformer 早期应用于自然语言处理(Natural Language Processing，NLP)[17]领域。2020 年后，Transformer 在目标分类[12]与检测[18]中开始展现出优异的性能，其自注意力机制和交叉注意力机制[19]可以自适应地对多模态信息之间的交互进行学习[17,20]。所以相对于传统的基于传感器标定的硬性编码数据联合，基于交叉注意力机制的自适应柔性联合有望解决毫米波雷达、激光雷达与相机的数据联合问题。针对上述问题，本文提出一种基于 Transformer 的毫米波雷达/激光雷达/相机融合 3D 目标检测方法。

所提出的方法由三个模块组成：相机模块对输入图像进行特征提取，同时生成初始 3D 目标预测，通过多层 Transformer 解码器的自注意力机制对初始 3D 目标预测与图像特征的交互进行学习，生成视觉更新的 3D 目标预测。雷达模块将毫米波雷达点云与激光雷达点云作为输入，对连续的 5 帧毫米波雷达点云进行融合处理，然后将融合后的毫米波雷达点云与激光雷达点云进行数据融合，输出总雷达点云。最后对总雷达点云进行特征提取及位置编码。在融合模块中，相机模块的输出与雷达模块的输出作为输入数据进入三层 Transformer 解码器，Transformer 交叉注意力机制对雷达特征和视觉更新的 3D 目标预测之间的交互进行迭代学习，实现毫米波雷达、激光雷达与相机的融合。本文有以下创新点：

1) 针对一种或两种传感器融合存在检测性能局限性的问题，本文提出基于 Transformer 框架将毫米波雷达、激光雷达和相机三种传感器进行融合的方法。三种传感器基于交叉注意力机制融合，通过各自的优势对其他性能局限性进行弥补，提高 3D 目标检测准确性与鲁棒性。

2) 针对单帧毫米波雷达点云稀疏的问题，本文提出点云时序融合算法。该算法将连续 5 帧毫米波雷达点云进行融合，增加点云对环境描述的详细性，为后续目标检测任务的准确性提供保障。

3) 针对毫米波雷达对细小目标、激光雷达对远距离目标出现的误检与漏检情况，本文提出点云加权融合算法对毫米波雷达点云与激光雷达点云进行数据融合。利用毫米波雷达的远距离点云探测性能与激光雷达的近距离点云精度性能进行互补，使探测距离及漏检率都得到优化。

1　网络介绍

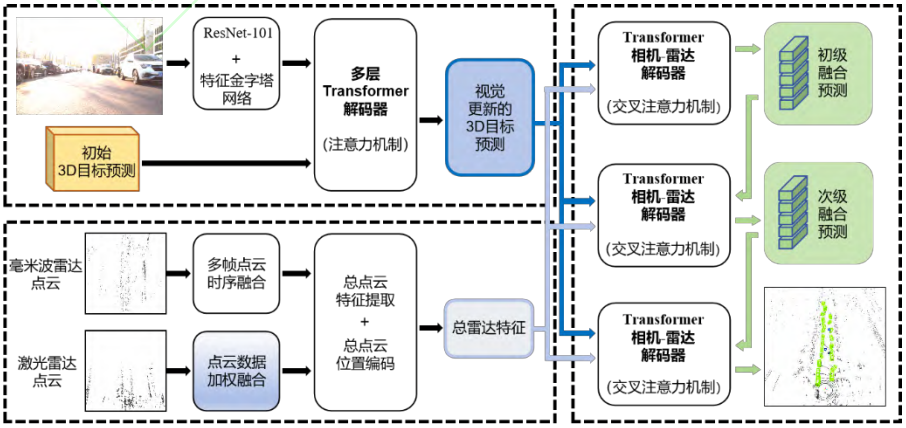


图 1　本文方法整体原理图

Fig. 1 The overall schematic diagram of the method in this paper

图 1 为本文方法的整体原理图。相机模块对输入图像进行特征提取，并与初始 3D 目标预测结合，生成视觉更新的 3D 目标预测；雷达模块对毫米波雷达点云数据进行多帧融合，再与激光雷达数据进行数据融合，进行位置编码及特征提取，输出总雷达特征。相机模块与雷达模块的输出进入到融合模块中，通过 Transformer 交叉注意力机制融合并输出目标检测结果。由于雷达视场与相机视场的宽度不一致，该方法整体工作在视场重合范围之内。所提方法各模块详细介绍如下。

1.1 相机模块

相机模块以 DETR3D[9]网络为主体，输入数据为相机图像和初始 3D 目标预测，采用 3D 转 2D 的由上到下(Top-Down)式思路[18,21]，其中的 ResNet-101[22]和特征金字塔网络(Feature Pyramid Network，FPN)[23]对图像数据进行多尺度特征提取。同时该模块在图像范围内生成初始 3D 目标预测，对图像特征进行索引。多层 Transformer 解码器对图像特征与初始 3D 目标预测的交互进行学习，输出一组 3D 空间中视觉更新的 3D 目标预测。视觉更新的 3D 目标预测作为输入数据进入后续 Fusion Network 中。下面将对 Transformer 解码器的工作过程进行介绍。

相机模块在输入图像视角内随机初始化 x 个初始 3D 目标预测 $\boldsymbol{p}_{\mathrm{re}}{}^{0}=\{\boldsymbol{p}_{1}^{0},\boldsymbol{p}_{2}^{0},...,\boldsymbol{p}_{x}^{0}\}$，其中上标 0 代表第一层解码器，下标为目标预测索引。第一层 Transformer 解码器通过注意力机制将 3D 目标预测的位置与对应的图像特征进行匹配，同时对这一过程中需要的转换矩阵及向量属性权重进行学习，生成该层的视觉更新的 3D 目标预测。另外，该模块从特征金字塔的每一层中对图像特征进行采样，并将每一层的特征总和作为更新图像特征 F_{i} 进入下一层解码器。

后续的 Transformer 解码器依次将上一层的视觉更新的 3D 目标预测与更新图像特征结合，生成该层的视觉更新的 3D 目标预测 $\boldsymbol{p}_{\mathrm{re}}{}^{n}=\{\boldsymbol{p}_{1}^{n},\boldsymbol{p}_{2}^{n},...,\boldsymbol{p}_{x}^{n}\}$，其中 n 为当前层，$\boldsymbol{p}_{\mathrm{re}}{}^{n}$ 为第 $(n+1)$ 层 Transformer 解码器的输入数据。最后一层 Transformer 解码器输出的视觉更新的 3D 目标预测将作为输入数据进入融合模块。

1.2 雷达模块

雷达模块的输入数据为毫米波雷达点云与激光雷达点云。针对毫米波雷达点云稀疏性，本文提出点云时序融合算法对多帧毫米波雷达点云进行叠加。针对毫米波雷达对细小目标、激光雷达对远距离目标的误检与漏检问题，本文提出点云加权融合算法将预处理后的毫米波雷达点云与激光雷达点云进行数据融合，将二者的点云性能进行弥补，最终得到总雷达点云。随后该模块对总雷达点云进行点位编码及特征提取，形成总雷达特征进入融合模块。

1.2.1 点云时序融合算法

针对毫米波雷达点云的稀疏性，本文提出点云时序融合算法。通过对不同时间点的点云进行时序融合，逐帧地合成更具信息量的点云，从而实现对动态目标的更加精确的描述和更高的检测性能。该算法采用数据级融合方法，首先，对连续 2 帧的点云数据进行时空对齐，针对不同位置的点云信息进行提取，对各点位的信息进行权重的计算并加权叠加，以充分重合 2 帧点云的信息。随后，再对后续 3 帧点云进行陆续融合，将连续 5 帧点云叠加的结果转换为当前帧。相比于简单的点云拼接，该方法能够有效增强毫米波雷达的结构信息。算法处理流程如图 2 所示。该算法通过时间同步、点云对齐、加权叠加、去重和滤波等步骤，对多帧点云数据进行融合和优化。接下来介绍算法的主要流程。

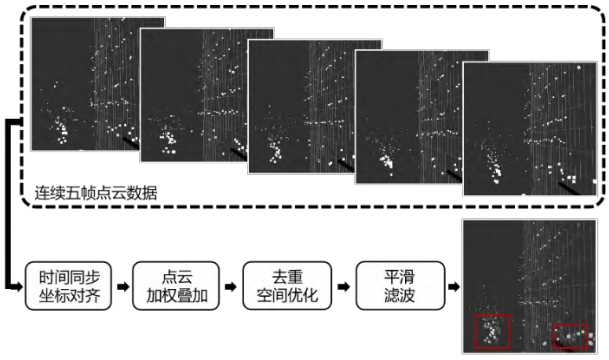


图 2　点云时序融合算法整体流程图

Fig. 2 The overall flow chart of point cloud timing fusion algorithm

为了确保不同时间获取的点云能够进行准确融合，确保不同时间采集的点云之间的时间一致性，避免因时间差异造成的空间错位问题，首先需要对各帧点云进行时间同步处理。对于输入数据中的一帧点云 $\boldsymbol{P}_i=\{(x_j,y_j,z_j)\}$，$i$ 表示第 i 帧，每一帧点云包含若干个三维坐标点 (x_j,y_j,z_j)，其中，j 表示第 j 个点的三维坐标。运用坐标变换矩阵 $\boldsymbol{T}_i$ 将每一帧点云 $\boldsymbol{P}_i$ 转换至同一全局坐标系，使得每一帧的点云数据得到正确的坐标变换，将雷达传感器的局部坐标系转换为全局坐标系。对第 i 帧点云进行坐标变换后，变换后的点云数据 $\boldsymbol{P}_i'$ 可通过下式得到：

$$\boldsymbol{P}_i^{'}=\boldsymbol{T}_i\boldsymbol{P}_i \tag{1}$$

其中，$\boldsymbol{P}_i'$ 表示第 i 帧点云在全局参考坐标系中的表示，$\boldsymbol{T}_i$ 为第 i 帧的坐标变换矩阵。

在时间同步与坐标对齐后，该算法对多帧点云进行加权叠加。由于每一帧点云来自不同的时间点，且不同帧点云之间可能包含不同的目标信息，因此需要为每一帧点云赋予一个权重 w_i，以反映其在最终融合结果中的贡献程度。该算法中权重与时间间隔相关，距离当前时刻的时间间隔越短的帧，其权重越大，则加权叠加公式为：

$$\boldsymbol{P}_{\text{fusion}}=\sum_{i=1}^{n}w_i\boldsymbol{P}_i^{'} \tag{2}$$

其中，w_i 是第 i 帧点云的权重，且满足归一化条件 $\sum_{i=1}^{n}w_i=1$。通过加权叠加，得到的点云 $\boldsymbol{P}_{\text{fusion}}$ 为五帧点云的加权平均结果。

由于不同帧的点云数据可能存在重复的点，尤其是在相对静态的环境中，多帧点云可能会多次检测到同一物体或区域。为了消除冗余点云，提高点云的质量，需要进行去重处理。若两点之间的空间距离小于预定的阈值 $\grave{o}$，则认为这两点是重复的，保留其中一个。通过欧氏距离度量点与点之间的距离：

$$d(p_1,p_2)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2} \tag{3}$$

若 $d(p_1,p_2)<\grave{o}$，则去除点 p_2，保留点 p_1。去重后的点云数据 $\boldsymbol{P}_{\text{unique}}$ 可以表示为：

$$\boldsymbol{P}_{\text{unique}}=\text{RemoveDuplicates}(\boldsymbol{P}_{\text{fusion}},\grave{o}) \tag{4}$$

其中，$\grave{o}$ 为去重阈值，控制去重的严格度。

在去重处理之后，得到的点云数据可能依然包含一定的噪声，尤其是在动态场景中。为了进一步提高点云的质量和稳定性，需要对点云进行平滑处理。该算法采取均值滤波的方法，通过计算每个点的邻域点的均值来平滑该点：

$$\boldsymbol{p}_{\text{smoothed}}=\frac{1}{n}\sum_{i=1}^{n}p_i \tag{5}$$

其中，p_i 为点云中与点 p 相邻的点，n 是邻域内点的数量。滤波后的点云数据 $\boldsymbol{p}_{\text{smoothed}}$ 经过平滑处理后，可以更好地去除噪声并增强点云的精度。

1.2.2　点云加权融合算法

针对毫米波雷达对细小目标、激光雷达对远距离目标的误检与漏检问题，本文提出点云加权融合算法，将毫米波雷达点云与激光雷达点云进行数据融合。毫米波雷达的最远探测范围在 200~300m，激光雷达的最远探测范围在 100~200 m，所以毫米波雷达在前方探测距离上的表现要优于激光雷达，并且在恶劣天气条件下表现更稳定；而激光雷达具备更高的点云精度和角分辨率，适用于精确目标识别。所以将二者的点云进行融合、优势进行互补，可以有效提高后续特征提取与位置编码的信息量与准确率。该算法运算量小，有利于保证网络的效率。融合后的总雷达点云可以详细的分辨近距离的目标，同时面对远距离目标时也展现出良好的检测性能。

该算法采用数据级融合策略，将毫米波雷达点云与激光雷达点云信息进行加权融合。根据毫米波点云与激光雷达点云所包含距离、速度、位置等信息进行权重的计算，并根据权重进行加权求和，得到新的点云信息作为融合后的点云。毫米波雷达与激光雷达通过时间同步与空间校准完成融合数据的预处理。基于最近邻关联法完成二者的数据关联。经过空间校准后，毫米波雷达点云数据中的一个点为 $P_r(x_r,y_r,z_r)$，激光雷达点云数据中的一个点为 $P_l(x_l,y_l,z_l)$，计算他们之间的欧几里得距离 D 并设置关联阈值 D_{th}。本文中关联阈值 D_{th} 为雷达距离测量分辨率。若 $D\le D_{\text{th}}$，则认为这两个点相关联，属于同一目标的观测点。

对于关联后的点，将数据的各个属性值与通过权重加权融合的方法进行融合。毫米波雷达在距离测量属性上的标准差为 σ_{radar}，激光雷达在测量距离属性上的标准差为 σ_{lidar}，则定义毫米波雷达在距离测量属性上的权重为 $w_{\text{radar}}=\sigma_{\text{lidar}}^2/(\sigma_{\text{lidar}}^2+\sigma_{\text{radar}}^2)$，激光雷达的权重为 $w_{\text{lidar}}=\sigma_{\text{radar}}^2/(\sigma_{\text{radar}}^2+\sigma_{\text{lidar}}^2)$，其中 $w_{\text{radar}}+w_{\text{lidar}}=1$。同时定

义毫米波雷达的距离测量属性值为 a_{radar}，激光雷达的属性值为 a_{lidar}。所以融合后的属性值为 $A_{\text{fusion}} = w_{\text{radar}} a_{\text{radar}} + w_{\text{lidar}} a_{\text{lidar}}$。同理，对于两个雷达的其他属性值，也根据权重进行加权融合。通过上述加权融合算法，保持了毫米波雷达与激光雷达原始数据的完整性和真实性，使融合后的数据对于检测目标有着更准确的表示或估计。

1.2.3 雷达模块原理

经过上述算法得到的总雷达点云具有 36 个通道，其中涵盖了自身车辆坐标 x，y，z，目标在 X 轴与 Y 轴上的速度 v_x 和 v_y，自身车辆运动补偿速度 v_{xc} 和 v_{yc}，虚警概率 P_{f}，时间偏移通道及其他通道。雷达模块将上述内容转化为独热编码，通过多层感知器得到点云特征 $\boldsymbol{F}_{\text{R\&L}} \in \mathrm{R}^{M\times C}$ 和点云位置编码 $\boldsymbol{P}_{\text{R\&L}} \in \mathrm{R}^{M\times C}$。其中，$M$ 和 C 分别代表点云数量和特征通道数量。$\boldsymbol{F}_{\text{R\&L}} \in \mathrm{R}^{M\times C}$ 与 $\boldsymbol{P}_{\text{R\&L}} \in \mathrm{R}^{M\times C}$ 交互形成总雷达特征 $\boldsymbol{F}_{\text{ea}_{\text{R\&L}}} = (\boldsymbol{F}_{\text{R\&L}} + \boldsymbol{P}_{\text{R\&L}}) \in \mathrm{R}^{M\times C}$ 进入后续融合模块。

1.3 融合模块

为了有效融合可见光图像与雷达点云，本文在相机模块中采用 DETR3D 网络进行图像特征提取，并利用初始目标预测进行索引；在雷达模块中，本文对点云特征进行提取。在融合模块中，我们采用 Transformer 结构进行跨模态特征交互，以增强 3D 目标检测能力。相比于简单的数据拼接，该方法能够有效提升对小目标和远距离目标的感知能力。

视觉更新的 3D 目标预测与总雷达特征进入融合模块后，通过级联在一起的三个 Transformer 解码器，其交叉注意力机制对视觉更新的 3D 目标预测和总雷达特征的交互进行学习，使二者进行匹配。图 3 展示解码器的工作流程。第一层解码器的输入参数是 Query、Key 和 Value，其中 Query 是视觉更新的 3D 目标预测 $\boldsymbol{p}_{\text{re}_{\text{img}}} \in \mathrm{R}^{N\times C}$，Key 和 Value 是总雷达特征 $\boldsymbol{F}_{\text{ea}_{\text{R\&L}}} \in \mathrm{R}^{M\times C}$。在进行注意力计算时，先计算 Query 与所有 Key 的相似度得分，然后运用 softmax 函数对这些得分进行归一化处理，得到注意力权重分布。该分布作为权重对所有的 Value 向量进行加权求和，得到每个 Query 位置的上下文向量。通过 Query 搜索与其最相关的 Key，并从中抽取对应的 Value 作为更新当前状态所需的有效信息。对于每一个时间步 t，其 Query 向量为 $\boldsymbol{Q}_t$，Key 矩阵为 $\boldsymbol{K}$，Value 矩阵为 $\boldsymbol{V}$，则注意力得分计算方式如下：

$$\text{Attention}(\boldsymbol{Q}_t, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}_t \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\text{k}}}}\right)\boldsymbol{V} \tag{6}$$

其中，d_{k} 是 Key 向量的维度。式中将 d_{k} 作为分母，是为了稳定梯度和防止数值过大。每个时间步的最终输出是对应 Query 位置与整个序列其他位置交互后的加权和。这种设计使得网络可以根据当前 Query 的关注焦点对不同位置信息的重视程度进行动态调整。

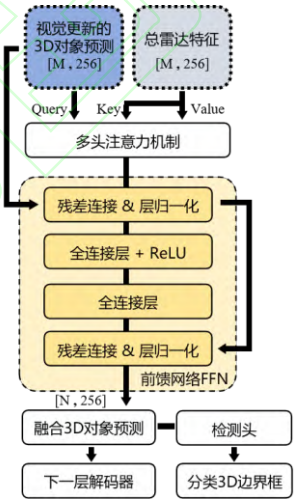


图 3 解码器工作流程图

Fig. 3 Decoder work flow chart

交叉注意力层根据输入信息得到注意力得分矩阵 $\boldsymbol{Y}_1$。$\boldsymbol{Y}_1$ 矩阵共有 N 行，其中，第 i 行中共有 M 个元素，它们表示第 i 个视觉更新的 3D 目标预测与所有雷达特征（共 M 个）之间的注意力得分。这些注意力得分的和为 1。注意力分数是评估融合网络性能的指标之一，它体现了视觉更新的 3D 目标预测与总雷达特征的关联程度。为了总雷达特征与视觉更新预测融合，融合模块定义了注意力加权雷达特征，其表示为 $\boldsymbol{F}^{*}_{\text{R\&L1}} = (\boldsymbol{Y}_1 \boldsymbol{F}_{\text{R\&L}}) \in \mathrm{R}^{N\times C}$。如图 3 所示，这些加权雷达特征与视觉更新预测相结合，通过前馈网络（Feed-Forward Network，FFN）$\boldsymbol{\Phi}_{\text{FFN1}}$ 进行增强。最终第一层解码器输出初级融合预测。第二层和第三层 Transformer 解码器有着与第一层解码器类似的工作方式。区别是它们的输入不再是视觉更新预测，而是上一级解码器输出

的融合预测。在两个解码器之后，应用两组 FFN 来执行边界框预测，并将第三层解码器输出的边界框确定为最终融合预测结果。

2　实验结果与分析

本实验在自制数据集上对本文方法的性能进行了测试，并采用 nuScenes 数据集检测基准对该方法进行评估。实验车及传感器安装情况如图 4 所示。RoboSense RS-LIDAR-16 激光雷达安装在车辆顶部，ARS 548 毫米波雷达安装在车辆前方，HIKROBOT MV-CE050-30UC 相机安装在车辆的前方。三个传感器对场景内的信息进行同步采集。数据集场景环境为常规城市道路，部分道路两侧停有车辆，道路场景信息丰富，行人、汽车数量较多。

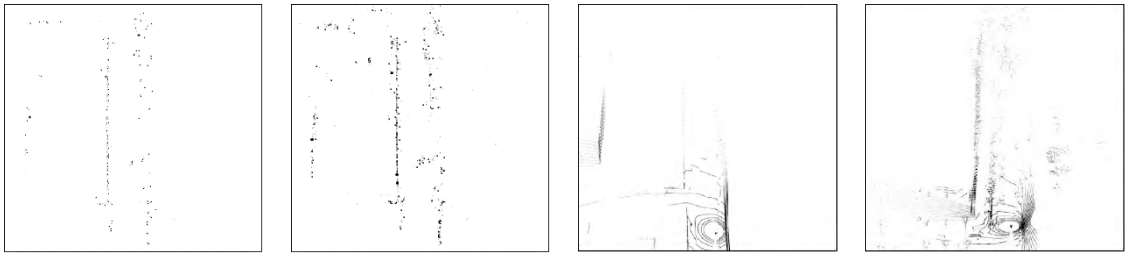
图 4　实验车及传感器安装情况

Fig. 4 Experimental vehicle and sensor installation

2.1　点云融合实验

毫米波雷达通过快慢时间维度的傅里叶变换得到目标的距离多普勒图像，然后采用恒虚警率技术与数字波束得到目标的角度等信息，最终得到单帧毫米波雷达点云。由于单帧毫米波雷达点云数据稀疏，无法对环境特征与目标信息进行准确的描述，本文提出点云时序融合算法，对连续 5 帧点云进行叠加并变成当前帧。点云叠加实验结果如图 5 所示。图 5(a)为单帧毫米波雷达点云，场景中点云数目为 200 个左右。环境中的道路信息非常模糊，图像中 35 m 外的目标点云呈现无规律性。图 5(b)为经过多帧叠加之后的点云图像，场景中的点云数目为 1200 个左右。通过对比叠加前后的图像，场景中 5 m 到 35 m 内路边停放的每辆车平均的点云数目为 120 个左右。场景中的点云数量明显增加，特征描述更加清晰，有效地提高了后续目标检测的可靠性。

针对毫米波雷达近距离小目标与激光雷达远距离目标的探测能力不足的问题，本文提出点云加权融合算法对预处理后的毫米波雷达点云与激光雷达点云进行数据融合，形成总雷达点云。激光雷达点云如图 5(c)所示。在该环境中，激光雷达点云在 35 m 内具有良好的点云密度及点云描述详细性。对于 35 m 之外的目标，激光雷达点云数量逐渐减少，无法对目标进行探测。融合后的总雷达点云如图 5(d)所示。车辆前方场景特征变得更加明显，感知范围得到有效增加。总雷达点云对于 50 m 内的目标有着准确的点云描述，整个环境内的特征信息变得更加完整，保证了后续目标检测任务的准确性与鲁棒性。



（a）单帧毫米波雷达点云　（b）多帧叠加点云　（c）激光雷达点云　（d）总雷达点云

图 5　点云融合实验结果对比

Fig. 5 Comparison of point cloud fusion experimental results

2.2 目标检测实验

本实验对比了本文算法与 SOTA 方法之间的目标检测性能。其中：Pointpillar[24]算法是基于激光雷达的 3D 目标检测算法，它通过将稀疏点云投影到柱状体结构中；CenterFusion[11]算法通过融合图像和毫米波雷达信息，检测 2D 目标中心并将其与雷达点关联；DeepFusion[25]通过深度特征交叉融合策略在 Transformer 架构下整合激光雷达和摄像头数据。结果如表 1 所示。表 1 第二列表示使用的传感器，其中，C 代表相机，L 代表激光雷达，R 代表毫米波雷达。通过对 nuScenes 数据级官方评估指标的对比可以看出，本文方法整体性能要优于现有的多传感器融合目标检测方法。得益于毫米波雷达、激光雷达与相机三种传感器性能的相互补充，相对于其他基于相机、激光雷达、激光雷达/相机融合、毫米波雷达/相机融合的目标检测方法，本文方法有着更高的平均均值精度(mean Average Precision，mAP)与 NuScenes 检测评分(NuScenes Detection Score，NDS)。另外，本文方法的平均均值速度误差(mean Average Velocity Error，mAVE)明显下降。因为激光雷达点云与毫米波雷达点云进行融合，增强了点云性能，使点云的测速能力得到了增强。毫米波雷达有着优异的测速性能，加上激光雷达点云的辅助，对速度的测量误差变小，所以 mAVE 误差降低。表 1 中的数据体现了本文方法在性能方面的先进性，验证了本文方法对多传感器融合目标检测任务的有效性。

表 1　与 SOTA 方法对比实验结果

Tab. 1 Comparison of experimental results with SOTA method

方法	传感器	NDS ↑	mAP ↑	mAVE ↓
DETR3D[9]	C	47.5	40.9	0.839
Pointpiller[24]	L	44.8	30.1	0.312
CenterFusion[11]	RC	44.3	31.6	0.612
DeepFusion[25]	LC	48.9	39.8	0.489
本文方法	**RLC**	**51.8**	**42.5**	**0.443**

在 nuScenes 数据集中，平均精度(Average Precision，AP)的计算不是基于传统的 IoU 阈值匹配，而是基于在地平面上的 2D 中心距离 d 来计算。nuScenes 定义了四个距离阈值，范围从 0.5 m 到 4.0 m。针对不同中心距离评价指标，表 2 展示了本文方法与其他 SOTA 方法，针对汽车类别检测的对比结果。如表 2 所示，本文方法在 4 个指标上 AP 均有所提高。

表 2　与 SOTA 方法的 NuScenses 中心距离指标实验对比结果

Tab. 2 Comparison with SOTA methods using nuScenes center distance evaluation metrics

方法	AP （d=0.5 m）	AP（d=1.0 m）	AP （d=2.0 m）	AP（d=4.0 m）
CenterFusion[11]	18.43	48.54	73.67	83.55
DeepFusion[25]	22.78	53.55	74.92	83.98
本文方法	23.43	54.02	75.36	84.54

为了直观地展现出本文方法与 SOTA 方法的性能效果对比，本实验在自制数据集上设计了本文方法与 SOTA 方法的实验结果对比。实验结果如图 6 所示。在实验结果图中，绿色框是对车辆的检测结果，蓝色框是行人的检测结果。结果中漏检的目标用红色框标出，误检的目标用黑色框标出。图 6 第二列为 CenterFusion 的实验结果。由于 CenterFusion 是基于毫米波雷达/相机的目标检测方法，其在检测距离上有着良好的表现。但是对于部分行人与车辆目标，CenterFusion 有漏检与误检的现象。图 6 第三列为本文方法的实验结果。可见光图像与点云分别进行特征提取，两种特征进入 Transformer 交叉注意力机制中，通过注意力权重加权融合，对环境中的目标进行检测与分类。另外，在雷达点云进行目标检测的基础上，可见光图像协助点云对目标的种类进行分类，将重叠的行人与车辆目标区分开。在行人目标的检测方面，本文方法有着准确的检测结果。在保证检测距离的同时，提高了检测的准确性。

图 6　与 SOTA 方法对比实验结果图

Fig. 6 Comparison of experimental results with SOTA method

2.3　消融实验

为了评估本文方法中每个组件的有效性，本实验设计将本文方法的各组件分别去掉作为消融实验。具体结果如表 3 所示。Baseline 为单帧毫米波雷达点云与相机融合的方法。Baseline 与本文方法相比，由于单帧毫米波雷达点云稀疏且分辨率低，导致误检率与漏检率较高，所以检测结果指标较低。方法 II 为缺少了多帧点云叠加的方法，该方法与本文方法相比，由于缺少了多帧毫米波雷达点云的远距离点云描述，激光雷达点云不能对远距离的目标进行检测，所以方法 II 对于远距离目标有着较高的漏检率，其各项指标均低于本文方法。方法 III 为缺少了激光雷达点云的方法，该方法与本文方法相比，由于缺少了激光雷达点云的近距离详细点云描述，仅依靠毫米波雷达点云不能准确地对小目标进行准确检测，所以方法 III 的行人目标检测结果不佳，性能指标不如本文方法。经过对各个组件有效性的验证，本文方法均表现出优异的性能。

表 3　消融实验结果

Tab. 3 Ablation experimental results

方法	多帧毫米波雷达点云叠加	激光雷达点云	NDS ↑	mAP ↑	mAVE ↓
Baseline			40.5	32.6	0.456
方法 II		✓	43.4	33.2	0.466
方法 III	✓		43.5	34.2	0.449
本文方法	✓	✓	43.9	35.4	0.471

为了直观展示本文方法中各个组件的作用，对上述五种方法的实验结果进行对比。实验结果如图 7 所示。其中绿色框代表车辆，蓝色框代表行人。红色框代表漏检的目标，黑色框代表误检的目标。图 7 第四列为本文方法的实验结果。第一列为 Baseline 的目标检测结果。Baseline 的误检率及漏检率均较高，无法得到准确的环境信息。第二列为方法 II 的目标检测结果。方法 II 去除了多帧毫米波雷达点云，与本文方法相比，由于缺少了多帧毫米波雷达点云对远距离目标的点云描述，仅依靠激光雷达无法对远距离目标进

行准确检测，出现漏检及误检情况。第三列为方法 III 的目标检测结果。方法 III 去除了激光雷达点云，与本文方法相比，缺少激光雷达的近距离详细点云，无法对行人等小目标进行准确的检测。所以对于行人及重叠目标，方法 III 存在漏检的情况。综上所述，本文方法的各组件对于提高目标检测准确率及降低误检率、漏检率方面有着显著的贡献。

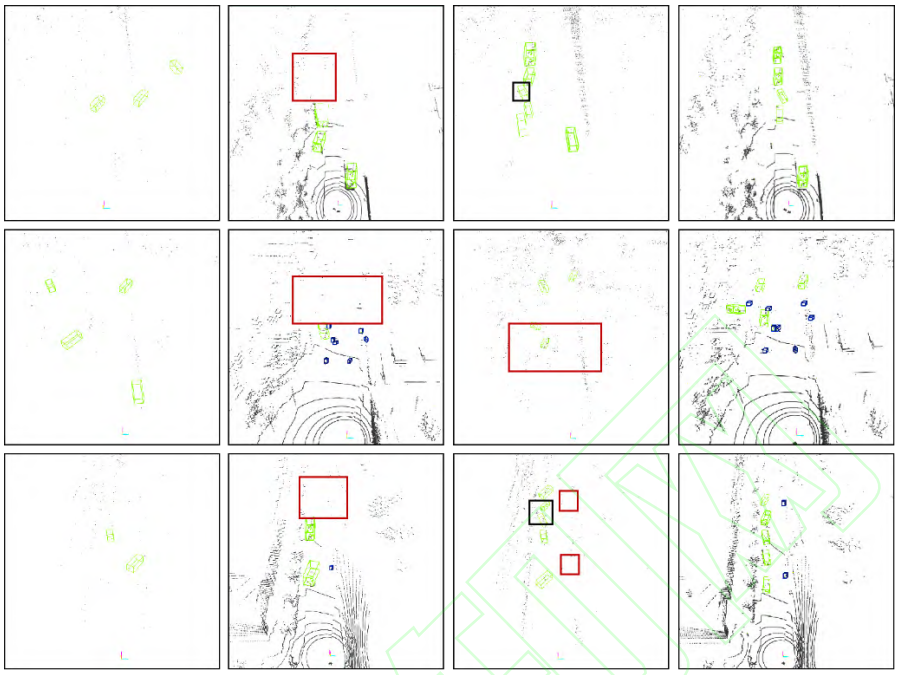


图 7　消融实验结果图

Fig. 7 Ablation experimental results

3　结　论

为了提高环境感知的准确性和可靠性，本文提出一种基于 Transformer 的毫米波雷达/激光雷达/相机融合的 3D 目标检测方法。该方法对多帧毫米波雷达点云进行融合，并将处理后的毫米波雷达点云与激光雷达点云融合得到总雷达点云。该方法的 Transformer 框架对视觉更新的 3D 目标预测与总雷达特征之间的交互进行学习，输出 3D 目标检测结果。本文在常规城市道路环境自制数据集上进行测试，并采用 nuScenes 数据集检测指标对该方法进行评估，指标结果体现了本文方法的检测性能优势，有效地提高了 3D 目标检测任务的准确性与鲁棒性。

参考文献（References）

[1] SHI S, GUO C, JIANG L, et al. Pv-rcnn: Point-voxel feature set abstraction for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 10529-10538.

[2] SHI S, WANG X, LI H. Pointrcnn: 3D object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE,2019: 770-779.

[3] 刘丰宇,程向红,李艺恒,等.基于惯性/仿生视觉/激光雷达的智能感知无人系统[J].飞控与探测,2024,7(3):22-30.
LIU F Y, CHENG X H, LI Y H, et al. Intelligent perception unmanned system based on inertial/bionic vision/lidar [J]. Flight Control & Detection, 2024,7(3):22-30（in Chinese）.

[4] YIN T, ZHOU X, KRAHENBUHL P. Center-based 3D object detection and tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 11784-11793.

[5] BRAZIL G, LIU X. M3D-rpn: Monocular 3D region proposal network for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019: 9287-9296.

[6] CHEN Y, TAI L, SUN K, et al. Monopair: Monocular 3D object detection using pairwise spatial relationships[C]//Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 12093-12102.

[7] SHEN C, ZHAO X, WU X, et al. Multi-aperture visual velocity measurement method based on biomimetic compound-eye for UAVs[J]. IEEE Internet of Things Journal, 2023，11（7）： 11165 - 11174.

[8] 姜忠旭,高磊,关智聪,等.基于 YOLOv8 的遥感图像舰船目标检测算法[J].飞控与探测,2024,7(3):56-66.

JIANG Z X, GAO L, GUAN Z C, et al. Ship target detection algorithm in remote sensing images based on YOLOv8 [J]. Flight Control and Detection,2024,7(3):56-66（in Chinese）.

[9] WANG Y, GUIZILINI V C, ZHANG T, et al. Detr3D: 3D object detection from multi-view images via 3D-to-2D queries[C]//Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022: 180-191.

[10] 全刚，王龙，郑昊鹏，等．单脉冲雷达避盲对测距的影响分析[J/OL]. 飞控与探测,1-7[2025-03-04].http://kns.cnki.net/kcms/detail/10.1567.TJ.20250210.0954.002.html.

QUAN G, WANG L, ZHENG H P, et al. Analysis of the impact of monopulse radar blind-avoidance on distance measurement [J/OL]. Flight Control and Detection, 1-7[2025-03-04].http://kns.cnki.net/kcms/detail/10.1567.TJ.20250210.0954.002.html（in Chinese）.

[11] NABATI R, QI H. Centerfusion: Center-based radar and camera fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Snowbird, UT, USA: IEEE, 2021: 1527-1536.

[12] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, WA, USA: IEEE,2020: 2446-2454.

[13] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 7345-7353.

[14] 关世伟, 李志, 王佳俊, 等. 基于毫米波雷达和相机融合的无人碾压机施工障碍物快速精准感知方法 [J]. 水利学报, 2024, 55 (11): 1404-1416.

GUAN S W, LI Z, WANG J J, et al. A fast and accurate perception method for construction obstacles of unmanned roller compacted machine based on millimeter wave radar and camera fusion[J]. Journal of Hydraulic Engineering, 2024, 55 (11): 1404-1416（in Chinese）.

[15] 刘威,许勇,方娟,等.基于雷达和视觉融合的多模态空中手写体识别研究[J/OL].计算机科学,1-11[2025-03-03].http://kns.cnki.net/kcms/detail/50.1075.tp.20241101.1104.002.html.

LIU W, XU Y, FANG J, et al. Research on multi-modal aerial handwriting recognition based on radar and vision fusion [J/OL]. Computer Science, 1-11. [2025-03-03]. http://kns.cnki.net/kcms/detail/50.1075.tp.20241101.1104.002.html（in Chinese）.

[16] 李健浪,吴新电,陈灵,等.基于 4D 毫米波雷达与视觉融合的三维目标检测算法[J/OL].计算机工程,1-15[2025-03-03].https://doi.org/10.19678/j.issn.1000-3428.0070113.

LI J L, WU X D, CHEN L, et al. A 3D target detection algorithm based on 4D millimeter-wave radar and vision fusion [J/OL]. Computer Engineering, 1-15. [2025-03-03].https://doi.org/10.19678/j.issn.1000-3428.0070113（in Chinese）.

[17] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 4604-4612.

[18] XU D, ANGUELOV D, JAIN A. Pointfusion: Deep sensor fusion for 3D bounding box estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 244-253.

[19] 李沂杨,陆声链,王继杰,等.基于 Transformer 的 DETR 目标检测算法研究综述[J/OL].计算机工程,1-20[2025-03-03].https://doi.org/10.19678/j.issn.1000-3428.0069312.

LI X Y, LU S L, WANG J J, et al. Review of DETR target detection algorithm based on Transformer [J/OL]. Computer Engineering, 1-20. [2025-03-03].https://doi.org/10.19678/j.issn.1000-3428.0069312（in Chinese）.

[20] SHEN C, WU Y, QIAN G, et al. Intelligent bionic polarization orientation method using biological neuron model for harsh conditions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024，46(3): 1234-1245.

[21] LIU X, TANG J, SHEN C, et al. Brain-like position measurement method based on improved optical flow algorithm[J]. ISA transactions, 2023, 143: 221-230.

[22] SHEN C, XIONG Y, ZHAO D, et al. Multi-rate strong tracking square-root cubature Kalman filter for MEMS-INS/GPS/polarization compass integrated navigation system[J]. Mechanical Systems and Signal Processing, 2022, 163: 108146.

[23] 马郑凯, 周林立, 梁兴柱. 改进特征金字塔网络的小目标检测[J]. 电光与控制, 2024, 31(12): 48-54.
MA Z K, ZHOU L L, LIANG X Z. Improved feature pyramid network for small target detection [J]. Electro-Optics and Control, 2024, 31 (12): 48-54（in Chinese）.

[24] LANG A H, VORA S, CAESAR H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 12697-12705.

[25] LI Y, YU A W, MENG T, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 17182-17191.