基于多传感器融合的协同感知方法

王秉路[①②] 靳 杨[①] 张 磊[③] 郑 乐[②] 周天飞*[④]

[①](西安建筑科技大学信息与控制工程学院 西安 710399)

[②](北京理工大学信息与电子学院 北京 100081)

[③](西北工业大学自动化学院 西安 710129)

[④](北京理工大学计算机学院 北京 100081)

摘要：该文提出了一种新的多模态协同感知框架，通过融合激光雷达和相机传感器的输入来增强自动驾驶感知系统的性能。首先，构建了一个多模态融合的基线系统，能有效地整合来自激光雷达和相机传感器的数据，为后续研究提供了可比较的基准。其次，在多车协同环境下，探索了多种流行的特征融合策略，包括通道级拼接、元素级求和，以及基于Transformer的融合方法，以此来融合来自不同类型传感器的特征并评估它们对模型性能的影响。最后，使用大规模公开仿真数据集OPV2V进行了一系列实验和评估。实验结果表明，基于注意力机制的多模态融合方法在协同感知任务中展现出更优越的性能和更强的鲁棒性，能够提供更精确的目标检测结果，从而增加了自动驾驶系统的安全性和可靠性。

关键词：自动驾驶；协同感知；3D目标检测；多模态融合；智能交通系统

中图分类号：TN957.51 **文献标识码**：A **文章编号**：2095-283X(2024)01-0087-10

DOI: 10.12000/JR23184

引用格式：王秉路, 靳杨, 张磊, 等. 基于多传感器融合的协同感知方法[J]. 雷达学报(中英文), 2024, 13(1): 87–96. doi: 10.12000/JR23184.

Reference format: WANG Binglu, JIN Yang, ZHANG Lei, *et al*. Collaborative perception method based on multisensor fusion[J]. *Journal of Radars*, 2024, 13(1): 87–96. doi: 10.12000/JR23184.

Collaborative Perception Method Based on Multisensor Fusion

WANG Binglu[①②] JIN Yang[①] ZHANG Lei[③] ZHENG Le[②] ZHOU Tianfei*[④]

[①](*College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an* 710399, *China*)

[②](*School of Information and Electronics, Beijing Institute of Technology, Beijing* 100081, *China*)

[③](*School of Automation, Northwestern Polytechnical University, Xi'an* 710129, *China*)

[④](*School of Computer Science & Technology, Beijing Institute of Technology, Beijing* 100081, *China*)

Abstract: This paper proposes a novel multimodal collaborative perception framework to enhance the situational awareness of autonomous vehicles. First, a multimodal fusion baseline system is built that effectively integrates Light Detection and Ranging (LiDAR) point clouds and camera images. This system provides a comparable benchmark for subsequent research. Second, various well-known feature fusion strategies are investigated in the context of collaborative scenarios, including channel-wise concatenation, element-wise summation, and transformer-based methods. This study aims to seamlessly integrate intermediate

收稿日期：2023-10-04；改回日期：2023-12-10；网络出版：2023-12-27

*通信作者： 周天飞 ztfei.debug@gmail.com *Corresponding Author: ZHOU Tianfei, ztfei.debug@gmail.com

基金项目：中国博士后科学基金(2022M710393, 2022TQ0035)

Foundation Items: China Postdoctoral Science Foundation (2022M710393, 2022TQ0035)

责任主编：刘凡 Corresponding Editor: LIU Fan

representations from different sensor modalities, facilitating an exhaustive assessment of their effects on model performance. Extensive experiments were conducted on a large-scale open-source simulation dataset, *i.e.*, OPV2V. The results showed that attention-based multimodal fusion outperforms alternative solutions, delivering more precise target localization during complex traffic scenarios, thereby enhancing the safety and reliability of autonomous driving systems.

Key words: Autonomous driving; Collaborative perception; 3D object detection; Multimodal fusion; Intelligent transportation systems

1 引言

协同感知是自动驾驶领域中的一个关键问题，它允许自动驾驶车辆通过车载传感器收集环境信息并通过车对车(Vehicle-to-Vehicle, V2V)无线通信技术与其他车辆进行实时共享，从而实现更强大和全面的环境感知能力[1–3]。协同感知任务的目标是利用多个车辆作为移动传感器网构建高精度的多点观测场景表征，以增强车辆在车队协同、交通流优化、自动驾驶辅助和自动驾驶等应用场景下的感知能力[4]。相对于传统的单车智能，多车协同感知可以更好地应对复杂的交通场景，提高自动驾驶系统的决策准确性和行驶安全性。

早期的研究主要基于单一传感器模式，如仅使用激光雷达[5–14]或仅使用相机[15,16]进行环境感知。然而，这些单模态方法未能充分利用两种传感器的互补优势，限制了感知系统的性能。例如，基于纯激光雷达的方法可能会忽略相机传感器捕捉到的细粒度视觉细节，这些细节对于一些特殊情境的目标检测和识别至关重要。此外，相机可以捕捉到颜色、纹理和形状等物体的视觉特征，这些特征对于识别特定类别的目标非常有帮助，如识别交通标志或行人。另外，基于纯相机的方法虽然能够提供丰富的语义信息，但通常缺乏对于精确目标定位至关重要的准确深度信息，这导致了在三维环境感知方面的一些挑战，如避免碰撞或进行高精度的车道保持，而激光雷达能够提供准确的距离和深度测量。此外，在低光或恶劣天气条件下，相机可能受到限制，而激光雷达(Light Detection and Ranging, LiDAR)能够继续提供有用的几何信息，因此在这些场景中具有独特的优势[17]。

为了克服单传感器的局限性，近年来，研究人员开始探索基于多传感器融合的感知方法[18–23]。这些方法旨在将LiDAR和相机的优势结合起来，以获得更全面、准确的环境感知。PointPainting[19]首次将图像上的语义信息附加到激光雷达点上实现不同模态的点级融合。BEVFusion[20,21]探索了适合多模态特征融合的统一表示，即将LiDAR模态和相机模态转换到一个共享的鸟瞰(Bird's Eye View, BEV)空间实现融合。

先前的研究已经证明了融合LiDAR和相机数据的优势，包括改进的目标检测、增强的场景理解以及在具有挑战性的环境条件下的稳健性能。然而，当涉及多车协同感知任务时，在多模态融合方面的探索仍然非常有限。最近提出的HM-ViT[24]是多车协同感知在多模态融合方面的初期探索，它假设每个车辆只能获取任意一种模态，并基于此设定来实现不同车辆之间的异构模态融合。

尽管现有文献中已有大量关于多模态的研究，但在协同感知领域中，对于多模态交互的深入探讨仍显不足。为了更全面地利用不同类型传感器的互补优势，本文构建了一个多模态融合的基线系统，作为实现协同感知的基础框架。该基线系统采用了一种集成激光雷达和相机数据的方法，通过将两种传感器的数据融合到同一表示空间中，能够利用激光雷达提供的精确深度信息和相机提供的丰富视觉特征，实现更准确、更稳健的环境感知。在构建该基线系统时，本文考虑了多种不同的融合策略，并进行了深入的分析。首先，采用了通道级拼接和元素级求和方法，这两种方法简单直观，能够将来自不同传感器的特征直接拼接在一起，形成一个统一的特征表示。虽然这种方法在某些情况下能够取得良好的性能，但它也有一定的局限性，主要表现在不能充分考虑到不同传感器数据之间的相关性。为了解决这一问题，本文进一步探索了基于注意力机制的融合方法，这种方法能够自适应地调整不同传感器数据的权重，从而更好地捕捉它们之间的相关性。通过多头自注意力机制，所提算法能够在不同的传感器特征之间建立复杂的关联，从而实现更精细的融合。在OPV2V[8]数据集上的实验结果表明，这种基于注意力机制的融合方法相比传统的融合方法，在协同感知任务中展现出更优越的性能和更强的鲁棒性。

2 传统协同感知范式回顾

考虑了一个具有N个智能网联车的协同感知场

景，其中每个车辆配有车载传感器(LiDAR或相机)获取局部观测数据(3D点云或图像)并通过无线通信网络进行交互和协作。具体来说，假设$\boldsymbol{A}=\{a_1,a_2,\cdots,a_N\}$为所有$N$个车辆的集合，集合中的第$i$个车辆$a_i$为中心车辆，则与其进行协作的其他车辆集合定义为$\boldsymbol{J}=\{a_j\}_{j\in[N]\backslash\{i\}}$。传统的协同感知范式主要由特征提取、特征融合以及检测头网络组成。假设车辆a_i的局部观测数据为$\boldsymbol{O}_i$，首先，特征提取负责根据$\boldsymbol{O}_i$提取相应的中间层特征表示$\boldsymbol{F}_i$：

$$\boldsymbol{F}_i=\text{Backbone}(\boldsymbol{O}_i) \tag{1}$$

然后，车辆a_i将自身的特征$\boldsymbol{F}_i$与来自其他车辆的特征$\{\boldsymbol{F}_j\}_{j\in\boldsymbol{J}}$通过特征融合获得更全面的特征表示$\boldsymbol{H}_i$：

$$\boldsymbol{H}_i=\text{Fusion}(\boldsymbol{F}_i,\{\boldsymbol{F}_j\}_{j\in\boldsymbol{J}}) \tag{2}$$

最后，将融合后的特征$\boldsymbol{H}_i$送入检测头网络生成预测结果$\boldsymbol{Y}_i$：

$$\boldsymbol{Y}_i=\text{DetectionHead}(\boldsymbol{H}_i) \tag{3}$$

然而，现有的协同感知范式存在一些局限性。首先，模态的单一性忽略了其他模态所提供的有用信息。其次，现有特征融合策略都相对简单，通常是简单的平均或者加权平均，这可能会导致某些重要的特征在融合过程中被淡化或丢失。这些局限性为进一步的研究和优化提供了机会。

3　多模态融合的协同感知方法

本节，对所提出的多模态融合协同感知框架进行了概述。图1展示了该框架的核心组成部分，涵盖多模态特征提取、多模态特征融合以及检测头网络。各组成部分共同协作，实现LiDAR与相机数据的有效整合，充分挖掘二者的互补性，以提高V2V场景下的感知准确性。

3.1　多模态特征提取

为了捕捉和保留来自不同模态的独特线索，本文使用单独的分支进行特征提取并生成统一的BEV表示。

对于多视角图像数据，本文采用CaDDN (Categorical Depth Distribution Network)[25]架构，该架构包含4个主要模块：编码器、深度估计、体素变换和折叠，确保从输入图像中捕捉到尽可能丰富和准确的信息。为了将2D图像特征和3D点云特征这两种异构特征进行融合，需要显式地预测图像特征中每个像素的深度来将2D平面提升到3D空间，最终转换到统一的BEV空间。以车辆a_i为例，首先，编码器模块对原始的输入图像$I^i\in\mathbb{R}^{h\times w\times 3}$进行初步的特征抽取，生成维度为$X\times Y\times D$的图像特征$\boldsymbol{F}_I^i$，该过程可以表示为

$$\boldsymbol{F}_I^i=\text{Encoder}(I^i)\in\mathbb{R}^{X\times Y\times D} \tag{4}$$

然后，深度估计模块为每个像素预测出一个深度概率分布$\boldsymbol{P}$，可以表示为

$$\boldsymbol{P}=\text{DepthPred}(\boldsymbol{F}_I^i)\in\mathbb{R}^{X\times Y\times D_{\text{depth}}} \tag{5}$$

该分布反映了该像素的深度信息。之后，体素转换模块将先前抽取的特征从2D投影到3D空间。它根据所有可能的深度分布和图像的校准矩阵，生成的相应的3D体素特征：

$$\boldsymbol{V}=\text{Lift}(\boldsymbol{F}_I^i,\boldsymbol{P},\text{CalibMat})\in\mathbb{R}^{X'\times Y'\times Z'} \tag{6}$$

最后，折叠模块将3D体素特征合并到一个高度平面上

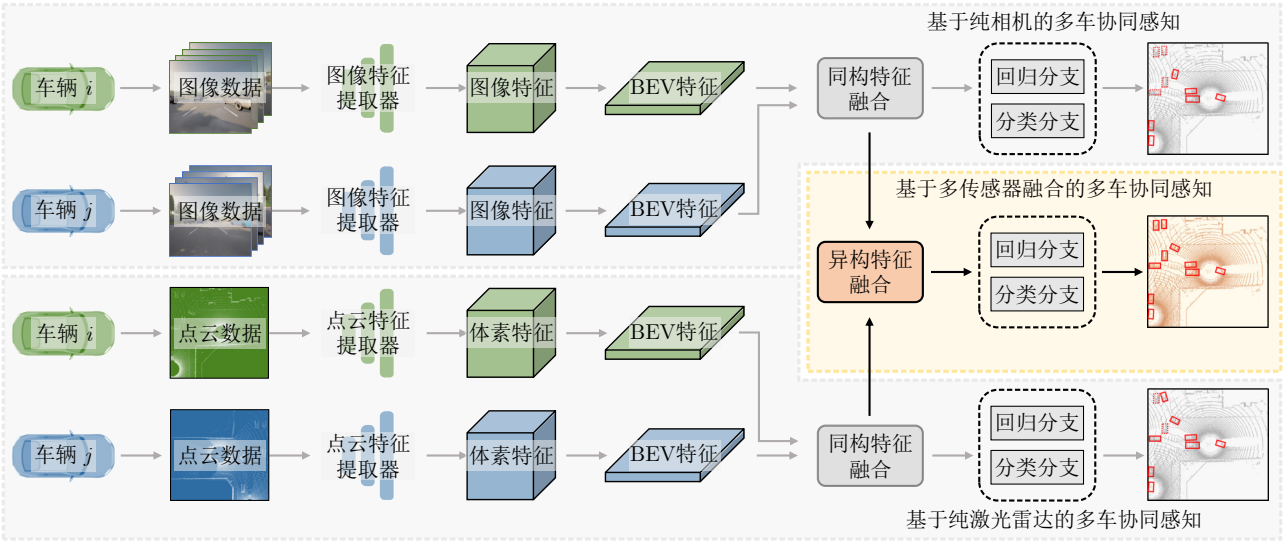


图 1 多传感器融合的协同感知框架

Fig. 1 Multisensor fusion collaborative perception framework

$$\boldsymbol{B}_I^i = \text{Collapse}(\boldsymbol{V}) \in \mathbb{R}^{H\times W\times C} \tag{7}$$

其中，$\boldsymbol{B}_I^i$为生成的BEV特征图，H和W表示图像BEV网格的高度和宽度，C表示通道数。该过程可以有效地获取语义丰富的视觉特征。

在处理点云数据时，由于其特有的稀疏性与三维结构，直接利用传统的3D卷积网络很可能引发计算和内存的巨大开销。为了有效而高效地从这种数据中提取特征，本文将PointPillar[26]作为点云数据的特征提取器。首先，对于给定的3D点$\boldsymbol{p}\in\mathbb{R}^3$，其在柱状坐标系中的位置可以被定位为$\boldsymbol{p}=(i,j,l)$。这里，$i$和$j$分别是二维网格的$x$和$y$坐标，而$l$表示其垂直方向上的高度。然后，将这三维空间划分为一系列均匀间隔的柱状结构。形式上，这种划分可以表示为

$$\boldsymbol{p}' = \left(\left\lfloor\frac{i}{W_{\text{pillar}}}\right\rfloor, \left\lfloor\frac{j}{H_{\text{pillar}}}\right\rfloor, l\right) \tag{8}$$

其中，W_{pillar}和H_{pillar}分别表示柱状体在x和y方向上的宽度和高度。通过这种方式可以将复杂的3D数据转换为2D的结构，每一个柱子内的点共享相同的高度信息。随后，所有柱状体沿其高度方向被压平，生成一个伪图像。对于柱状体内的所有3D点，其对应的3D特征为$\boldsymbol{F}_{\text{3D}}\in\mathbb{R}^{N_p\times C_p}$(其中，$N_p$是柱体中点的数量，$C_p$是点云特征的维度)被转化为2D柱体特征$\boldsymbol{F}_{\text{pillar}}\in\mathbb{R}^{1\times C_p'}$

$$\boldsymbol{F}_{\text{pillar}} = \phi(F_{\text{3D}}) \in \mathbb{R}^{1\times C_p'} \tag{9}$$

其中，$\phi(\cdot)$是转换函数，用于将柱体内所有的点转换为该柱体的整体表示。最后，通过一系列二维卷积对$\boldsymbol{F}_{\text{pillar}}$进行进一步的特征编码与整合，得到维度为$H\times W\times C$的BEV特征$\boldsymbol{B}_L^i$，其维度与图像BEV特征相同。

3.2 多模态特征融合

随着从激光雷达和相机中提取的BEV特征的获取，特征融合环节成为关键。在这一阶段，各车辆首先对自身的BEV特征进行压缩编码，然后向中心车辆发送。当中心车辆成功接收来自所有其他车辆的BEV特征后，将这些多模态信息进行有策略的融合，生成一个更为全局和详尽的场景表示。该融合过程主要涉及两个关键步骤：同构特征融合与异构特征融合。其中同构特征融合主要处理来自相同传感器类型的信息，而异构特征融合则旨在结合不同类型传感器的信息，充分挖掘不同模态之间的互补性。

3.2.1 同构模态特征融合

在多模态间的特征融合中，同构特征的融合显然比异构特征更为直观和简单，因为它们来自相同的数据源，共享相似的特性和分布。针对这一特点，本文为图像BEV特征和点云BEV特征分别设计了不同的融合策略，以更直接的方式整合同一模态下的多源信息。

首先，受文献[16]的启发，本文采用了元素级的最大化操作来融合不同车辆的图像BEV特征。其背后的原理是：在多车协同的环境中，某些车辆观察到的某些特征可能比其他车辆更为明显或清晰。而通过比较每个车辆的图像BEV特征，并采用元素级的最大值，可以确保最终融合的特征包含了所有车辆观察到的最显著特点。具体来说，当中心车辆收到其他车辆发送的图像BEV特征$\{\boldsymbol{B}_I^j\}_{j\in\boldsymbol{J}}\in\mathbb{R}^{(N-1)\times H\times W\times C}$时，会逐元素地将其与自身的图像BEV特征$\boldsymbol{B}_I^i\in\mathbb{R}^{H\times W\times C}$进行逐元素比较，然后选取最大值作为融合结果。该过程可表示为

$$\boldsymbol{B}_I^{\text{fused}} = \max\left(\boldsymbol{B}_I^i, \left\{B_I^j\right\}_{j\in\boldsymbol{J}}\right) \in \mathbb{R}^{H\times W\times C} \tag{10}$$

其中，$\max(\cdot)$操作是逐元素执行的，这种方式确保融合后的特征继承了多个车辆中最为显著和丰富的部分。

对于点云BEV特征，本文遵循文献[8]利用自注意力机制进行融合。点云BEV特征的融合首先涉及构建本地图的构建。本地图连接了来自不同车辆但处于相同空间位置的特征向量。在这个过程中，每个特征向量被视为一个节点，两个特征向量之间的边用于连接来自不同车辆的相同空间位置的特征向量，这样的连接不仅有助于融合来自不同数据源的信息，同时也为后续的自注意力机制处理提供了基础。

然后，对该本地图应用自注意力机制进行点云BEV特征融合。自注意力机制的核心在于其能够赋予不同特征向量以不同的权重，这些权重反映了特征之间的相互依赖关系。在点云BEV特征的情境下，这意味着每个特征点不仅被其自身属性所定义，还受到周围特征点的影响。具体而言，对于每个车辆的点云BEV特征$\boldsymbol{B}_L^i\in\mathbb{R}^{H\times W\times C}$，将该二维特征图展开为维度为$M\times C$的一维特征向量，其中$M=H\times W$。接着，该特征向量会被转换成查询($Q$)、键($K$)和值($V$)3部分，进而通过自注意力公式计算得到更新后的特征向量$\tilde{\boldsymbol{B}}_L^i$。这一过程可以描述为

$$\tilde{\boldsymbol{B}}_L^i = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \in \mathbb{R}^{M\times C} \tag{11}$$

其中，d_k是键向量的维度，$\text{softmax}(\cdot)$函数则确保所有权重加起来为1。最后将更新后的特征向量变

换为原始维度$H \times W \times C$。按照以上方式更新所有车辆的BEV特征并堆叠起来得到融合结果$\boldsymbol{B}_L^{\text{fused}} \in \mathbb{R}^{H\times W\times 3C}$。

通过多源同构模态特征的融合，所得到的融合特征更好地整合和表达了来自多个车辆相同传感器的观测信息，从多个角度捕获了环境细节的丰富性和多样性。同时，这也为后续进行异构特征融合奠定了坚实基础。深度融合不同传感模态之间的互补信息可以进一步丰富环境描述，充分发挥各类传感资源的价值。

3.2.2 异构模态特征融合

在异构模态融合阶段，本文整合了不同模态的融合BEV特征，以利用激光雷达和相机数据之间的互补信息。本文采用3种常见的融合策略：通道级拼接、元素级求和以及Transformer融合，以有效地组合模态并生成综合表示。具体来说，给定融合后的图像特征$\boldsymbol{B}_I^{\text{fused}} \in \mathbb{R}^{H\times W\times C}$和点云特征$\boldsymbol{B}_L^{\text{fused}} \in \mathbb{R}^{H\times W\times 3C}$，融合过程如下：

(1) 通道级拼接融合。通道级拼接是一种直观且广泛应用的特征融合方法。这种方法主要依赖于特征的空间一致性，即相同的空间位置上来自不同模态的特征被认为是相关的。在通道级拼接融合中，首先将图像和点云特征沿通道维度拼接，得到一个维度为$H \times W \times 4C$的特征张量$\boldsymbol{B}^{\text{cat}}$。然后将拼接后的特征送入两个卷积层进行进一步处理。其中，第1个卷积层用于捕捉空间关系并从拼接数据中提取有价值的特征。随后，第2个卷积层将通道维度降低到C来进一步细化融合特征。

(2) 元素级求和融合。元素级求和旨在保持空间结构的同时融合两种模态的互补信息。在逐元素求和融合模块中，首先将LiDAR特征$\boldsymbol{B}_L^{\text{fused}}$送入一个1×1卷积层对通道维度进行降采样。然后，对图像特征和变换后的LiDAR特征进行逐元素相加，得到融合特征$\boldsymbol{B}^{\text{fused}} \in \mathbb{R}^{H\times W\times C}$。这种元素及求和的特征融合策略确保了两种模态对最终融合结果的贡献是相等的。

(3) Transformer融合。近年来，Transformer[27]架构在自然语言处理领域取得了很大的成功，尤其是在捕获序列数据中的长距离依赖关系方面。在多模态融合中，Transformer因其卓越的性能和对长距离依赖关系的处理能力而受到广泛关注[28]。特别是，其内部的自注意机制为不同模态特征之间的相互作用和交互提供了一个强大的框架。在异构模态特征融合的过程中，每种模态的特征都可视为一个“序列”，其中每个“元素”都代表着空间信息的一部分。

首先，将图像和点云的BEV特征沿通道维度连接，形成一个联合特征张量

$$\boldsymbol{B}^{\text{cat}} = \text{concat}(\boldsymbol{B}_I^{\text{fused}}, \boldsymbol{B}_L^{\text{fused}}) \tag{12}$$

考虑到原始的Transformer的结构不包含关于元素位置的任何信息，需要为此联合特征添加位置编码，确保模型能够识别特征在空间上的位置

$$\bar{\boldsymbol{B}}^{\text{cat}} = \boldsymbol{B}^{\text{cat}} + \text{PositionEncoding}(\boldsymbol{B}^{\text{cat}}) \tag{13}$$

接着，为了捕获两种模态之间的复杂交互，本文引入了多头自注意力机制。这种机制使模型能够为不同模态的每个部分分配不同的权重，多头的设计可以使模型从多个角度或来捕捉不同模态特征之间的相关性

$$\boldsymbol{B}^{\text{att}} = \text{MultiHeadAttention}(\bar{\boldsymbol{B}}^{\text{cat}}) \tag{14}$$

然后，通过前馈网络进一步强化模型的非线性处理能力

$$\boldsymbol{B}^{\text{FFN}} = \text{FeedForwardNetwork}(\boldsymbol{B}^{\text{att}}) \tag{15}$$

最后，为确保模型的稳定性并维持特征规模，在每一步后都采用了残差连接和层归一化技术得到最终的融合特征

$$\boldsymbol{B}^{\text{fused}} = \text{LayerNorm}(\bar{\boldsymbol{B}}^{\text{cat}} + \boldsymbol{B}^{\text{FFN}}) \tag{16}$$

整体来说，这种基于Transformer的融合策略允许算法充分考虑和利用两种模态之间的复杂交互和依赖关系，从而为下游任务提供一个高度丰富和代表性的特征表示。

3.3 检测头网络

检测头网络负责根据融合特征来预测目标的类别和位置。该网络包含3个反卷积层，这些层负责上采样特征图，从而为后续的类别和位置预测提供更细粒度的信息。之后是检测头中的两个分支：类别预测分支和边界框回归分支。类别预测分支为每个锚框输出一个分数，指示该锚框内是否存在某一特定类别的对象，以及存在的概率。这些输出分数经过激活函数处理后，可以转化为各个类别的概率分布。另一方面，边界框回归分支则负责细化目标的位置信息。它为每个锚框预测4个值，这4个值代表中心位置的偏移量以及框的宽度和高度的变化。通过这些预测值，可以校正锚框的位置，使其更紧密地围绕目标对象。综合两个分支的输出，检测头网络输出每个锚框中目标类别和调整后的位置信息。

4 实验

4.1 实验设置

本文实验是在OPV2V[8]数据集上进行的。OPV2V数据集汇集了由仿真框架OpenCDA[29]以及CARLA[30]

模拟器所模拟的LiDAR点云与RGB图像的大量数据。数据集总计11464帧，主要分为两个子集：CARLA默认城镇子集和Culver City数字城镇子集。其中，CARLA默认城镇子集占10914帧，按6764帧、1980帧和2170帧的比例分为训练、验证和测试3部分。该子集覆盖了不同复杂程度的多种场景，为协同感知模型提供了充足的训练与评估数据。相对而言，Culver City子集只有550帧，但其目标在于评估模型在现实世界场景的泛化表现，特别是那些对模型感知能力构成挑战的实际城市环境。

在实现细节方面，所提算法基于PyTorch框架，并在配有24 GB RAM的NVIDIA RTX 4090 GPU的PC上进行训练。在训练过程中，随机选择了一组能够在场景中建立通信的车辆，并规定每个车辆的通信范围为70 m。同时，点云的范围被设定为沿x,y与z轴的[–140.8, 14.8]×[–40, 40]×[–3, 1]。体素的分辨率为0.4 m。为了提升训练数据的多样性，采用了一系列数据增强技术，如随机翻转、±0.05范围内缩放以及±45°范围内旋转。利用Adam优化器对模型进行训练，设定模型的初始学习率为0.002，批量大小为2。此外，利用了基于验证损失的早停策略以防止模型过拟合。本文算法的详细架构和其他参数如图2所示。

性能评价指标方面，本文采用了标准的评价指标来评估算法的性能，即平均精度(Average Precision, AP)。平均精度是一种常用的目标检测性能指标，用于衡量算法在不同交并比(Intersection over Union, IoU)阈值下的准确性。在计算平均精度时，需事先设定交并比IoU的阈值，即预测框和真值框的重合面积占预测框和真值框面积总和的比例，如果大于设定阈值则认为检测正确。本文分别计算模型在IoU阈值为0.5和0.7时的AP值，即AP@0.5和AP@0.7。这两个指标能够提供对算法在不同严格程度下的性能评估，从而更全面地衡量算法在目标检测任务中的表现。

根据不同的异构模态融合策略，本文构建了

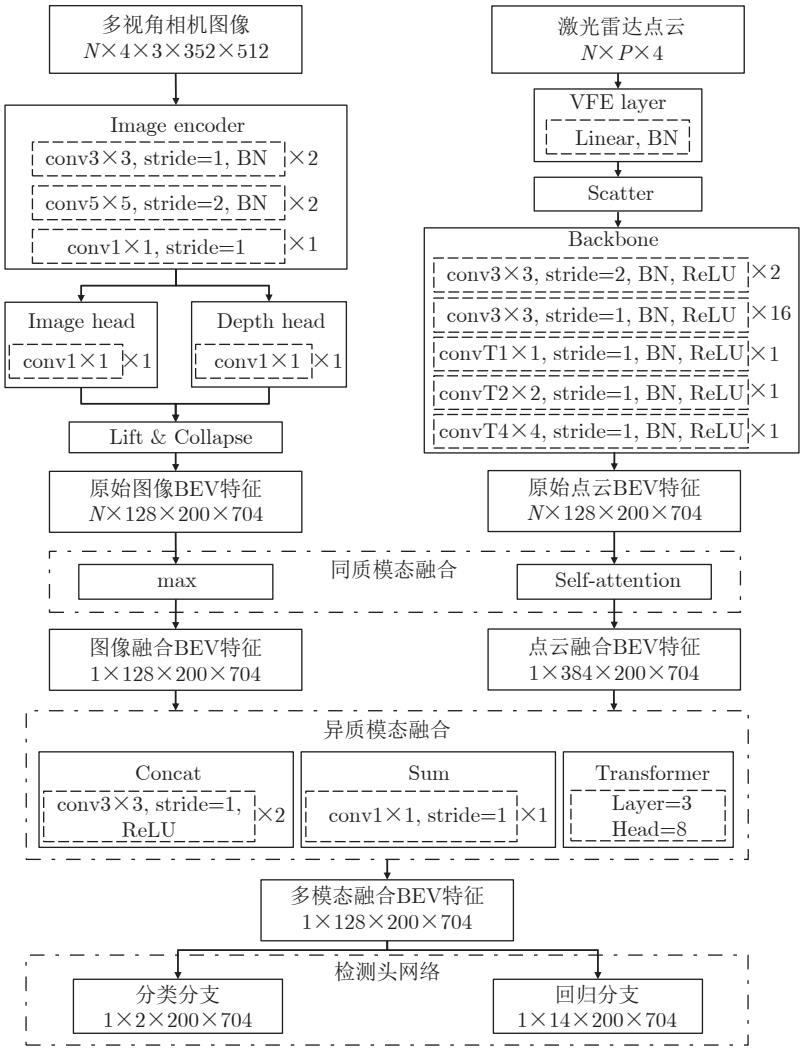


图 2 模型详细架构与参数细节

Fig. 2 Detailed model architecture and parameter specifics

3个版本的多模态融合模型：“Ours-C”表示通道级拼接融合，“Ours-S”对应于元素级求和融合，“Ours-T”代表Transformer融合。此外，为了更全面地验证所提多模态融合模型的效果，本文算法还与多种现有先进算法进行了对比，包括Cooper[5], F-Cooper[6], V2VNet[7], AttFuse[8]以及CoBEVT[15]。

除了考虑所有车辆都可以同时获得点云和图像两种模态的数据这种一般场景以外，本文还考虑了更加现实和复杂的异构模态场景。在这种异构模态场景下，每个车辆只能获取点云和图像中的任意一种模态，用于模拟所提算法在各种模态缺失情况下的性能表现。

4.2 定量实验

在表1中，本文将多种SOTA算法在OPV2V数据集上的表现进行了综合对比。首先，所有的协同感知算法都在不同场景的不同指标上显著优于单车感知(No Fusion)，这表明多车之间的相互协作有利于彼此更好地感知周围环境。其次，可以发现前期融合(Early Fusion)策略在大多数情况下均表现优于后期融合(Late Fusion)。这表明在不考虑通信带宽的情况下，数据在初期的融合能够更好地保留原始场景信息。值得注意的是，本文算法采用不同融合策略的3种实现版本(即元素级求和融合(Ours-S)、通道级拼接融合(Ours-C)和Transformer融合(Ours-T))的性能在所有对比算法中均具有较强竞争力，尤其是使用了Transformer架构的多模态融合模型Ours-T，在两种IoU阈值下均取得更高的AP分数。尽管在Default子集中，Ours-T在AP@0.7指标上略低于CoBEVT[15]，但在更具挑战性的Culver City子集中，其在AP@0.5和AP@0.7两个指标上均领先于其他所有对比算法。

表 1　与SOTA算法的综合性能对比(%)

Tab. 1　Comprehensive performance comparison with SOTA algorithms (%)

算法	Default		Culver city	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7
No Fusion	67.9	60.2	55.7	47.1
Early Fusion	89.1	80.0	82.9	69.6
Late Fusion	85.8	78.1	79.9	66.8
V2VNet[7]	89.7	82.2	86.8	73.3
Cooper[5]	89.1	80.0	82.9	69.6
F-Cooper[6]	88.7	79.1	84.5	72.9
AttFuse[8]	89.9	81.1	85.4	73.6
CoBEVT[15]	**91.4**	**86.2**	85.9	77.3
Ours-S	89.5	82.6	86.7	76.4
Ours-C	91.1	85.0	87.0	78.1
Ours-T	**91.4**	85.2	**88.6**	**78.8**

表2展示了本文算法在异构模态场景中的性能比较。其中，Camera-only表示所有车辆只能获取图像数据。而LiDAR-only表示所有车辆只能访问点云数据。此外，还考虑了Camera-only与LiDAR-only混合的情况：Hybrid-C表示所有车辆中一半只能获取图像数据(包括中心车辆)，另一半只能获取点云数据；与之不同的是，Hybrid-C表示一半数量的车辆只能获取点云数据。从表2可以看出，纯图像模式(Camera-only)下的性能明显较低，特别是在Culver City场景中的AP@0.7，只有8.6%。这可能是由于单一的图像模态无法很好地提供3D检测所需要的信息，因为根据2D图像预测在3D空间中的深度信息本身就存在一定的不确定性。而纯点云模式(LiDAR-only)的性能则明显优于Camera-only，主要是因为3D点云可以提供准确的场景深度测量，因此更适用于3D目标检测任务。进一步考虑到异构模态的混合情况，Hybrid-C和Hybrid-L都表现出了相对较好的性能。特别是在中心车辆只能获取点云数据的情况下(Hybrid-L)，其性能接近LiDAR-only，说明中心车辆的数据模态在V2V协同检测中起到了关键的作用。而对于Hybrid-C，尽管其性能略低于Hybrid-L，但仍然明显优于Camera-only，这表明即使在一半车辆只能获取图像数据的情况下，点云数据的存在仍然能大大增强系统的检测性能。表2的实验结果验证了所提算法在不同的异构模态场景下都具有相对稳健的性能，特别是在模态数据丢失的情况下。此外，这也再次突显了点云数据在V2V协同检测任务中的重要性，以及中心车辆对整体检测性能的影响。

定位误差是现实场景中需要考虑的复杂因素之一。为了分析模型对定位误差的鲁棒性，本文从高斯分布中分别采样坐标噪声($\sigma_{xyz} \in [0, 1.0]$ m)和角度噪声($\sigma_{\text{heading}} \in [0°, 1.0°]$)，并将其添加到准确的定位数据上来模拟定位误差。图3给出了所提多模态融合算法(Ours-T)与其他两种单模态方案(LiDAR-only, Camera-only)在不同定位误差情况下的性能对比。可以看出，随着定位噪声的增加，3种模型的性能整体均趋于下降，尤其是纯点云模式

表 2　所提算法不同异构模态场景下的性能对比(%)

Tab. 2　Performance comparison of the proposed algorithm under different heterogeneous modal scenarios (%)

算法	Default		Culver city	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7
Camera-only	43.9	28.1	19.0	8.6
LiDAR-only	**90.9**	**82.9**	**85.9**	**75.4**
Hybrid-C	70.7	58.1	58.9	44.5
Hybrid-L	87.8	78.6	76.6	63.6

(LiDAR-only)对定位误差最敏感：当误差增加到0.2时，性能下降超过4%；误差为0.4时，下降超过10%。相比之下，纯图像模式(Camera-only)受定位误差较小，性能下降平缓。其主要原因是LiDAR数据提取的三维几何特征对坐标轴的偏移比图像提取的视觉特征更加敏感。此外，由于多模态融合模型Ours-T中包含一定的冗余信息使其不会过于依赖某种单一模态，因此其在保持较高性能的同时对于定位误差具有较强的容忍能力。

4.3 定性实验

图4展示了3种不同方案在OPV2V数据集上的

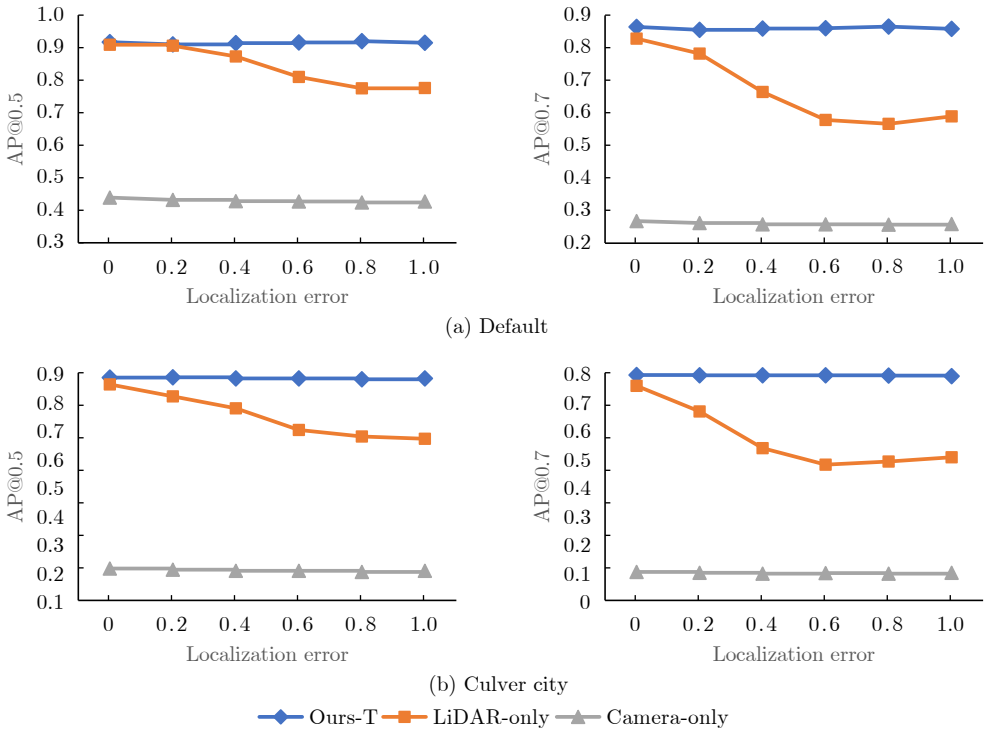
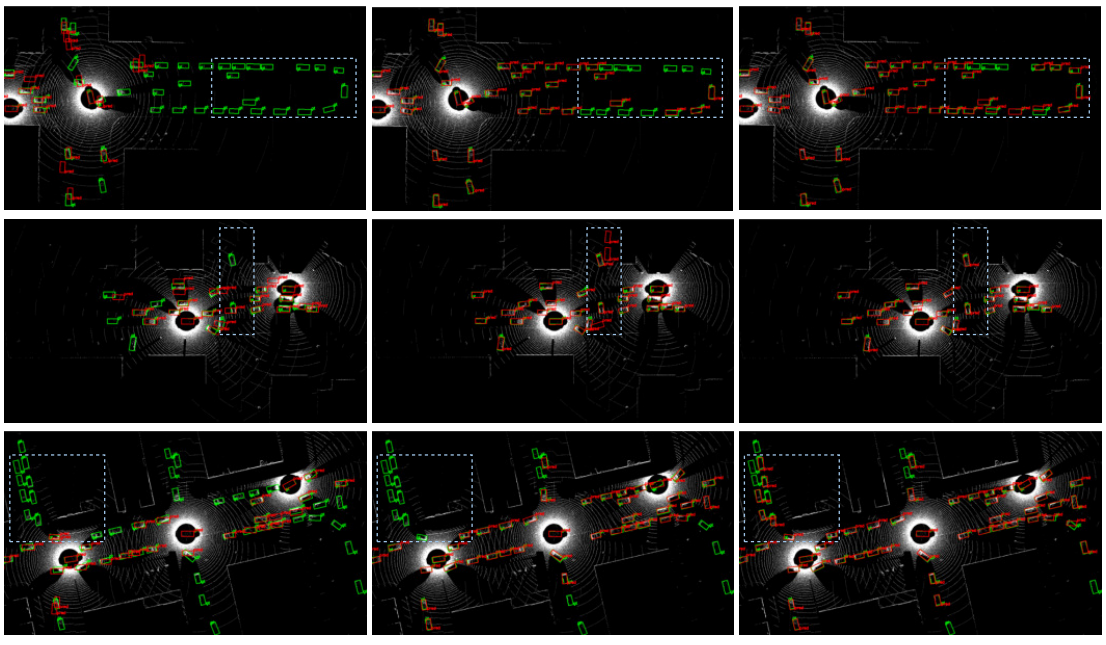


图 3 定位误差对模型性能的影响

Fig. 3 Impact of positioning error on model performance



(a) Camera-only (b) LiDAR-only (c) Ours-T

图 4 不同模型检测结果可视化对比

Fig. 4 Visualization comparison of detection results from different models

可视化检测结果。如图4所示，纯相机和纯激光雷达模型会在不同程度上受到语义模糊和不确定性等因素的影响，从而导致漏检和误检等问题。相比之下，本文提出的多传感器融合模型在检测精度和鲁棒性方面有了显著改进。通过利用激光雷达和相机模式的互补优势，融合模型有效地解决了单个传感器的局限性，实现了更精确、更可靠的物体检测。

5　结语

本文提出并实现了一种集成激光雷达和相机数据的多模态融合协同感知基线系统，该系统突破了单一传感器模式的限制，有效地融合了两种传感器的优势，为实现更精确、更稳健的环境感知奠定了基础。特别地，本研究首次深入探讨了几种先进的融合策略在多模态协同感知任务中的适用性和有效性，尤其是引入了基于注意力机制的融合方法，这一创新策略在提高感知系统准确性和鲁棒性方面表现出色。这种方法通过多头自注意力机制，在不同传感器特征之间建立了复杂的关联，提高了融合的精细度，并在OPV2V数据集上展现出卓越的性能和鲁棒性。此外，在更具挑战性的异构模态场景下的实验再次验证了多模态融合的有效性，这也为现实世界中实际应用提供了一种经济可行的方案。未来的工作将集中在进一步优化融合策略，并探索V2V协同感知中更多传感器模态的应用潜力，以进一步推动自动驾驶和智能交通系统的发展。

利益冲突 所有作者均声明不存在利益冲突

Conflict of Interests The authors declare that there is no conflict of interests

参 考 文 献

[1] LIU Si, GAO Chen, CHEN Yuan, *et al*. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception[EB/OL]. https://arxiv:abs/2308.16714, 2023.

[2] HAN Yushan, ZHANG Hui, LI Huifang, *et al*. Collaborative perception in autonomous driving: Methods, datasets, and challenges[J]. *IEEE Intelligent Transportation Systems Magazine*, 2023, 15(6): 131–151. doi: 10.1109/MITS.2023.3298534.

[3] REN Shunli, CHEN Siheng, and ZHANG Wenjun. Collaborative perception for autonomous driving: Current status and future trend[C]. 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control, Singapore, Singapore, 2023: 682–692. doi: 10.1007/978-981-19-3998-3_65.

[4] 上官伟, 李鑫, 柴琳果, 等. 车路协同环境下混合交通群体智能仿真与测试研究综述[J]. 交通运输工程学报, 2022, 22(3): 19–40. doi: 10.19818/j.cnki.1671-1637.2022.03.002.
SHANGGUAN Wei, LI Xin, CHAI Linguo, *et al*. Research review on simulation and test of mixed traffic swarm in vehicle-infrastructure cooperative environment[J]. *Journal of Traffic and Transportation Engineering*, 2022, 22(3): 19–40. doi: 10.19818/j.cnki.1671-1637.2022.03.002.

[5] CHEN Qi, TANG Sihai, YANG Qing, *et al*. Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds[C]. 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, USA, 2019: 514–524. doi: 10.1109/ICDCS.2019.00058.

[6] CHEN Qi, MA Xu, TANG Sihai, *et al*. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds[C]. 4th ACM/IEEE Symposium on Edge Computing, Arlington, USA, 2019: 88–100. doi: 10.1145/3318216.3363300.

[7] WANG T H, MANIVASAGAM S, LIANG Ming, *et al*. V2VNet: Vehicle-to-vehicle communication for joint perception and prediction[C]. 16th European Conference on Computer Vision, Glasgow, UK, 2020: 605–621. doi: 10.1007/978-3-030-58536-5_36.

[8] XU Runsheng, XIANG Hao, XIA Xin, *et al*. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication[C]. 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, USA, 2022: 2583–2589. doi: 10.1109/ICRA46639.2022.9812038.

[9] XU Runsheng, XIANG Hao, TU Zhengzhong, *et al*. V2x-ViT: Vehicle-to-everything cooperative perception with vision transformer[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 107–124. doi: 10.1007/978-3-031-19842-7_7.

[10] LI Yiming, REN Shunli, WU Pengxiang, *et al*. Learning distilled collaboration graph for multi-agent perception[C]. 34th International Conference on Neural Information Processing Systems, Virtual Online, 2021: 29541–29552.

[11] LI Yiming, ZHANG Juexiao, MA Dekun, *et al*. Multi-robot scene completion: Towards task-agnostic collaborative perception[C]. 6th Conference on Robot Learning, Auckland, New Zealand, 2023: 2062–2072.

[12] QIAO Donghao and ZULKERNINE F. Adaptive feature fusion for cooperative perception using LiDAR point clouds[C]. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 1186–1195. doi: 10.1109/WACV56688.2023.00124.

[13] ZHANG Zijian, WANG Shuai, HONG Yuncong, *et al*. Distributed dynamic map fusion via federated learning for intelligent networked vehicles[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021: 953–959. doi: 10.1109/ICRA48506.2021.9561612.

[14] WANG Binglu, ZHANG Lei, WANG Zhaozhong, *et al*. CORE: Cooperative reconstruction for multi-agent perception[C]. IEEE/CVF International Conference on

Computer Vision, Paris, France, 2023: 8710–8720.

[15] XU Runsheng, TU Zhengzhong, XIANG Hao, *et al*. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers[C]. 6th Conference on Robot Learning, Auckland, New Zealand, 2022: 989–1000.

[16] HU Yue, LU Yifan, XU Runsheng, *et al*. Collaboration helps camera overtake LiDAR in 3D detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 9243–9252. doi: 10.1109/CVPR52729.2023.00892.

[17] 党相卫, 秦斐, 卜祥玺, 等. 一种面向智能驾驶的毫米波雷达与激光雷达融合的鲁棒感知算法[J]. 雷达学报, 2021, 10(4): 622–631. doi: 10.12000/JR21036.
DANG Xiangwei, QIN Fei, BU Xiangxi, *et al*. A robust perception algorithm based on a radar and LiDAR for intelligent driving[J]. *Journal of Radars*, 2021, 10(4): 622–631. doi: 10.12000/JR21036.

[18] CHEN Xiaozhi, MA Huimin, WAN Ji, *et al*. Multi-view 3D object detection network for autonomous driving[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6526–6534. doi: 10.1109/CVPR.2017.691.

[19] 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, USA, 2020: 4603–4611. doi: 10.1109/CVPR42600.2020.00466.

[20] LIANG Tingting, XIE Hongwei, YU Kaicheng, *et al*. BEVFusion: A simple and robust LiDAR-camera fusion framework[C]. 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 10421–10434.

[21] LIU Zhijian, TANG Haotian, AMINI A, *et al*. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 2023: 2774–2781. doi: 10.1109/ICRA48891.2023.10160968.

[22] JIAO Yang, JIE Zequn, CHEN Shaoxiang, *et al*. MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 21643–21652. doi: 10.1109/CVPR52729.2023.02073.

[23] PRAKASH A, CHITTA K, and GEIGER A. Multi-modal fusion transformer for end-to-end autonomous driving[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 7073–7083. doi: 10.1109/CVPR46437.2021.00700.

[24] XIANG Hao, XU Runsheng, and MA Jiaqi. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer[EB/OL]. https://arxiv: abs/2304.10628, 2023.

[25] READING C, HARAKEH A, CHAE J, *et al*. Categorical depth distribution network for monocular 3D object detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 8551–8560. doi: 10.1109/CVPR46437.2021.00845.

[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, USA, 2019: 12689–12697. doi: 10.1109/CVPR.2019.01298.

[27] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[28] 郭帅, 陈婷, 王鹏辉, 等. 基于角度引导Transformer融合网络的多站协同目标识别方法[J]. 雷达学报, 2023, 12(3): 516–528. doi: 10.12000/JR23014.
GUO Shuai, CHEN Ting, WANG Penghui, *et al*. Multistation cooperative radar target recognition based on an angle-guided transformer fusion network[J]. *Journal of Radars*, 2023, 12(3): 516–528. doi: 10.12000/JR23014.

[29] XU Runsheng, GUO Yi, HAN Xu, *et al*. OpenCDA: An open cooperative driving automation framework integrated with co-simulation[C]. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, USA, 2021: 1155–1162. doi: 10.1109/ITSC48978.2021.9564825.

[30] DOSOVITSKIY A, ROS G, CODEVILLA F, *et al*. CARLA: An open urban driving simulator[C]. 1st Annual Conference on robot learning, Mountain View, USA, 2017: 1–16.

作 者 简 介

王秉路，博士，副教授，主要研究方向为多模态信息融合。

靳　杨，硕士生，主要研究方向为计算机视觉和深度学习。

张　磊，博士生，主要研究方向为计算机视觉和深度学习。

郑　乐，博士，教授，主要研究方向为雷达目标跟踪和雷达成像。

周天飞，博士，教授，主要研究方向为图像处理、深度学习和机器学习。

(责任编辑：于青)