智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于特征流的点云目标检测方法

陆军, 邹康成, 李杨

引用本文:
陆军, 邹康成, 李杨. 基于特征流的点云目标检测方法[J]. 智能系统学报, 2026, 21(1): 146-155.
LU Jun, ZOU Kangcheng, LI Yang. Feature flow-based point cloud object detection method[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 146-155.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202503005

DOI: 10.11992/tis.202503005

基于特征流的点云目标检测方法

陆军, 邹康成, 李杨
(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: 针对现有激光雷达点云三维目标检测方法因点云稀疏性导致的场景信息缺失与目标漏检问题, 本文提出一种基于特征流的单阶段三维目标检测算法, 该算法通过多帧时空特征融合与动态对齐机制优化检测性能。首先, 构建门控网络驱动的多帧融合框架, 利用可变形注意力机制协同时空特征提取模块, 实现跨帧特征的动态对齐, 抑制未对齐特征融合导致的误检; 其次, 设计时空特征引导的可变形注意力机制, 通过目标运动信息预测特征偏移与权重, 提升稀疏点云的特征匹配精度; 最后, 设计层级式特征流提取模块, 结合多尺度特征提取与渐进融合策略, 增强场景表征能力。实验结果表明, 所提算法在 NuScenes 验证集上的平均精度均值达到 63.73%, 较体素基准方法提升 4.51%, 其中摩托车、自行车等小目标检测精度提升超过 14%。消融实验结果表明, 多帧互补机制使远距离目标 (>50 m) 召回率提升 16.2%, 遮挡场景漏检率降低 11.8%。本研究为自动驾驶领域稀疏点云三维检测提供了有效方案。

关键词: 激光雷达点云; 目标检测; 特征流; 特征对齐; 时序特征融合; 可变形注意力机制; 鸟瞰视角表示; 多帧点云融合

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2026)01−0146−10

中文引用格式: 陆军, 邹康成, 李杨. 基于特征流的点云目标检测方法 [J]. 智能系统学报, 2026, 21(1): 146–155.

英文引用格式: LU Jun, ZOU Kangcheng, LI Yang. Feature flow-based point cloud object detection method[J]. CAAI transactions on intelligent systems, 2026, 21(1): 146–155.

Feature flow-based point cloud object detection method

LU Jun, ZOU Kangcheng, LI Yang
(College of Intelligent Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: Aiming at the problem of missing scene information and missing target detection caused by the sparsity of point cloud in the existing 3D target detection method of lidar point cloud, this paper proposes a single-stage 3D target detection algorithm based on feature flow, and the algorithm optimizes the detection performance through multi-frame spatio-temporal feature fusion and dynamic alignment mechanism. Firstly, a multi-frame fusion framework driven by gated network is constructed. The deformable attention mechanism is used to cooperate with the spatio-temporal feature extraction module to realize the dynamic alignment of cross-frame features and suppress the false detection caused by unaligned feature fusion. Secondly, a deformable attention mechanism guided by spatio-temporal features is designed to predict feature offset and weight through target motion information, so as to improve the feature matching accuracy of sparse point clouds. Finally, a hierarchical feature flow extraction module is designed to enhance the scene representation ability by combining multi-scale feature extraction and progressive fusion strategy. Experiments show that the proposed algorithm achieves 63.73% mAP on the NuScenes verification set, which is 4.51% higher than the voxel benchmark method, and the detection accuracy of small targets such as motorcycles and bicycles is improved by more than 14%. Ablation experiments show that the multi-frame complementary mechanism increases the recall rate of long-distance targets (>50 m) by 16.2%, and reduces the missed detection rate of occlusion scenes by 11.8%. This study provides an effective solution for three-dimensional detection of sparse point clouds for autonomous driving.

Keywords: lidar point cloud; object detection; feature flow; feature alignment; temporal feature fusion; deformable attention mechanism; bird's-eye view; multi-frame point cloud fusion

深度学习[1]技术的快速发展推动了计算机视觉[2]、语音识别[3]、自然语言处理[4]等领域的突破性进展。以目标检测与语音识别为例, 基于深度学习的方法在准确率与跨场景泛化能力上已超越传统人工系统方法。随着算法优化与工程化落地, 该技术已规模化应用于自动驾驶及语音交互领域。其中, 自动驾驶技术[5-7]的发展依赖于环境感知[8-9]模块, 其通过多传感器 (视觉摄像头、激光雷达、毫米波雷达) 融合, 实时提取道路结构信息 (车道线、交通标志) 与动态障碍物状态 (车辆、行人), 为决策控制提供输入。激光雷达点云[10]数据包含了丰富的空间位置信息, 尤其是光学传感器图像数据中缺失的深度信息, 这使得点云数据在目标检测任务中能够实现更精确的定位。然而, 点云数据本身的稀疏性与无序性也为处理带来了一定的挑战。目前, 基于激光雷达点云的三维目标检测在自动驾驶领域的应用仍面临严峻挑

收稿日期: 2025−03−04.
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.

战。核心问题在于点云的稀疏性, 这导致特征提取时信息不足, 容易造成漏检并降低检测精度。因此, 针对因点云稀疏性导致场景信息不足、目标漏检率高的问题, 本文提出一种基于特征流的单阶段三维目标检测算法来提升点云的三维目标检测精度。

当前, 基于点云的三维目标检测算法[11] 根据处理方式的不同主要可分为 3 类: 基于投影的点云目标检测[12]、基于体素的点云目标检测[13], 及直接基于点处理的点云目标检测[14]。在基于投影的点云目标检测方法中, 三维点云被投影到各种二维平面上, 转化为卷积神经网络可处理的二维图像。例如, 3DFCN(fully convolutional network)[15] 将点云投影到正前方的二维平面, 生成点云的前视图, 并保留每个点的深度信息, 随后在前视图上应用全卷积网络[16], 利用卷积特征图预三维检测框。PIXOR(real-time 3D object detection from point clouds)[17] 网络则将点云投影至地面, 转化为鸟瞰图[18] 进行处理, 从而更高效地利用三维空间数据。然而, 投影过程会破坏空间相关性, 并丢失点云数据中点与点之间的结构信息, 这种信息丢失会降低模型在三维边界框回归时的性能, 尤其是在检测行人等小目标时, 影响更为显著。基于体素的点云目标检测方法则在一定程度上保留了点云的空间结构, 使得卷积神经网络能够处理转换后的点云。VoxelNet[19]、PointPillars[20] 和 PV-RCNN(point-voxel region-based convolutional neural network)[21] 是这一方法的代表性算法。然而, 由于体素分辨率的限制, 体素化过程中会丢失部分点云信息, 分辨率越低, 信息丢失越多, 从而导致检测精度下降。PointNet[22] 网络的出现有效解决了点云无序性的问题, 而 PointNet++[23-24] 在 PointNet 的基础上进一步改进, 兼顾了全局特征与局部特征的提取。PointRCNN[25] 网络是首个基于点的端到端点云目标检测算法, 其骨干特征提取网络采用了 PointNet++。针对单帧点云数据因遮挡及远距离等因素导致的场景信息缺失问题, 本文在 CenterPoint[26-30] 网络的基础上, 设计了一种基于特征流的三维点云目标检测算法, 详细阐述了改进后的点云目标检测网络结构, 即特征流提取模块, 其包括时空特征提取、特征对齐、门控特征融合等模块, 最后实现了无锚框中心检测头[31] 的回归参数及损失函数。

1　点云目标检测方法概述

本文将对基于体素形式点云的三维点云目标检测算法进行研究, 并以 CenterPoint 为基础网络结构对其进行改进。CenterPoint 网络是二维目标检测网络 CenterNet 在点云领域的三维实现, 是首个提出中心检测头并将三维点云目标表示为点来进行检测与跟踪的网络, 与其他基于体素化点云的目标检测方法相比, 其检测精度显著提高。由于激光雷达在数据采集过程中面临遮挡、反射以及远距离等问题, 从而导致场景中部分区域点云的分布极其稀疏。这使得一些检测目标包含的点数过少, 难以提供足够的特征信息, 从而导致这些目标无法被正确检测, 这是漏检的主要成因。本文将针对这一问题对 CenterPoint 网络进行改进, 提出了基于特征流的点云目标检测网络。该网络通过多帧点云信息互补完善场景信息, 使用带速度头的无锚框中心检测头实现三维目标位置与速度的检测, 在原算法的基础上提高了三维目标检测的精度, 网络整体结构如图 1 所示。

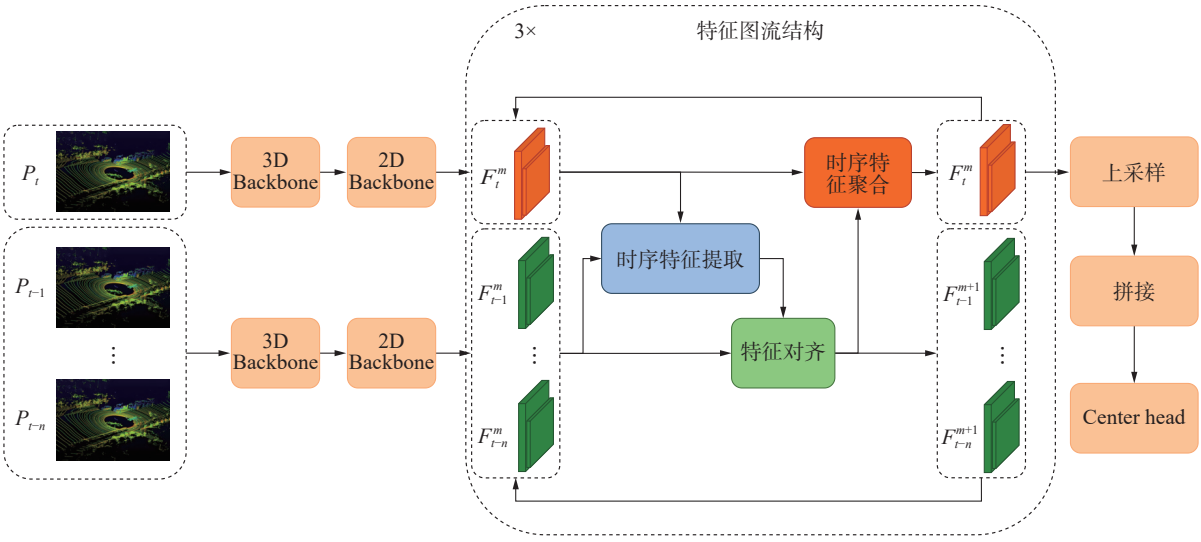


图 1　基于特征流的点云目标检测网络结构

Fig. 1　Point cloud target detection network structure based on feature flow

在特征流处理模块中，本文通过融合多时刻点云的鸟瞰视图 (Bird's eye view, BEV)[20] 特征图来丰富检测场景信息并解决点云稀疏性引起的特征缺失问题。由于检测目标并非一直保持静止状态，直接融合多帧点云特征可能因为位置偏差导致误检，因此本文在特征融合之前先进行特征对齐操作。本文特征流提取模块为 3 个阶段：在初始阶段，该模块提取不同帧点云之间的时空特征，以捕捉目标在时间序列中的动态变化信息；在第 2 阶段中，特征流提取模块基于时空特征引导的多尺度可变形注意力机制，将历史时刻的点云特征与当前时刻的点云特征对齐；第 3 阶段中，基于门控融合网络完成多帧点云特征的稳定融合。此外，为进一步提升性能，本文通过叠加 3 次特征流提取模块来优化特征对齐和融合的效果。

2　点云特征提取

本文算法对点云采取拥有更高检测精度的体素格形式的划分方式。本文算法采取的体素特征编码方式为 meanVFE(voxel feature encoding)。与 RGB 图像中像素信息非常密集不同，体素化后点云中空体素占据了点云的大部分空间，使用普通三维卷积来提取点云特征需要消耗大量的计算资源与时间。本文算法使用 SECOND(sparsely embedded convolutional detection) 网络的体素特征提取模块作为点云特征提取网络，该网络通过配合使用子流形稀疏卷积 (SubMConv3d) 和普通空间卷积 (SparseConv-3d) 大大地提高了点云特征提取的效率。

点云经过体素特征提取后，特征保持三维结构 (C, X, Y, Z)。为适配二维检测头，需将 Z 轴特征通道与原始通道合并，将张量重塑为二维形式 ($C\times Z$, X, Y)，以保留高度方向的信息。其中 C 表示经过特征提取后体素的特征维度，X、Y、Z 则代表特征提取后三维点云特征的大小。此外，压缩处理可增强点云特征的高度方向感受野并且加快网络的训练与推理速度。本文点云特征体取网络结构如图 2 所示。

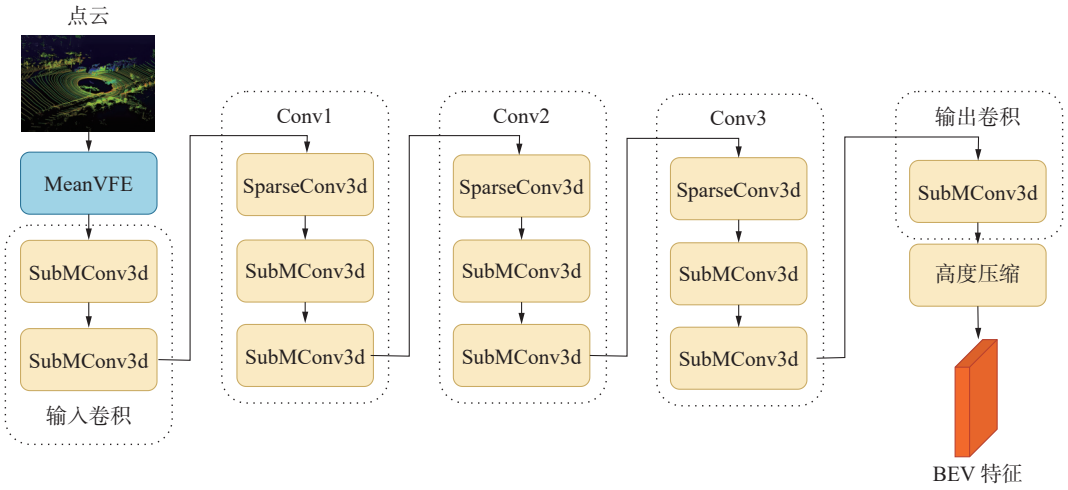


图 2　点云特征提取网络

Fig. 2　Point cloud feature extraction network

3　特征流提取模块

与独立的目标检测任务相比，实际自动驾驶场景在空间与时间上往往是连续的，不同时刻的点云信息可以通过相互补充来接近真实场景。因此，融合邻近时间帧的点云信息可增强上下文表征，这有助于更精确地识别和定位检测目标，同时还能在一定程度上弥补因点云稀疏性而导致的漏检问题。本文增加的特征流提取模块是一种基于多帧时序点云数据的特征处理框架，通过融合历史帧与当前帧的时空特征，解决单帧点云稀疏性导致的场景信息缺失问题。其核心目标是通过动态对齐与加权融合，增强三维目标检测的特征表达能力。特征流模块包含时空特征提取、特征对齐和门控特征融合 3 个子模块。

3.1　时空特征提取

在点云的采集过程中，由于激光雷达与场景中物体的运动，点云数据在时间序列中呈动态变化，将目标当前帧所处的空间位置与历史帧所处空间位置进行比对，检测是否存在差异。在这种情况下，直接将历史帧的特征图与当前帧特征图进行融合可能由于目标位置偏差进而导致特征叠加不准确，引发拖尾现象。拖尾现象发生时，历史帧目标所在位置会在融合后特征图上产生高响应，导致检测器误认为目标仍处于之前时刻的位置，从而产生误检。因此，有必要在特征融合之

前对帧间 BEV 特征进行对齐。在特征流提取模块中, 本文从当前帧与历史帧点云 BEV 特征之间提取时空特征 $\boldsymbol{M}_t$ 用于引导后续的特征对齐操作, 时空特征提取网络如图 3 所示。

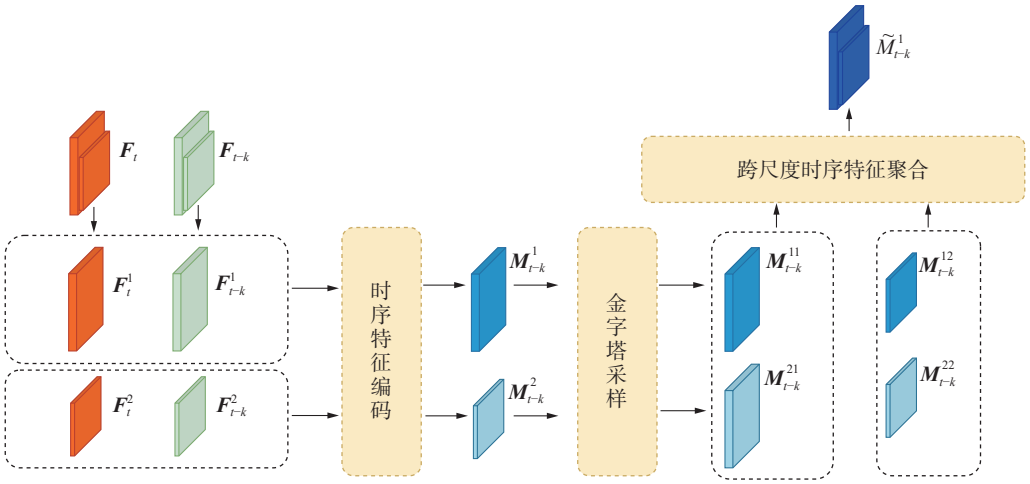


图 3　时空特征提取网络

Fig. 3　Spatial-temporal feature extraction network

点云时空特征 $\boldsymbol{M}$ 的提取过程共分为两步, 首先单独提取不同尺度的时空信息。假设当前帧与历史帧对应的多尺度特征图集合分别为 $\boldsymbol{F}_t$ 和 $\boldsymbol{F}_{t-k}$, 特征提取方式为

$$\boldsymbol{M}^s_{t-k} = \mathrm{Conv}_{3\times3}(\boldsymbol{F}^s_t - \boldsymbol{F}^s_{t-k}) \tag{1}$$

在单个尺度的时空特征提取过程中, 假设该尺度大小为 S, 通过将当前帧中对应尺度的特征图 $\boldsymbol{F}^s_t$ 减去历史帧对应特征的特征图 $\boldsymbol{F}^s_{t-k}$ 得到特征差 $(\boldsymbol{F}^s_t - \boldsymbol{F}^s_{t-k})$, 之后对特征差进行卷积操作来得到对应尺度的时空特征 $\boldsymbol{M}^s_{t-k}$。对每个尺度的特征图进行同样的处理后, 可以得到未融合的多尺寸时空特征集合 $\boldsymbol{M}_{t-k}$。

在点云时空特征提取的第 2 步中, 需要对第 1 步中获得的不同尺度的时空特征进行交叉融合, 具体融合方式表示为

$$\tilde{\boldsymbol{M}}^{s'}_{t-k} = \mathrm{Conv}_{1\times1}\left(\left[\boldsymbol{M}^{1s'}_{t-k} \quad \boldsymbol{M}^{2s'}_{t-k} \quad \cdots \quad \boldsymbol{M}^{Ss'}_{t-k}\right]\right) \tag{2}$$

为了融合不同尺度的时空特征, 首先对时空特征进行金字塔采样, 为每个尺度的时空特征生成相同形式的特征集合。

3.2　特征对齐

在特征对齐模块中, 本文提出一种时空特征引导的可变形注意力机制 (deformable attention)。该机制通过动态预测采样点偏移量, 降低计算复杂度并支持多尺度特征对齐, 从而提升模型收敛速度与资源效率。本文算法在原可变性注意力机制基础上根据使用场景做出调整, 使用包含物体运动信息的时空特征来预测采样位移与注意力权重。

在原本的可变形注意力机制中, 采样位移、注意力权重甚至 Value 值均由同一输入特征计算得到, 考虑到提取的时空特征中包含了目标在过去一段时间内的运动信息及具体应用场景, 使用时空特征来预测采样位移与注意力权重对于特征对齐操作更为合理, 具体计算过程为

$$\Delta^{sh}_{t-k}(p_s) = \boldsymbol{W}'_h \tilde{\boldsymbol{M}}^s_{t-k}(p_s) \tag{3}$$

$$A^{sh}_{t-k}(p_s) = \mathrm{Soft\,max}\left(\boldsymbol{W}''_h \tilde{\boldsymbol{M}}^s_{t-k}(p_s)\right) \tag{4}$$

式中: p_s 为在尺度 s 上的特征图上的二维坐标。p_s 处的时空特征 $\tilde{\boldsymbol{M}}^s_{t-k}(p_s)$ 在经过线性层 $\boldsymbol{W}_h$ 处理后得到注意力范围 $\Delta^{sh}_{t-k}(p_s)$。在本文算法中, 每个注意力头在每个尺度的特征图上都需要采样 4 个点, 而 $\Delta^{sh}_{t-k}(p_s)$ 中包含了多组二维坐标偏移量, 将坐标偏移量与 p_s 相加后即可得到参与注意力计算的采样点坐标位置。同理, 时空特征 $\tilde{\boldsymbol{M}}^s_{t-k}(p_s)$ 经线性层 W_h 处理后得到注意力权重 $\boldsymbol{A}^{sh}_{t-k}(p_s)$, $\boldsymbol{A}^{sh}_{t-k}(p_s)$ 中包含了多个采样点对应的注意力权重值。

Value 是历史帧 BEV 特征图中的语义特征表示, Value 值通过历史帧特征获取, 获取过程表示为

$$V^{sh}_{t-k}(p_s) = \boldsymbol{W}_h \boldsymbol{F}^s_{t-k}(p_s) \tag{5}$$

其中 p_s 处的 Value 值 $V^{sh}_{t-k}(p_s)$ 由历史帧 p_s 处特征 $\boldsymbol{F}^s_{t-k}(p_s)$ 经线性层 $\boldsymbol{W}_h$ 后得到。因注意力范围 $\Delta^{sh}_{t-k}(p_s)$ 预测的位置偏移量可能为非整数, 无法直接从采样点坐标获取对应的 Value 值。为此, 本文采用双线性差值法, 利用距离采样点最近的 4 个特征图坐标点的 Value 值计算采样点的 Value 值。图 4 给出了单一尺度的多头可变形注意力网络结构。在多尺度可变形注意力中, 某位置的最终注意力结果为不同尺度和注意力头的注意力结果之和。为确保多尺度注意力采样点坐标的统一性, 二维坐标采用归一化形式表示。多尺度注意力计算过程用公式表示为

$$z_{t-k}^{s}(p_s)=\sum_{h=1}^{H}\boldsymbol{W}_h'''\cdot\left(\sum_{i=1}^{L}\sum_{j=1}^{K}\boldsymbol{A}_{t-k}^{shij}(p_s)\cdot V_{t-k}^{ih}\left(\psi_i(p_s)+\Delta_{t-k}^{shij}(p_s)\right)\right) \tag{6}$$

式中：h 为注意力头序号，i 为特征图序号，j 为采样点序号。多尺度可变性注意力中每个头在 L 个特征图上均采样 K 个特征，本文算法从多尺度特征图中采取 2×4 个特征，此时注意力权重应该满足 $\sum_{i=1}^{L}\sum_{j=1}^{K}\boldsymbol{A}_{t-k}^{shij}=1$。

注意力计算结果经过处理后可得到 p_s 处的对齐后的历史帧特征 $\overline{\boldsymbol{F}}_{t-k}^{s}$，具体处理方式为

$$\overline{\boldsymbol{F}}_{t-k}^{s}(p_s)=\mathrm{FFN}\left(\mathrm{LN}\left(\mathrm{dropout}\left(z_{t-k}^{s}(p_s)\right)+\boldsymbol{F}_{t-k}^{s}(p_s)\right)\right) \tag{7}$$

式中：FFN 表示前馈神经网络，LN 表示层归一化，dropout 表示随机失活。

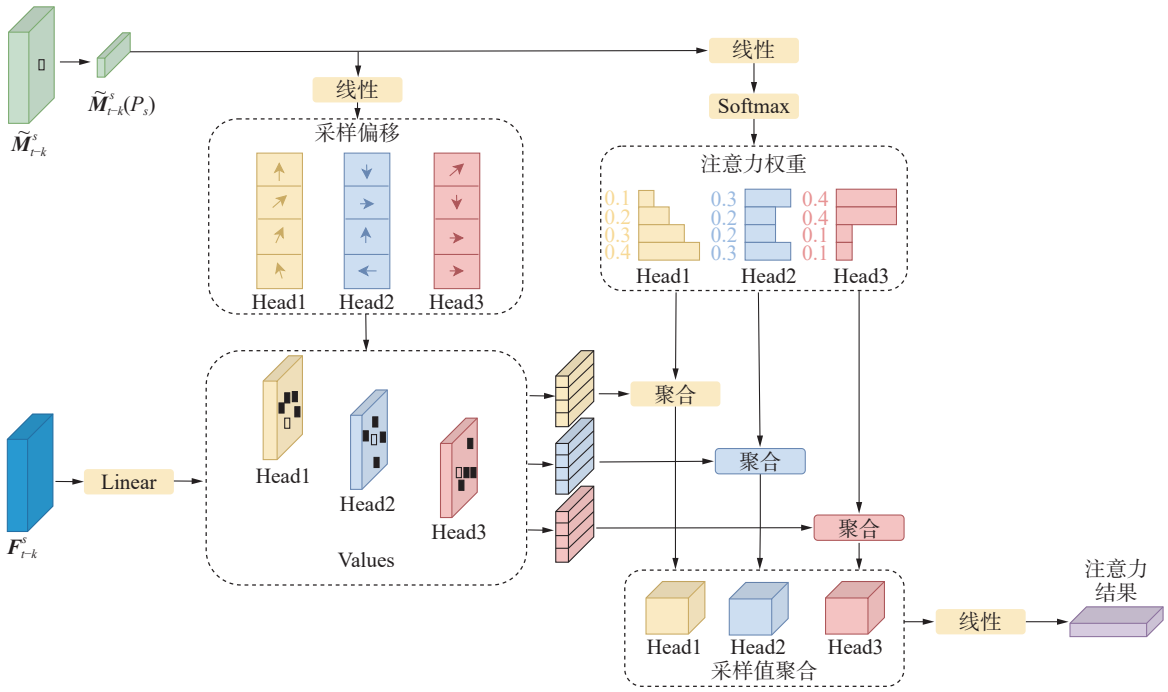


图 4　单一尺度可变形注意力网络结构

Fig. 4　Single-scale deformable attention network structure

3.3　特征融合

本文使用门控融合网络融合不同时序的特征。门控融合网络可以通过学习控制信息集成方式的门控机制来融合不同来源的信息，门控机制决定每个信息来源对最终输出贡献多少权重，通过网络动态地强调或弱化特定信息源，从而实现更好的表示学习并捕获输入之间的复杂依赖关系并提供稳定的融合效果。门控融合结构如图 5 所示。

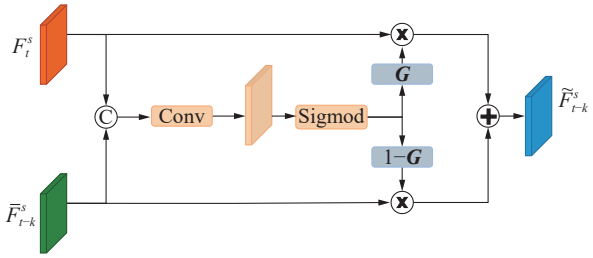


图 5　门控融合网络结构

Fig. 5　Gated fusion network structure

在门控融合网络中，首先将当前帧特征图与对齐后的历史帧特征图在通道维度进行拼接，然后通过一个卷积层将拼接后特征的特征维度降低，之后使用一个 Sigmoid 层计算模态权重 $\boldsymbol{G}$。权重张量 $\boldsymbol{G}$ 与当前帧特征在长、宽、通道数上相等，使用权重张量能够控制当前帧与历史帧特征在特征融合中的贡献程度，通过学习能够提供稳定的融合效果。具体融合过程用公式表示为

$$\boldsymbol{G}_{t-k}^{s}=\sigma\left(\mathrm{Conv}_{3\times3}\left(\left[\boldsymbol{F}_{t}^{s},\overline{\boldsymbol{F}}_{t-k}^{s}\right]\right)\right) \tag{8}$$

$$\tilde{\boldsymbol{F}}_{t-k}^{s}=\boldsymbol{G}_{t-k}^{s}\otimes\boldsymbol{F}_{t}^{(l),s}+\left(1-\boldsymbol{G}_{t-k}^{s}\right)\otimes\overline{\boldsymbol{F}}_{t-k}^{s} \tag{9}$$

式中：⊗表示特征加权操作，在当前帧与历史帧点云特征分别与对应权值逐元素相乘得到加权特征后，两个加权特征的和即为历史帧对应的融合特征。

在得到 $t-k$ 时刻的融合特征后，通过融合所有历史时刻的融合特征即可得到当前时刻的融合特征集合，具体融合方式为

$$\tilde{\boldsymbol{F}}_{t}^{s}=\mathrm{Conv}_{3\times3}\left(\left[\tilde{\boldsymbol{F}}_{t-1}^{s}\quad\tilde{\boldsymbol{F}}_{t-2}^{s}\quad\cdots\quad\tilde{\boldsymbol{F}}_{t-N+1}^{s}\right]\right) \tag{10}$$

4　检测目标回归

本文网络使用无锚框中心检测头对目标位置进行回归。与基于锚框的检测头相比，中心检测

头无需预设锚框, 只需要对不同尺度的特征图的目标中心点和宽高进行回归, 减少了耗时和算力。同时也可以避免一些由于锚框设置不合理导致的漏检或重复检测问题。

4.1　中心检测头

特征流提取模块输出的融合 BEV 特征图经共享卷积后, 输入至中心检测头的两个分支: 中心热力图分支和参数回归分支。中心热力图分支: 输出通道数与检测类别数一致, 每个通道表示对应类别的检测结果, 热力图值反映目标中心存在的概率。训练时, 标注框中心点热力值为 1, 周围采用二维高斯核扩展, 以增强前景信息。若某点位于多个高斯核范围内, 取最大值作为热力值。针对点云稀疏性, 调整高斯半径以确保热力图信息充足。参数回归分支: 回归三维检测框的中心点偏移量、高度、尺寸及朝向角等参数。由于 BEV 特征经过下采样和高度压缩, 中心坐标需通过偏移量精细回归。此外, 特征流提取模块融合多时序点云信息, 新增速度回归分支, 用于预测目标在俯视图下的速度。

4.2　损失函数

本文算法的回归损失为中心检测头各分支回归损失的加权和, 具体计算方式为

$$L_{\text{total}} = \boldsymbol{w}_{\text{cls}} L_{\text{cls}} + \boldsymbol{w}_{\text{reg}} \sum L_{\text{reg}} \tag{11}$$

式中: L_{cls}代表中心热力图分支损失, L_{reg}代表各参数分支损失, $\boldsymbol{w}_{\text{cls}}$、$\boldsymbol{w}_{\text{reg}}$为人为设置的各损失权重。上述各参数分支的回归损失均采用 L1 正则化损失函数计算。

5　目标检测算法试验

5.1　实验环境

本文算法单块英伟达图像处理卡 (GPU) 进行训练与验证, 显存大小为 48 GB。基于 PyTorch 与 OpenPCDet 深度学习框架实现, 相关软件信息如表 1 所示。

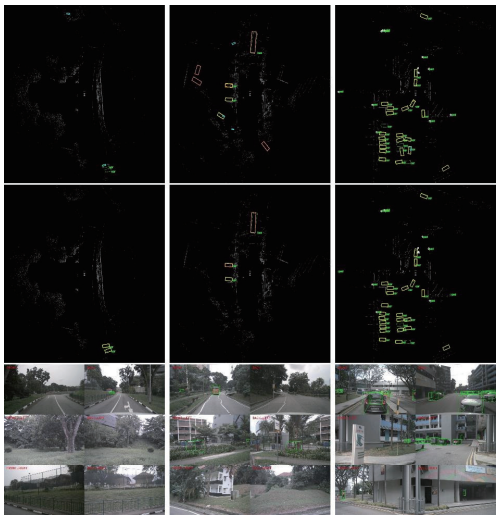
表 1　软件环境

Table 1　Software environment

软件名称	版本号
操作系统	Ubuntu 18.04
Cuda	11.1
cudatoolkit	11.1
Python	3.8
PyTorch	1.8.1
OpenPCDet	0.6

5.2　实验结果与分析

为了验证算法性能及可行性, 本文在 NuScenes[32] 验证集上分别对基准网络 CenterPoint 网络和本文检测网络进行了测试, 并对小汽车、行人及自行车等 10 类目标的检测结果进行了可视化, 三维目标检测对比如图 6 所示。其中第 1 行为 CenterPoint 网络的检测结果, 第 2 行为本文网络的检测结果, 最后一行为检测场景的 RGB 图像。可视化结果中带有类别标注的绿色框为真实标注框, 其中小汽车检测框为黄色, 卡车检测框为橙色, 行人检测框为白色。



(a) 场景 a　(b) 场景 b　(c) 场景 c

图 6　三维目标检测对比

Fig. 6　Comparison of three-dimensional target detection

根据可视化结果可对算法的检测性能进行评估。在场景 a 中, 配备激光雷达的采集车沿道路行进, 途中经过两辆小汽车的正前方。随着采集车与小汽车之间的距离不断拉大, 点云中数据越来越稀疏。这导致基准网络无法从稀疏的点云中提取充分特征, 进而准确检测目标。然而, 本文提出的算法通过利用检测帧之前的点云数据来丰富当前时刻的点云特征, 使得即使在目标特征稀疏的情况下也能保留足够的小汽车信息, 实现精确的目标回归。在场景 b 中, 位于点云场景边缘部分的检测目标, 距离传感器距离较远导致被捕获到的点云较少。在这种情况下, 基准网络容易将信息缺失的待检测目标与背景混淆产生误检。而本文算法通过结合多时刻的场景信息, 有效避免这种误检。在场景 c 中, 面对众多待检测目标, 目标之间的相互遮挡导致了部分目标信息存在一定的丢失, 导致基准网络和本文算法均存在误检和漏检问题。然而, 相比于基准网络, 本文提出的算法在减少误检与漏检方面表现得更为出色。更为关键的是, 对于那些成功检测到的目标, 本文算法凭借其更加丰富的特征信息, 能够实现更精准的边框回归, 确保了检测结果的高质量。

除基准网络外，本文还将本文算法与其他主流的点云三维目标检测网络在 NuScenes 验证集上的性能进行了比较。为了保证比较的公平性，主要比较对象为基于纯激光点云输入的三维目标检测方法，平均精度对比结果如表 2 所示，具体指标为平均检测精度 (mean average precision, mAP)、平均位置误差 (mean average translation error, mATE)、平均尺度误差 (mean average scale error, mASE)、平均朝向误差 (mean average orientation error, mAOE)、平均速度误差 (mean average velocity error, mAVE)、平均属性误差 (mean average attribute error, mAAE)、NuScenes 综合检测指标 (NuScenes detection score, NDS)。具体各类目标的检测精度如表 3 所示。

表 2　检测算法平均精度对比结果

Table 2　Average accuracy comparison results of the detection algorithm　%

算法名称	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	NDS↑
PointPillar	44.69	33.85	26.00	31.91	28.79	20.26	33.85
Second	50.56	31.20	25.52	**26.32**	26.22	20.38	62.32
CenterPoint-Pillar	50.03	31.13	26.04	42.92	23.90	19.14	60.70
CenterPoint-Voxel	59.22	28.80	**25.43**	37.27	21.55	**18.24**	66.48
VoxelNext	60.52	30.10	25.23	40.57	21.69	18.56	66.65
TranFusion-LiDAR	63.51	28.35	25.60	34.81	20.12	19.36	**70.12**
CenterPoint++	62.86	29.52	25.73	36.23	20.88	19.54	68.92
本文算法	**63.73**	**27.89**	25.65	33.40	**19.49**	19.67	69.26

注：加粗表示结果在该列最好。

表 3　检测算法各类精度对比结果 (AP↑)

Table 3　Comparison results of various accuracy of detection algorithms　%

算法名称	小汽车	卡车	公交车	拖车	施工车	行人	摩托车	自行车	交通锥	障碍物
PointPillar	81.3	49.9	63.4	35.3	12.1	72.4	29.4	6.0	47.0	49.8
Second	81.5	51.6	66.7	37.4	14.8	77.7	42.4	17.0	57.3	59.3
CenterPoint-Pillar	82.0	50.9	64.5	37.1	14.6	76.3	43.8	18.1	56.9	56.2
CenterPoint-Voxel	84.9	57.4	70.8	38.1	16.9	85.1	59.0	42.0	69.8	68.3
VoxelNext	83.9	55.5	70.5	38.1	21.1	84.6	62.8	50.0	69.4	**69.4**
TranFusion-LiDAR	85.2	**58.8**	70.5	40.1	20.3	83.7	67.4	52.3	70.1	65.8
CenterPoint++	84.0	55.2	69.8	39.5	19.8	82.5	65.1	48.7	68.9	63.2
本文算法	**85.9**	56.0	**73.1**	**42.3**	**23.4**	**85.5**	**73.1**	**58.9**	**74.6**	62.6

注：加粗表示结果在该列最好。

从表 3 可以看出，本文算法网络在多种类别目标上的检测精度均高于基准网络。特别是对于摩托车与自行车等在点云场景中体积较小，表征不充分的小目标，本文算法通过融合多帧点云数据以增强特征表达，从而显著提高了检测精度，其 AP 值较基准网络分别提高了 14.1 和 16.9 百分点，本文网络 mAP 较基准网络提高了 4.51 百分点。与其他采用体素激光雷达点云作为输入的主流目标检测算法相比，本文算法在 mAVE 和 mAOE 上显著优于基于激光点云输入的主流检测方法 (包括 CenterPoint-Voxel、VoxelNext、TransFusion-LiDAR 及 CenterPoint++)，mAVE 降低 3.0~6.7 百分点，mAOE 降低 4.0~8.2 百分点。这源于可变形注意力机制对运动目标的精准对齐能力，而 TransFusion-LiDAR[33] 依赖全局 Transformer 计算，CenterPoint++[34] 仅用线性插值融合时序特征，均未显式解决动态目标的位置偏移问题，因此本文算法在动态目标对齐、小目标增强、异形目标建模三方面均超越近些年高竞争力基线。尤其值得注意的是，纯点云输入条件下，本文以更低计算成本达到与多模态方法 (TransFusion) 相近的综合检测分数 (NDS)，凸显特征流设计的高效性。未来将进一步探索模块轻量化，适配实时自动驾驶系统。

5.3　消融实验

本文网络在基准网络的基础上加入了特征流提取模块来融合多帧点云特征，而特征流提取模块中包含时空特征提取模块，特征对齐模块及门控融合模块 3 个部分。为了验证各个模块在网络中发挥的作用及有效性，本文设计了消融实验。以原 CenterPoint 算法作为基准网络模型，分别比较在基准网络中叠加不同模块的目标检测结果，

实验结果如表 4 所示, 其中 1 表示在网络中加入该模块, 0 表示删除该模块。

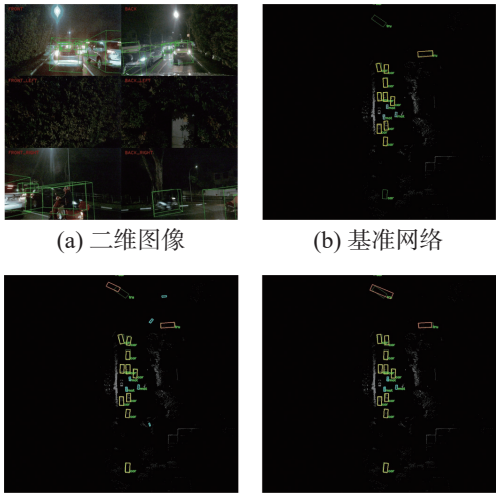
表 4　算法消融实验对比结果

Table 4　Algorithm ablation experiment comparison results　%

方法	特征对齐	时空特征提取	门控融合	NDS↑	mAP↑
A	0	0	0	66.30	58.76
B	1	0	0	67.95	61.85
C	0	1	0	68.20	61.12
D	0	0	1	67.33	60.34
E	1	1	0	68.60	61.85
F	1	0	1	68.54	62.49
G	0	1	1	68.22	61.24
H	1	1	1	**69.26**	**63.73**

基于消融实验结果, 本文提出的特征流模块中时空特征提取、特征对齐与门控融合三者的交互作用可归纳如下: 首先, 各模块独立应用时均能提升检测性能, 其中特征对齐模块贡献最大 (方法 B 的 mAP 较基线提升 3.09%), 时空特征提取 (方法 C 提升 2.36%) 与门控融合 (方法 D 提升 1.58%) 次之, 验证了模块的基础有效性。其次, 双模块协同效应呈现显著差异: 特征对齐与门控融合组合 (方法 F, mAP 62.49%) 提升幅度 (3.73%) 超越单模块理论叠加值 (4.67%), 体现二者在误差抑制与信息互补上的强协同性; 而时空特征提取与门控融合组合因缺乏对齐引导, 性能增益低于预期 (理论值 3.94% vs 实际值 2.48%), 表明时空特征的有效性高度依赖对齐模块的动态校正。最终, 三模块联合 (方法 H, mAP 63.73%) 虽存在误差累积导致实际增益 (4.97%) 低于理论叠加 (7.03%), 但其非线性协同机制仍显著优于最优双模块组合 (提升 1.24%), 形成“运动感知→精准对齐→自适应融合”的闭环优化链。实验表明, 特征对齐模块是协同效应的核心枢纽, 优化其精度可释放时空特征与门控融合的联合潜力, 为后续算法设计提供关键方向。

特征对齐及特征融合模块的消融实验可视化对比结果如图 7 所示。图 7(a) 为检测场景的 RGB 图像, 图 7(b) 为基线方法 CenterPoint 的检测结果。基线方法未能检测到场景边缘的前方卡车及后方小汽车, 同时误将右前方卡车检测为汽车。加入门控特征融合模块后, 如图 7(c) 所示, 前方卡车与后方小汽车被成功检测, 右前方卡车标签也正确, 但前方卡车的回归位置更接近历史帧位置, 且新增了误检结果。再次加入时空特征提取及特征对齐模块后, 如图 7(d) 所示, 误检现象明显改善, 前方卡车的回归位置也更加精确。表 5 给出了网络性能随特征流提取模块数量的变化情况, 进一步验证了特征流提取模块的可叠加性能。



(a) 二维图像　(b) 基准网络　(c) 基准网络+门控融合网络　(d) 本文算法

图 7　消融实验检测对比

Fig. 7　Ablation experiment detection comparison

表 5　特征流提取模块叠加对比结果

Table 5　Feature flow module superposition comparison results　%

特征流层数	mAP↑	NDS↑
1	61.53	67.89
2	62.91	68.76
3	**63.73**	**69.26**

由于内存限制, 本文将特征流层数加到最多三层。从表中可以看出, 网络的检测精度随着特征流提取模块层数的增加而提高。这证实了特征流提取模块能够有效的对齐多帧点云 BEV 特征并提供稳定的特征融合效果。

表 6 给出了随着输入点云帧数变化, 网络检测性能的相应变化情况。从表 6 可以看到, 随着输入点云帧数的增加, 网络的检测准确率先是不断提高, 但当输入点云帧数增至 4 帧时, 准确率反而开始降低。NuScenes 数据集中关键帧采样频率为 2 Hz, 4 帧点云输入在时间上的跨度已经达到 2 s, 目标在空间上的跨度也已经达到一个较大的范围。这一现象表明, 虽然特征流提取模块能通过对齐 BEV 特征来缓解特征位置偏差问题, 从而增强当前帧的点云特征, 但其处理特征位置变化的能力是有限的。如果某个目标在空间内移动距离较远, 特征流模块提取不能通过整合多帧点云特征有效丰富其信息。

表 6　网络输入点云帧数对比结果

Table 6　Network input point cloud frame number comparison　%

输入点云帧数	mAP↑	NDS↑
1	58.76	66.30
2	63.06	69.11
3	**63.73**	**69.26**
4	62.15	68.51

6　结束语

本研究聚焦于自动驾驶场景中的激光雷达点云目标检测方法，针对点云数据的稀疏性导致场景中部分目标区域点云数量不足，造成目标特征提取不充分，引发漏检，从而显著降低检测精度的问题，本文提出了一种基于特征流的点云目标检测算法，通过在基准网络中引入特征流处理模块，实现了多帧点云特征的融合，利用多帧点云信息的互补性完善了场景信息。为了确保特征融合的稳定性，算法采用门控融合网络动态调节不同帧点云的权重。此外，为避免因多帧点云特征位置误差导致的误检问题，在特征融合前利用可变形注意力机制实现多帧点云特征的位置对齐。同时，通过提取包含运动信息的点云帧间时空特征，进一步引导特征对齐的准确性，从而提升对齐精度。

实验结果表明，相较于基线方法 CenterPoint-Voxel，本文算法在 NuScenes 验证集上的平均检测精度 (mAP) 提升 4.51%。这一改进为自动驾驶环境感知中的目标检测任务提供了新的思路与方法。未来研究将进一步优化算法性能，并探索其在更复杂场景下的适用性。

参考文献：

[1] HERRMANN L, KOLLMANNSBERGER S. Deep learning in computational mechanics: a review[J]. Computational mechanics, 2024, 74(2): 281−331.

[2] ZHAO Xia, WANG Limin, ZHANG Yufei, et al. A review of convolutional neural networks in computer vision[J]. Artificial intelligence review, 2024, 57(4): 99.

[3] KHEDDAR H, HEMIS M, HIMEUR Y. Automatic speech recognition using advanced deep learning approaches: a survey[J]. Information fusion, 2024, 109: 102422.

[4] TORFI A, SHIRVANI R A, KENESHLOO Y, et al. Natural language processing advancements by deep learning: a survey[EB/OL]. (2020−03−02)[2025−03−04]. https://arxiv.org/abs/2003.01200.

[5] UŠINSKIS V, MAKULAVIČIUS M, PETKEVIČIUS S, et al. Towards autonomous driving: technologies and data for vehicles-to-everything communication[J]. Sensors, 2024, 24(11): 3411.

[6] 徐向阳, 胡文浩, 董红磊, 等. 自动驾驶汽车测试场景构建关键技术综述[J]. 汽车工程, 2021, 43(4): 610−619.
XU Xiangyang, HU Wenhao, DONG Honglei, et al. Review of key technology for autonomous vehicle test scenario construction[J]. Automotive engineering, 2021, 43(4): 610−619.

[7] FAN Lili, WANG Junhao, CHANG Yuanmeng, et al. 4D mmWave radar for autonomous driving perception: a comprehensive survey[J]. IEEE transactions on intelligent vehicles, 2024, 9(4): 4606−4620.

[8] LEI Han, WANG Baoming, SHUI Zuwei, et al. Automated lane change behavior prediction and environmental perception based on SLAM technology[EB/OL]. (2024−04−06)[2025−03−04]. https://arxiv.org/abs/2404.04492.

[9] XIE Jing, ABBASS K, LI Di. Advancing eco-excellence: Integrating stakeholders' pressures, environmental awareness, and ethics for green innovation and performance[J]. Journal of environmental management, 2024, 352: 120027.

[10] LI Ying, MA Lingfei, ZHONG Zilong, et al. Deep learning for LiDAR point clouds in autonomous driving: a review[J]. IEEE transactions on neural networks and learning systems, 2021, 32(8): 3412−3432.

[11] 李佳男, 王泽, 许廷发. 基于点云数据的三维目标检测技术研究进展[J]. 光学学报, 2023, 43(15): 1515001.
LI Jianan, WANG Ze, XU Tingfa. Three-dimensional object detection technology based on point cloud data[J]. Acta optica sinica, 2023, 43(15): 1515001.

[12] JHALDIYAL A, CHAUDHARY N. Semantic segmentation of 3D LiDAR data using deep learning: a review of projection-based methods[J]. Applied intelligence, 2023, 53(6): 6844−6855.

[13] POUX F, BILLEN R, POUX F, et al. Voxel-based 3D point cloud semantic segmentation: unsupervised geometric and relationship featuring vs deep learning methods[J]. ISPRS international journal of geo-information, 2019, 8(5): 213.

[14] XU Xiaobin, ZHANG Lei, YANG Jian, et al. Object detection based on fusion of sparse point cloud and image information[J]. IEEE transactions on instrumentation and measurement, 2021, 70: 2512412.

[15] LIU Ruihua, NAN Haoyu, ZOU Yangyang, et al. AS-3DFCN: automatically seeking 3DFCN-based brain tumor segmentation[J]. Cognitive computation, 2023, 15(6): 2034−2049.

[16] WANG Jianfeng, SONG Lin, LI Zeming, et al. End-to-end object detection with fully convolutional network[C]//2021

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2021: 15844−15853.

[17] NGUYEN D A, HOANG K N, NGUYEN N T, et al. Enhancing indoor robot pedestrian detection using improved PIXOR backbone and Gaussian heatmap regression in 3D LiDAR point clouds[J]. IEEE access, 2024, 12: 9162−9176.

[18] XIE Enze, YU Zhiding, ZHOU Daquan, et al. M^2BEV: multi-camera joint 3D detection and segmentation with unified birds-eye view representation[EB/OL]. (2022−04−11)[2025−03−04]. https://arxiv.org/abs/2204.05088.

[19] 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. Piscataway: IEEE, 2023: 21674−21683.

[20] Vision and pattern Recognition. 2023: 21674−21683.

[21] SHI Shaoshuai, GUO Chaoxu, JIANG Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10526−10535.

[22] CHARLES R Q, HAO Su, MO Kaichun, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 77−85.

[23] QI C R, YI Li, SU Hao, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[EB/OL]. (2017−06−07)[2025−03−04]. https://arxiv.org/abs/1706.02413.

[24] SHI Shaoshuai, WANG Xiaogang, LI Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 770−779.

[25] QIAN Guocheng, LI Yuchen, PENG Houwen, et al. PointNeXt: revisiting PointNet++ with improved training and scaling strategies[EB/OL]. (2022−06−09)[2025−03−04]. https://arxiv.org/abs/2206.04670.

[26] YANG Zetong, SUN Yanan, LIU Shu, et al. 3DSSD: point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11037−11045.

[27] YIN Tianwei, ZHOU Xingyi, KRHENBUHL P. Center-based 3D Object Detection and Tracking[EB/OL]. (2020−06−19)[2025−03−04]. https://arxiv.org/abs/2006.11275.

[28] ABBAS W, SHABBIR M, LI Jiani, et al. Resilient distributed vector consensus using centerpoint[J]. Automatica, 2022, 136: 110046.

[29] HU Yaoqi, NIU Axi, SUN Jinqiu, et al. Dynamic center point learning for multiple object tracking under Severe occlusions[J]. Knowledge-based systems, 2024, 300: 112130.

[30] WANG Hai, TAO Le, CAI Yingfeng, et al. CenterPoint-SE: a single-stage anchor-free 3-D object detection algorithm with spatial awareness enhancement[J]. IEEE transactions on intelligent transportation systems, 2023, 24(10): 10760−10773.

[31] 刘小波, 肖肖, 王凌, 等. 基于无锚框的目标检测方法及其在复杂场景下的应用进展[J]. 自动化学报, 2023, 49(7): 1369−1392.
LIU Xiaobo, XIAO Xiao, WANG Ling, et al. Anchor-free based object detection methods and its application progress in complex scenes[J]. Acta automatica sinica, 2023, 49(7): 1369−1392.

[32] 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. Piscataway: IEEE, 2020: 11618−11628.

[33] 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. Piscataway: IEEE, 2022: 1080−1089.

[34] WU Hai, WEN Chenglu, SHI Shaoshuai, et al. Virtual sparse convolution for multimodal 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 21653−21662.

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、机械臂控制。编写著作 5 部, 发表学术论文 80 余篇。E-mail: lujun0260@sina.com。

邹康成, 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: z127577@163.com。

李杨, 硕士研究生, 主要研究方向为点云目标检测, 跟踪, 机器视觉, 图像处理。E-mail: liyang142857@126.com。