中图法分类号： **文献标识码：** **文章编号：** 1006-8961(XXXX)XX-0001-17

论文引用格式： Zhang Xingwang, Wang Junfan, Miao Qiheng, Dong Zhekang, He Zhiwei, Ma Guojin. 3D object detection for dynamic traffic scenarios[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.250543. (张兴旺, 王俊帆, 缪其恒, 董哲康, 何志伟, 马国进. 面向动态交通场景的3D目标检测[J/OL]. 中国图象图形学报,XXXX:1-17. DOI: 10.11834/jig.250543.)[DOI:10.11834/jig.250543]

面向动态交通场景的3D目标检测

张兴旺[1,2],王俊帆[1,2],缪其恒[3],董哲康[1,2],何志伟[1,2],马国进[1,2]

1. 杭州电子科技大学电子信息学院， 杭州 310018; 2. 浙江省全省智能汽车电子研究重点实验室， 杭州 310018; 3. 浙江华锐捷技术有限公司， 杭州 310051

摘 要：目的 在动态交通场景下,现有纯视觉3D交通目标检测仍面临两大瓶颈:一是固定的体素采样策略难以适应多变场景;二是时序信息利用不足导致模型对遮挡目标感知受限。针对上述问题,提出了一种面向动态交通场景的3D目标检测框架。**方法** 首先,设计了自适应体素特征采样策略,通过端到端的流程评估场景复杂度(包括特征裁剪和统计量提取等),从而动态选择采样方式,实现了不同场景采样的自适应优化。其次,提出了时序分组融合模块,通过将连续多帧鸟瞰图(bird′s-eye-view,BEV)特征分组融合,实现了高效的时空信息建模,增强了对动态目标的鲁棒性。**结果** 在权威公开数据集上的实验表明,所提方法在轻量化配置(ResNet50)下相较于基线模型提高了0.8%的平均精度均值(mean average precision,mAP)和0.8%的nuScenes综合检测指标(nuScenes detection score,NDS),优于Fast-BEV等主流方法;在高性能配置(ResNet101)下进一步提升了1.6%的mAP和2.3%的NDS。消融实验验证了自适应体素特征采样策略和时序分组融合模块各自的有效性,可视化结果也表明该方法在遮挡和动态场景下具有更完整、更精准的检测能力。**结论** 本文通过自适应采样策略与时序分组融合模块,有效提升了纯视觉3D目标检测在动态交通场景下的性能与适应性。未来工作将聚焦于增强短时序建模能力和引入更细粒度的局部自适应机制。https://www.scidb.cn/s/rA7ZFf

关键词： 3D目标检测;鸟瞰图;时序融合;自适应采样;动态交通场景

3D object detection for dynamic traffic scenarios

Zhang Xingwang[1,2], Wang Junfan[1,2], Miao Qiheng[3], Dong Zhekang[1,2], He Zhiwei[1,2], Ma Guojin[1,2]

1. School of Electronic and Information Engineering, Hangzhou Dianzi University, Hangzhou 310018, China; 2. Zhejiang Provincial Key Laboratory of Intelligent Vehicle Electronics Research, Hangzhou 310018, China; 3. Zhejiang Huaruijie Technology Co., Ltd, Hangzhou, 310051, China

Abstract: Objective The advancement of autonomous driving technology imposes stringent demands on precise environmental perception. While pure vision-based Bird′s-Eye View (BEV) 3D object detection has garnered significant attention due to its cost-effectiveness, it still confronts two major bottlenecks in dynamic traffic scenarios. Firstly, existing methods often inadequately exploit temporal information when processing multi-frame images, leading to insufficient perception capabilities for moving objects and occluded areas. Secondly, the prevalent fixed voxel sampling strategies struggle to adapt to the variability of scene complexity, failing to balance computational efficiency in sparse scenes with detection accu-

收稿日期：2025-10-30;**修回日期：**2026-03-10

基金项目： 浙江省优秀青年基金项目(LZYQ25F020005); 浙江省“领雁”科技计划项目(2024C01143); 浙江省“尖兵”科技计划项目(2023C01132)

Supported by: Zhejiang Provincial Outstanding Youth Fund Project (Grant No. LZYQ25F020005); Zhejiang Provincial "Leading Goose" Science and Technology Program (2024C01143); Zhejiang Provincial "Vanguard" Science and Technology Program (2023C01132).

racy in dense, crowded environments. This paper aims to address these critical limitations by constructing an efficient and robust pure vision-based 3D detection framework specifically designed for dynamic traffic conditions. **Method** This paper proposes a novel framework for 3D object detection in dynamic traffic scenes, introducing two core technical innovations that address key challenges in temporal modeling and computational efficiency. The Temporal Grouping Fusion Module revolutionizes temporal feature integration by processing consecutive BEV features through a grouped architecture inspired by Res2Net principles. This module partitions temporal sequences into multiple groups, where features within each group are concatenated and compressed via shared 1×1 convolutions to generate compact representations. The breakthrough lies in the residual connections established between groups, allowing subsequent groups to directly inform preceding ones before applying shared 3×3 convolutions. This sophisticated design captures fine-grained temporal dependencies at local levels while enabling comprehensive information propagation across the entire temporal sequence. The architecture significantly enhances the model's capacity to model complex dynamic interactions, particularly in handling scene transitions and occlusion patterns, while maintaining parameter efficiency through weight sharing across all processing groups. Complementing this, the Adaptive Voxel Feature Sampling Strategy introduces an intelligent sampling mechanism governed by a comprehensive scene complexity assessment pipeline. The system begins by processing multi-view image features through a meticulous normalization procedure involving clamping, rectification, and maximum-value normalization. It then computes both global feature intensity (mean) and local texture complexity (standard deviation) statistics, which are intelligently fused using optimized weighting coefficients. The framework incorporates temporal consistency through historical reference integration and exponential moving average smoothing, producing robust complexity scores that reliably classify scenes into 'simple' or 'complex' categories. This classification dynamically orchestrates the sampling strategy, seamlessly switching between computationally efficient Fast-Ray transformations for straightforward scenarios and sophisticated deformable attention mechanisms for challenging environments. In the integrated framework, multi-camera images are used as input. Surround-view images first pass through a backbone network to extract multi-scale 2D features, which are then fed into a Feature Pyramid Network (FPN) and an Adaptive Voxel Feature Sampling Strategy respectively. In the Adaptive Voxel Feature Sampling Strategy, a scoring function is employed to calculate the complexity of the current frame, thereby adaptively selecting the sampling method. Secondly, the outputs of the two branches are lifted from multi-view 2D perspective features to the 3D space through geometric projection in the 2D-3D module to generate a BEV (Bird's-Eye View) representation. Finally, after decoding by the Temporal Grouping Fusion Module and the 3D detection head, the final object detection results are obtained. The Adaptive Voxel Feature Sampling Strategy effectively balances the trade-off between detection accuracy and computational efficiency. In addition, the Temporal Grouping Fusion Module further enhances the model's stability in dynamic traffic environments and its perception capability for occluded objects. **Result** Extensive experiments were conducted on the large-scale public autonomous driving dataset, nuScenes. The evaluation adheres to the official nuScenes metrics, including mean Average Precision (mAP) and the comprehensive nuScenes Detection Score (NDS). Under a lightweight configuration using ResNet-50 as the backbone network and an input resolution of 256×704, our proposed method achieved a mAP of 36.2% and an NDS of 49.5%. This performance surpasses contemporary state-of-the-art pure vision methods such as Fast-BEV (35.4% mAP, 48.7% NDS), and demonstrates substantial improvements over other mainstream methods like BEVDet and BEVDepth. Under a higher-performance configuration employing a ResNet-101 backbone and an input resolution of 900×1600, our method's performance was further elevated, reaching 42.9% mAP and 55.8% NDS, ranking it among the top-performing approaches in the comparison. Ablation studies systematically validated the individual and collective contributions of the two core modules. Removing either module led to a noticeable performance drop, confirming that both the temporal grouping fusion and the adaptive sampling strategy are pivotal to the overall success. Qualitative results from visualization further corroborated the quantitative findings, demonstrating that our method produces more complete detections and more accurately aligned 3D bounding boxes, particularly in challenging scenarios involving occlusions and fast-moving objects, compared to the baseline model. **Conclusion** This paper addresses two fundamental challenges in vision-based 3D object detection for dynamic traffic scenes: the effective utilization of temporal information and the adaptive adjustment of feature sampling under varying scene complexities. To this end, we propose a unified and efficient framework integrating a Temporal Grouping Fusion Module and an Adaptive Voxel

Feature Sampling Strategy. The proposed temporal grouping design enables fine-grained feature interaction and multi-scale information propagation across consecutive BEV representations, ensuring consistent spatiotemporal perception even in complex and occluded environments. Meanwhile, the adaptive sampling mechanism dynamically aligns computational effort with scene complexity, thereby achieving a more balanced trade-off between efficiency and accuracy. Comprehensive experiments conducted on the nuScenes benchmark confirm the superior generalization and robustness of our framework. The proposed approach not only surpasses existing pure vision-based methods in both mAP and NDS but also demonstrates remarkable scalability across different backbone architectures and resolutions. More importantly, it maintains real-time inference capability, showing great promise for deployment in resource-constrained autonomous driving systems. Beyond quantitative results, qualitative visualization further illustrates the framework's ability to preserve object integrity, refine 3D box alignment, and maintain temporal consistency during rapid motion or heavy occlusion. In future work, we plan to enhance the framework by introducing region-aware adaptive mechanisms and shorter temporal fusion pipelines for low-latency applications. Moreover, integrating self-supervised temporal learning and multi-modal knowledge distillation could further improve the robustness and cross-domain adaptability of the system. Ultimately, this research provides a feasible, scalable, and cost-effective perception paradigm, contributing to the broader goal of realizing safe, intelligent, and human-centered autonomous driving technology.

Key words: 3D Object Detection; Bird's-Eye View; Temporal Fusion; Adaptive Sampling; Dynamic Traffic Scenario

+中图法分类号:(TP391) **文献标识码**:A **文章编号**:1006-8961(年)-

0 引 言

近年来,随着国家大力推动智能交通与自动驾驶的发展,交通感知技术已成为智慧出行体系中的核心支撑环节。基于纯视觉的3D目标检测凭借其低部署成本和丰富的语义感知优势,已成为智能交通感知系统的重点攻关方向(Pu 等, 2025)。如图1所示,相比于缺乏深度信息的2D检测(Zhao 等, 2024a; Wang 等, 2024a; Wang 等, 2024b; Xi 等, 2025)和成本高昂的激光雷达(Mei 等, 2024; Chae 等, 2024; Baur 等, 2024; Zhao 等, 2025; Zhou 等, 2024b)方案,纯视觉3D检测方法能够在鸟瞰图(Bird's-Eye-View, BEV)空间中直接恢复目标的精确位置与尺度,为下游规划决策提供完整的空间表征。尽管近些年基于BEV的感知算法(Li 等, 2024a; Li 等, 2023a)取得了显著进展,但在应对复杂多变的交通场景时,传统方法在时序信息利用与特征采样效率两个关键维度上仍存在一些瓶颈。

首先,现有的特征采样机制缺乏对场景复杂度的自适应能力,无法满足不同场景对采样策略的合理适配。无论是基于固定的体素采样(Wang 等, 2021; Liu 等, 2022; Liu 等, 2023; Yang 等, 2023),还是依赖2D检测器反向指导3D查询(Cheng 等, 2024),目前大多方法采用静态采样策略。例如,Fast-BEV(Huang 等, 2023)仅通过预计算的静态查找表将图像特征沿相机射线投影至体素,这种方式虽提升了推理速度,却难以应对环境的动态变化。然而,交通场景具有高度的动态变化性,驾驶者上一时刻可能行驶在复杂场景,下一时刻可能在简单场景。BEVCon(Leng 等, 2025)虽然引入了密集的实例级特征池化与尺度感知机制来细化表征,但在空

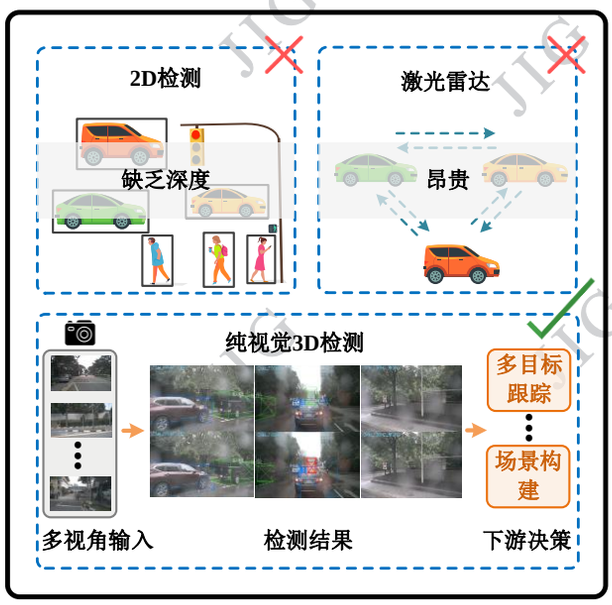


图1 纯视觉3D目标检测优势示意图

Fig. 1 Schematic Diagram of Advantages of Vision-only 3D Object Detection

旷的简单路况下，这种全场景的高密度计算会导致大量算力浪费在背景区域。因此，固定的采样范式无法在不同场景下动态平衡计算效率与检测精度。

其次，如何高效地利用时序历史信息，是提升遮挡、运动物体检测精度的关键。当前大多数处理多帧BEV特征的方法陷入了“性能与效率”的博弈。一类方法仅采用简单的特征拼接或局部卷积（Lang等，2019；Yang等，2024；Yin等，2021；Mao等，2023），例如（Huang等，2023）仅通过简单的空间对齐将历史帧特征在通道维度上进行直接拼接，这类方法对运动目标的感知能力较弱，且在处理目标相互遮挡等复杂情况时，使得模型的感受野受限，很难看到更大范围的动态变化。另一类方法试图引入状态空间模型来增强全局感知（You等，2025），虽然在一定程度上捕捉了长期依赖，但其核心机制需要将二维BEV特征“离散序列化”以适应一维序列处理。同时，（Feng等，2025）采用基于Mamba和线性注意力的长时时序感知模块来捕捉长期时序依赖，这类方法不可避免地破坏了特征原本的2D空间结构与几何一致性，并且增加了参数量和训练时间。

针对上述两大瓶颈，本文提出了一种面向动态交通场景的3D目标检测框架。该框架包含两个针对性的核心模块：自适应体素特征采样和时序分组融合模块。自适应体素特征采样策略利用场景感知驱动从而动态选择采样机制。该策略能够根据场景的复杂度自适应选择采样方式，在简单场景下节省算力，在复杂场景下保障精度，实现对不同场景采样方式的动态适配。时序分组融合模块通过基于分组卷积的时空建模机制对连续的BEV特征帧进融合。不同于简单拼接和破坏空间结构的序列化方法，该模块在保持BEV特征2D几何结构的同时，通过分组交互有效聚合多帧历史信息，显著增强了模型对运动目标和遮挡场景的推断能力。在公开权威数据集nuScenes上的实验结果表明，所提出方法在保持较高精度的同时具备一定的应用潜力，可为后续视觉感知系统在自动驾驶等场景中的落地提供参考价值。本文的主要贡献总结如下：(1)设计了一个时序分组融合模块，通过融合多帧BEV特征，进一步增强了对运动目标和遮挡场景的感知能力；(2)提出了一种自适应体素特征采样策略，实现了采样机制随场景复杂度的灵活切换，有效平衡了检测精度与计算开销。(3)本文方法在复杂天气与动态场景下均表现出优越的鲁棒性。在公开权威数据集上，本方法相较于基线模型取得了更优的性能表现，mAP提升至42.9%，NDS达到55.8%。

1 相关工作

本文从2个方向阐述近些年来纯视觉3D交通目标检测研究现状，分别是基于固定采样的纯视觉3D交通目标检测、基于时序信息的纯视觉3D交通目标检测。

1.1 基于固定采样的纯视觉3D交通目标检测

LSS（Philion等，2020）首次提出了具有开创性的“Lift-Splat-Shoot”框架。该方法为每个像素预测离散的深度分布，并利用相机的内外参数将像素沿深度方向投影到3D空间以形成“视锥”点云。随后，将特征铺展至BEV栅格中，在BEV特征图上采用2D卷积网络完成目标检测。FCOS3D（Wang等，2021）将经典的2D检测器FCOS扩展至3D领域，在2D特征图的每个位置直接回归目标的3D属性，实现端到端的3D目标检测。DETR3D（Wang等，2022）通过在3D空间中定义一组可学习的目标查询，利用相机内外参将其反投影到多视角图像中进行特征采样，实现端到端的3D目标检测。BEVDepth（Li等，2023b）通过利用激光雷达点云生成深度真值来辅助深度预测网络训练。PETR（Liu等，2022）通过构建位置编码实现从图像域到BEV域的隐式映射，避免了显式几何投影过程。FastBEV（Huang等，2023）基于预计算优化策略，提前建立图像像素与BEV栅格之间的投影对应关系，使推理阶段仅需在BEV有效区域进行特征采样，从而显著减少冗余计算并提升整体效率。

1.2 基于时序信息的纯视觉3D交通目标检测

时序信息的引入对于解决纯视觉感知中的动态物体遮挡问题至关重要。现有的基于时序的方法大致可分为基于BEV特征的稠密融合与基于稀疏查询的流式融合两类。

在早期的探索中，BEVDet4D（Huang等，2022）率先将多帧特征融合引入基于LSS的检测框架，通过简单的特征拼接与空间对齐，显著提升了速度预测的准确性，确立了时序融合的基准。在此基础上，基于BEV特征的稠密融合方法进一步关注如何高效提取时空特征。BEVFormer（Li等，2022）提出了

时空注意力机制，利用可学习的BEV查询通过时间自注意力从历史BEV特征中获取信息，大幅增强了模型在复杂场景下的鲁棒性。随后，BEVNeXt（Li等，2024b）针对注意力机制计算量大的问题，设计了更高效的时序融合模块以替代复杂的Transformer结构，并强化了BEV编码器，在保持高性能的同时提升了推理效率。SOLOFusion（Park等，2022）则进一步探索了长短时记忆特征的平衡，利用高分辨率的短时特征与低分辨率的长时特征互补。

另一类方法则转向了更高效的基于稀疏查询的流式融合。PETR系列（Liu等，2022）摒弃了显式的BEV特征构建，直接在3D位置编码特征上进行查询更新。StreamPETR（Wang等，2023）在此基础上提出了流式感知范式，将上一帧的目标查询（Object Query）作为当前帧的先验信息进行传播与更新。这种方法避免了重复计算历史特征，能够以极低的计算成本实现长时序信息的利用，显著提升了运动目标的检测稳定性。

1.3 小结

当前纯视觉3D交通目标检测的方法虽然取得了显著进展，但仍存在两个关键瓶颈：(1)时序信息的利用仍不充分；(2)特征采样策略较为固定，缺乏自适应性。

2 3D交通目标检测框架

本文提出的3D交通目标检测框架如图2所示，总框架以多摄像头图像作为输入，环视图像首先通过骨干网络提取多尺度2D特征，随后分别送入特征金字塔网络和自适应体素特征采样策略。在自适应体素特征采样策略中，通过得分函数来计算当前帧的复杂度从而自适应选择采样方法。其次，两个分支的输出在2D-3D模块中通过几何投影将多视角的2D透视特征提升至3D空间，以生成BEV表征。最后经过时序分组融合模块和3D检测头解码，得到最终的目标检测结果。自适应体素特征采样策略有效平衡了检测精度与计算效率之间的权衡。此外，时序分组融合模块进一步提升了模型在动态交通环

境中的稳定性和对遮挡目标的感知能力。

2.1 时序分组融合模块

在3D纯视觉交通目标检测中，时序融合是提升动态场景感知精度的核心环节，通过关联多帧BEV特征，可有效补偿单帧视觉数据在深度估计、

运动轨迹建模上的固有局限。然而传统时序融合方法仍存在显著设计缺陷，多数方案直接进行通道拼接或简单加权平均处理多帧特征，这些融合方法难以适配动态交通场景的复杂需求。

本文提出的时序分组融合模块如图3所示，对四帧连续的BEV特征（$\boldsymbol{B}_{t-3}$, $\boldsymbol{B}_{t-2}$, $\boldsymbol{B}_{t-1}$, $\boldsymbol{B}_t$）进行时序分组融合。$\boldsymbol{B}_t \in \mathbf{R}^{C\times H\times W\times Z}$表示当前帧的BEV特征。整体流程如下：

首先，将连续的四帧BEV特征按照相邻的两帧分为2组，得到分组后的特征$\boldsymbol{G}_1$和$\boldsymbol{G}_2$：

$$\begin{aligned}\boldsymbol{G}_1 &= [\boldsymbol{B}_{t-2}, \boldsymbol{B}_{t-3}]\\ \boldsymbol{G}_2 &= [\boldsymbol{B}_t, \boldsymbol{B}_{t-1}]\end{aligned} \tag{1}$$

对于特征$\boldsymbol{G}_1$和$\boldsymbol{G}_2$，先在通道维度进行拼接，并通过共享的1×1卷积进行降维，得到：

$$\begin{aligned}\boldsymbol{B}'_1 &= \boldsymbol{K}^{1\times 1}(\boldsymbol{G}_1)\\ \boldsymbol{B}'_2 &= \boldsymbol{K}^{1\times 1}(\boldsymbol{G}_2)\end{aligned} \tag{2}$$

式中$\boldsymbol{B}'_1$和$\boldsymbol{B}'_2$为经过1×1卷积后的输出特征，$\boldsymbol{K}^{1\times 1}$为共享的1×1卷积操作。

随后，对$\boldsymbol{B}'_2$施加3×3卷积得到$\boldsymbol{B}''_2$。

$$\boldsymbol{B}''_2 = \boldsymbol{K}^{3\times 3}(\boldsymbol{B}'_2) \tag{3}$$

式中$\boldsymbol{K}^{3\times 3}$为3×3卷积操作。

同时，对$\boldsymbol{B}'_1$和$\boldsymbol{B}''_2$进行残差相加，再通过3×3卷积，得到输出$\boldsymbol{B}''_1$：

$$\boldsymbol{B}''_1 = \boldsymbol{K}^{3\times 3}\left(\boldsymbol{B}'_1 + \boldsymbol{B}''_2\right) \tag{4}$$

公式(3)和(4)实现了跨组的时序信息传递，能够增强网络对动态场景变化的建模能力。最终，时序融合BEV表征$\tilde{\boldsymbol{B}}$表示为：

$$\tilde{\boldsymbol{B}} = \boldsymbol{K}^{1\times 1}\left(\left[\boldsymbol{B}''_1, \boldsymbol{B}''_2\right]\right) \tag{5}$$

式中$[\boldsymbol{B}''_1, \boldsymbol{B}''_2]$为两组融合特征在通道维度的拼接结果；$\tilde{\boldsymbol{B}}$为最终的时序融合BEV表征。

2.2 自适应体素特征采样策略

如图4所示，“动态交通场景”不仅指目标的运动，还指场景复杂度的变化。固定的采样策略会造成算力浪费或导致感知精度下降（信息获取不足），不仅带来安全隐患而且无法为时序分组融合模块提供高质量的体素信息。自适应体素特征采样策略通

过场景复杂度判断来动态切换采样策略，适配动态场景的多变性，平衡效率与精度。

自适应体素特征采样策略示意图如下所示：

基于图5，整个过程主要由特征裁剪、归一化、统计特征得分以及指数移动平均（exponential mov-

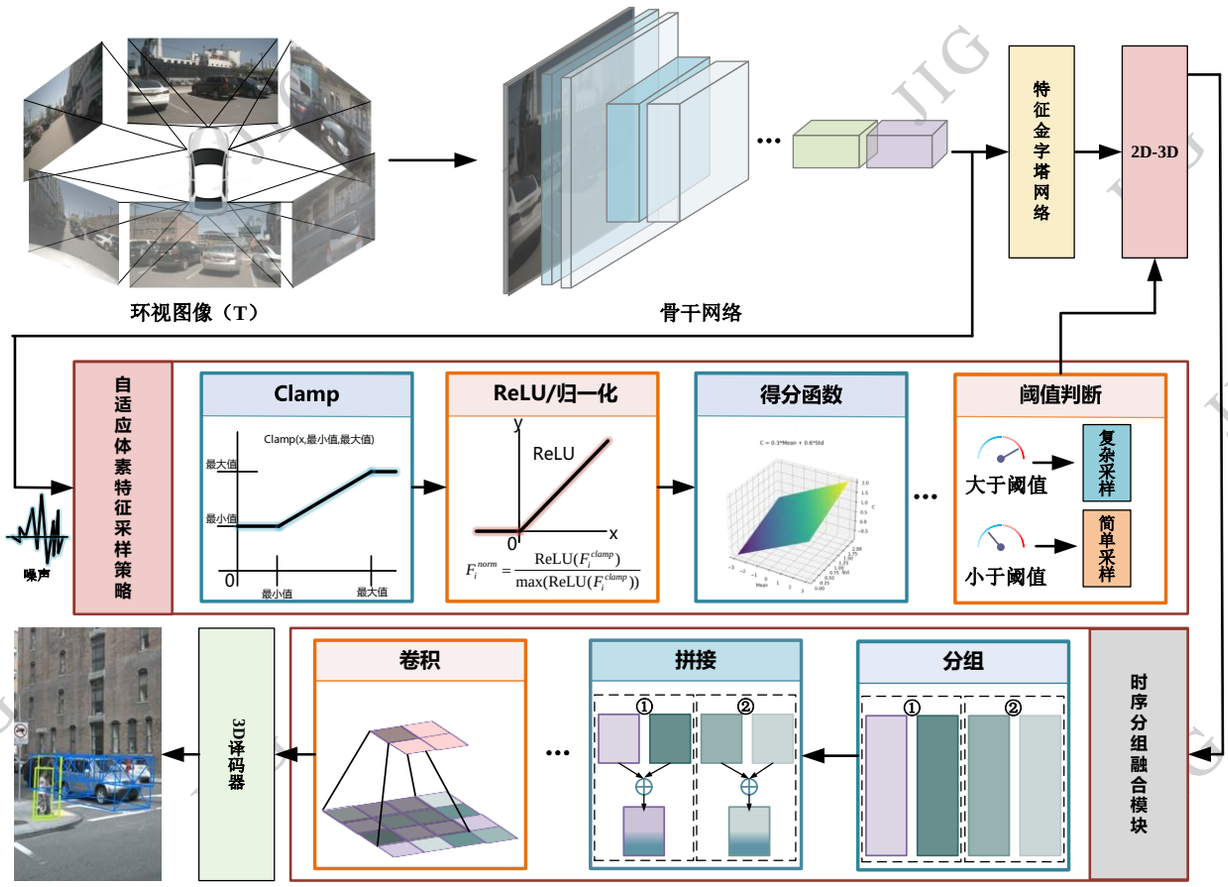


图2　提出的3D交通目标检测框架

Fig. 2　The proposed 3D traffic object detection framework

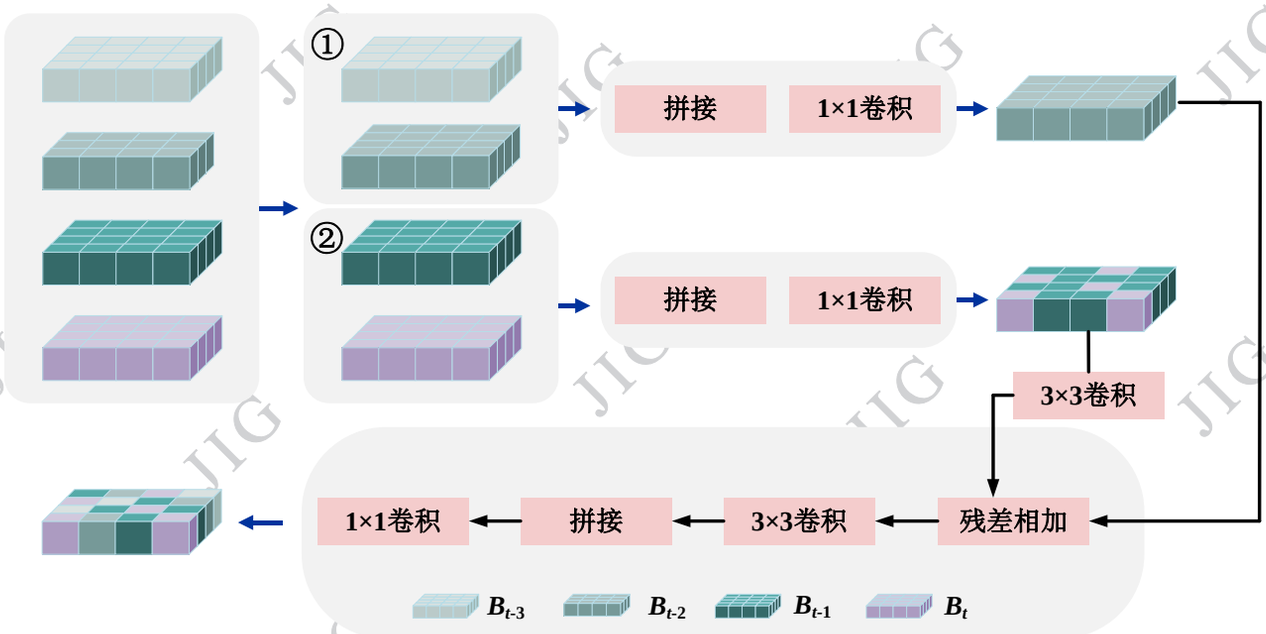


图3　时序分组融合模块

Fig. 3　Temporal grouping fusion module

ing average，EMA）四个核心步骤构成，具体流程如下：

首先，将当前输入的六张环视图像通过骨干网络提取2D特征向量 $\boldsymbol{F}_i \in \mathbf{R}^N$，然后采用Clamp操作抑

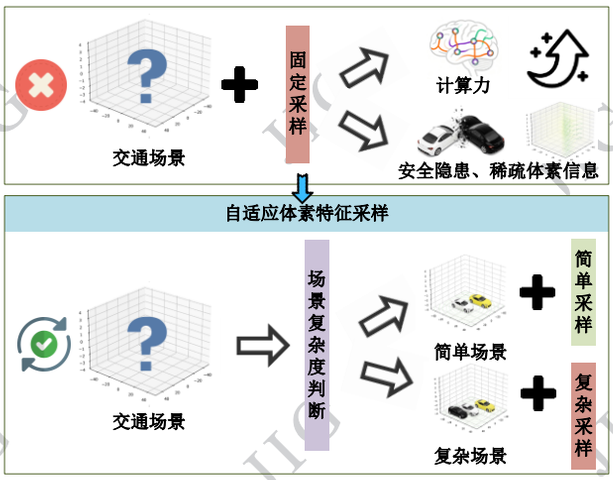


图4　自适应体素特征采样在交通场景中的优化示意图

Fig. 4　Schematic diagram of adaptive voxel feature sampling optimization in traffic scenarios

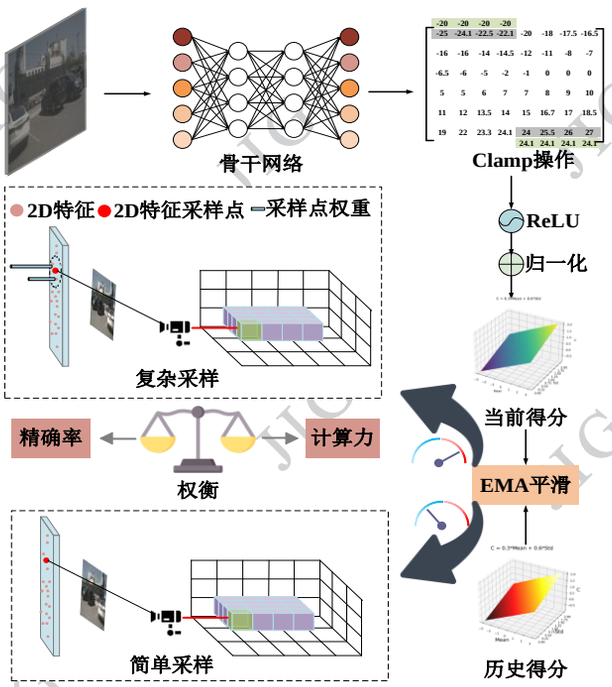


图5　自适应体素特征采样策略

Fig. 5　Adaptive voxel feature sampling strategy

制极端值。通过设定分位数阈值$(q_{\min}, q_{\max})$，$q_{\min}$=5%，$q_{\max}$=95%，将特征裁剪到合理范围，以避免异常点对统计量的干扰。

$$\boldsymbol{F}_i^{\text{clamp}} = \text{Clamp}(\boldsymbol{F}_i, q_{\min}, q_{\max}) \tag{6}$$

式中，$\boldsymbol{F}_i^{\text{clamp}}$为第$i$张输入图像的特征向量；$q_{\min}$为截断的下限值，$q_{\max}$为截断的上限值；Clamp为对输入特征进行截断，确保所有特征值位于设定分位数阈值中。

随后，对裁剪后的特征进行非线性处理与归一化操作。首先使用ReLU激活函数去除负值，以确保复杂度不受无效负值的影响；然后除以该图像特征中的最大值，以实现多帧特征的归一化。

$$\boldsymbol{F}_i^{\text{norm}} = \frac{\text{ReLU}(\boldsymbol{F}_i^{\text{clamp}})}{\max(\text{ReLU}(\boldsymbol{F}_i^{\text{clamp}}))} \tag{7}$$

式中，$\boldsymbol{F}_i^{\text{norm}}$为归一化后的特征，归一化后，每一张图像特征均映射至[0,1]区间。

在此基础上，本文通过均值与标准差作为客观指标，用以表征每张图像的整体强度水平与离散程度。

$$\text{mean}_i = \frac{1}{K}\sum_{j=1}^{K} F_{ij}^{\text{norm}}, \tag{8}$$

$$\text{std}_i = \sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(F_{ij}^{\text{norm}} - \text{mean}_i\right)^2} \tag{9}$$

式中，mean_i归一化特征的均值，反映特征整体的激活强度。std_i归一化特征的标准差，衡量特征分布的离散程度，$F_{i,j}^{\text{norm}}$表示第i张图像的特征图中的第j个特征值，K表示该特征图中的特征值的个数。

为了兼顾全局与局部信息，将均值和标准差线性融合作为得分函数（复杂度），其中均值反映整体特征强度，而标准差能够度量局部差异与纹理复杂度。

$$C_i = \alpha \cdot \text{mean}_i + \beta \cdot \text{std}_i \tag{10}$$

式中，α与β为在数据集上学习得到的最优权重，用于平衡均值与标准差对复杂度的贡献；C_i为第i张图片的复杂度。

随后，对六张输入图像的复杂度进行平均计算，得到当前帧的复杂度$C_{\text{raw-images}}$。

$$C_{\text{raw-images}} = \frac{1}{6}\sum_{i=1}^{6} C_i \tag{11}$$

式中，$C_{\text{raw-images}}$为六张相机图像取平均后的当前帧复杂度。

假设最近三帧的历史复杂度结果为$\{H_1, H_2, H_3\}$，其均值作为历史复杂度参考，随后与当前帧的复杂度进行加权融合。

$$C_{\text{history}} = \frac{H_1 + H_2 + H_3}{3} \tag{12}$$

$$C_{\text{raw}} = \lambda \cdot C_{\text{raw-images}} + (1 - \lambda) \cdot C_{\text{history}} \tag{13}$$

式中，$\lambda \in [0,1]$控制历史与当前复杂度的相对权重。当λ较大时，模型更敏感于当前场景变化；当λ较小时，模型则更注重平滑与稳定性。C_{raw}为融合后的复

杂度，C_{history}历史帧的复杂度。

在得到短期融合结果后，进一步采用EMA对复杂度时间序列进行平滑，以抑制瞬时波动并增强全局趋势的连续性。

$$C_{\text{EMA}}(t)=\gamma\cdot C_{\text{raw}}(t)+(1-\gamma)\cdot C_{\text{EMA}}(t-1) \tag{14}$$

式中，γ为平滑系数，决定当前观测值与历史平滑值的相对贡献。$C_{\text{EMA}}(t)$为时刻t的平滑复杂度分数。$C_{\text{raw}}(t)$为时刻t融合后的复杂度。$C_{\text{EMA}}(t-1)$为前一帧EMA平滑值(初始时为0)。

最后，我们将平滑后的复杂度结果与预设阈值$(\tau_{\min},\tau_{\max})$进行比较，完成对场景复杂度的判定。

$$\text{Scene}=\begin{cases}\text{Simple}, & C_{\text{EMA}}(t)<\tau_{\min}\\ \text{Complex}, & C_{\text{EMA}}(t)\geqslant\tau_{\max}\end{cases} \tag{15}$$

式中Simple和Complex这两个单词分别表示场景简单和场景复杂。

为了更直观清晰地体现整个策略流程，具体的算法伪代码如表1所示：

表1　自适应体素特征采样策略伪代码

Table 1　Pseudo-Code of the Adaptive Voxel Sampling Strategy

算法：自适应体素特征采样策略伪代码
Input: Current image features $F_1, F_2, F_3, F_4, F_5, F_6$; historical EMA values H_1, H_2, H_3
Result: Final smoothed complexity value C_{EMA}
for each feature map F_i **do**
Clamp F_i within quantile range (5%–95%)
Apply ReLU and normalize to [0,1]
Calculate complexity $C_i=\alpha\cdot\text{mean}(F_i)+\beta\cdot\text{std}(F_i)$
end for
Compute overall complexity of current frame C_{cur} by fusing $C_1\sim C_6$
Fuse with historical EMA values and apply EMA smoothing:
$C_{\text{EMA}}(t)=\gamma\cdot C_{\text{cur}}(t)+(1-\gamma)\cdot C_{\text{EMA}}(t-1)$
return C_{EMA}

3　实　验

3.1　实施细节

3.1.1　数据集

本文在nuScenes数据集（Caesar等，2020）上对所提出的方法进行了广泛的评估。nuScenes是一个多模态数据集，涵盖了1000个不同的场景，采集设备包括一台激光雷达、六个环视摄像头和五个毫米波雷达，并以2Hz的频率进行标注。本文方法的评估基于nuScenes官方指标体系，包括：平均精度均值(mAP)、平均平移误差(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)，以及最终的nuScenes检测得分(NDS)。

3.1.2　训练细节

本文方法在4张4090GPU上训练20个epoch。采用AdamW优化器，初始学习率设为1e-3，权重衰减设置为1e-2。学习率调度器使用“PolyLR”策略逐步降低学习率，并在前1000次迭代中使用“warmup”策略进行预热。为了保证对比实验的公平性，本文均采用类平衡分组采样(class-balanced grouping and sampling，CBGS)的训练策略（Zhu等，2019）。除非有明确说明否则实验设置都是在输入尺寸为256×704，骨干网络为ResNet50下进行。

3.2　与其他方法进行总体对比

如表2所示，在ResNet50骨干网络上，本文方法在nuScenes验证集上取得了36.2%的mAP和49.5%的NDS，均优于LST-BEV（NDS0.484，mAP0.321）等方法。在ResNet101骨干网络上，本文方法取得了0.558的NDS和0.429的mAP，领先于BEVCon、MambaBEV和BEVFormer-small-QAF2D等优秀模型。mAP和NDS的提升验证了自适应体素特征采样策略的有效性。该策略通过对多帧图像特征进行裁剪、归一化、复杂度评分及EMA平滑，使模型能准确评估动态场景的复杂度，进而动态分配计算资源，确保了在各种复杂度的场景下均能保持高精度的检测水平。

其次，在所有对比方法中，本文的方法在mAVE指标中均达到了最佳水平(分别为0.355和0.328)，显著优于BEVCon和Fast-BEV等模型。mAVE数据指标的降低验证了时序分组融合模块的有效性，模型通过跨组时序信息传递机制，极大地增强了对动态交通场景的建模能力，从而实现了更精确的速度估计和整体检测性能的提升。

3.3　消融实验

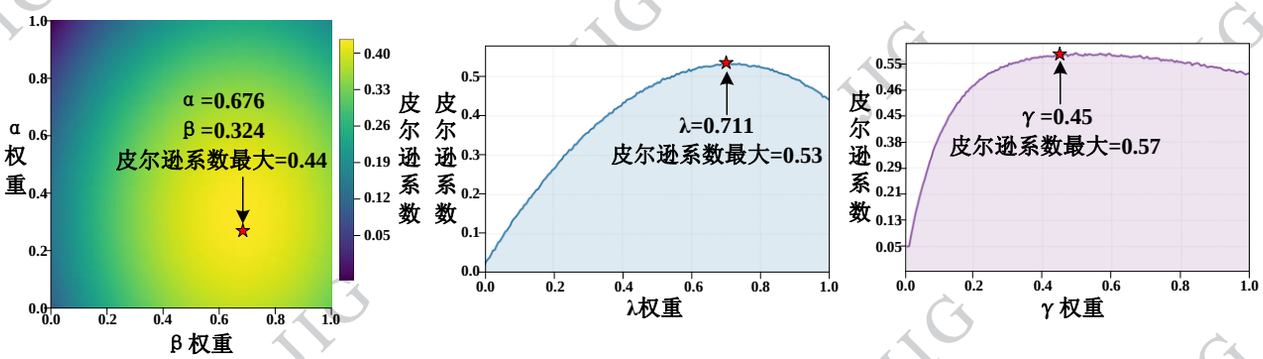
为具体分析面向动态交通场景的3D目标检测框架中各模块的性能表现，本文以nuScenes数据集为实验载体，对自适应体素特征采样策略和时序分组融合模块在基线模型中进行替换性消融

表2　主流方法对比实验

Table 2　Comparison Experiments with Mainstream Methods

/注:加粗字体表示各列最优结果

方法	会议	年份	骨干网络	输入尺寸（像素）	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	NDS ↑
BEVDet	—	2022	ResNet50	256×704	0.286	0.724	0.278	0.590	0.873	0.247	0.372
BEVDepth	AAAI	2023	ResNet50	256×704	0.351	**0.639**	**0.267**	0.479	0.428	**0.198**	0.475
Fast-BEV	NeurIPS	2023	ResNet50	256×704	0.354	0.656	0.281	**0.384**	0.361	0.217	0.487
LST-BEV	—	2025	ResNet50	256×704	0.362	—	—	—	—	—	0.484
本文的			ResNet50	256×704	**0.362**	0.671	0.288	0.391	**0.355**	0.226	**0.495**
FCOS3D	ICCV	2021	ResNet101	900×1600	0.321	0.754	**0.260**	0.486	1.331	**0.158**	0.395
PETR	ECCV	2022	ResNet101	900×1600	0.370	0.711	0.264	0.383	0.865	0.201	0.442
BEVFormer	ECCV	2022	ResNet101	900×1600	0.416	0.673	0.274	0.372	0.394	0.198	0.517
Fast-BEV	NeurIPS	2023	ResNet101	900×1600	0.413	0.584	0.279	**0.311**	0.329	0.206	0.535
BEVFormer-small-QAF2D	—	2024	ResNet101	736×1280	0.397	0.70.3	0.274	0.369	0.404	0.213	0.502
MambaBEV	—	2025	ResNet101	900×1600	0.410	0.688	0.275	0.375	0.423	0.203	0.508
BEVCon	—	2025	ResNet101	900×1600	0.424	0.674	0.274	0.357	0.354	0.183	0.528
本文的			ResNet101	900×1600	**0.429**	**0.566**	0.281	0.344	**0.328**	0.215	**0.558**



(a) α和β (b) λ (c) γ((a)α和β; (b) λ; (c)γ)

图6　α、β、λ和γ与真值标签对应的皮尔逊系数

Fig. 6　Pearson correlation coefficients between α, β, λ, γ and the ground truth labels

实验。

3.3.1　自适应体素特征采样策略性能

基于2.2自适应体素特征采样策略，本文将详细介绍有关公式中参数的作用机理与取值策略。α与β作为计算复杂度的调节超参数，分别控制了全局特征强度与局部纹理差异对最终复杂度评分的贡献比重。均值表征了场景特征图的整体激活水平，而标准差则敏锐捕捉了物体边缘、纹理变化等高频信息。参数λ决定了模型对当前时刻观测值的信赖程度，确保模型在保持对动态环境快速响应，并且具备足够的抗干扰鲁棒性，避免因单帧异常导致的复杂度评分跳变。参数γ在模型中主要起到低通滤波器的作用，能够在有效滤除复杂度评分在时间维度上的高频抖动的同时，避免因过度平滑而导致系统对场景切换的感知滞后，从而为下游计算资源分配策略提供稳定、连续的决策依据。

本文使用nuScenes数据集的真值标签和对应参数的公式来计算皮尔逊系数(pearson correlation coefficient，PCC)，PCC越大说明两者之间的相关性越强，从而确认α、β、λ和γ的选择。具体地，如图6(a)所示，遍历α和β的取值，并计算出对应的PCC，实验结果表明，在α为0.676，β为0.324时，最大PCC为0.44；同样地，如图6(b)，(c)所示，当λ为0.711时，最大PCC为0.53；当γ为0.45时，最大PCC为0.57。通过皮尔逊相关系数驱动的参数选择过程能够有效刻画模型参数与真实性能之间的统计相关性，从而为实验中参数设置的合理性提供了可靠的统计依据。

如表3所示，本文参考(Liu等，2024)使用的动态场景复杂度量化(dynamic scenario complexity quantification，DSCQ)方法来设定场景复杂度阈值，并依据分位数在不确定性量化、风险评估及分布特征匹配等领域的广泛应用(Lo等，2012；Zwart等，2025；Lu等，2025)，将固定阈值设置为分位数在25%的数值，该分位点通过对nuScenes数据集进行遍历统计后获得。通过阈值的设定将复杂度划分为两个区间，将复杂度在0-25%的区间设定为简单场景，这类场景通常是交通参与者稀疏、交互行为较少的路况；将复杂度在25%-100%的区间设定为复杂场景。复杂度阈值的设定为后续自适应采样的选择提供了基础。

本文中，自适应采样模块的消融实验结果如表4所示。需要说明的是，为更充分体现自适应采样模块的场景适配性，本实验选用包含不同动态交通场景信息的图像数据。

基于表4，当骨干网络为ResNet50时，引入自

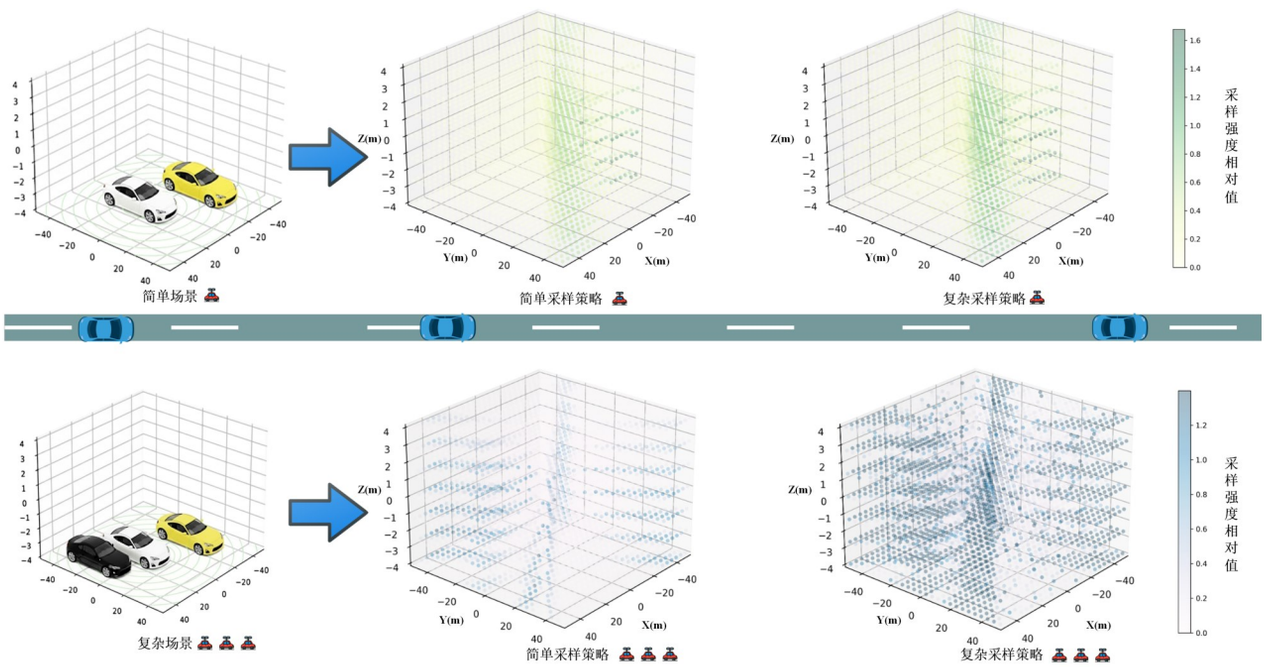


图7　不同场景下不同采样策略体素信息图

Fig. 7　Voxel information map translation with different sampling strategies in various scenario

表3　复杂度阈值设定实验

Table 3　Experiment on the Setting of Complexity Threshold

复杂度范围	复杂度平均值	复杂度分类
0–25%	0.3542	简单场景
25–100%	0.6012	复杂场景

表4　自适应采样消融实验结果

Table 4　Results of Ablation Study on Adaptive Sampling

简单采样	自适应采样	骨干网络	mAP	NDS
√	—	ResNet50	0.354	0.487
—	√	ResNet50	**0.359**	**0.492**
√	—	ResNet101	0.413	0.535
—	√	ResNet101	**0.421**	**0.543**

注：加粗字体表示各列最优结果，“√”表示采用，“—”表示未采用。

适应体素特征采样后，模型mAP从35.4提升至35.9，NDS从48.7提升至49.2；骨干网络替换为ResNet101时，引入自适应体素特征采样后，模型的mAP从41.3提升至42.1，NDS从53.5提升至54.3。实验结果表明，自适应体素特征采样策略可有效适配动态场景的多变性，同时保持较高的感知精度。

图7为“不同场景下不同采样策略体素信息图”，清晰呈现了不同采样策略在简单与复杂场景下的体素信息差异。其中3D空间中体素点的颜色和数目代表采样获取2D信息的强和弱。上半部分针对仅含白色和黄色两辆车的简单场景，左侧是场景示意图，中间和右侧分别为简单采样策略、复杂采样策略的体素信息图，从采样强度相对值的颜色条和体素点信息分布可见，体素信息效果接近，说明简单场景下简单采样可在保证信息完整性的前提下替代复杂采样；下半部分针对包含黑色、白色、黄色三辆车的复杂场景，左侧是场景示意图，中间为简单采样策略的体素图中，采样点稀疏且采样强度相对值低，信息含量匮乏，而右侧复杂采样策略的体素图中，采样点密集且采样强度相对值高，信息丰富度远远超过简单采样策略。整体而言，在简单场景下复杂与简单采样方法的体素信息表现几乎无差异，而在复杂场景下简单采样的信息含量则显著低于复杂采样。直观的验证了自适应体素特征采样策略在动态交通3D目标检测的合理和适配性。

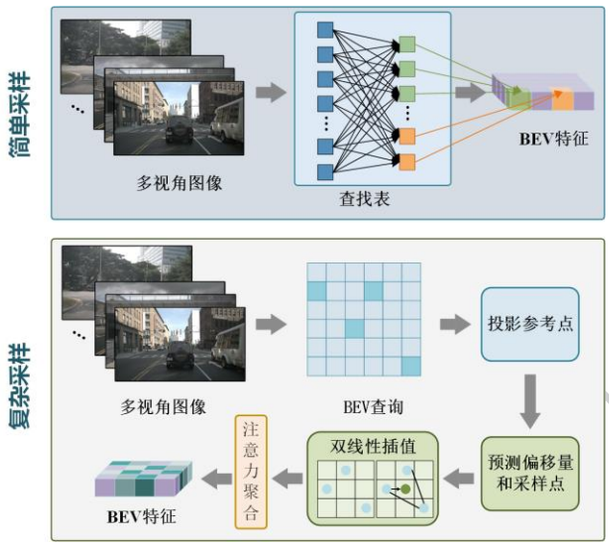


图8　简单和复杂采样原理图

Fig. 8　Schematic Diagrams of Simple and Complex Sampling

为了进一步验证自适应采样策略的有效性，如图8的采样原理所示，简单采样策略通过预计算查找表实现了高效的静态映射，避免了复杂采样中基于可变形注意力机制的大量动态浮点运算，从而显著降低了计算延迟。实验结果表明自适应采样策略在精度与计算开销之间实现了更加合理的权衡；如表5所示，本文对比了单独使用两种不同的采样策略，尽管单独的简单采样在mAP和NDS指标上都略低于复杂采样，但是在CPU上的推理延迟上简单采样远远低于复杂采样。

表5　单独检测不同采样方法实验结果

Table 4　Experimental Results of Separately Detecting Different Sampling Methods

简单采样	复杂采样	mAP	NDS	CPU(ms)
√	–	0.354	0.487	**7.0**
–	√	**0.369**	**0.497**	1078

注：加粗字体表示各列最优结果。

表6　时序分组融合不同分组实验结果

Table 6　Experimental Results of Different Groups /in Temporal Grouping Fusion

使用帧数	分组大小	mAP	NDS
4	1	0.352	0.480
4	2	**0.357**	**0.489**
4	3	0.355	0.482

注：加粗字体表示各列最优结果。

3.3.2　时序分组融合模块

分组大小决定了需要处理的BEV数量，为了验证不同分组大小对检测结果的影响，本文对时序分组融合模块中设置分组大小的合理性进行了实验验证。如表6所示，在使用帧数为4帧的基础上，分别进行了一帧一组、两帧一组、三帧一组的实验测试，其中“两帧一组”的策略取得了最优性能（0.357 mAP，0.489 NDS）。实验结果表明“两帧一组”的策略能够在短时局部性和长时感受野之间达到平衡。

表7　时序分组融合模块和不同方法对比

Table 7　Comparison between the Temporal Grouping /Fusion Module and Different Methods

查询传播模块	自注意力模块	时序分组融合模块	mAP	NDS
√	–	–	**0.361**	0.481
–	√	–	0.360	0.488
–	–	√	0.357	**0.489**

注：加粗字体表示各列最优结果。

为了验证“时序分组融合模块”的有效性，本文分别使用StreamPETR查询传播机制与BEVFormer时序自注意力机制进行对比实验。如表7所示，虽然时序分组融合模块在mAP上略低于StreamPETR的查询传播机制，但是在NDS指标上达到了0.489，均优于StreamPETR和BEVFormer。NDS是一个多维度的综合指标，有效地证明了时序分组融合模块在捕捉物体运动状态和精细化定位方面具有更全面的优势。除此之外，如图9所示，根据nuScenes官方提供的可见性标签将nuScenes的验证集划分为四个子集，从0-40%（代表严重遮挡）到80-100%（代表几乎无遮挡）。为了公平地比较召回率，所有方法的预测框均为300。实验结果表明，在0-40%严重遮挡区间，时序分组融合模块达到了56%的召回率，相比于BEVFormer的50%和StreamPETR的54%，分别提升了6个和2个百分点。实验结果验证了时序分组融合模块在严重遮挡情况下的有效性。

本文中，时序分组融合模块的消融实验结果如表8所示。需补充说明的是，为精准验证时序分组融合模块的功能表现，本实验选取涵盖运动目标及遮挡场景的交通图像数据。

基于表8，当骨干网络为ResNet50时，引入时序分组模块后，模型mAP从35.4提升至35.7，NDS从48.7提升至48.9；骨干网络替换为ResNet101时，模型mAP从41.3提升至42.0，NDS从53.5提升至54.1。实验结果表明时序分组融合模块能够进一步增强模型对动态目标和被遮挡物体的检测能力。

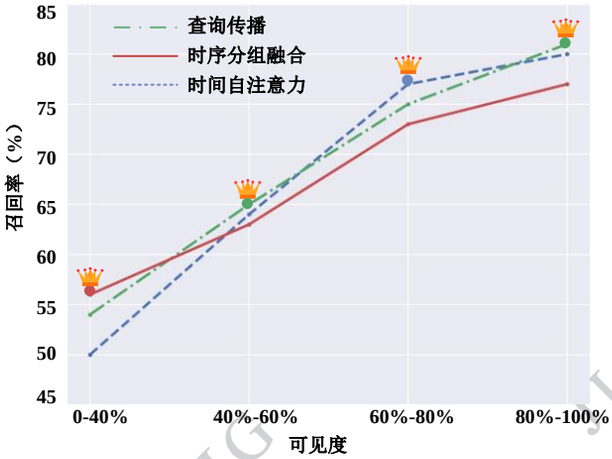


图9　时序分组融合模块对比不同方法的召回率实验

Fig. 9　Experiment on the Recall Rate of the Temporal Grouping Fusion Module Compared with Different Methods

表8　时序分组融合消融结果

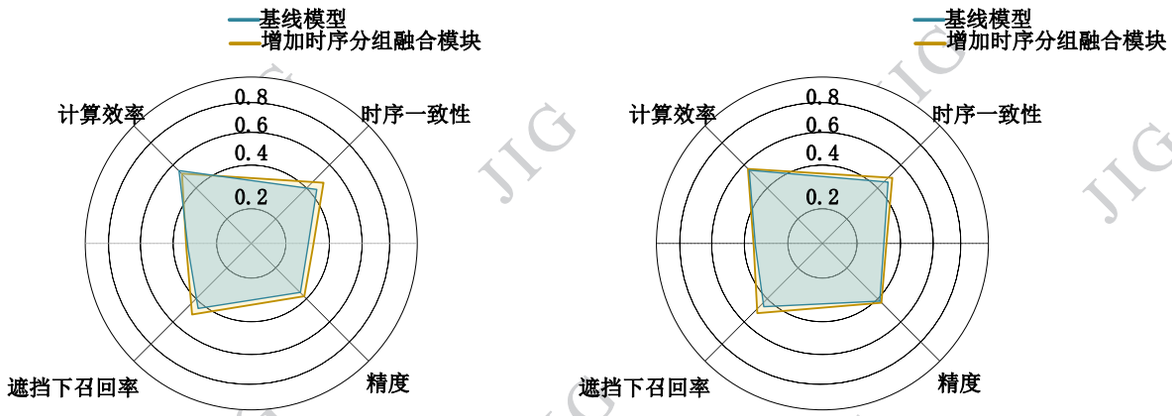
Table 8　Results of Ablation on Temporal Grouped Fusion

简单拼接	时序融合	骨干网络	mAP	NDS
√	—	ResNet50	0.354	0.487
—	√	ResNet50	**0.357**	**0.489**
√	—	ResNet101	0.413	0.535
—	√	ResNet101	**0.420**	**0.541**

注：加粗字体表示各列最优结果，“√”表示采用，“—”表示未采用。

为了更清晰展示时序分组融合模块的有效性，本文增加计算效率、时序一致性、遮挡下召回率3个指标和基线模型的对比实验。实验结果（图10）表明：相较于基线模型的简单拼接，时序分组融合模块在时序一致性、遮挡下召回率和检测精度方面表现更优，说明其在多帧时序特征聚合中具有更强的语义保持与信息融合能力。

进一步，为了验证时序分组融合模块的时效性，本文对比了不同时序融合方法在视图转换部分的延迟，实验结果如图11所示。实验结果表明：在不同BEV特征大小（K1、K2、K3、K4）下，加入时序分组融合模块并不会引入过多延迟。



(a)ResNet50骨干网络 (b)ResNet101骨干网络

((a) ResNet50 backbone network; (b) ResNet101 backbone network)

图10　时序分组融合模块消融实验对比

Fig. 10　Comparison of ablation experiments on the temporal grouping fusion module

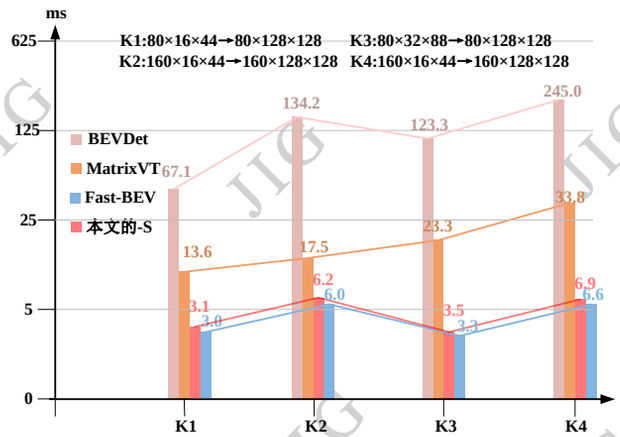


图11　时序分组融合延迟对比

Fig. 11　Comparison of latency for the temporal grouping fusion

3.3.3　可视化

为直观对比本文方法与基线模型的目标检测性能，图12和图13展示了在nuScenes数据集上的可视化结果。其中，图12对比了在多云、阴天、雨天和夜间等典型复杂场景下，本文方法与基线模型的目标检测效果。结果显示，本文方法在复杂天气条件下能够更精准地定位目标，且在遮挡情况下漏检率显著降低；图13从多视角和BEV视角两个层面，展示了本文方法与基线模型在2D图像到3D空间感知的对比结果。实验结果表明，本文方法在BEV检测中，无论是目标定位精度还是空间完整性，均优于基线模型。

4　总　结

本文针对纯视觉3D交通目标检测中时序建模不足和固定采样策略局限性的问题进行了深入研究。为克服现有方法对运动目标与遮挡场景感知能力弱、计算资源分配不灵活等问题，本文创新性地提出了时序分组融合模块和自适应体素特征采样策略。其中，时序分组融合模块实现了跨帧BEV特征的高效聚合，在保持计算开销可控的同时，有效增强了对运动目标和遮挡场景的鲁棒感知能力。自适应体素特征采样策略通过对场景复杂度的判断，能够在稀疏与密集场景间自适应切换采样方法。基于nuScenes数据集的实验结果表明，本文方法在精度和效率上均优于现有主流的3D交通目标检测方法。尽管如此，本文方法仍存在一些局限性：一方面，时序分组融合模块在多帧输入时效果显著，但在仅依赖两帧特征时，其优势并不明显，表明该模块的时序建模能力仍依赖较长的历史信息；另一方面，自适应体素特征采样策略仅基于自车环视图的复杂度进行判断，缺乏对局部区域动态变化的细粒度建模。未来工作将进一步探索先进技术，从而进一步

缓解现有方法的局限性。

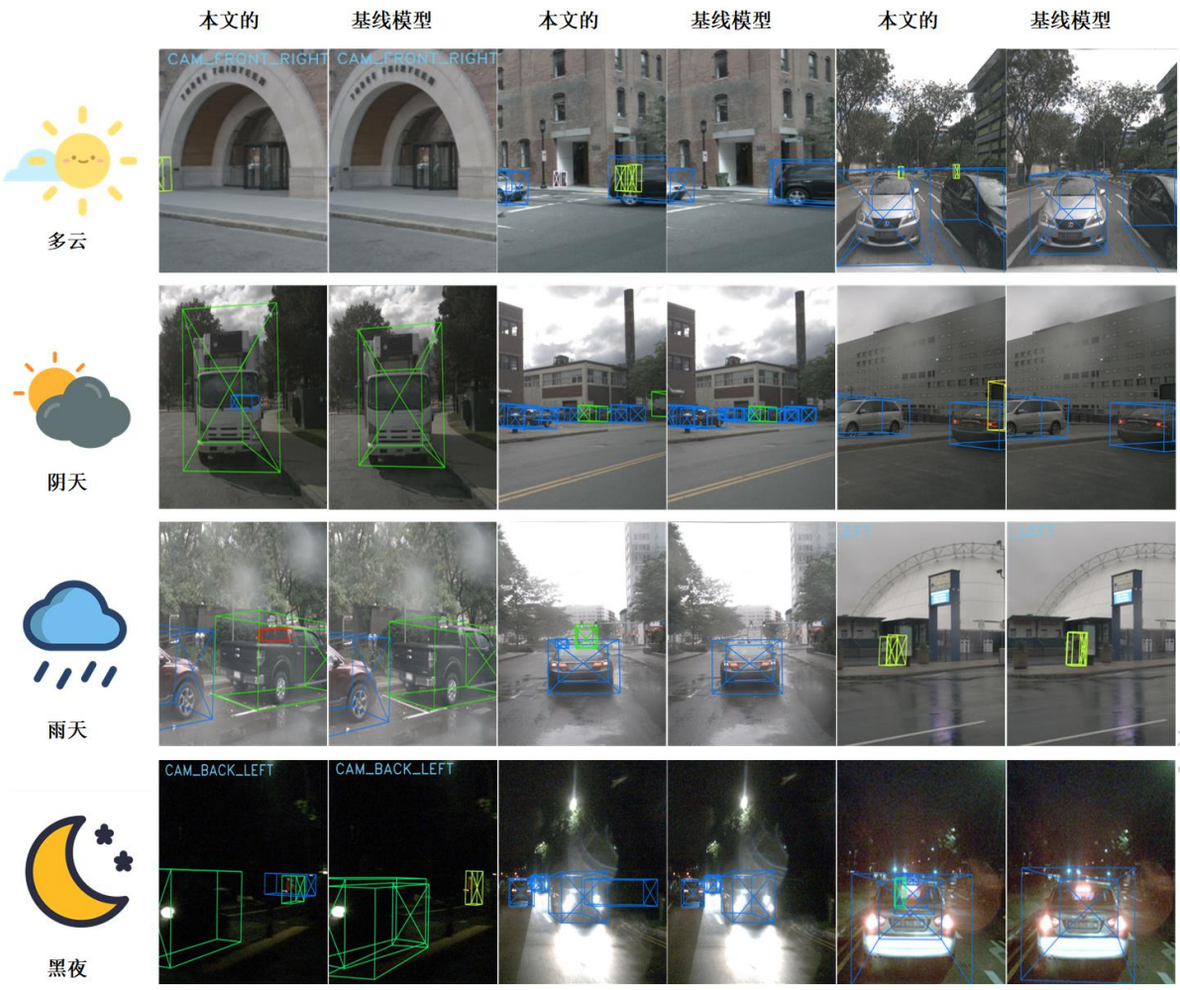


图12　本文方法和基线模型在nuScenes数据集可视化

Fig. 12　The visualization of the method in this paper and the baseline model on the nuScenes dataset

图13　环视图像可视化结果

Fig. 13　Visualization results of panoramic images

参考文献(References)

Baur S A, Moosmann F and Geiger A. 2024. Liso: lidar-only self-supervised 3d object detection//Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland: 253-270 [DOI: 10.1007/978-3-031-73016-0_15]

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. Nuscenes: a multimodal dataset for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 11621-11631 [DOI: 10.48550/arXiv.1903.11027]

Chae Y, Kim H, Oh C, Kim M and Yoon K J. 2024. Lidar-based all-weather 3d object detection via prompting and distilling 4d radar//Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland: 368-385 [DOI: 10.1007/978-3-031-72992-8_21]

Cheng E, Ji H and Liang P. 2024. Enhancing 3D object detection with 2D detection-guided query anchors[EB/OL]. [2025-11-13]. https://doi.org/10.48550/arXiv.2403.06093

Feng Q, Zhao C, Liu P, Zhang Z, Jin Y and Tian W. 2025. LST-BEV: generating a long-term spatial-temporal bird's-eye-view feature for multi-view 3D object detection. Sensors, 25(13): 4040 [DOI: 10.3390/s25134040]

Huang B, Li Y, Liang F, Xie E, Wang L, Shen M, et al. 2023. Fast-BEV: towards real-time on-vehicle bird′s-eye view perception[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2301.07870

Huang J and Huang G. 2022. Bevdet4d: exploit temporal cues in multi-camera 3d object detection[EB/OL]. [2026-01-27]. https://arxiv.org/pdf/2203.17054.pdf

Lang A H, Vora S, Caesar H, Zhou L, Yang J and Beijbom O. 2019. Pointpillars: fast encoders for object detection from point clouds//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE: 12689-12697 [DOI: 10.1109/CVPR.2019.01298]

Leng Z, Yang J, Ren Z and Zhou B. 2025. BEVCon: advancing bird's eye view perception with contrastive learning. IEEE Robotics and Automation Letters, 10(4) [DOI: 10.48550/arXiv.2508.04702]

Li C G, Chen G, Hou Z X, Huang K, Zhang W. 2024a. Survey of 3D object detection algorithms for autonomous driving. Journal of Image and Graphics, 29(11): 3238-3264 (李昌财，陈刚，侯作勋，黄凯，张伟. 2024a. 自动驾驶中的三维目标检测算法研究综述. 中国图象图形学报，29(11)：3238-3264) [DOI: 10.11834/jig.230779]

Li X Y, Ye Z H, Wei S K, Chen Z, Chen X T, Tian Y G, et al. 2023a. 3D object detection for autonomous driving from image: a survey ——benchmarks, constraints and error analysis. Journal of Image and Graphics, 28(06): 1709-1740 (李熙莹，叶芝桧，韦世奎，陈泽，陈小彤，田永鸿，等. 2023a. 基于图像的自动驾驶3D目标检测综述——基准、制约因素和误差分析. 中国图象图形学报，28(06)：1709-1740) [DOI: 10.11834/jig.230036]

Li Y, Ge Z, Yu G, Yang J, Wang Z, Shi Y, et al. 2023b. Bevdepth: acquisition of reliable depth for multi-view 3d object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press: 1477-1485 [DOI: 10.1609/aaai.v37i2.25233]

Li Z, Lan S, Alvarez J M, Li Z, Mo G, Li W, et al. 2024b. Bevnext: reviving dense bev frameworks for 3d object detection//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 20113-20123 [DOI: 10.1109/CVPR52733.2024.01901]

Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. 2022. Bevformer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers//Proceedings of the European Conference on Computer Vision (ECCV). Tel-Aviv, Israel: Springer Nature Switzerland: 22-40 [DOI: 10.48550/arXiv.2203.17270]

Liu H, Teng Y, Lu T, Wang H and Wang L. 2023. Sparsebev: high-performance sparse 3d object detection from multi-camera videos//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 18534-18544 [DOI: 10.1109/ICCV51070.2023.01703]

Liu Y, Wang T, Zhang X and Sun J. 2022. Petr: position embedding transformation for multi-view 3d object detection//Proceedings of the European Conference on Computer Vision (ECCV). Tel-Aviv, Israel: Springer Nature Switzerland: 531-548 [DOI: 10.48550/arXiv.2203.05625]

Liu T, Wang C, Yin Z, Mi Z, Xiong X and Guo B. 2024. Complexity quantification of driving scenarios with dynamic evolution characteristics. Entropy, 26(12): 1033 [DOI: 10.3390/e26121033]

Lu D, Du H, Wu Z and Yang S. 2025. Risk assessment in autonomous driving: a comprehensive survey of risk sources, methodologies, and system architectures. Autonomous Intelligent Systems, 5(1): 24 [DOI: 10.1007/s43684-025-00112-1]

Lo E, Soleilhac E, Martinez A, Lafanechère L and Nadon R. 2012. Intensity quantile estimation and mapping—a novel algorithm for the correction of image non-uniformity bias in HCS data. Bioinformatics, 28 (20): 2632-2639 [DOI: 10.1093/bioinformatics/bts491]

Mao J, Shi S, Wang X and Li H. 2023. 3D object detection for autonomous driving: a comprehensive survey. International Journal of Computer Vision, 131(8): 1909-1963 [DOI: 10.1007/s11263-023-01790-1]

Mei C, He H, Liu Y and Guo Z. 2024. SEGT: a general spatial expansion group transformer for nuScenes lidar-based object detection task[EB/OL]. [2025-11-13]. https://doi.org/10.48550/arXiv.2412.09658

Philion J and Fidler S. 2020. Lift, splat, shoot: encoding images from

arbitrary camera rigs by implicitly unprojecting to 3d//Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, United Kingdom: Springer International Publishing: 194-210 [DOI: 10.48550/arXiv.2008.05711]

Pu F, Wang Y, Deng J and Yang W. 2025. Monodgp: monocular 3D object detection with decoupled-query and geometry-error priors// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE: 6520-6530 [DOI: 10.48550/arXiv.2410.19590]

Park J, Xu C, Yang S, Keutzer K, Kitani K, Tomizuka M, et al. 2022. Time will tell: new outlooks and a baseline for temporal multi-view 3d object detection[EB/OL]. [2026-01-27]. https://arxiv.org/pdf/2210.02443.pdf

Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. 2024a. Yolov10: real-time end-to-end object detection//Advances in Neural Information Processing Systems 37 (NeurIPS 2024. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc.: 107984-108011 [DOI: 10.48550/arXiv.2405.14458]

Wang C Y, Yeh I H and Liao H Y M. 2024b. Yolov9: learning what you want to learn using programmable gradient information//Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland: 1-21 [DOI: 10.1007/978-3-031-72751-1_1]

Wang S, Liu Y, Wang T, Li Y and Zhang X. 2023. Exploring object-centric temporal modeling for efficient multi-view 3d object detection//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE: 3598-3608 [DOI: 10.1109/ICCV51070.2023.00335]

Wang T, Zhu X, Pang J and Lin D. 2021. Fcos3d: fully convolutional one-stage monocular 3d object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE: 913-922 [DOI: 10.1109/ICCVW54120.2021.00107]

Wang Y, Guizilini V C, Zhang T, Wang Y, Zhao H and Solomon J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries//Proceedings of the Conference on Robot Learning. Cambridge, MA, USA: PMLR: 180-191 [DOI: 10.48550/arXiv.2110.06922]

Xi X, Huang Y, Luo R and Qiu Y. 2025. OW-OVD: unified open world and open vocabulary object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE: 25454-25464 [DOI: 10.1109/CVPR52734.2025.02370]

Yang C, Chen Y, Tian H, Yan Z, Liu J, Huang Z, et al. 2023. Bevformer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE: 17830-17839 [DOI: 10.1109/CVPR52729.2023.01710]

Yang H, Zhang S, Huang D, Wu X, Zhu H, He T, et al. 2024. Unipad: a universal pre-training paradigm for autonomous driving// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 15238-15250 [DOI: 10.1109/CVPR52733.2024.01443]

Yin T, Zhou X and Krahenbuhl P. 2021. Center-based 3d object detection and tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Virtual: IEEE: 11784-11793 [DOI: 10.1109/CVPR46437.2021.01161]

You Z, Wang N, Wang H, Zhao Q and Wang J. 2025. MambaBEV: an efficient 3D detection model with Mamba2 [EB/OL]. [2025-11-13]. https://doi.org/10.48550/arXiv.2410.12673

Zhao S, Xia Q, Guo X, Zou P, Zheng M, Wu H, et al. 2025. SP3D: boosting sparsely-supervised 3D object detection via accurate cross-modal semantic prompts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE: 29374-29384 [DOI: 10.1109/CVPR52734.2025.02735]

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. 2024a. Detrs beat yolos on real-time object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA: IEEE: 16965-16974 [DOI: 10.1109/CVPR52733.2024.01605]

Zhou H, Qi H G, Deng Y Q, Li J J, Liang H, Miao J. 2024b. 3D object detection and classification combined with point cloud depth information. Journal of Image and Graphics, 29(08): 2399-2412 (周昊,齐洪钢,邓永强,李娟娟,梁浩,苗军.2024.融合点云深度信息的3D目标检测与分类.中国图象图形学报,29(08):2399-2412) [DOI: 10.11834/jig.230568]

Zhu B, Jiang Z, Zhou X, Li Z and Yu G. 2019. Class-balanced grouping and sampling for point cloud 3d object detection [EB/OL]. [2026-03-07]. https://arxiv.org/abs/1908.09492

Zwart P H, Varga T, Qafoku O and Sethian J A. 2025. Behind the noise: conformal quantile regression reveals emergent representations[EB/OL]. [2026-01-27]. https://arxiv.org/pdf/2505.08176.pdf

作者简介

张兴旺,男,硕士研究生,主要研究方向为3D目标检测。E-mail: 241040063@hdu.edu.cn

王俊帆,女,助理研究员,主要研究方向3D目标检测。E-mail: wangjunfan@hdu.edu.cn

缪其恒,男,博士,主要研究方向为智能交通、目标检测。E-mail: qiheng.miao@hirige.com

董哲康,男,教授,主要研究方向为智能交通、目标检测。E-mail: englishp@hdu.edu.cn

何志伟,男,教授,主要研究方向为目标追踪、目标检测。E-

mail：zwhe@hdu.edu.cn
马国进，男，教授级高工，主要研究方向为3D目标检测。E-mail：magj@hdu.edu.cn