OEE | 光电工程 Opto-Electronic Engineering

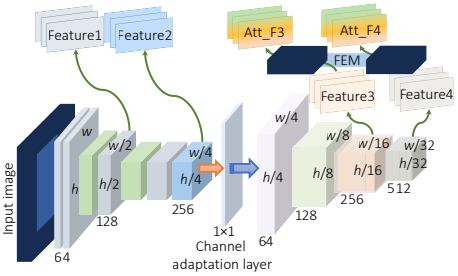
Article

2026 年 6 月, 第 53 卷, 第 6 期

混合编码器融合注意力机制的单目三维车道线检测

郑　奎, 黄影平*, 罗　鑫

上海理工大学光电信息与计算机工程学院, 上海 200093



摘要： 三维车道线检测是自动驾驶领域的关键技术，旨在从二维图像中恢复车道线在三维空间中的几何形态。受三维信息的不确定性与复杂场景下表征歧义的限制，现有基于卷积神经网络的方法常面临特征分辨率与语义聚合之间的权衡：浅层特征具备更强的几何细节保真度，而深层特征语义判别性强但易出现空间量化损失与定位偏移。为此，本文提出一种新型三维车道线检测网络 HA-Lane Net：1) 设计了一种混合编码器，利用 VGG 浅层捕获细粒度纹理，结合 ResNet 深层提取高阶语义，通过通道适配与分级耦合构建均衡的特征层次；2) 在输出层上引入投影前通道注意力机制，对车道相关通道进行自适应重标定，以增强跨视角一致性并抑制背景干扰。在 Apollo-Sim3D 数据集上的实验结果表明，该方法有效缓解了跨尺度与跨视角的不一致性。在不引入 Transformer 等复杂结构的前提下，HA-Lane Net 的性能指标相较于 3DLaneNet 与 Gen-Lane Net 均有显著提升。

关键词： 三维车道线检测；混合编码器；注意力机制；自动驾驶

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250359 | CSTR: 32245.14.oee.2026.250359 | 中图分类号：TP391.4 | 文献标识码：A

郑奎, 黄影平, 罗鑫. 混合编码器融合注意力机制的单目三维车道线检测 [J]. 光电工程, 2026, **53**(6): 250359

Zheng K, Huang Y P, Luo X. HA-Lane Net: Hybrid feature encoding and attention for monocular 3D lane detection[J]. *Opto-Electron Eng*, 2026, **53**(6): 250359

1　引　言

车道线检测是无人驾驶车辆实现路径规划与横向控制的核心基础。相较于仅能提供图像平面坐标的传统二维检测，三维车道线检测致力于重建车道线的三维几何属性，提供对道路环境更精确的理解，是无人驾驶技术深入发展的必然需求[1]。

当前，基于单目视觉的三维车道线检测已成为主流研究方向，其中基于鸟瞰图 (bird's-eye-view, BEV) 为代表的方法表现尤为突出。此类方法多遵循“图像特征提取-空间视角变换-结构化预测”的范式：利用卷积神经网络提取前视 (front view，FV) 特征，通过投影变换将其映射至 BEV 空间以消除透视效应，从而将三维检测简化为 BEV 平面上的二维检测与高度回归任务。尽管该范式在理论上具有可行性，但在实际应用中仍面临以下瓶颈：1) 特征编码器表征能力受限：现有基线模型多采用通用卷积网络，在面对光照剧变、遮挡及恶劣天气时，难以提取具有高判别性的深层语义特征。2) 通道信息冗余与干扰：标准卷积缺乏对特征通道间依赖关系的显式建模，导致关键的车道线语义信息易被背景噪声淹没，降低了特征的信噪比。3) 视角变换的误差级联效应：空间映射对输入特征质量极度敏感。前视图像中的局部模糊或特征缺失经几何变换后会被显著放大，直接削弱了 BEV 视角下车道线空间定位与几何拟合的精度。

收稿日期：2025-11-21；修回日期：2026-03-13；录用日期：2026-03-20；网络出版日期：2026-06-25

基金项目：国家自然科学基金面上项目：智能汽车环境感知多模态数据融合方法研究 (62276167)

*通信作者：黄影平，huangyingping@usst.edu.cn

版权所有©2026 中国科学院光电技术研究所

针对上述挑战，本文的主要贡献如下：

1) 设计了一种混合编码器：协同融合 VGG 网络的浅层细粒度纹理特征与 ResNet 网络的深层高阶语义，并引入余弦退火策略优化训练，构建了兼顾局部细节与全局语义的均衡特征表示。

2) 引入注意力机制优化：在编码器深层嵌入 Squeeze-and-Excitation (SE) 模块，通过“压缩-激励-重标定”机制显式建模通道依赖关系，有效增强车道线相关特征响应并抑制了背景噪声。

3) 在 Apollo-Sim3D 数据集上的实验证明，本文方法在复杂挑战场景下的车道线坐标检测回归精度及稳定性均取得显著提升。

2 相关工作

实际道路场景具有高度动态性，受光照变化、遮挡及恶劣天气等因素影响，传统二维车道线检测方法仅能在图像平面提供像素级位置信息，缺乏真实的三维几何信息，导致其在坡度路面等复杂场景中性能不佳。为满足自动驾驶对高精度三维感知的需求，主要借助单目视觉、3D 激光雷达及多传感器融合获取所需的三维信息[1]。

2.1 基于单目视觉的方法

单目三维车道线检测旨在从单帧二维图像中重构缺失的深度信息。根据空间表征方式的不同，现有算法主要可分为基于 BEV 和不依赖 BEV 两类[1]。

2.1.1 基于 BEV 的方法

此类方法通过相机内外参将前视特征重投影至俯视图平面，使非线性的三维几何结构在 BEV 空间内呈现为规则的二维布局[2]。3DLaneNet[3] 作为开创性工作，首次构建图像视角与 BEV 视角的双分支网络并引入逆透视映射实现特征域转换，以锚点形式输出连续三维车道线。随后，Gen-Lane Net[4] 通过解耦图像语义分割与几何推理，并以几何引导锚点替代稠密标签，降低了对三维标注的依赖。为进一步提升复杂形状的表达能力，3D-LaneNet+[5] 引入半局部无锚点表示，而 PersFormer[6] 将 Transformer 融入特征变换，利用相机参数与前视图局部区域生成表达力更强的 BEV 特征。此外，BEV-LaneDet[7] 则提出虚拟相机机制以有效规避几何畸变，CLGO[8] 则通过几何先验约束估计相机姿态，确保投影的一致性。Chen 等[9] 提出了一种分解式交叉注意力机制，通过统一建模 BEV 表示与车道线表示，实现二者同步学习，从而兼顾精度与计算效率。针对动态场景，Bi^{2}Lane[10] 利用双向时序优化进一步缓解了遮挡导致的信息缺失。

2.1.2 不依赖 BEV 的方法

相比于依赖显式 BEV 表征的方法，BEV-free 方案侧重前视特征与三维空间之间的直接交互。SALAD[11] 证明了在无需相机参数的前提下，借助于伪深度图可实现从二维平面到三维空间的坐标恢复。PVALane[12] 则利用先验引导生成稀疏锚点进行特征对齐。另一类趋势是直接在三维空间建模，如 CurveFormer[13] 将车道线表示为有序三维锚点构成的曲线查询，通过 Transformer 迭代优化预测结果。随后，Anchor3DLane[14] 提出以三维锚点 (按俯仰/偏航定义的射线) 作为结构化表示，并将其投影到前视图上以进行分类与回归，在此基础上 LATR[15] 进一步将查询机制细化为“点级”与“车道级”，并引入三维动态地面位置嵌入与跨注意力结构融合多尺度语义与空间定位信息。

2.2 基于 3D 激光雷达及多模态融合的方法

三维激光雷达 (LiDAR) 通过测量脉冲飞行时间获取高精度点云，不受光照变化影响，具有极佳的几何精度与鲁棒性[16]。Huang 等[17] 通过利用 RANSAC 算法结合点密度分类可实现结构化道路边界的提取。然而，LiDAR 设备成本高昂且对路面反射特性敏感。

为综合视觉的语义优势与雷达的几何优势，研究者们提出了多模态融合策略。M^2-3DLaneNet[18] 在 BEV 空间内融合图像特征与雷达深度信息，提升了远距离定位精度。LiLaDet[19] 将点云几何特征嵌入 BEV 表征实现高效融合，DV-3DLane[20] 则通过双视图融合和三维可变形注意力机制，进一步增强了模型在动态和复杂场景下的鲁棒性。

总体来看，LiDAR 可提供高精度三维信息，多模态融合方法又兼顾几何与语义，但受成本与计算复杂度限制，单目方法仍是目前研究主流。随着硬件性能提升和 LiDAR 成本下降，多模态融合方法有望在未来获得更广泛地应用。

3 本文方法

3.1 几何变换与锚点表示

基于锚点的车道线检测方法的核心思想在于：在图

像感兴趣区域内预定义一组具有固定几何含义的锚点，将连续的车道线结构建模问题转化为对离散锚点属性的联合回归与分类问题，从而显著降低建模复杂度[21]。该思想已在 3D-LaneNet[3] 中得到验证，但其直接在图像平面进行三维回归易受视角变化与透视畸变影响，导致几何稳定性不足。为此，本文采用 Gen-Lane Net[4] 提出的几何引导锚点表示，通过显式几何建模在虚拟俯视视角 (BEV) 中构建锚点坐标系，以削弱视角变化对空间推理的干扰。如图 1 所示，定义相机中心 C 及车辆坐标系 (x, y, z)。其中相机高度为 h，俯仰角为 θ，并以地面为参考平面建立相机与车辆坐标之间的几何关系。

如图 2 左所示，在该几何约束下，相机中心 C、三维车道点 P 在虚拟俯视图中的投影点 P' 满足共线关系。由相似三角形原理，可推导出从 BEV 平面到车辆坐标系的显式几何变换关系：$x = x_v \cdot (1 - z/h)$；$y = y_v \cdot (1 - z/h)$。该变换表明，在已知 BEV 平面位置与高度信息的条件下可恢复唯一对应的三维空间坐标。

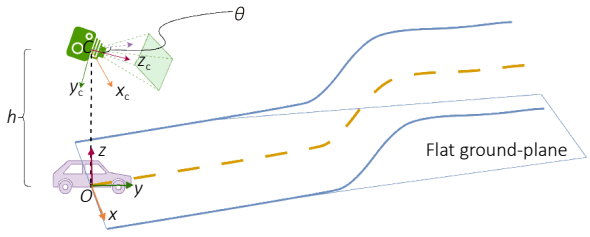


图 1 相机设置与车辆坐标系示意图

Fig. 1 Schematic diagram of camera setup and vehicle coordinate system

基于此，本文在 BEV 中空间中沿纵向预定义一组等间距锚线 $y_1, y_2, \cdots, y_K$。如图 2 右所示，在每条锚线上设置锚点，网络预测每个锚点对应的几何属性 $\{(x_k, z_k, v_k)\}_{k=1}^{K}$。其中 x_k 表示锚点的水平偏移量；z_k 为该位置的真实路面高度；v_k 为可见性指示变量。最终，一条完整的三维车道线可表示为 $\{(x_k, z_k, v_k)\}_{k=1}^{K}$ 及其存在概率 p，即 $Anchor_{\dim} = 3K + 1$。该锚点表示法通过显式引入高度与可见性约束，有效缓解了车道线的起伏和遮挡条件下的三维回归不稳定问题，并为后续端到端学习提供了结构化、物理一致的几何先验。

3.2 网络架构

本文提出的 HA-LaneNet (hybrid-encoder with pre-IPM channel attention) 总体架构如图 3 所示，主要由混合编码器、基于 IPM 的几何映射模块、BEV 特征融合模块以及车道线回归头构成。输入图像首先经混合编码器提取兼具多尺度语义与几何细节的特征表示，并通过通道注意力机制增强特征；随后通过几何变换将图像特征映射至 BEV 空间，实现跨视角特征对齐；在 BEV 空间内，多尺度特征进一步融合以增强结构一致性，最终由回归分支输出三维车道线表示。该端到端架构在保持几何约束一致性的同时，有效提升了复杂场景下三维车道线检测的稳定性与精度。

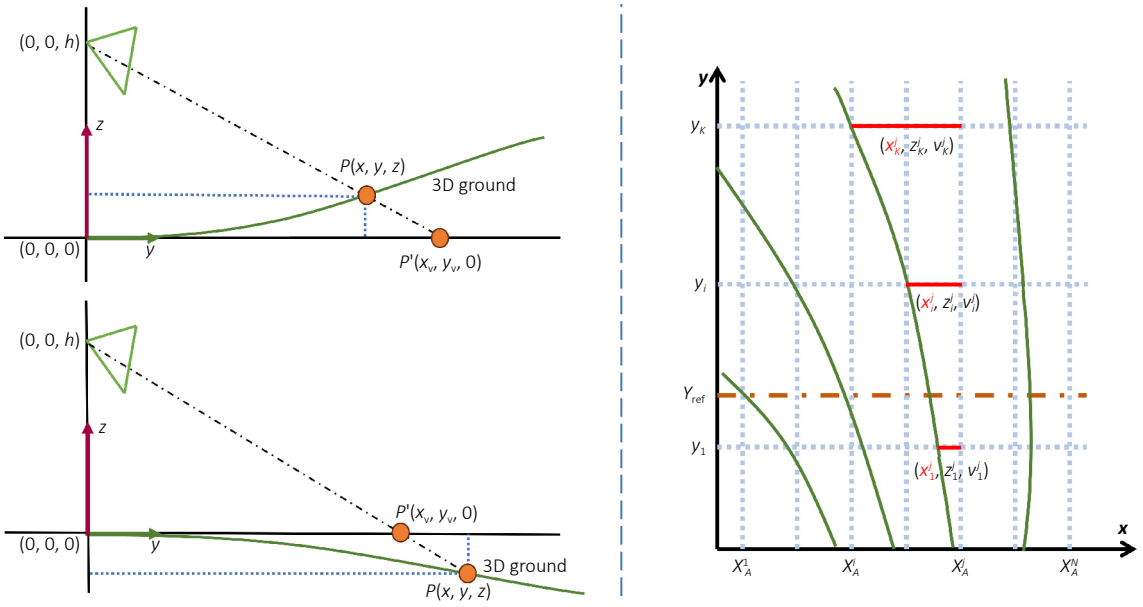


图 2 几何变换与锚点表示图

Fig. 2 Geometric transformation and anchor point representation diagram

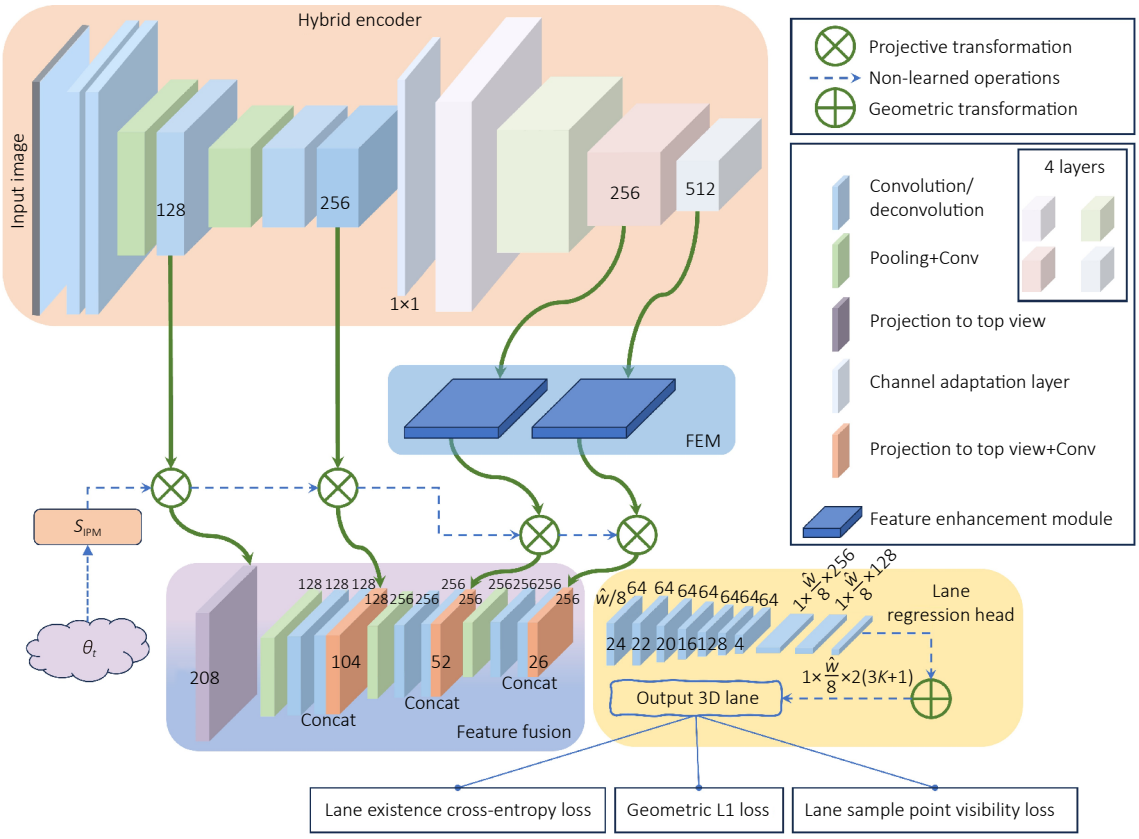


图 3 网络整体结构示意图

Fig. 3 Schematic diagram of the overall network structure

3.2.1 混合编码器

针对三维车道线检测任务对精细几何建模与高层语义理解的双重需求，本文设计了一种融合浅层与深层特征的混合编码器 (hybrid encoder)，以实现几何与语义信息的协同表达。如图 4 所示，浅层分支采用 VGG16[22] 的前 7 层结构，通过连续3×3卷积提取高分辨率的局部纹理与边缘信息，并在 1/4 与 1/8 的空间尺度处输出特征以保留细粒度几何结构。深层分支基于 ResNet34[23]，仅保留 layer3 与 layer4 的输出以增强高层语义表征能力。为实现跨层特征对齐，本文在融合路径中引入1×1卷积进行通道映射，并结合 Network-in-Network[24] (NIN) 结构完成通道维度重整与特征压缩。该混合编码策略在保证空间分辨率的同时增强了语义判别性，为后续 IPM 投影与 BEV 特征构建提供了一致且可投影的特征基础。

3.2.2 采用注意力机制的特征增强

为进一步提升编码特征的表达质量，本文在混合编码器的高层特征输出处引入由 Hu 等[25] 提出的基于通道注意力 SE 模块，对语义特征进行自适应重标定。

如图 5 所示，将 SE 模块嵌入 ResNet 的 layer3、layer4 输出之后。针对输入特征图 X，经卷积变换 $F_{\mathrm{tr}}(\cdot,\theta)$得到初始特征$U$。随后，通过压缩操作$F_{\mathrm{sq}}(\cdot)$ (即全局平均池化) 对U的每个通道进行空间维度压缩，获得$1\times1\times C$的通道级统计描述；接着进行激励操作 $F_{\mathrm{ex}}(\cdot,W)$，利用包含参数W且降维率$r=16$的两层全连接网络建立通道依赖，并通过 Sigmoid 函数生成通道权重；最后，通过尺度重标定操作$F_{\mathrm{scale}}(\cdot)$，利用该权重对原特征$U$逐通道重加权，得到增强特征$\tilde{X}$。该机制能够显式强化与车道线结构相关的判别性响应，同时抑制冗余背景，在不显著增加计算开销的前提下增强网络的特征选择能力。

3.2.3 损失函数设计

为同时监督车道线的存在性、几何结构与可见性，本文采用多任务联合优化策略构建端到端损失函数。设模型对第 i 条车道的预测结果为$\hat{A}_i=\{\hat{p}_i,\hat{v}_i,\hat{x}_i,\hat{z}_i\}$，对应的真值标签为$\mathcal{A}_i^{\mathrm{gt}}=\{p_i,v_i,x_i,z_i\}$。其中$p_i\in0,1$表示车道线存在标签，$v_i$为逐点的可见性掩膜。总损失函数定义为分

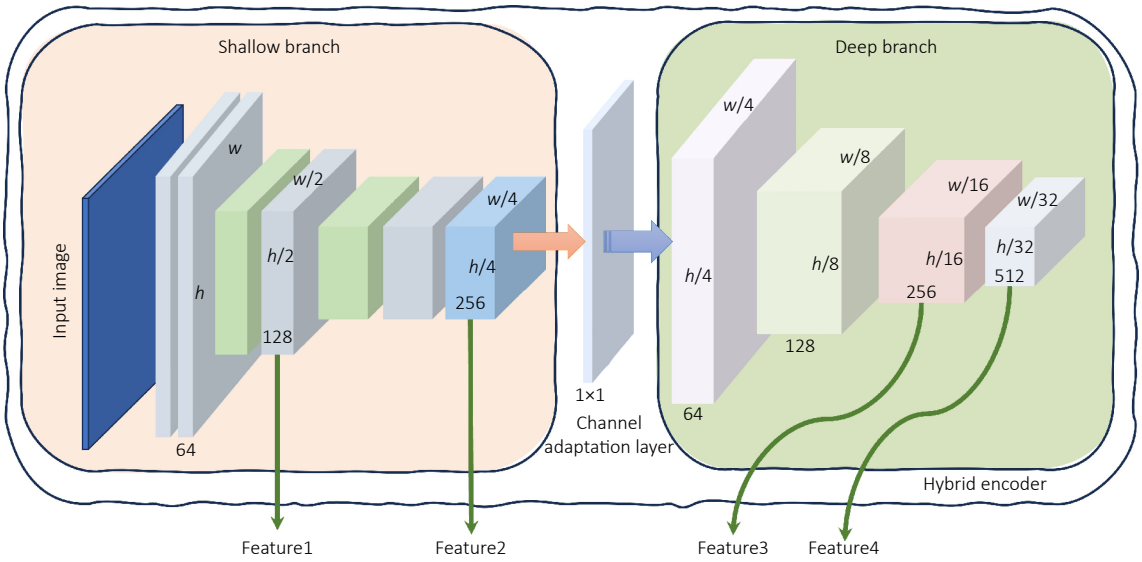


图 4 混合编码器结构示意图

Fig. 4 Schematic diagram of the hybrid encoder structure

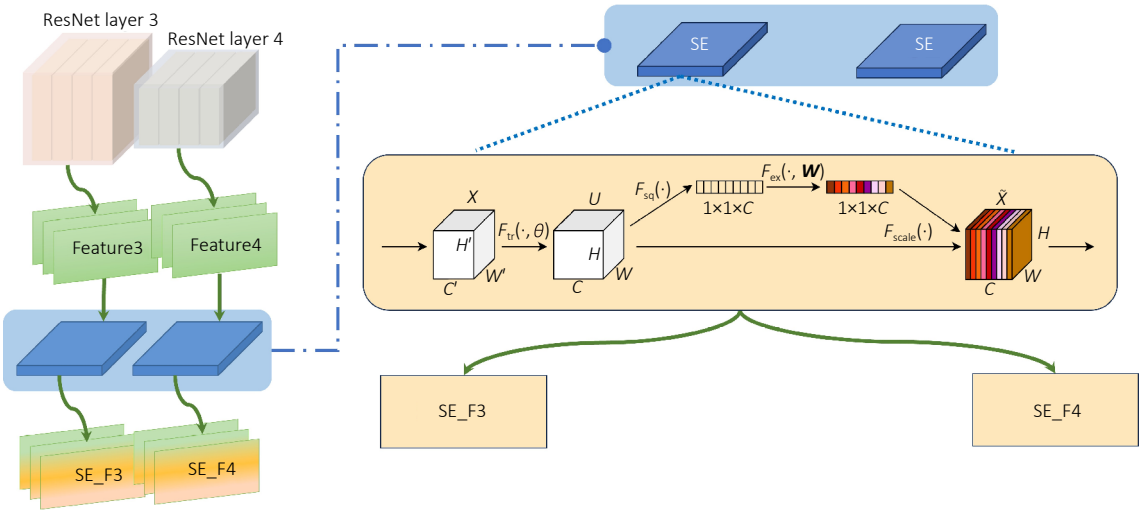


图 5 注意力模块添加的位置与结构示意图

Fig. 5 Schematic diagram of the location and structure of the attention module

类损失L_{cls}、几何回归损失L_{geo}与可见性损失L_{vis}的加权求和。

$$L=-\sum_i \underbrace{\ell_{\text{bce}}(\hat{p}_i,p_i)}_{L_{\text{cls}}}+\sum_i p_i \cdot \left(\underbrace{|v_i\cdot(\hat{x}_i-x_i)|_1+|v_i\cdot(\hat{z}_i-z_i)|_1}_{L_{\text{geo}}}+\underbrace{|\hat{v}_i-v_i|_1}_{L_{\text{vis}}} \right), \tag{1}$$

其中：L_{cls}采用二元交叉熵监督车道线存在性；L_{geo}基于L1范数对车道线在BEV坐标系下的水平偏移x与高度z进行回归，并显式引入v_i以避免不可见点对几何建模的干扰；L_{vis}则用于约束模型对车道线采样点可见性的预测准确性。该联合损失设计在保证结构连续性的同时提升了模型对遮挡与远距离车道线的感知能力。

4 实验与结果分析

4.1 数据集与评估指标

为验证本文算法的有效性，实验采用与Gen-LaneNet[4]一致的Apollo-Sim3D仿真数据集。该数据集包含10500帧高分辨率(1920 pixel×1080 pixel) RGB图像及对应的三维车道线标注。数据集涵盖了标准场景、罕见

场景 (如急转弯、陡坡) 以及视觉变化场景 (如极端光照、雨雾天) 三大类别，各场景样本分布如表 1 所示。本文采用F_1分数及平均精度 (average precision，AP) 作为核心性能指标。F_1定义为：

$$F_1 = \frac{2\times P\times R}{P+R}, \tag{2}$$

$$P = \frac{TP}{TP+FP}, \tag{3}$$

$$R = \frac{TP}{TP+FN}, \tag{4}$$

其中：真正例 (TP) 的判定遵循 Gen-Lane Net 设定三维匹配流程：对预测值与真值 (GT) 进行等间距采样，计算对应采样点间的三维欧氏距离。当采样点中满足距离阈值$d_{\max} = 1.5$ m的比例不低于 75% 时，该预测车道线被判定为TP。通过对 P-R 曲线进行积分计算得到AP：

$$AP = \sum_{k=1}^{N} P(k)\cdot \Delta R(k), \tag{5}$$

式中：$P(k)$为第 k 个预测的精确率；$\Delta R(k)$为召回率增量；N为预测结果数量。

表 1 合成数据集分布
Table 1 Distribution of synthetic datasets

Scene	Training set size	Test set size
Balanced scene	5992	1496
Rare subset	5992	320
Visual variations	3968	472

此外，为进一步评估几何回归的精度，本文分别统计了近距离 ($0 \leqslant y \leqslant 40$ m) 与远距离 ($40\text{ m} \leqslant y \leqslant 100$ m) 范围内的横向误差E_X与纵向误差E_Z。

4.2 实验参数设置及模型训练过程

本实验模型基于 PyTorch 框架开发，硬件环境为 NVIDIA RTX 4090D GPU (24 GB)，操作系统为 Ubuntu 22.04，开发环境为 cpython 3.12，编辑器为 Neovim。在数据预处理阶段，将输入图像统一调整为 360 pixel×480 pixel。模型在 BEV 空间构建了尺寸为 208 pixel×108 pixel的空间网格，其对应的三维物理范围设为：x 轴：$[-10,10]$ m；y 轴：(3, 5, 10, 15, 20, 30, 40, 50, 65, 80, 100) m；z 轴：$[-1,1]$ m。为降低远景区域的透视畸变影响，设置前向投影参考高度$Y_{\text{ref}} = 5$ m。网络权重采用正态分布随机初始化，使用 Adam 优化器进行端到端训练。初始学习率为2×10^{-4}，并引入余弦退火的策略进行动态调整，公式如下：

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})\left[1+\cos\left(\frac{\pi t}{\tau_{\max}}\right)\right], \tag{6}$$

其中：η_t为当前学习率；t 为当前 epoch，总训练周期$\tau_{\max} = 60$。训练过程中批大小设为 8，最终学习率衰减至1×10^{-6}。

4.3 定量对比分析与计算效率评估

4.3.1 消融实验

为了系统验证混合编码器、注意力机制以及学习策略对模型性能的贡献，本文基于视觉变化场景构建了消融实验方案。实验以几何锚点结构作为基线，通过控制变量法对各核心组件进行组合测试。评估指标涵盖的AP与F_1分数及坐标预测误差，AP与F_1用于衡量检测精度，值越大说明性能越好；而横向误差E_X与纵向误差E_Z则数值越低表示预测越精确。核心组件消融结果如表 2 所示。

表 2 视觉变化场景消融实验结果
Table 2 Ablation experiment results of visual change scenes

Geometric anchor	Hybrid encoder	Attention mechanism	Learning strategy	AP/% ↑	F_1/% ↑
√				84.0	81.9
√	√			86.5	84.4
√		√		86.8	84.7
√	√	√		86.9	84.9
√		√	√	89.8	87.9
√	√		√	90.4	88.5
√	√	√	√	91.5	89.5

基线网络的AP为 84%，通过对比发现，单独引入混合编码器或注意力机制时，模型性能虽有小幅提升，但均未达到显著的增益效果，这表明架构调整需配合特定的学习策略方能充分释放其表征潜力。当引入相应的学习策略后，注意力机制的AP从 86.8% 跃升至 89.8%，体现了训练策略对结构效能的关键驱动作用。进一步的组合测试显示，同时集成混合编码器与注意力机制并辅以优化策略时，模型达到了 91.5% 的最高的AP与 89.5% 的F_1分数。这一结果有力地验证了混合编码器提供的层次化语义特征与注意力机制对关键信息的强化表达具有显著的互补性，两者的耦合能够系统性地增强模型在复杂视觉环境下的检测精度。

降维率 r 是通道注意力机制中平衡计算开销与特征表达能力的核心参数，本文对此进行了详细的参数敏感性实验，如表 3 所示。在基于基线网络采用的 VGG 模

型中，随着 r 的增大，模型性能呈现出先升后降的趋势：当$r=8$时取得阶段性最优表现 (AP为 87.7%)，而当 r 进一步增至 32 时，AP出现剧烈下滑，这归因于过高的压缩率导致了关键语义信息的丢失。相比之下，引入混合编码器的模型表现出更强的稳健性与更高的精度上限。实验数据表明，当$r=16$时，HA-Lane Net 模型在保持高精度的同时，横向预测误差E_X/F降至最低值 0.0610，实现了检测效率与精细化表征的最佳平衡。综上所述，降维率的选择需权衡冗余信息抑制与特征完整性保留，本文最终选取$r=16$作为 HA-Lane Net 架构的最优降维参数，以最大化提升模型在复杂场景下的坐标预测准确性。

4.3.2 检测性能对比分析

在合成数据集上完成模型训练后，对模型进行了全面测试，并与本文主要参考方法进行对比分析。实验采用合成数据集常用的多项评价指标，包括AP、F_1等，结果如表 4 与表 5 所示。其中，“L”表示本文采用的新学习策略，“T”则表示将混合编码器中的 ResNet34 替换为 ResNet18。AP与F_1用于衡量检测精度，值越大说明性能越好；而横向误差E_X与纵向误差E_Z则数值越低表示预测越精确。由于合成数据集包含三类不同场景，本文主要以视觉变化最大的场景表 5 作为评估基准。

表 3　视觉变化场景注意力机制降维率的敏感性分析表

Table 3　Sensitivity analysis of the dimensionality reduction ratio of the attention mechanism under visual variation scenarios

Reduction ratio/r	Methods	AP/%↑	F_1/%↑	E_x/F (m)↓	E_x/N (m)↓	E_z/F (m)↓	E_z/N (m)↓
4	VGG+SE	86.9	84.9	0.0913	0.601	0.0192	0.293
8	VGG+SE	**87.7**	**85.8**	0.1110	0.581	0.0226	0.306
16	VGG+SE	86.8	84.7	0.1100	0.561	0.0211	0.305
32	VGG+SE	81.1	79.3	0.1200	0.585	0.0244	0.271
8	Hybrid+SE	88.8	86.7	0.0688	0.470	0.0134	0.273
16	Hybrid+SE	**91.5**	**89.4**	0.0610	0.451	0.0140	0.276
16-8	Hybrid+SE	89.5	87.4	0.0748	0.484	0.0154	0.296

表 4　在 Apollo-Sim3D 数据集的标准与罕见场景上与其他方法的比较

Table 4　Comparison with other methods on standard and rare scenarios of the Apollo-Sim3D dataset

Scene	Methods	AP/%↑	F_1/%↑	E_x/F (m)↓	E_x/N (m)↓	E_z/F (m)↓	E_z/N (m)↓
Balanced scene	3DLaneNet[3]	89.3	86.4	0.0680	0.477	0.0150	0.202
	3DLaneNet_ext (baseline)	91.7	89.7	0.0669	0.453	0.0133	0.241
	Gen-LaneNet[4]	89.2	87.3	0.0636	0.534	0.0127	0.247
	LS-3DLane[1]	89.6	89.4	0.0834	0.480	0.0217	0.271
	3DLaneNet (DA)[6]	93.2	91.0	0.0820	0.439	0.0110	0.242
	GenLaneNet (DA)[6]	92.4	90.3	0.0800	0.473	0.0110	0.247
	Reconstruct from Top[26]	93.8	91.9	0.0490	0.387	0.0080	0.213
	CLGo[10]	94.2	91.9	0.0610	0.361	0.0290	0.250
	Ours (hybrid+se_L)	93.8	91.8	0.0440	0.383	0.0080	0.230
Rare subset	3DLaneNet[3]	74.6	72.0	0.1660	0.855	0.0390	0.521
	3DLaneNet_ext (baseline)	82.4	81.2	0.1550	0.882	0.0316	0.622
	Gen-LaneNet[4]	77.9	76.9	0.1470	0.977	0.3170	0.633
	LS-3DLane[1]	79.9	78.1	0.1730	0.952	0.0469	0.705
	3DLaneNet (DA)[6]	85.8	84.1	0.2890	0.925	0.0250	0.625
	GenLaneNet (DA)[6]	83.2	81.7	0.2830	0.915	0.0280	0.653
	Reconstruct from Top[26]	85.2	83.7	0.1260	0.903	0.0230	0.625
	CLGo[10]	88.3	86.1	0.1470	0.735	0.0710	0.609
	Ours (hybrid+se_L)	86.6	84.7	0.1090	0.828	0.0219	0.616

表 5 在 Apollo-Sim3D 数据集的视觉变化场景上与其他方法的比较

Table 5 Comparison with other methods on visual variation scenarios of the Apollo-Sim3D dataset

Methods	*AP*/%↑	F_1/%↑	E_x/F (m)↓	E_x/N (m)↓	E_z/F (m)↓	E_z/N (m)↓
3DLaneNet[3]	74.9	72.5	0.1150	0.601	0.0320	0.230
3DLaneNet_ext (baseline)	84.0	81.9	0.1070	0.583	0.0249	0.328
Gen-LaneNet[4]	87.2	85.3	0.0740	0.538	0.0150	0.232
LS-3DLane[1]	89.0	87.3	0.0848	0.529	0.0321	0.330
3DLaneNet (DA)[6]	87.4	85.4	0.1180	0.559	0.0180	0.290
GenLaneNet (DA)[6]	88.5	86.8	0.1040	0.544	0.0160	0.294
Reconstruct from Top[26]	92.1	89.9	0.0600	0.446	0.0110	0.235
CLGo[10]	89.2	87.3	0.0840	0.464	0.0450	0.312
Ours (ResNet34)	84.5	82.5	0.1070	0.573	0.0250	0.331
Ours (hybrid)	86.5	84.4	0.1160	0.604	0.0178	0.311
Ours (vgg+se)	86.6	84.7	0.1090	0.561	0.0210	0.305
Ours (hybrid+se)	86.9	84.9	0.0896	0.564	0.0187	0.288
Ours (vgg+se_L)	89.8	87.9	0.0575	0.462	0.0135	0.271
Ours (hybrid_L)	90.4	88.5	0.0634	0.457	0.0133	0.258
Ours (hybrid+se_L_T)	91.2	89.2	0.0700	0.473	0.0140	0.263
Ours (hybrid+se_L)	91.5	89.4	0.0610	0.451	0.0140	0.276

从多组对比实验结果可见，本文方法较 3DLaneNet[3] 与 Gen-Lane Net[4] 均表现出显著提升。然而，与部分当前先进方法 (如 CLGo[10]) 相比，整体性能相近，但在某些特定场景下略有不足。出现这一差异的主要原因在于，本文采用的逆透视映射依赖地面平坦性假设，而实际环境中地面往往存在起伏或非理想几何形态，可能导致特征投影精度下降，进而影响最终检测效果。

从表 4 可以看出，在标准场景与罕见场景中，本文方法相较于基线模型均取得了约 2%~4% 的性能提升，且在罕见场景中表现更为突出。在表 5 的视觉变化场景对比中，模型性能提升可达约 7.5%，表明本文方法在复杂视觉条件下具有更强的鲁棒性。

4.3.3 算法复杂度与实时性分析

为验证 HA-Lane Net 在车载嵌入式平台上的工程化部署潜力，本节对模型的计算效率进行了定量评估。评估指标包括推理速度 (frames per second, FPS)、参数量 (parameters, Params) 以及浮点运算量 (floating point operations per second, FLOPs) 与乘加运算量 (multiplication and accumulation operations, MACs)。实验在 NVIDIA RTX 4090D 平台上进行，对比对象包括基线模型以及目前先进的基于 Transformer 的模型 LATR[15]。如表 6 所示，本方法在 Apollo-Sim3D 数据集的三个场景下的推理速度上表现优异，平均达到 111.62 f/s，远高于 LATR 的 35.53 f/s，完全满足自动驾驶系统实时感知的需求 (> 30 f/s)。

在计算开销方面，如表 7 所示，虽然由于混合编码器的引入使得参数量 (28.88 M) 略高于基线模型 (20.60 M)，但其单帧 MACs 从基线的 60.523 G 优化至 52.308 G，降低了约 13.6%。实验结果表明，本方法通过混合编码架构优化了特征提取流程，在不牺牲运行速度的前提下，实现了更高效的计算开销，在保持高帧率的同时优化了单位推理成本。

表 6 推理速度对比表

Table 6 Speed comparison of different methods

Scene	LATR[15]	Baseline	HA-Lane Net
Balanced scene	34.8550	156.670	**176.580**
Rare subset	35.2967	59.620	**64.530**
Visual variations	36.4233	89.265	**93.735**

表 7 复杂度对比表

Table 7 Computational complexity comparison

Methods	Params/M	GFLOPs	GMACs
Baseline	20.60	121.046	60.523
HA-Lane Net	28.88	104.616	52.308

4.4 典型场景下可视化结果分析

4.4.1 注意力机制特征图可视化分析

为直观验证 SE 模块对关键特征的增强作用，本节

利用通道维L_2能量投影，对网络特征图进行了可视化处理。记模块前后的特征图为：$F \in \mathbb{R}^{C\times H\times W}$。其中，$C$ 表示通道数，H,W表示空间分辨率。令$F_c(x,y)$表示第 c 个通道在空间位置(x,y)处的激活响应。采用如下公式对通道维做平方和，得到每个空间位置的总体响应强度。

$$H(x,y)=\sum_{c=1}^{C}F_c(x,y)^2\,. \tag{7}$$

如图 6 所示，图中上半部分为视觉变化场景，从左至右依次为：基线模型与添加 SE 后的能量图，Top-5% 注意力迁移图以及叠加在原图上的示意图。在迁移图与叠加图中，红色表示引入 SE 后新增的关注区域，蓝色表示基线原有的关注区域，紫色则代表两者的共同关注区域。在基线模型中，由于缺乏通道间的依赖建模，高响应区域往往弥散在路面阴影、路缘等非目标区域。而引入 SE 注意力机制后，特征能量显著向车道线物理结构汇聚，背景噪声干扰得到了有效抑制。通过右侧的响应迁移图，可以清晰观察到，红色区域精准覆盖了远端模糊的车道走势。这表明注意力机制通过自适应重标定通道权重，在 BEV 空间映射前实现了特征的“提纯”，为后续的三维坐标回归奠定了可靠的特征基础。而图 6 下半部分展示了在罕见场景中基线模型与 HA-Lane Net 的特征可视化对比。由于不同网络结构的高层特征分辨率存在差异，本文通过特征分辨率匹配算法，选取面积与长宽比最接近的层与混合编码器的高层特征进行公平对齐。实验结果进一步验证了通道注意力机制在复杂场景下的稳健性与有效性。

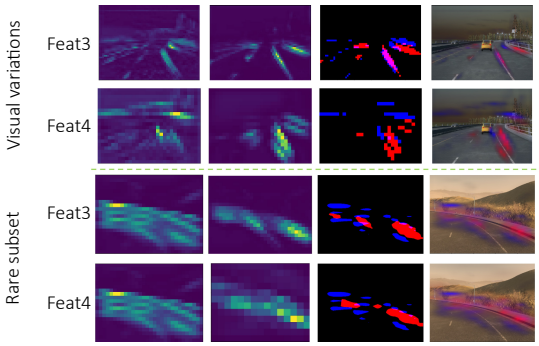


图 6 不同场景下注意力机制特征可视化对比图

Fig. 6 Comparative visualization of attention-based feature representations across different scenarios

4.4.2 车道线检测结果的可视化分析与定性评价

图 7-9 展示了本文方法与基线模型 3DLaneNet_ext，以及 Gen-Lane Net[4] 在多种复杂及罕见场景下的检测性能对比，其中蓝色曲线表示真实标注 (GT)，红色曲线为模型预测结果。评测场景涵盖了夜间低照度、跨时段光照显著变化以及坡道地形等典型环境。在夜间低照度场景 (图 7) 中，由于输入图像亮度较低且纹理信息匮乏，

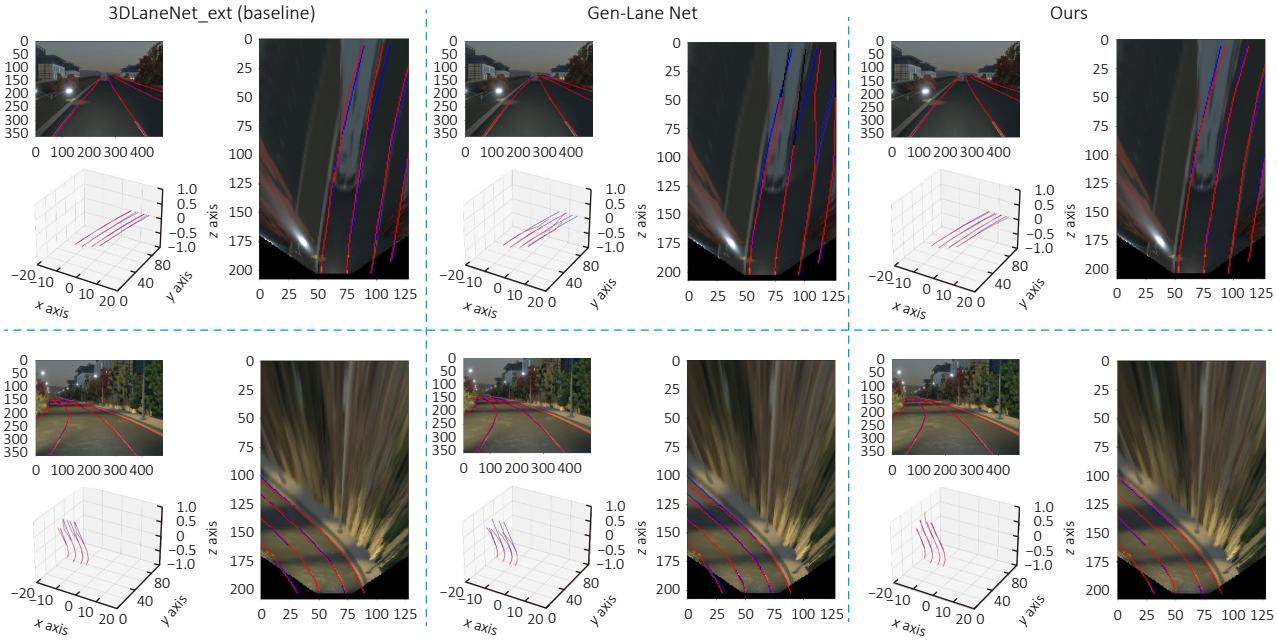


图 7 罕见场景下夜晚车道线检测直道与弯道效果对比图

Fig. 7 Comparison of lane detection performance on straightaways and curves at night in rare scenarios

伴随高光与噪声干扰，基线模型在弯道路段表现出明显的曲率拟合不足与预测漂移。尽管 Gen-Lane Net 能够恢复部分车道结构，但仍存在拓扑不连续现象。相比之下，本方法凭借混合编码器对微弱几何纹理的提取能力，有效避免了基线模型的曲率漂移，在直道与弯道场景下均能稳定拟合真实标注，预测曲线连续且平滑，展现出在低光照条件下更优异的特征保持能力与几何推理能力。在跨时段变光照弯道场景 (图 8) 中，上午与下午的光照差异导致阴影形态变化、局部过曝及路面反射特性显著不同。基线模型在高亮区域出现轻微定位偏移，而 Gen-Lane Net 在阴影边界附近易产生结构断裂。与之形成对比，本方法在波动光照条件下均保持了稳定的车道线走

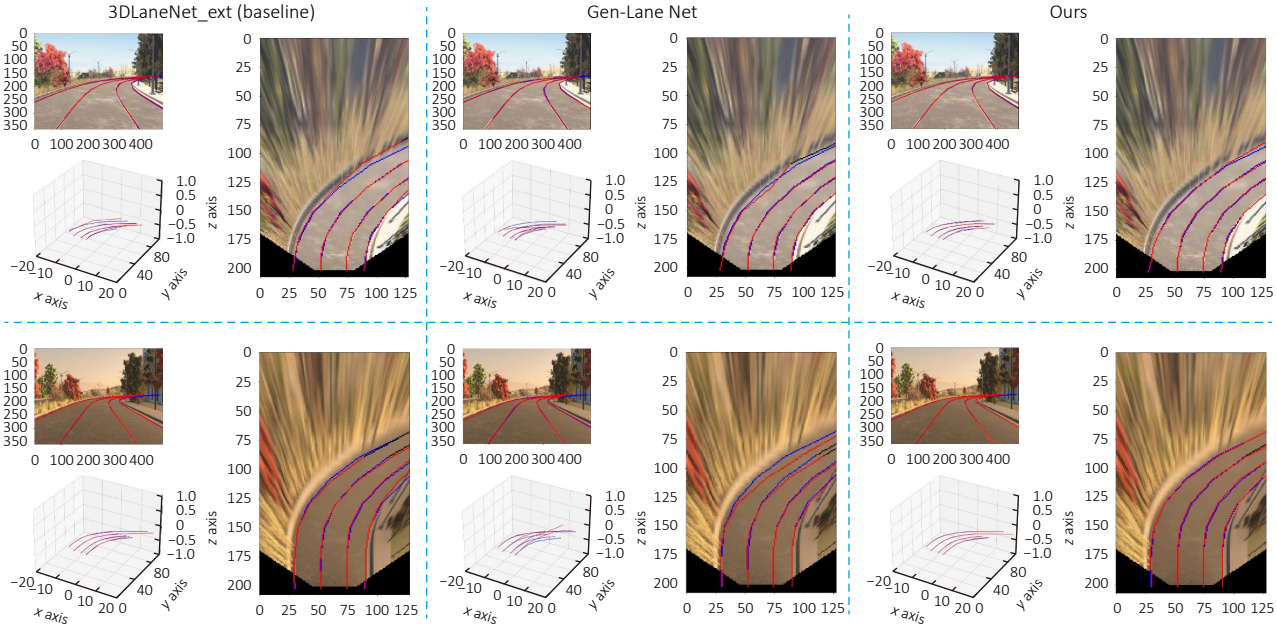


图 8 罕见场景下上午与下午车道线检测弯道效果对比图

Fig. 8 Comparison of lane line detection curve performance in rare scenarios (morning and afternoon)

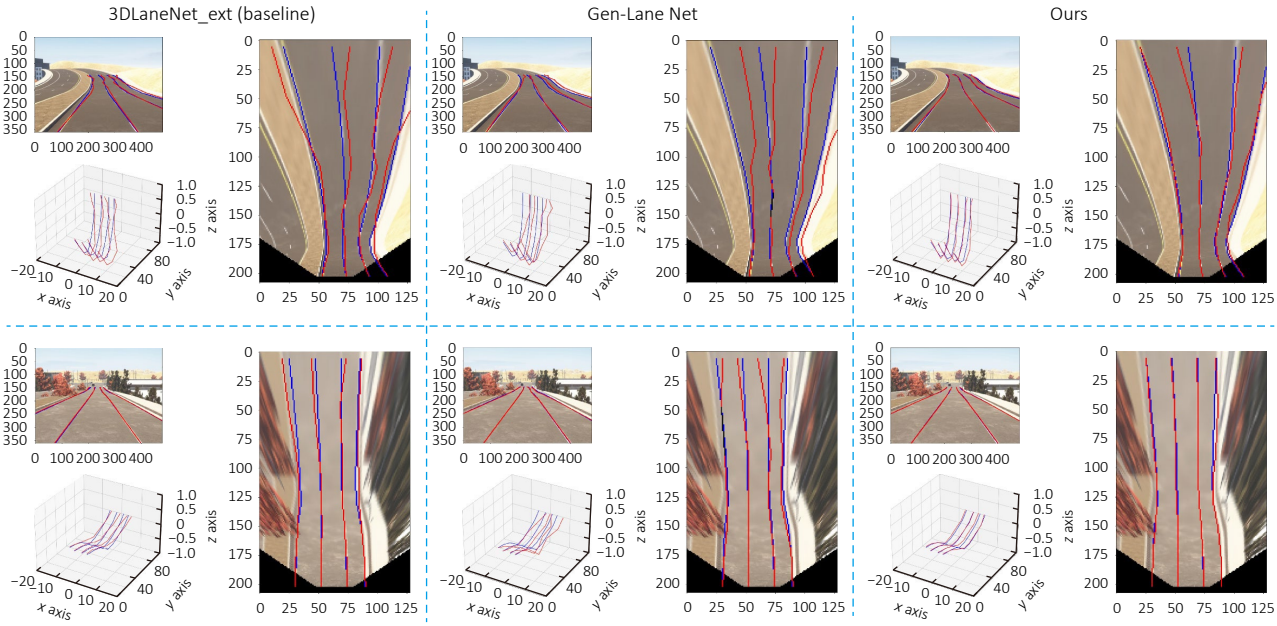


图 9 罕见场景下早晨车道线检测坡道效果对比图

Fig. 9 Comparison of lane line detection ramp performance in rare scenarios during the morning

势估计，预测曲线与 GT 的匹配精度更高，验证了模型具备良好的光照鲁棒性与跨时段泛化能力。针对坡道路面场景 (图 9) 中，路面坡度引发了显著的透视形变，增加了车道线三维空间结构表征的难度。基线模型与 Gen-Lane Net 均出现了不同程度的结构扭曲或曲率偏差，而本方法能够准确恢复车道线的三维空间形态，在重建视图中呈现出更高的平滑性与几何一致性，说明本方法在面临纵向坡度与复杂三维几何变换时，仍具备稳建的空间建模能力。

综上所述，图 7-9 的可视化定性分析表明，本方法在各类极端场景中均表现出更高的稳定性、鲁棒性以及车道结构恢复精度，较基线方法具有显著的技术优势。

5 总　结

本文针对单目三维车道线检测中编码器语义表征能力受限及特征通道冗余的问题，提出了一种轻量化的网络结构改进方法。通过构建融合 VGG 与 ResNet 特征提取优势的混合编码器，实现了浅层几何纹理信息与深层语义特征的协同建模；同时，在深层特征输出阶段引入 SE 模块，对通道特征进行自适应重标定，从而增强车道相关特征响应并抑制背景噪声干扰。实验结果表明，在不显著增加计算开销的条件下，该方法能够提升 BEV 空间特征表示能力并提高三维回归的稳定性。在 Apollo-Sim3D 数据集上的实验结果进一步验证了所提出方法在复杂场景下的有效性和鲁棒性，为后续多模态感知融合与时序建模研究提供了参考。

作者贡献

郑奎：研究资料的收集、分析，实验的设计与完成，论文写作；黄影平、罗鑫：指导论文选题、论文修改。

利益冲突

所有作者声明无利益冲突

参考文献

1. Luo X, Huang Y P, Cui J Y, et al. Deep learning-based lane detection for intelligent driving: a comprehensive survey of methods, datasets, challenges and outlooks[J]. *Neurocomputing*, 2025, **650**: 130795.
2. He X, Guo H Y, Zhu K, et al. Monocular lane detection based on deep learning: a survey[EB/OL]. arXiv: 2411.16316, 2024. https://doi.org/10.48550/arXiv.2411.16316.
3. Garnett N, Cohen R, Pe'er T, et al. 3D-LaneNet: end-to-end 3D multiple lane detection[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), 2019: 2921–2930. https://doi.org/10.1109/ICCV.2019.00301.
4. Guo Y L, Chen G, Zhao P T, et al. Gen-LaneNet: a generalized and scalable approach for 3D lane detection[C]//*Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, 2020: 666–681. https://doi.org/10.1007/978-3-030-58589-1_40.
5. Efrat N, Bluvstein M, Oron S, et al. 3D-LaneNet+: anchor free lane detection using a semi-local representation[EB/OL]. arXiv: 2011.01535, 2020. https://doi.org/10.48550/arXiv.2011.01535.
6. Chen L, Sima C H, Li Y, et al. PersFormer: 3D lane detection via perspective transformer and the openlane benchmark[C]//*Proceedings of the 17th European Conference on Computer Vision*, Tel Aviv, Israel, 2022: 550–567. https://doi.org/10.1007/978-3-031-19839-7_32.
7. Wang R H, Qin J, Li K Y, et al. BEV-LaneDet: an efficient 3D lane detection based on virtual camera via key-points[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023: 21611–21620. https://doi.org/10.1109/CVPR52729.2023.00103.
8. Liu R J, Chen D P, Liu T, et al. Learning to predict 3D lane shape and camera pose from a single image via geometry constraints[C]//*Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022: 1765–1772. https://doi.org/10.1609/aaai.v36i2.20069.
9. Chen Z Y, Smith-Miles K, Du B, et al. An efficient transformer for simultaneous learning of BEV and lane representations in 3D lane detection[EB/OL]. arXiv: 2306.04927, 2023. https://doi.org/10.48550/arXiv.2306.04927.
10. Li C X, Hu Y H, Zheng Z W, et al. Bi^{2}Lane: Bi-Directional temporal refinement with Bi-level feature aggregation for 3D lane detection[C]//*Proceedings of 2024 IEEE International Conference on Robotics and Automation*, Yokohama, Japan, 2024: 2894–2900. https://doi.org/10.1109/ICRA57147.2024.10610794.
11. Yan F, Nie M, Cai X Y, et al. ONCE-3DLanes: building monocular 3D lane detection[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, 2022: 17122–17131. https://doi.org/10.1109/CVPR52688.2022.01663.
12. Zheng Z W, Zhang X M, Mou Y Q, et al. PVALane: prior-guided 3D lane detection with view-agnostic feature alignment[C]//*Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024: 7597–7604. https://doi.org/10.1609/aaai.v38i7.28592.
13. Bai Y F, Chen Z R, Fu Z J, et al. CurveFormer: 3D lane detection by curve propagation with curve queries and attention[C]//*Proceedings of 2023 IEEE International Conference on Robotics and Automation*, London, United Kingdom, 2023: 7062–7068. https://doi.org/10.1109/ICRA48891.2023.10161160.
14. Huang S F, Shen Z W, Huang Z J, et al. Anchor3DLane: learning to regress 3D anchors for monocular 3D lane detection[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023: 17451–17460. https://doi.org/10.1109/CVPR52729.2023.01674.
15. Luo Y R, Zheng C D, Yan X, et al. LATR: 3D lane detection from monocular images with transformer[C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023: 7907–7918. https://doi.org/10.1109/ICCV51070.2023.00730.
16. Bi J P, Song Y C, Jiang Y H, et al. Lane detection for autonomous driving: comprehensive reviews, current challenges, and future predictions[J]. *IEEE Trans Intell Transp Syst*, 2025, **26**(5): 5710−5746.
17. Huang J, Choudhury P K, Yin S, et al. Real-time road curb and lane

detection for autonomous driving using LiDAR point clouds[J]. *IEEE Access*, 2021, **9**: 144940−144951.
18. Luo Y R, Yan X, Zheng C D, et al. M^2-3DLaneNet: exploring multi-modal 3D lane detection[EB/OL]. arXiv: 2209.05996, 2022. https://doi.org/10.48550/arXiv.2209.05996.
19. Zhao R K, Heng Y W, Wang H, et al. Advancements in 3D lane detection using LiDAR point clouds: from data collection to model development[C]//*Proceedings of 2024 IEEE International Conference on Robotics and Automation*, Yokohama, Japan, IEEE, 2024: 5382–5388. https://doi.org/10.1109/ICRA57147.2024.10610087.
20. Luo Y R, Cui S G, Li Z. Dv-3DLane: end-to-end multi-modal 3D lane detection with dual-view representation[C]//*Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.
21. Luo X, Huang Y P, Liang Z M. Axial attention-guided anchor classification lane detection[J]. *Opto-Electron Eng*, 2023, **50**(7): 230079.
罗鑫, 黄影平, 梁振明. 轴注意力引导的锚点分类车道线检测[J]. 光电工程, 2023, **50**(7): 230079.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//*Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA, 2015.
23. He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.
24. Lin M, Chen Q, Yan S C. Network in network[Z]. arXiv: 1312.4400, 2014. https://doi.org/10.48550/arXiv.1312.4400.
25. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018: 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
26. Li C G, Shi J, Wang Y, et al. Reconstruct from top view: a 3D lane detection approach based on geometry structure prior[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, 2022: 4369–4378. https://doi.org/10.1109/CVPRW56347.2022.00483.

作者简介



郑奎 (1999-)，男，硕士研究生，主要从事计算机视觉的研究。

E-mail：2049482458@qq.com



【通信作者】黄影平 (1966-)，男，教授，从事汽车电子、计算机视觉的研究。

E-mail：huangyingping@usst.edu.cn

HA-Lane Net: Hybrid feature encoding and attention for monocular 3D lane detection

Zheng Kui, Huang Yingping*, Luo Xin

School of Opto-Electronic and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract

Objective Three-dimensional lane detection provides essential geometric information for autonomous driving systems, supporting vehicle localization, trajectory planning, and lateral control. Compared with traditional two-dimensional lane detection, three-dimensional lane detection aims to recover the spatial structure of lane markings directly from monocular images. This task remains challenging because monocular imagery lacks explicit depth cues and is affected by perspective distortion, scale variation, and complex road environments. Recent studies have increasingly adopted bird's-eye-view representations that transform front-view image features into a top-down spatial layout. Such representations simplify geometric reasoning and facilitate the estimation of lane structures in three-dimensional space. However, existing convolutional neural network architectures still face difficulties in simultaneously preserving detailed lane geometry and extracting high-level semantic information. Shallow layers capture fine textures and structural boundaries, whereas deeper layers emphasize semantic abstraction and contextual understanding. Excessive dependence on deep features may introduce localization offsets after spatial projection, while shallow features alone are insufficient for reliable detection under challenging conditions such as illumination changes, occlusions, or background interference. A network architecture capable of integrating geometric detail with semantic representation is therefore important for improving monocular three-dimensional lane detection.

Methods To address this issue, a monocular three-dimensional lane detection framework named HA-Lane Net was developed. The framework focuses on constructing balanced feature representations through a hybrid encoding strategy together with an attention-based

Foundation item: National Nature Science Foundation of China (62276167)

*Correspondence: Huang Y P, E-mail: huangyingping@usst.edu.cn

feature refinement mechanism.

The backbone of the proposed framework is a hybrid encoder that combines complementary properties of two classical convolutional architectures. Shallow layers derived from VGG were used to capture detailed textures and structural patterns in lane regions. These features preserve spatial cues that are important for accurate geometric reconstruction. In parallel, deeper layers based on ResNet were employed to extract semantic representations and contextual information from complex road scenes. To integrate these heterogeneous features effectively, channel adaptation and hierarchical coupling mechanisms were introduced. These operations align feature dimensions and progressively fuse shallow geometric cues with deep semantic representations across multiple network stages. As a result, the encoder constructs a multi-scale feature hierarchy that maintains geometric precision while retaining strong semantic discrimination.

Results and Discussions Feature refinement was further introduced before spatial projection through a pre-projection channel attention mechanism. Inspired by the squeeze-and-excitation principle, this module recalibrates feature responses by strengthening channels related to lane structures and suppressing background signals. Applying attention prior to the front-view-to-bird's-eye-view projection improves feature consistency across viewpoints and reduces the influence of image blur, lighting variation, and environmental clutter. This design helps preserve structural lane information during the spatial transformation process.

Training stability was improved using a cosine annealing learning-rate schedule, which facilitates smoother optimization of the multi-task loss function. The framework was trained and evaluated on the Apollo-Sim3D dataset, which contains diverse simulated driving scenes with accurate three-dimensional lane annotations. During inference, the network directly predicts spatial coordinates and shapes of lane markings from monocular images, enabling reconstruction of lane geometry in three-dimensional space without relying on additional sensors.

Results and Discussions Experiments on the Apollo-Sim3D dataset show that HA-Lane Net improves three-dimensional lane detection performance compared with representative monocular approaches such as 3DLaneNet and Gen-Lane-Net. The hybrid encoder effectively integrates geometric detail with semantic context, enabling more accurate estimation of lane geometry.

The pre-projection attention mechanism further improves robustness by highlighting lane-related feature responses and suppressing irrelevant background information. This targeted refinement enhances feature consistency during the transformation from front-view images to bird's-eye-view representations and reduces localization errors caused by occlusion, illumination variation, or scene complexity. Consequently, the network produces more stable detection results across different driving scenarios.

Quantitative comparisons indicate improved accuracy in three-dimensional lane coordinate estimation relative to the reference methods used for comparison. Detection results also show increased stability across diverse road environments. These improvements arise from the combined effect of hybrid feature encoding and attention-based feature refinement, which strengthen lane structure representation while mitigating background noise. In addition, the proposed framework maintains a relatively efficient architecture. Performance improvements are achieved without introducing Transformer-based components or other computationally intensive modules, allowing the network to retain manageable complexity for real-time perception tasks.

Conclusions A monocular three-dimensional lane detection framework named HA-Lane Net has been presented for improved lane perception in autonomous driving environments. The proposed approach integrates a hybrid encoder that combines shallow VGG layers with deep ResNet layers, enabling balanced extraction of geometric detail and semantic representation. A pre-projection channel attention mechanism further refines lane-related features before spatial transformation, improving robustness against environmental interference.

Experimental evaluation on the Apollo-Sim3D dataset shows that the proposed framework improves three-dimensional lane detection performance compared with representative monocular approaches such as 3DLaneNet and Gen-Lane-Net. The hybrid encoding strategy together with attention-based refinement enables more reliable estimation of lane geometry while maintaining a relatively efficient network structure. These results suggest that the proposed framework provides a practical solution for monocular three-dimensional lane perception and offers useful insights for future research on autonomous driving scene understanding.

Keywords: 3D lane detection; hybrid encoder; attention mechanism; autonomous driving

Zheng K, Huang Y P, Luo X. HA-Lane Net: Hybrid feature encoding and attention for monocular 3D lane detection[J]. *Opto-Electron Eng*, 2026, **53**(6): 250359; DOI: 10.12086/oee.2026.250359

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文