基于改进 FCOS3D 的复杂交通场景单目三维目标检测方法

刘宇航[1] 黎润霖[1] 阴明旭[1] 刘 瑞[1] 文 强[1] 赵伟琼[1]
LIU Yuhang LI Runlin YIN Mingxu LIU Rui WEN Qiang ZHAO Weiqiong

摘 要 单目三维目标检测是自动驾驶感知系统中的关键任务，其精度和鲁棒性直接影响车辆对复杂交通环境的理解。然而，现有基于 FCOS3D 的检测方法在密集交通、目标遮挡及小目标检测等方面仍存在性能瓶颈。基于此，文章提出了一种基于 Swin Transformer 和动态调节损失函数的改进方法，以提升 FCOS3D 在复杂交通场景中的检测能力。首先，采用 Swin Transformer 替换原有的 ResNet 主干网络，通过层级化窗口注意力机制增强特征提取能力，使模型在远距离和小目标检测任务中表现更优；其次，引入基于类别难度的动态调节因子，对分类损失进行优化，以提高模型对易混淆目标的关注度，降低目标误分类率。实验基于 nuScenes 数据集中复杂场景进行验证，结果表明，相较于 FCOS3D 和 PETR，该方法在多个关键指标上均取得了显著提升，mAP 提高至 38.7%，mATE、mAVE 等误差指标均有所降低。

关键词 单目三维目标检测；FCOS3D；Swin Transformer；动态调节损失；自动驾驶

doi：10.3969/j.issn.1672-9528.2025.07.047

0 引言

随着智能网联汽车技术的飞速发展，自动驾驶系统在复杂交通环境中的应用日益增多。为了确保自动驾驶的安全性和可靠性，目标检测尤其是车辆检测成为关键技术之一。传统的目标检测方法通常依赖于昂贵且复杂的传感器，如激光雷达和毫米波雷达等。然而，这些传感器不仅成本较高，而且对天气和环境条件的依赖性较强。相比之下，单目三维目标检测方法通过使用传统的 RGB 摄像头来获取图像信息，具有成本低、系统简洁、易于集成等显著优势。因此，单目三维目标检测成为近年来智能交通和自动驾驶领域的重要研究方向[1]。

单目三维目标检测能够利用深度学习模型，从单幅 RGB 图像中推断出物体的三维位置和尺寸。相较于传统的基于 2D 图像的检测方法，单目三维目标检测不仅能够识别物体的位置，还能够获取物体的深度信息，提供更加丰富的场景理解。尽管单目三维目标检测具有诸多优势，但其在复杂环境中的应用仍面临挑战，特别是在处理遮挡、光照变化以及小目标等问题时，现有方法的性能仍有提升空间。

FCOS3D[2] 作为一种基于单目图像的端到端三维目标检测方法，通过完全卷积的架构有效地解决了传统方法中的复杂度问题，并在标准数据集上取得了优异的检测结果。然而，在复杂交通场景中，尤其是在背景杂乱、光照不均和目标遮挡的情况下，FCOS3D 的性能仍然受到一定的限制。为此，本文提出了一种基于 FCOS3D 的改进方法，通过优化网络结构和损失函数来提升其在复杂交通环境中的检测性能。

本文首先将 Swin Transformer[3] 作为主干网络替换掉原有的 ResNet 结构，以增强特征提取能力，特别是在复杂背景和小目标的处理上。其次，本文在分类损失函数中引入基于类别难度的动态调节因子，通过调整 Focal Loss 中的 γ 值，使得模型能够更关注那些难以分类的目标，尤其是那些受遮挡、光照不良或位于复杂背景中的目标。实验结果表明，所提出的方法在 NuScenes 数据集的复杂交通场景中，显著提升了模型的分类准确性和鲁棒性，尤其在处理稀有目标和复杂场景时，表现出了更好的检测性能。本文的研究为单目三维车辆检测提供了新的思路，并为自动驾驶系统的安全性提升做出了贡献。

1 单目三维目标检测算法概述

单目三维目标检测是从单幅 RGB 图像中估计物体的三维空间位置和尺寸的任务，作为计算机视觉领域的重要研究方向，近年来随着深度学习的发展，已取得了显著进展。相较于传统的 2D 目标检测，单目三维目标检测不仅要求检测物体的位置，还需要推测物体的深度信息，这对自动驾驶、智能交通等应用领域尤为重要。

1. 吉利学院智能科技学院 四川成都 641423
[基金项目] 四川省大学生创新训练计划项目（202412802027）；四川省大学生创新训练计划项目（S202412802324X）

1.1 单目三维目标检测的挑战

单目三维目标检测面临的主要挑战在于如何从单一的二维图像中推断出深度信息。与激光雷达（LiDAR）等传感器不同，RGB 图像本身缺乏直接的深度信息，检测模型必须依赖于图像中的纹理、形状以及上下文信息来进行三维重建。此外，目标的遮挡、光照变化、不同尺度和角度的目标都会给三维检测带来困难，这要求模型具备较强的特征学习能力和鲁棒性。

1.2 单目三维目标检测方法的分类

根据算法的工作原理，单目三维目标检测方法大致可以分为以下几类：

（1）基于深度回归的方法

该类方法通过回归网络直接从输入的单目图像中估计物体的三维位置、尺寸和朝向等信息。这类方法的优点是简单直观，适用于端到端的训练与推理 [4]。代表性方法包括 Mono3D 等。尽管这些方法计算高效，但通常存在精度不足的问题，特别是在复杂环境和小目标检测中。

（2）基于 2D-3D 匹配的方法

这类方法通常首先通过 2D 检测算法（如 Faster R-CNN）检测出物体在图像中的二维边界框，然后通过投影模型将这些二维信息映射到三维空间 [5]。这类方法通常会使用一些先验知识，如相机内参和场景几何结构，以此来推断三维信息。尽管这些方法在精度上有所提升，但需要额外的几何约束，并且对于复杂的场景仍然具有较大的挑战。

（3）基于深度学习的端到端方法

近年来，基于深度学习的端到端方法逐渐成为主流。这类方法通过设计专门的神经网络架构，直接从输入的 RGB 图像中提取特征，并同时进行目标检测和三维位置估计 [6]。FCOS3D 即其中的一种代表方法，其通过完全卷积网络结构实现端到端的三维目标检测，能够较好地解决复杂场景中的目标检测问题。

1.3 FCOS3D 算法概述

FCOS3D 是一种基于完全卷积的端到端三维目标检测方法。采用“单阶段”的检测策略，能够同时预测物体的三维边界框信息和目标类别。该方法以卷积神经网络（CNN）为主干网络，通常使用 ResNet[7] 或 VGG 作为特征提取器，并通过特征金字塔网络（FPN）实现多尺度特征的融合。在目标检测任务中，FCOS3D 通过一个基于深度回归的分支来预测物体的三维位置和朝向。为提升检测精度，还引入了 Focal Loss 作为分类损失函数，有效应对类别不平衡的问题。

尽管 FCOS3D 在标准数据集上表现良好，但在处理复杂场景时，特别是面对遮挡、光照变化和小目标时，仍然存在一定的局限性。因此，如何进一步提升其在复杂交通场景中的检测能力，成为本文研究的主要目标之一。

1.4 Transformer 在目标检测中的应用

近年来，Transformer 结构以其强大的全局特征建模能力和灵活性，在计算机视觉领域得到了广泛应用。与传统的卷积神经网络不同，Transformer 利用自注意力机制，可以有效捕获全局上下文信息，从而弥补卷积网络在建模长距离依赖关系上的不足 [8]。

在目标检测领域，Transformer 最早由 DETR[9]（Detection Transformer）引入，该方法通过编码器 - 解码器结构直接从图像中学习目标特征，并通过注意力机制进行目标定位与分类。尽管 DETR 在全局特征建模上表现出色，但其训练收敛速度较慢，并且对小目标的检测效果仍需改进。

为解决 DETR 的不足，研究者们进一步提出了一些基于 Transformer 改进的目标检测方法。例如，Deformable DETR 通过引入可变形注意力机制，大幅提升了小目标的检测效果和训练收敛速度。此外，Transformer 与卷积网络的结合也逐渐成为一种趋势，如 Swin Transformer，通过将图像分割为多个窗口（window）并在局部窗口内计算注意力，既保留了 Transformer 的全局建模能力，又显著降低了计算复杂度。Swin Transformer 已被广泛应用于各种视觉任务中，包括目标检测和分割。相较于传统卷积网络，Swin Transformer 在特征提取上具有更强的表达能力，尤其是在复杂背景和多尺度目标处理方面表现优异。因此，本文将 Swin Transformer 引入到 FCOS3D 的主干网络中，替代原有的 ResNet 结构，以提升模型对复杂交通场景的适应能力。

2 单目三维目标检测算法

本文提出了一种基于 FCOS3D 的改进单目三维目标检测算法，通过优化主干网络结构和分类损失函数，提升了模型在复杂交通场景中的检测性能。具体而言，本文采用 Swin Transformer 替换原有的 ResNet 主干网络，同时改进了分类损失函数，引入了基于类别难度的动态调节因子，以更精准地检测目标。

2.1 算法整体框架

本文改进后的算法整体框架如图 1 所示。与 FCOS3D 类似，改进后的模型为完全卷积的单阶段检测器，包含主干网络、特征金字塔网络（FPN）和检测头 3 个主要模块。模型的输入为单幅 RGB 图像，输出为多任务预测，包括目标的二维边界框、三维位置、尺寸、方向和类别信息。

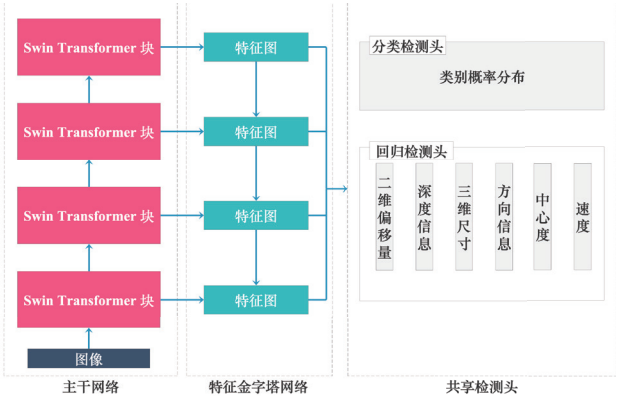


图 1 改进算法框架示意图

具体而言，检测任务输出包含以下回归目标：

（1）二维偏移量：$(\Delta x, \Delta y)$ 表示从当前点到目标二维中心点的偏移；

（2）深度信息：d 表示目标在相机坐标系下的深度；

（3）三维尺寸：(w, l, h) 分别表示目标的宽度、长度和高度；

（4）方向信息：θ 表示目标的朝向角；

（5）中心度（center-ness）：c 表示当前点是否接近目标中心，用于抑制低质量预测；

（6）速度：v 表示目标的运动速度；

（7）类别概率分布：表示目标属于不同类别的概率分布。

2.2 主干网络改进

FCOS3D 原始模型使用 ResNet 作为主干网络，虽然具有较好的特征提取能力，但局部感受野的限制使其在处理复杂背景和小目标检测时表现不足。本文采用 Swin Transformer 替代 ResNet 作为主干网络，通过其层级化窗口注意力机制（window-based multi-head self-attention, W-MSA）和移位窗口注意力机制（shifted window attention）增强特征提取能力。

Swin Transformer 的特征提取过程包括以下关键步骤：

（1）窗口划分与注意力计算，输入特征图被划分为多个大小为 $M\times M$ 的窗口，每个窗口内独立计算注意力：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

式中：$\boldsymbol{Q}=X\boldsymbol{W}_{Q}$、$\boldsymbol{K}=X\boldsymbol{W}_{K}$、$\boldsymbol{V}=X\boldsymbol{W}_{V}$ 为输入特征的线性变换；d 为注意力头的维度。窗口划分将计算复杂度从全局的 $O(N^2)$ 降低到局部的 $O(M^2)$，其中 N 为输入特征图像素总数；M 为窗口大小。

（2）移位窗口操作，为了跨窗口建模全局特征，Swin Transformer 在连续层中交替使用窗口注意力和移位窗口注意力。移位窗口将窗口边界的特征与相邻窗口连接，用公式表示为：

$$X_{\text{shifted}} = \text{Shift}(X, \Delta) \tag{2}$$

式中：Δ 为偏移步长。

（3）多尺度特征融合，主干网络输出的多尺度特征图通过特征金字塔网络（FPN）进一步融合，用于检测不同大小的目标。

2.3 分类损失函数改进

FCOS3D 原始分类损失函数使用 Focal Loss，其公式为：

$$\mathcal{L} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{3}$$

式中：α_t 是平衡因子，p_t 是预测的类别概率；γ 是调节因子，用于降低易分类样本的权重。然而，固定的 γ 无法动态适应复杂场景中不同类别的分类难度。

为此，本文引入基于类别难度的动态调节因子，改进后的分类损失函数为：

$$\mathcal{L}_{\text{class}} = -\alpha_t(1-p_t)^{\gamma_t}\log(p_t) \tag{4}$$

其中，动态调节因子 γ_t 定义为：

$$\gamma_t = \exp(-\beta\cdot\text{accuracy}_t) \tag{5}$$

式中：$\text{accuracy}_t = \dfrac{\text{TP}_t}{\text{TP}_t+\text{FN}_t}$ 表示类别 t 的分类准确率；β 表示超参数控制动态调节强度，准确率变化曲线如图 2 所示。

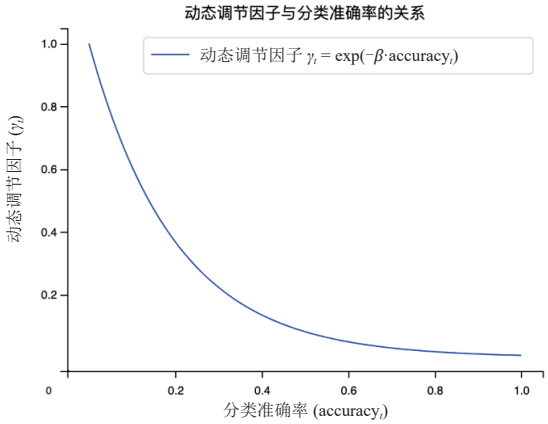


图 2 动态调节因子 γ_t 随分类准确率变化的曲线图

通过动态调整 γ_t，模型能够更关注分类难度高的类别，提高复杂场景中的检测精度。

3 实验

3.1 实验配置

本研究采用 NuScenes 数据集中的部分数据进行实验，该数据集是当前自动驾驶领域常用的三维感知基准数据集，包含多模态感知数据和详细的标注信息。为验证本文方法在复杂交通场景中的有效性，从中筛选了包含复杂交通场景的单目 RGB 图像及其配套标注。本文采用 NuScenes 的评价指标，包括以下几项：

平均精度（mean average precision, mAP）：衡量检测结

果中边界框与真实框之间的匹配程度，反映模型在定位目标时的准确性。mAP 的计算基于鸟瞰视角中边界框的重叠率（intersection over union, IoU），并对不同的物体类别进行平均。

平均平移误差（mean average translation error, mATE）：用于衡量检测到的目标与真实目标之间的平移误差，单位为 m，表示模型在定位目标中心点时的准确性。

平均尺度误差（mean average scale error, mASE）：用于评估目标边界框的尺度估计误差，衡量模型在预测物体尺寸上的准确性。

平均朝向误差（mean average orientation error, mAOE）：用于衡量目标方向估计的误差，单位为 rad，反映模型在方向估计上的精确程度。

平均速度误差（mean average velocity error, mAVE）：用于评估目标速度的预测与真实速度之间的误差，单位为 m/s，适用于动态物体的检测。

平均属性误差（mean average attribute error, mAAE）：衡量目标属性的估计误差，如目标是否在移动或目标的具体类型。

NuScenes 综合检测得分（nuScenes detection score, NDS）：作为一个综合性指标，通过结合上述各项误差和 mAP 得出，用于全面评估模型在检测精度和鲁棒性方面的表现。

实验在以下硬件配置下进行：

CPU：Intel® Xeon® Platinum 8331A，50 核

内存：250 GB

GPU：1 张 NVIDIA A100 SXM4 80 GB

3.2 实验结果

为验证本文提出的基于 FCOS3D 的改进方法在复杂交通场景中的有效性，本文将其与现有的两种方法（FCOS3D 和 PETR）进行对比。实验结果如表 1 所示，展示了不同方法在多个关键指标上的表现。

表 1 各方法评价指标对比

方法	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
FCOS3D	0.358	0.69	0.249	0.452	1.434	0.124	0.428
PETR	0.377	0.746	0.271	0.488	0.906	0.212	0.426
本方法	0.387	0.623	0.265	0.512	1.234	0.211	0.432

在 mAP（平均精度）方面，本文方法取得了 0.387 的结果，明显优于 FCOS3D 的 0.358 和 PETR 的 0.377。表明改进后的模型能够更好地匹配预测的边界框与真实目标，从而提升了目标检测的精度。对于误差指标，mATE（平均平移误差）的结果为 0.623，较 FCOS3D（0.69）和 PETR（0.746）更小，表明本文方法在目标位置估计上的表现更加准确。mASE（平均尺度误差）的结果为 0.265，尽管比 FCOS3D（0.249）略大，但优于 PETR（0.271），显示了该方法在物体尺寸估计方面的较强能力。mAVE（平均速度误差）的结果为 1.234，相较于 FCOS3D（1.434）较小，尽管仍大于 PETR（0.906）。这一结果表明，虽然本文方法在速度预测方面有所改善，但在动态物体检测方面，PETR 依然优于本文方法。mAAE（平均属性误差）的结果为 0.211，接近 PETR（0.212），但高于 FCOS3D（0.124）。该指标反映了本文方法在目标属性识别方面的稳定性和准确性，虽然在精度上略低于 PETR，但仍表现出较强的能力。

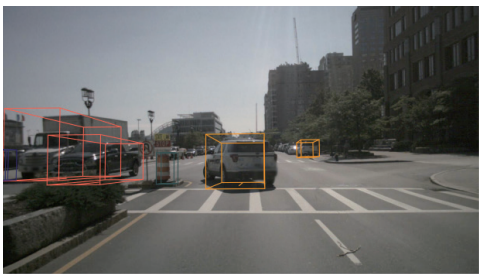
综合考虑所有评估指标，NuScenes 综合检测得分（NDS）为 0.432，超越了 FCOS3D（0.428）和 PETR（0.426），证明了本文方法在整体检测性能和鲁棒性方面的优势。通过与 FCOS3D 和 PETR 的对比，本文方法在目标检测精度、位置估计、尺度估计和综合检测得分等多个重要指标上均表现优异，验证了 Swin Transformer 作为主干网络与基于类别难度的动态调节因子引入的有效性，显著提升了 FCOS3D 在复杂交通环境中的检测能力。

为进一步评估本文方法在复杂交通场景中的检测性能，图 3 展示了 FCOS3D 与本文方法在相同场景下的前视图检测结果。通过对比两种方法的检测框，可以观察到本文方法在目标完整性、漏检率和目标姿态估计等方面均表现出优势。

在图 3（a）中，FCOS3D 的检测结果存在明显的目标漏检问题。最明显的例子是图像左侧的大型货车未被 FCOS3D 正确识别，导致目标检测不完整。此外，在密集交通环境下，FCOS3D 的检测框存在一定程度的重叠，部分目标的三维边界框形状畸变，导致目标的定位精度下降。远处的小目标（如右侧的行人）虽能被检测到，但边界框存在一定偏移，与真实目标的位置不完全匹配。

相较之下，图 3（b）展示的本文方法在相同场景下的检测结果明显改善。首先，本文方法成功检测到了左侧的大型货车，表明所提出的改进网络在特征提取和目标完整性方面具有更高的鲁棒性；其次，在密集交通环境下，本文方法的检测框更加清晰，减少了目标间的重叠问题，使得目标的三维边界框更加准确。此外，在远处的小目标检测方面，本文方法的检测框对目标位置的对齐更加精准，特别是在行人检测方面，明显减少了检测偏移现象。这一改进可以归因于 Swin Transformer 在多尺度特征提取上的优势，使得模型能够更有效地感知场景中的远处目标。此外，本文方法所引入的基于类别难度的动态调节因子进一步提升了模型在复杂背景和遮挡场景下的检测能力。

（a）FCOS3D



（b）本文方法

图 3 前视角检测结果对比图

整体而言，视觉对比实验进一步验证了本文方法在复杂交通环境中的有效性。与 FCOS3D 相比，改进后的方法在目标检测完整性、位置精度、小目标检测和目标方向估计方面均具有更优的表现。这些改进使得本文方法在自动驾驶感知系统中的应用更加可靠，能够有效提升复杂交通环境下的目标检测能力。

4 结语

本文针对 FCOS3D 在复杂交通场景中的单目三维目标检测性能局限性，提出了一种基于 Swin Transformer 和动态调节损失函数的改进方法。通过引入 Swin Transformer 主干网络，增强了模型对多尺度目标的特征提取能力，提升了在小目标和远距离目标检测方面的表现。同时，基于类别难度的动态调节因子有效优化了分类损失，使模型更加关注难以检测的目标，从而提高了检测精度和鲁棒性。

实验结果表明，相较于 FCOS3D 和 PETR，本文方法在 NuScenes 数据集的多个关键指标上均取得了更优的性能。在 mAP、mATE、mAVE 等指标上表现突出，特别是在目标检测精度和目标位置估计方面展现了显著的提升。此外，通过视觉对比实验分析，本文方法在目标检测完整性、目标框精度以及目标方向估计等方面均明显优于 FCOS3D，尤其在密集交通环境和远距离目标检测任务中表现优异。尽管本文方法在多个方面取得了改进，仍存在一定的局限性。例如，在动态目标的速度估计方面仍未超越 PETR，表明未来仍需进一步优化运动预测模块。此外，本文方法虽然提升了复杂场景下的检测性能，但计算复杂度相较于 FCOS3D 略有增加，未来可以考虑更高效的特征提取结构，以平衡精度与计算开销。

本文的研究为单目三维目标检测提供了一种有效的优化方案，为自动驾驶系统在复杂交通环境下的感知能力提升提供了技术支持。未来的研究方向将聚焦于进一步优化网络结构，提升实时检测能力，并探索如何在多传感器融合框架下，结合深度信息和时序信息，提高目标检测的稳定性和泛化能力。

参考文献：

[1] 李昌财，陈刚，侯作勋，等．自动驾驶中的三维目标检测算法研究综述 [J]. 中国图象图形学报 ,2024,29(11):3238-3264.

[2] WANG T, ZHU X G, PANG J M, et al. FCOS3D: fully convolutional one-stage monocular 3D object detection [EB/OL].(2021-09-24)[2024-05-25].https://doi.org/10.48550/arXiv.2104.10956.

[3] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision Transformer using shifted windows [EB/OL]. (2021-08-17)[2025-01-19]. https://doi.org/10.48550/arXiv.2103.14030.

[4] 赵文俊．深度线索引导的单目三维目标检测方法研究 [D]. 淮南：安徽理工大学 ,2024.

[5] 刘青，李伟，余少勇，等．结合深度信息引导和多尺度通道注意力机制的单目三维目标检测算法 [J]. 山东大学学报 (理学版), 2025,60(1):63-73.

[6] 张续坤，温显斌．基于特征信息强化的单目三维目标检测模型 [J/OL]. 天津理工大学学报 ,1-9[2025-02-13].http://kns.cnki.net/kcms/detail/12.1374.N.20241216.1739.046.html.

[7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016[2024-05-15].https://ieeexplore.ieee.org/document/7780459.DOI: 10.1109/CVPR.2016.90.

[8] 柳长源，高阁君，刘金凤．采用深度感知 Swin Transformer 的单目三维目标检测方法 [J/OL]. 北京工业大学学报 ,1-9[2025-02-13].http://kns.cnki.net/kcms/detail/11.2286.T.20240911.2106.002.html.

[9] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers [C]//Computer Vision–ECCV 2020: 16th European Conference. NewYork: ACM, 2020: 213-229.

【作者简介】

刘宇航（2001—），男，北京人，本科在读，研究方向：智能驾驶感知，email: 2963403673@qq.com。

黎润霖（1993—），男，四川成都人，硕士研究生，助教，研究方向：智能驾驶感知，email: arnold0824@hotmail.com。

（收稿日期：2025-02-13 修回日期：2025-07-09）