计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009819] [CSTR: 32024.14.csa.009819]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于融合采样和深尺约束的单目 3D 目标检测①

孙虎成, 臧　可

(青岛大学 自动化学院, 青岛 266071)
通信作者: 孙虎成, E-mail: sunhucheng2022@163.com

摘　要: 针对单目图像中不同深度目标的尺度差异所导致的单目 3D 目标检测算法精度不佳的问题, 提出一种基于融合采样和深尺约束的检测算法. 首先, 为增强采样特征对不同尺度目标的表征能力, 构建多尺度特征融合模块 (muti-scale fusion module, MFM), 通过分层聚合和迭代聚合对不同层级、不同尺度的特征进行融合采样, 从而提高对目标隐式尺度特征的提取能力. 此外, 构造深度尺度相关化模块 (depth-scale correlation module, DSCM), 利用深度与尺度之间的线性投影约束将不同尺度的目标补偿式放缩至同一特征水平, 以此平衡模型对不同距离目标的关注度. 基于 KITTI 数据集和 Waymo 数据集的定量结果表明, 所提出的算法相较于同类算法在多种难度下的整体平均精度 AP_{3D} 分别提升了 1.56 个百分点和 3.07 个百分点, 验证了算法的有效性及泛化性, 同时基于两类数据集的定性结果验证了该算法显著缓解了目标尺度差异对检测性能造成的影响.

关键词: 3D 目标检测; 融合采样; 多尺度特征; 可变形卷积; 关注度平衡; 尺度放缩

引用格式: 孙虎成,臧可.基于融合采样和深尺约束的单目 3D 目标检测.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9819.html

Monocular 3D Object Detection Based on Fused Sampling and Depth-scale Constraints

SUN Hu-Cheng, ZANG Ke

(College of Automation, Qingdao University, Qingdao 266071, China)

Abstract: Aiming at the poor accuracy of monocular 3D object detection algorithms caused by the scale differences of objects with different depths in monocular images, a detection algorithm based on fused sampling and depth-scale constraints is proposed. Firstly, to enhance the ability of the sampled features to represent objects at different scales, a multi-scale fusion module (MFM) is constructed. It fuses the sampled features at different levels and scales through hierarchical aggregation and iterative aggregation, thereby improving the ability to extract implicit scale features of the objects. In addition, a depth-scale correlation module (DSCM) is constructed. It uses the linear projection constraint between depth and scale for compensatory scaling of objects at different scales to the same feature level, balancing the model's focus on objects at different distances. Quantitative results based on the KITTI dataset and Waymo dataset show that for both types of datasets, the proposed algorithm improves the overall average accuracy AP_{3D} by 1.56 percentage points and 3.07 percentage points, respectively, compared to similar algorithms under multiple difficulties, which verifies the effectiveness and generalization of the algorithm. Meanwhile, qualitative results based on the two datasets validate that the algorithm significantly mitigates the impact of the object scale differences on detection performance.

Key words: 3D object detection; fused sampling; multi-scale features; deformable convolution; attention balancing; scale deflation

① 基金项目: 山东省自然科学基金 (ZR2023QF089)
收稿时间: 2024-09-24; 修改时间: 2024-11-07; 采用时间: 2024-11-12; csa 在线出版时间: 2025-02-28

城市车辆检测任务在交通管理、道路规划和军事监察等领域中发挥着重要作用[1]. 而 3D 目标检测是关乎道路信息化网络构建以及自动驾驶系统安全的一项基础性环境感知技术, 在智能交通领域中, 对道路场景中的 3D 目标进行实时检测对于保障汽车的安全行驶具有重要意义. 根据输入模态的差异, 3D 目标检测方法大致可分为 3 类: 基于单模态点云的检测方法、基于单模态图像的检测方法、基于点云和图像多模态融合的检测方法[2]. 考虑到图像色彩信息丰富, 同时相较于激光雷达, 图像传感器具有成本低廉且易于部署的优势, 因此目前基于单目图像的 3D 目标检测方法受到了工业界和学术界的诸多关注.

由于单目图像缺乏显式的深度信息, 因此从图像中获取目标距离是一个不适定问题[3]. 为解决单目图像深度信息缺失的问题, PatchNet[4]和 DDMP-3D[5]通过引入额外深度信息的方式进行辅助深度估计. CaDDN[6]尝试将图像平面过渡为视锥体, 进而向体素坐标系转换的方法提升模型的定位精度. Kinematic-3D[7]引入相邻帧的图像, 利用运动学方法来提取动态场景的特征, 提高模型对目标的定位性能. 尽管这些方法在一定程度上缓解了单目图像深度信息缺失的问题, 但对深度的估计极度依赖于额外数据的输入, 这就造成网络结构冗杂以及推理速度慢的问题. MonoFlex[8]提出一种检测结构来建模两个相邻对象间的几何关键点, 通过编码的方式捕获对象间的几何特征. MonoDTR[9]提出一个深度感知模块, 搭配特有的位置编码器对目标特征进行学习. MonoCon[10]将 3D 边界框关键点回归作为辅助任务, 通过隐式监督的方式提升了 3D 目标的检测精度. GUPNet[11]利用立体约束来弥补图像中深度信息缺乏的问题. 该类方法虽然有效, 但学习深度信息的能力主要依靠模块化的深度估计网络, 未与整体网络形成有效联动, 易造成不精确的目标定位结果.

针对单目图像中目标尺度差异影响模型检测性能的问题, Mono3D[12]使用大量特征, 通过语义分割、对象轮廓和位置先验的方法生成 3D 锚框, 之后通过特定损失函数评估这些特征, 以适应对相关参数的学习. 3D-RCNN[13]依赖于在 2D 边界框中检测到的特征, 并利用外部数据将 2D 信息匹配到 3D 特征中, 通过比较损失来恢复 3D 实例的形状和姿势. M3D-RPN[14]提出了一个独立的网络来同时生成 2D 和 3D 目标候选框. 但是该类方法[15,16]忽略了目标尺度与深度间的几何关系, 通过尺度先验以及维度匹配的方法对目标尺寸进行推理, 低效的特征匹配策略将直接影响模型对小目标的检测精度.

基于已有研究的局限性, 本文提出一种基于融合采样和深尺约束的单目 3D 目标检测算法来提升对道路上 3D 目标的检测性能. 针对目标尺度差异导致的算法检测精度不佳的问题, 本文在 DLA-34[17]特征提取网络的基础上构建了具有分层聚合网络 (hierarchical fusion network, HFN) 和迭代聚合网络 (iterative fusion network, IFN) 的多尺度特征融合模块, 该模块通过融合不同层级、不同尺度的特征提高了模型对目标特征的提取和表征能力. 此外, 受透视投影原理的启发, 本文根据目标尺度与深度之间的线性几何约束构建深度尺度相关化模块, 引入单目图像缺少的细粒度深度信息的同时以尺度放缩的方式提高目标尺寸推理的准确性与可靠性, 有效联动了整体网络的推理过程. 基于 KITTI 数据集和 Waymo 数据集的实验证明了所提出算法的有效性.

1 算法原理

1.1 算法整体框架

本文所提出算法的整体框架如图 1 所示, 该算法由多尺度特征融合模块 (multi-scale fusion module, MFM)、深度尺度相关化模块 (depth-scale correlation module, DSCM)、多任务检测头这 3 个部分组成. 网络输入为单目 RGB 图像$I \in \mathbb{R}^{H\times W\times 3}$, 首先单目图像$I$经过多尺度特征融合模块 MFM, 在可变形卷积网络采样目标实际尺寸的同时, 聚合全局特征并进行下采样得到原始特征F. 然后原始特征F到达深度尺度相关化模块 DSCM 后首先通过深度离散化网络 (depth discretization network, DDN) 得到细粒度的深度特征F_f. 之后原始特征F和深度特征F_f在尺度归一化网络 (scale normalization network, SNN) 中利用深度与尺度之间的几何约束生成归一化后的尺度特征F_s. 最后聚合特征F_a经过多任务检测模块完成训练过程. 图 1 中的推理结果图显示, 经过训练的模型可以准确定位目标, 有效完成道路场景中的 3D 目标检测任务.

1.2 多尺度特征融合模块

为解决采样特征表征能力不佳的问题, 本文构建多尺度特征融合模块作为主干特征提取网络. 网络结构如图 2 所示, 本文在 DLA-34 网络的基础上构造了分层聚合网络 HFN 和迭代聚合网络 IFN, 分别对不同

层级、不同尺度的特征进行特征聚合, 从而提高模型对目标的提取和表征能力. 此外, 为更进一步地捕获目标尺寸信息、拟合目标特征边界, 本文在 HFN 和 IFN 中采用了可变形卷积网络 (deformable convolutional network, DCN)[18]替代普通卷积操作以采样非规则、多尺度目标的隐式特征.

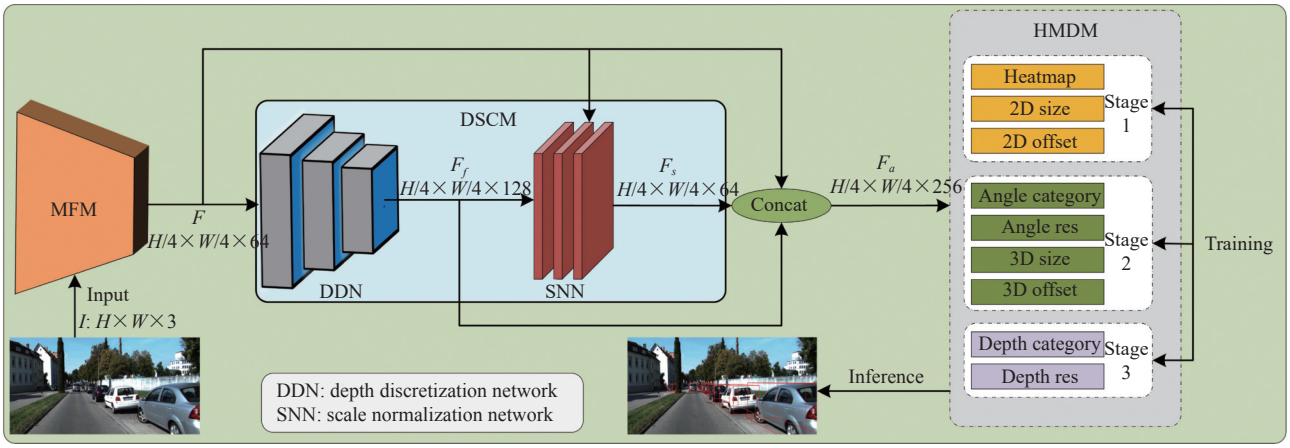

图 1 整体算法框架示意图

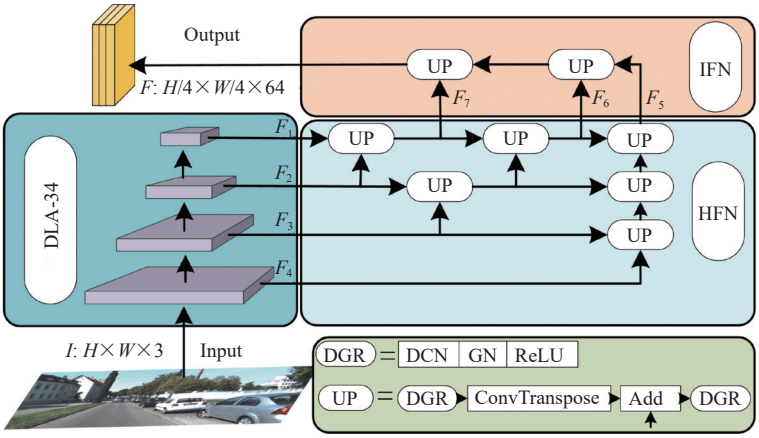

图 2 多尺度特征融合模块示意图

由图 2 可知, 一张单目 RGB 图像$I \in \mathbb{R}^{H\times W\times 3}$输入主干特征提取网络后, 首先经过 DLA-34 网络中进行不同尺度间的特征融合以及下采样, 进而得到 4 个不同感受野的层级特征: F_1–F_4, 其特征尺寸分别为输入图像I的 1/4、1/8、1/16、1/32, 特征层数分别为 64/128/256/512. 之后层级特征F_1–F_4, 在 HFN 中通过上采样模块聚合得到 3 个不同尺寸的聚合特征: F_5–F_7, 其特征尺寸分别为输入图像I的 1/4、1/8、1/16, 特征层数分别为 64/128/256. 最后在 IFN 中对层级聚合特征F_5–F_7进一步的进行迭代特征融合得到深层次聚合特征F.

此外, 为降低模型训练对数据批量大小的敏感度, 本文在多尺度特征融合模块中使用组归一化 (group normalization, GN)[19]方法替换批量归一化 (batch normalization, BN)[20]方法, 从而特征融合过程更具鲁棒性. 由此, 该特征融合网络充分联系了特征整体与局部的关系, 在可变形卷积网络的协助下, 多尺度特征融合模块在采样目标实际特征尺度的同时有效增强了特征的表征能力.

1.3 深度尺度相关化模块

如图 3 所示, 根据透视投影原理, 相同大小的目标距离相机位置越近在图像中所占像素越多、信息越丰富; 反之距离相机位置越远则目标在图像中所占像素越少、特征信息越模糊. 由此造成了模型针对近距离

大目标检测精度高而针对远距离小目标检测精度低的现象.

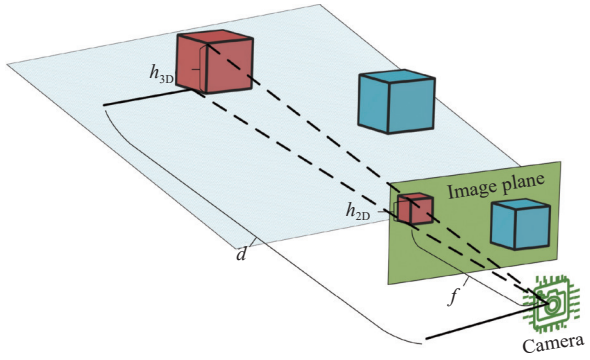

图 3 透视投影原理示意图

由图 3 可知, 根据相机部件与目标之间的线性投影约束, 相机焦距f与目标深度d之间的比例等同于图像平面中目标高度h_{2D}与现实场景中目标高度h_{3D}的比例, 即:

$$\frac{f}{d}=\frac{h_{2D}}{h_{3D}} \tag{1}$$

为解决尺度差异导致的模型对不同距离目标关注度不平衡的问题, 本文根据上述透视投影原理构建了深度尺度相关化模块 DSCM. 该模块由深度离散化网络 DDN 和尺度归一化网络 SNN 两部分组成.

1.3.1 深度离散化网络

为解决单目图像深度信息缺失的问题, 本文构建深度离散化网络来生成细粒度深度特征. 其结构如图 4 所示.

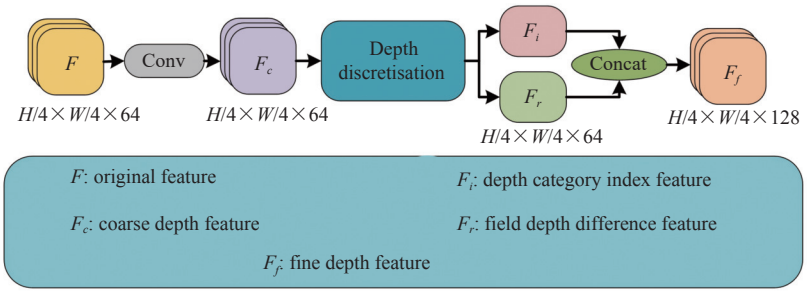

图 4 深度离散化网络

在该网络下, 主干特征提取网络生成的原始特征F首先通过若干卷积块生成粗粒度的深度特征F_c, 并以此为基础进行深度离散化. 本文所选择的离散化方法为线性递增离散化[21]. 该方法下, 深度类别标签所对应的深度范围随着距离的增加而线性递增. 公式如下:

$$index=\left\lfloor -0.5+\frac{1}{2}\times\sqrt{1+4\times\left(d_{\text{pred}}-d_{\min}\right)\times\frac{D(D+1)}{d_{\max}-d_{\min}}}\right\rfloor \tag{2}$$

其中, d_{pred}表示预测深度值; $index$表示对应于d_{pred}的深度类别标签, 表明d_{pred}位于标签为$index$的深度类别所覆盖的深度范围之内; $d_{\min}$表示最小检测距离; $d_{\max}$表示是最大检测距离; D表示指定的深度类别标签个数; 由$\lfloor\cdot\rfloor$表示向下取整可知, 深度离散化过程引入了量化误差, 因此为消除量化误差对深度推理造成的不良影响, 本文引入了领域深度差的概念, 其值由式 (3)、式 (4) 确定:

$$d_{index}=\frac{[2\times(index+0.5)]^2-1}{4}\times\frac{d_{\max}-d_{\min}}{D(D+1)}+d_{\min} \tag{3}$$

$$d_{\text{res}}=d_{\text{pred}}-d_{index} \tag{4}$$

其中, d_{index}表示反离散化标签深度值, 由式 (1) 的逆向计算求得; 通过对实际深度值d_{pred}与反离散化标签深度值d_{index}求差可得到领域深度差d_{res}, 该值实际等于离散化过程中在索引为$index$的深度类别领域中引入的量化误差.

通过以上离散化方法生成深度类别标签特征F_i以及领域深度差特征F_r, 并通过特征拼接得到细粒度的深度特征表示F_f, 由此引入了精细化的深度信息, 为后续的尺度归一化提供了条件.

1.3.2 尺度归一化网络

为提高模型对不同距离目标的检测能力, 本文根据尺度与深度间的投影几何约束构建尺度归一化网络, 放缩因距离不同而尺寸不同的目标至同一特征水平, 以此平衡模型对不同距离目标的关注度. 网络结构如图 5 所示.

该网络首先对细粒度深度特征F_f进行反拼接得到深度类别标签特征F_i以此作为依照深度进行尺度放缩的基准; 然后对F_i与f求商得到放缩比例特征F_p; 之后原始特征F对F_p进行像素级求积得到尺度归一化特

征F_s, 从而将特征图中不同距离的目标缩放到同一尺度. 这种反比例补偿式的尺度放缩将深度与尺度联系起来, 平衡模型对不同距离目标的关注度, 提高模型对不同尺度目标的检测能力, 在提升模型对近距离目标检测效果的同时也减少了远距离小目标误检和漏检现象的发生.

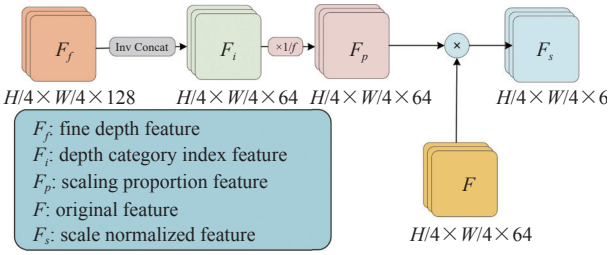

图 5 尺度归一化网络

1.4 损失函数

本文所提出的算法的损失函数分为 3 个部分: 2D 检测损失ℓ_{2d}、3D 检测损失ℓ_{3d}、深度检测损失ℓ_d. 在训练阶段本文使用分层级任务学习 (hierarchical task learning, HTL)[11]框架对模型进行渐进式训练, 稳定训练过程的同时保证参数推理的精度.

1.4.1 2D 检测损失函数

2D 检测损失包含目标位置损失ℓ_{loc}、目标二维尺寸损失ℓ_{2s}、目标二维偏移损失ℓ_{2o}, 设置该类损失目的是定位和预测目标正样本在单目图像中的二维位置及尺寸, 为后续三维检测提供较为准确的待检测目标.

考虑到单目图像中目标类别以及目标分布的非均衡性, 本文采用高斯焦点损失函数作为目标位置损失函数, 其公式为:

$$\ell_{loc}=\frac{-1}{N}\sum_{cxy}\begin{cases}\left(1-\hat{Y}_{cxy}\right)^{\alpha}\log\left(\hat{Y}_{cxy}\right), Y_{cxy}=1\\ \left(1-Y_{cxy}\right)^{\beta}\left(\hat{Y}_{cxy}\right)^{\alpha}\log\left(1-\hat{Y}_{cxy}\right), \text{else}\end{cases} \tag{5}$$

其中, N表示真实样本总个数; Y_{cxy}表示热力图真实标签值; $\hat{Y}_{cxy}$表示热力图模型预测值; α和β为超参数, 用于调整正负样本损失之间的比例, 平衡损失权重, 本文取$\alpha=2, \beta=4$.

针对目标二维尺寸损失ℓ_{2s}以及目标二维偏移损失ℓ_{2o}本文采用 L1 损失函数进行损失计算, 公式如下:

$$\ell_1=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{6}$$

其中, N为真实样本总个数; y_i表示样本 2D 边界框尺寸标签值; $\hat{y}_i$表示样本 2D 边界框尺寸预测值. 由上, 2D 检测损失可表示为:

$$\ell_{2d}=\ell_{loc}+\ell_{2s}+\ell_{2o} \tag{7}$$

1.4.2 3D 检测损失函数

3D 检测损失包含角度类别损失ℓ_{ac}、角度偏移损失ℓ_{ao}、目标 3D 边界框尺寸损失ℓ_{3s}、目标 3D 边界框偏移损失ℓ_{3o}.

针对角度类别损失ℓ_{ac}, 本文使用多分类交叉熵损失作为损失函数, 公式如下:

$$\ell_{ac}=\ell_{cross}=\frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}a_{ij}\log\left(\hat{a}_{ij}\right) \tag{8}$$

其中, N表示真实样本总个数; C表示角度类别个数; a_{ij}表示角度类别标签值; $\hat{a}_{ij}$表示角度类别预测值.

针对角度偏移损失ℓ_{ao}、目标三维尺寸损失ℓ_{3s}以及目标三维偏移损失ℓ_{3o}, 本文使用 L1 损失作为损失函数. 由上, 3D 检测损失可表示为:

$$\ell_{3d}=\ell_{ac}+\ell_{ao}+\ell_{3s}+\ell_{3o} \tag{9}$$

1.4.3 深度检测损失函数

深度检测损失ℓ_d分为深度类别标签损失ℓ_{dc}和领域深度差损失ℓ_{do}. 针对深度类别标签损失ℓ_{dc}, 本文采用拉普拉斯任意不确定性损失函数, 计算公式为:

$$\ell_{dc}=\frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{2}\exp(-vre_i)\left|d_i-\hat{d}_i\right|+vre_i\right) \tag{10}$$

其中, N表示真实样本总个数; vre_i表示异方差任意不确定性, 通过无监督的方式进行学习; d_i表示样本深度类别标签值; $\hat{d}_i$表示样本深度类别预测值. 针对领域深度差损失本文采用 L1 损失函数. 由上, 深度损失ℓ_d可表示为:

$$\ell_d=\ell_{dc}+\ell_{do} \tag{11}$$

1.4.4 总损失函数

算法总损失函数的定义如式 (12) 所示:

$$\ell_{total}=\sum_{i=1}^{\tau}w_i(t)\cdot\ell_i(t)=\alpha\ell_{2d}+\beta\ell_{3d}+\lambda\ell_d \tag{12}$$

其中, ℓ_{total}表示模型当前第t轮训练阶段的总损失; 总任务数为τ, 本文取$\tau=3$, 分别为 2D 检测损失ℓ_{2d}、3D 检测损失ℓ_{3d}以及深度检测损失ℓ_d; i表示 3 项任务中的第i项任务; $w_i(t)$表示第i项任务在当前第t轮训练阶段的

损失权重, 本文对应于α、β和λ; ℓ_i表示第i项任务当前阶段的损失函数.

1.4.5 分层级任务学习

为了消除误差放大效应对模型最终性能的影响, 本文遵循文献[11]的方法引入分层级任务学习框架HTL, 结构如图 6 所示. 该框架把模型训练过程分为不同的阶段, 损失权重$w_i(t)$与第i项任务的所有前项任务相关联, 只有在前项任务得到良好训练的基础上第i项任务才开始渐进式训练, 依照前项任务的当前评价动态自适应的控制模型损失权重, 保障训练过程的稳定性的同时消除了早期训练不充分导致的训练误差对后续学习过程的误导.

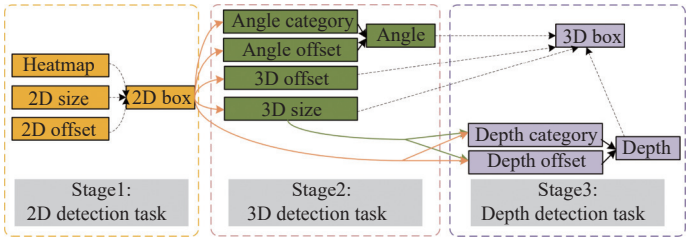

图 6 分层级任务学习框架

2 实验与结果分析

2.1 数据集与评价指标

为验证所提出模型的有效性及泛化性, 本文分别在 KITTI 数据集[22]和 Waymo 数据集[23]上进行模型的实验与评估.

KITTI 数据集为自动驾驶场景中的计算机视觉问题提供了多个测试基准. 针对 3D 目标检测任务, 该数据集提供 7 481 个训练样本和 7 518 个测试样本. 本文遵循文献[9–11]的数据集划分方法将 KITTI 数据集的训练样本划分为包含 3 712 个样本的训练集与包含 3 769 个样本的验证集. 根据 KITTI 数据集官方评估标准, 本文采用平均精度 AP_{3D} 和 AP_{BEV} 定量评估单目 3D 目标检测算法在不同难度等级 (easy、moderate、hard)、不同 3D 交并比 (0.5、0.7) 下针对汽车类别的检测性能.

Waymo 数据集是一个大规模、多样化和具有挑战性的自动驾驶数据集. 该数据集提供 798 个训练序列和 202 个测试序列. 本文遵循文献[24]的数据集划分方法, 通过对完整数据集抽样得到一个包含 52 386 个样本的训练集和一个包含 39 848 个样本的验证集. 根据 Waymo 数据集官方评估标准, 本文采用 AP_{3D} 和 APH_{3D} 定量评估单目 3D 目标检测算法在不同难度等级 (Level_1、Level_2)、不同 3D 交并比 (0.5、0.7)、不同距离 (0–∞ m、0–30 m、30–50 m、50–∞ m) 下针对车辆类别的检测性能. 此外为定性验证所提出算法的检测性能, 本文分别在 KITTI 数据集和 Waymo 数据集中将预测边界框投影到单目图像平面及鸟瞰图平面以供直观检验模型的检测效果.

2.2 实验环境及参数配置

本文所提出的检测模型在云服务器上进行训练, 处理器采用 15 个核心的 AMD EPYC 7543, 显卡选用显存为 24 GB 的 NVIDIA GeForce RTX 3090, 操作系统为 Ubuntu 20.04, 深度学习框架为 PyTorch, 版本为 1.10.0, Python 版本为 3.9, CUDA 版本为 11.3, CUDNN 版本为 8.2.1.

针对 KITTI 数据集, 本文通过顶部裁剪的方式裁剪图像顶端 100 行像素的背景噪声, 之后通过填充的方式将输入图像I的尺寸调整为 1248×288, 以此作为网络的输入进行训练. 针对 Waymo 数据集, 本文通过顶部裁剪的方式裁剪图像顶端 320 行像素的背景噪声, 之后通过填充的方式将输入图像I的尺寸调整为 1920×960, 以此作为网络的输入进行训练. 此外本文设置最小检测深度$d_{\min}$为 0, 最大检测深度$d_{\max}$为 72 m, 深度类别标签个数D为 72 个, 在网络训练时使用 Adam 优化器进行端到端的优化, 初始学习率设置为 0.01, 训练轮数设置为 150 轮, 采用分阶段调整学习率方式, 分别在第 90 轮和第 120 轮对学习率缩小 10 倍进行学习率更新, Batch size 设置为 8.

2.3 实验结果及分析

2.3.1 定量评估

为评估所提出的网络模型的 3D 目标检测性能, 本文在官方评价指标下将该模型与目前主流的单目 3D

目标检测模型在 KITTI 测试集上进行对比分析, 实验结果如表 1 所示. 为便于观察, 本文在同种评价指标下, 对检测性能排名第 1 的算法用粗体表示, 排名第 2 的算法用下划线表示.

表 1 不同算法基于 KITTI 测试集中车辆类别的检测性能对比

Method	Extra data	AP_{3D}@IoU=0.7 (%)			AP_{BEV}@IoU=0.7 (%)			Runtime (ms)
		Easy	Moderate	Hard	Easy	Moderate	Hard	
DDMP-3D[5]	Depth	19.71	12.78	9.80	28.08	17.89	13.44	180
MonoDTR[9]	Depth	21.99	15.39	12.73	28.59	20.38	17.14	37
MonoDistill[25]	Depth	22.97	16.03	13.60	31.87	22.59	19.72	40
CaDDN[6]	Lidar	19.17	13.41	11.46	27.94	18.91	17.19	630
PatchNet-C[26]	Lidar	22.40	12.53	10.64	—	—	—	—
DD3D[27]	Lidar	23.22	16.34	14.20	30.98	22.56	20.03	—
CMAN[28]	None	17.77	11.87	9.16	25.89	17.04	12.88	—
MonOAPC[26]	None	18.77	12	9.75	28.91	19.67	16.99	35
DEVIANT[24]	None	21.88	14.46	11.89	29.65	20.44	17.43	—
MonoFlex[8]	None	19.94	13.89	12.07	28.23	19.73	16.89	35
GUPNet[11]	None	22.26	15.02	13.12	30.29	21.19	18.20	34
MonoCon[10]	None	<u>22.50</u>	**16.46**	<u>13.95</u>	<u>31.12</u>	<u>22.10</u>	<u>19.00</u>	**26**
Ours	None	**23.82**	<u>16.29</u>	**14.36**	**32.65**	**23.08**	**20.11**	<u>32</u>
Improvement	—	+1.32	−0.17	+0.41	+1.53	+0.98	+1.11	+6

由表 1 可知, 本文提出的模型在 KITTI 测试集 3 个难度指标上取得了汽车类别的综合最佳检测性能. 与同类对比算法中综合性能排名第 1 的方法 MonoCon 相比, 所提出的模型在简单、中等、困难 3 个级别上 AP_{3D} 的精度分别提升了 1.32 个百分点、−0.17 个百分点、0.41 个百分点, AP_{BEV} 的精度分别提升了 1.53 个百分点、0.98 个百分点、1.11 个百分点, 因此在多种难度下的综合 AP_{3D} 和 AP_{BEV} 分别提升了 1.56 个百分点、3.62 个百分点, 验证了所提出算法的有效性. 由于 MFM 模块在初始采样阶段进行大量的上采样和特征融合操作以尽可能提取足够贴合目标实际尺寸的语义信息, 因此本文所提算法的检测速度略低于 MonoCon, 尽管如此, 对比其他算法在实时性方面仍有较强的竞争力. 此外, 本文所提出的算法在不使用任何额外数据的情况下, 多种难度下的综合 AP_{3D} 和 AP_{BEV} 优于使用深度图作为额外信息的 MonoDistill 算法和使用点云作为额外信息的 DD3D 算法, 有效证明了本文提出的深度尺度相关化模块在深度和尺度推理方面的显著作用.

为进一步验证所提出模型的检测性能, 本文针对不同的 IoU 阈值, 将所提出的算法与若干同类算法在 KITTI 验证集上进行性能对比, 结果如表 2 所示.

表 2 不同算法基于 KITTI 验证集中车辆类别的检测性能对比 (%)

Method	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7			AP_{3D}@IoU=0.5			AP_{BEV}@IoU=0.5		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
SVDM[29]	22.25	14.62	15.26	31.16	22.24	23.18	—	—	—	—	—	—
GUPNet[11]	22.76	16.46	13.72	31.07	22.94	19.75	57.62	42.33	37.59	61.78	47.06	40.88
CaDDN[6]	23.57	16.31	13.84	—	—	—	—	—	—	—	—	—
Ground-Aware[30]	23.63	16.16	12.06	—	—	—	60.92	42.18	32.02	—	—	—
MonOAPC[31]	24, 58	<u>16.94</u>	13.92	<u>34.40</u>	<u>24.85</u>	<u>21.11</u>	—	—	—	—	—	—
DEVIANT[24]	24.63	16.54	<u>14.52</u>	32.60	23.04	19.99	<u>61.00</u>	<u>46.00</u>	<u>40.18</u>	<u>65.28</u>	<u>49.63</u>	<u>43.50</u>
Ours	**24.68**	**18.76**	**15.01**	**34.62**	**25.11**	**21.14**	**64.73**	**46.31**	**40.20**	**69.11**	**52.34**	**45.28**

由表 2 可知, 在阈值为 0.7 的严格条件下, 模型 3 种难度级别的 AP_{3D} 分别提高了 0.05 个百分点、2.22 个百分点、0.49 个百分点, AP_{BEV} 分别提高了 2.02 个百分点、2.07 个百分点、1.15 个百分点. 在阈值为 0.5 的宽松条件下, 模型 3 种难度级别的 AP_{3D} 分别提高了 3.73 个百分点、0.31 个百分点、0.02 个百分点, AP_{BEV} 分别提高了 3.83 个百分点、2.71 个百分点、1.78 个百分点, 结果再次验证了本文所提出的算法的有效性. 在

对多尺度、不规则目标具有强大特征提取能力的多尺度特征融合模块的加持下, 模型依托于深度离散化网络提供的精细化深度特征、尺度归一化网络提供的距离相关的尺度特征极大提高了对单目图像中的目标进行精确 3D 检测与定位的能力.

为进一步验证本文所提出模型的泛化性能, 本文在官方评价指标下将该模型与目前主流的单目 3D 目标检测模型在 Waymo 验证集上进行对比分析, 实验结果如表 3 所示.

由表 3 可以观察到, 与同类算法相比, 本文所提出的算法在多种难度下均取得了更为优异的检测性能, 具体来说, 相较于同种指标, 整体 AP_{3D} 提升 3.07 个百分点, 整体 APH_{3D} 提升 2.4 个百分点, 与表 1、表 2 观察到的结果一致, 由此验证了本文所提出的算法具有较强的泛化能力. 然而算法在 50 m 以外的区域并未取得最佳的检测效果, 原因在于本文采用的深度预测方法是最终反向传播的方式对目标深度进行学习, 并未同 MonoJSG 算法一样在深度生成阶段进行辅助监督. 尽管如此, 本文所构建的深度离散化网络仍发挥了较强的深度预测功能, 对比其余算法, 本文算法仍在 50 米以外的距离取得了第 2 名的检测性能, 有效证明了算法对远离目标的定位能力.

表 3 不同算法基于 Waymo 验证集中车辆类别的检测性能对比

IoU_{3D}	Difficulty	Method	AP_{3D} (%)				APH_{3D} (%)			
			Overall	0–30 m	30–50 m	50–∞ m	Overall	0–30 m	30–50 m	50–∞ m
0.7	Level_1	MonoJSG[32]	0.97	4.65	0.55	**0.10**	0.95	4.59	0.53	**0.09**
		GUPNet[11]	2.28	6.15	0.81	0.03	2.27	6.11	0.80	0.03
		DEVIANT[24]	2.69	6.95	0.99	0.02	2.67	6.90	0.98	0.02
		Ours	**2.93**	**7.42**	**1.02**	0.03	**2.89**	**7.38**	**0.99**	0.03
0.7	Level_2	MonoJSG[32]	0.91	4.64	0.55	**0.09**	0.89	4.65	0.53	**0.09**
		GUPNet[11]	2.14	6.13	0.78	0.02	2.12	6.08	0.77	0.02
		DEVIANT[24]	2.52	6.93	0.95	0.02	2.50	6.87	0.94	0.02
		Ours	**2.71**	**7.22**	**0.99**	0.03	**2.67**	**7.04**	**0.97**	0.03
0.5	Level_1	MonoJSG[32]	5.65	20.86	3.91	**0.97**	5.47	20.26	3.79	**0.92**
		GUPNet[11]	10.02	24.78	4.84	0.22	9.94	24.59	4.78	0.22
		DEVIANT[24]	10.98	26.85	5.13	0.18	10.89	26.64	5.08	0.18
		Ours	**11.33**	**27.64**	**5.78**	0.24	**11.06**	**27.24**	**5.51**	0.24
0.5	Level_2	MonoJSG[32]	5.34	20.79	3.79	**0.85**	5.17	20.19	3.67	**0.82**
		GUPNet[11]	9.39	24.69	4.67	0.19	9.31	24.50	4.62	0.19
		DEVIANT[24]	10.29	26.75	4.95	0.16	10.20	26.54	4.90	0.16
		Ours	**10.96**	**27.12**	**5.43**	0.21	**10.76**	**26.98**	**5.43**	0.21

2.3.2 定性评估

为直观地展示所提出的算法的检测性能, 本文使用经过训练集训练的模型对 KITTI 验证集和 Waymo 数据集进行推理, 并将预测边界框投影到单目图像平面以及鸟瞰图平面以供直观检验模型的检测效果. 检测效果如图 7 所示.

图 7 为经过训练的模型针对 8 幅不同单目图像的检测效果图. 顶部 4 幅图像为针对 KITTI 数据集的检测效果图, 底部 4 幅图像为针对 Waymo 数据集的检测效果图. 每幅效果图左侧部分为 3D 目标预测框的在单目图像中效果示例, 右侧部分表示目标相对位置的 BEV 效果图. 从单目图像效果示例可以看出, 本文所提出的单目 3D 目标检测算法推理出的 3D 目标预测框对目标具有高度的包裹性, 表明模型对目标尺度及朝向具有较高的预测精度, 且由 BEV 视图可以看出目标相对位置基本准确, 表明模型在深度离散化网络提供的细粒度深度特征的加持下模型对目标深度的推理能力优秀. 此外, 从 BEV 效果图中可以观察到该算法能够准确检测出数据集中严重被遮挡的目标和未标注的远距离小目标, 这主要归因于本文提出的多尺度特征融合模块 MFM 以及深度尺度相关化模块 DSCM. 其中 MFM 在特征提取阶段由于采用了可变形卷积, 因而特征采样更贴合被遮挡目标的原始尺寸, 保障了模型对不同遮挡程度目标的检测性能. 而 DSCM 将不同距离的目标通过几何约束像素级地放缩到同一水平, 极大平衡了模型对不同距离目标的关注度, 减少了小目标误检及漏检现象的发生, 显著缓解了目标尺度差异对检测效果造成的影响. 由此, 实验结果定性地证明了所提出

算法的有效性.

2.4 消融实验

为验证本文所提出模型每个组成部分的有效性, 本小节通过控制变量的方式在 KITTI 验证集上进行消融实验, 记录由模型调整引起的 AP_{3D} 和 AP_{BEV} 的改变, 并进行分析评估.

图 7 KITTI 验证集中车辆类别检测效果图

2.4.1 多尺度特征融合模块

为验证多尺度特征融合模块 MFM 对检测性能带来的影响, 本文分别将 DLA-34 网络和 MFM 作为主干特征提取网络, 在其余参数相同的情况下进行对比实验, 结果如表 4 所示. 从表中可以看出, 多尺度特征融合模块 MFM 的检测精度更高, 原因在于 DCN 在进行卷积的同时在为卷积核引入了偏移向量, 更能适应多尺度、不规则的目标, 使得对特征的提取更贴近目标真实形状, 并且在两个多级特征融合模块 HFN 和 IFN 的帮助下, 模型能够显著聚合图像中的有效特征, 使得采样特征更具表征能力. 此外 GN 对数据批量大小不敏感, 降低了小批量数据训练对模型最终性能的影响, 因此模型训练更具鲁棒性.

表 4 多尺度特征融合模块对检测性能的影响 (%)

Optional module	AP_{3D}/AP_{BEV}		
	Easy	Moderate	Hard
DLA-34	23.83/33.15	17.61/24.48	14.45/19.98
MFM	24.68/34.62	18.76/25.11	15.01/21.14

2.4.2 深度离散化网络

该部分分别将深度回归 depth regression 和深度离散 depth discretization 作为深度推理方法来验证深度离散化网络 DDN 的作用. 结果如表 5 所示, 深度离散化模块提升了模型的检测性能, 这是由于该模块将连续的深度值离散为深度类别标签和领域深度差, 从而把一个具体的深度值划分到一个对应的深度类别之中, 同时解决了分类带来的量化误差, 在推理过程中通过对两者推理结果反离散化得到比深度回归更加精确的

深度结果, 从而提升了模型检测精度.

2.4.3 尺度归一化网络

该部分在所提出完整模型的基础上剔除了尺度归一化网络 SNN, 实验结果如表 6 所示. 从表中可以发现, 在缺失尺度归一化网络提供的尺度特征的情况下模型的检测性能有一定幅度的下降, 尤其针对困难目标的检测性能下降明显. 造成性能降低的主要原因是因为模型脱离了深度对目标尺度的监督, 因而对远距离小目标的识别能力下降, 由此得证本文所提出的尺度归一化网络在平衡模型对不同距离目标的关注度方面的有效作用.

表 5 深度离散化网络对检测性能的影响 (%)

Optional module	AP_{3D}/AP_{BEV}		
	Easy	Moderate	Hard
Depth regression	22.35/31.32	16.31/22.03	13.38/18.96
Depth discretization	24.68/34.62	18.76/25.11	15.01/21.14

表 6 尺度归一化网络对检测性能的影响 (%)

Optional module	AP_{3D}/AP_{BEV}		
	Easy	Moderate	Hard
Without SNN	22.76/32.02	16.79/22.98	13.57/19.01
With SNN	24.68/34.62	18.76/25.11	15.01/21.14

3 结语

为解决单目图像中不同距离目标的尺度差异影响 3D 目标检测算法性能的问题, 本文提出了一种基于融合采样和深尺约束的单目 3D 目标检测算法. 该算法在构建多尺度特征融合模块以增强模型对不同尺度特征的拟合能力的基础上, 通过透视投影原理构建目标尺度与深度之间的几何约束关系, 以尺度放缩的方式平衡了模型对不同距离目标的关注度. 基于 KITTI 数据集和 Waymo 数据集的检测结果验证了本文所提出算法的有效性. 然而尽管本文所提出的算法相较于同类算法有更好的表现, 但相较于点云, 基于单目图像的检测算法能够直接获得的物理信息有限, 因此需要更多的样本数据用于训练以更准确地预测待检测目标, 然而相关数据集难以获取的问题在一定程度上阻碍了技术的迭代, 因此在未来可考虑无监督学习对模型进行训练.

参考文献

1 余以春, 李明旭. 改进 YOLOv5s 的自动驾驶汽车目标检测. 计算机系统应用, 2023, 32(9): 97–105. [doi: 10.15888/j.cnki.csa.009198]

2 Mao JG, Shi SS, Wang XG, *et al*. 3D object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision, 2023, 131(8): 1909–1963. [doi: 10.1007/s11263-023-01790-1]

3 李熙莹, 叶芝桧, 韦世奎, 等. 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, 2023, 28(6): 1709–1740. [doi: 10.11834/jig.230036]

4 Hu SM, Zhang FL, Wang M, *et al*. PatchNet: A patch-based image representation for interactive library-driven image editing. ACM Transactions on Graphics (TOG), 2013, 32(6): 196.

5 Wang L, Du L, Ye XQ, *et al*. Depth-conditioned dynamic message propagation for monocular 3D object detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 454–463.

6 Reading C, Harakeh A, Chae J, *et al*. Categorical depth distribution network for monocular 3D object detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8555–8564.

7 Brazil G, Pons-Moll G, Liu XM, *et al*. Kinematic 3D object detection in monocular video. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 135–152.

8 Zhang YP, Lu JW, Zhou J. Objects are different: Flexible monocular 3D object detection. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 3288–3297.

9 Huang KC, Wu TH, Su HT, *et al*. MonoDTR: Monocular 3D object detection with depth-aware transformer. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4002–4011.

10 Liu XP, Xue N, Wu TF. Learning auxiliary monocular contexts helps monocular 3D object detection. Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 1810–1818.

11 Lu Y, Ma XZ, Yang L, *et al*. Geometry uncertainty projection network for monocular 3D object detection. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3091–3101.

12 Yan C, Salman E. Mono3D: Open source cell library for monolithic 3-D integrated circuits. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65(3):

1075–1085. [doi: 10.1109/TCSI.2017.2768330]

13 Kundu A, Li Y, Rehg JM. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3559–3568.

14 Brazil G, Liu XM. M3D-RPN: Monocular 3D region proposal network for object detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9286–9295.

15 孙延康, 王璇之, 封澳, 等. 基于多尺度融合和高阶交互的单目 3D 检测算法. 计算机技术与发展, 2024, 34(10): 38–45.

16 孙逊, 冯睿锋, 陈彦如. 基于深度与实例分割融合的单目 3D 目标检测方法. 计算机应用, 2024, 44(7): 2208–2215.

17 Yu F, Wang DQ, Shelhamer E, *et al*. Deep layer aggregation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2403–2412.

18 Dai JF, Qi HZ, Xiong YW, *et al*. Deformable convolutional networks. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 764–773.

19 Wu YX, He KM. Group normalization. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

20 Ioffe S, Szedegy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 448–456.

21 Tang YL, Dorn S, Savani C. Center3D: Center-based monocular 3D object detection with joint depth understanding. Proceedings of the 42nd DAGM German Conference on Pattern Recognition. Tübingen: Springer, 2021. 289–302.

22 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the 2012 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361.

23 Sun P, Kretzschmar H, Dotiwalla X, *et al*. Scalability in perception for autonomous driving: Waymo open dataset. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2446–2454.

24 Kumar A, Brazil G, Corona E, *et al*. DEVIANT: Depth EquiVarIAnt network for monocular 3D object detection. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 664–683.

25 Chong ZY, Ma XZ, Zhang H, *et al*. MonoDistill: Learning spatial features for monocular 3D object detection. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

26 Simonelli A, Bulò SR, Porzi L, *et al*. Are we missing confidence in pseudo-LiDAR methods for monocular 3D object detection? Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3225–3233.

27 Park D, Ambruş R, Guizilini V, *et al*. Is pseudo-lidar needed for monocular 3D object detection? Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3142–3152.

28 Cao YZH, Zhang H, Li YD, *et al*. CMAN: Leaning global structure correlation for monocular 3D object detection. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 24727–24737. [doi: 10.1109/TITS.2022.3205446]

29 Shi YG. SVDM: Single-view diffusion model for pseudo-stereo 3D object detection. arXiv:2307.02270, 2023.

30 Liu YX, Yuan YX, Liu M. Ground-aware monocular 3D object detection for autonomous driving. IEEE Robotics and Automation Letters, 2021, 6(2): 919–926. [doi: 10.1109/LRA.2021.3052442]

31 Yao HD, Chen J, Wang Z, *et al*. Occlusion-aware plane-constraints for monocular 3D object detection. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(5): 4593–4605. [doi: 10.1109/TITS.2023.3323036]

32 Lian Q, Li PL, Chen XZ. MonoJSG: Joint semantic and geometric cost volume for monocular 3D object detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1070–1079.

(校对责编: 王欣欣)