doi:10.3969/j.issn.1001-4616.2025.06.013

基于深度融合与噪声扰动增强的两阶段单目 3D 目标检测

章　友[1],伊　君[2],余　烨[1]

(1.合肥工业大学计算机与信息学院,安徽 合肥 230601)
(2.黄冈师范学院,湖北 黄冈 438000)

[**摘要**]　单目 3D 目标检测是自动驾驶系统中的重要技术之一,随着自动驾驶需求的增加,单目 3D 目标检测受到了越来越多的重视. 然而,从单幅图像中精确定位 3D 目标是一个极具挑战性的问题,一方面,深度信息估计的精度还有待提升,另一方面,现有方法通常采用 3D 与 2D 检测分支联合训练的策略,这种耦合方式限制了 2D 检测分支的性能优化. 为解决上述问题,本文提出了一种基于深度融合与噪声扰动增强的两阶段单目 3D 目标检测方法. 该方法设计了一种深度信息融合机制,通过评估不同深度估计结果的可靠性,采用自适应加权策略进行深度融合,显著提升了深度估计精度. 同时,提出了一种解耦式训练策略,将 3D 与 2D 检测分支独立训练,并在 2D 检测分支中引入噪声扰动,通过数据增强的方式强化 2D 特征提取能力,从而为 3D 检测提供更可靠的 2D 信息支持. 在 KITTI 数据集上对本文提出的模型进行了验证,结果显示该模型对车辆目标识别较好,对于数据集中不同难度的目标都取得了较好的效果.

[**关键词**]　3D 目标检测,噪声扰动,特征增强,深度融合
[**中图分类号**]O643/X703　[**文献标志码**]A　[**文章编号**]1001-4616(2025)06-0121-08

Two-Stage Monocular 3D Object Detection Based on Deep Fusion and Noise Perturbation Enhancement

Zhang You[1],Yi Jun[2],Yu Ye[1]

(1.School of Computer and Information,Hefei University of Technology,Hefei 230601,China)
(2.Huanggang Normal University,Huanggang 438000,China)

Abstract:Monocular 3D object detection is one of the important technologies in autonomous driving systems. With the increasing demand for autonomous driving, monocular 3D object detection has received more and more attention. However,accurately locating 3D objects from a single image is a very challenging problem. On the one hand,the accuracy of depth information estimation needs to be improved. On the other hand,existing methods usually adopt a strategy of joint training of 3D and 2D detection branches. This coupling method limits the performance optimization of the 2D detection branch. To solve the above problems,this paper proposes a two-stage monocular 3D object detection method based on depth fusion and noise perturbation enhancement. This method designs a depth information fusion mechanism. By evaluating the reliability of different depth estimation results,an adaptive weighting strategy is used for depth fusion, which significantly improves the depth estimation accuracy. At the same time,a decoupled training strategy is proposed to train 3D and 2D detection branches independently,introduce noise perturbations in the 2D detection branch,and enhance the 2D feature extraction capability by data enhancement,thereby providing more reliable 2D information support for 3D detection. The model proposed in this paper is verified on the KITTI dataset. The results show that the model has good vehicle target recognition and has achieved good results for targets of different difficulty levels in the dataset.

Key words:3D object detection,noise perturbation,feature enhancement,depth fusion

收稿日期:2025-03-11.
基金项目:安徽省自然科学基金面上资助项目(2308085MF216)、国家自然科学基金面上资助项目(62372153).
通讯作者:伊君,博士,讲师,研究方向:计算机视觉、多媒体技术与三维重建. E-mail:junyi@ hgnu.edu.cn

近年来,随着智能驾驶需求的日益增长,作为其核心技术之一的 3D 目标检测[1]取得了快速发展. 根据传感器类型的不同,现有 3D 目标检测方法主要可分为三类:基于激光雷达、立体视觉和单目视觉的解决方案. 尽管激光雷达[2]凭借其精确的深度测量能力在检测性能上具有显著优势,但其设备部署和计算成本较高,限制了其广泛应用. 立体视觉方案虽然在成本效益上有所改善,但仍面临硬件部署复杂、系统兼容性差等实际问题. 相比之下,基于单目视觉的 3D 目标检测方法凭借其硬件部署简单、成本低廉且易于集成的特点,已成为当前计算机视觉领域的研究热点. 该方法的核心挑战在于如何从单张 RGB 图像中准确估计目标的 3D 位置和姿态信息,其成功突破将为智能驾驶系统的普及提供重要的技术支撑.

基于单目图像的 3D 目标检测方法主要可分为两大类:第一类是仅依赖单目图像信息的纯图像方法,这类方法直接从 RGB 图像中估计目标的 2D 和 3D 信息,并基于这些信息推导出目标的 3D 边界框;第二类是基于额外信息辅助的方法,这类方法通过引入深度估计、几何约束等额外信息来增强模型对目标 3D 信息的理解能力,常用的辅助信息包括预训练模块、关键点标注以及 CAD 模型等.

基于纯图像的 3D 检测方法主要依赖于深度神经网络强大的特征提取能力,可分为两类:一是将成熟的 2D 检测器扩展至三维空间,二是直接设计端到端的 3D 检测器. 代表性研究工作包括:Duan 等[3]提出了一种无锚单目检测器 CenterNet,并将其扩展到基于图像的 3D 检测,其核心思想是以目标对象编码为中心点,并通过关键点估计定位目标位置,同时采用多分支并行结构预测目标的尺寸、方向等 3D 属性. Deep3DBox[4]则在传统 2D 检测器的基础上,增加了 3D 方向估计和边界框尺寸回归模块,直接通过网络预测确定 3D 边界框的空间参数. MonoMAE[5]增强了特征表示,使模型能够更好地捕捉空间关系和对象几何形状. WeakMono3D[6]结合了投影和多视图一致性,通过两次一致性损失来指导 3D 边界框的预测. 他们还提出了一个二维方向标签来取代点云数据上标记的三维旋转标签. MonoATT[7]实现了视频流的高效在线处理,显著提高了在具有挑战性的场景中的检测精度,如遮挡和不同的对象尺度,同时保持了适用于资源受限环境的计算效率. MonoPGC[8]引入了像素深度估计作为辅助任务,并设计了一个深度交叉注意力金字塔模块,将局部和全局深度几何知识注入视觉特征中.

尽管这些方法显著提升了单目 3D 检测的精度和鲁棒性,但在处理复杂场景(如目标遮挡、光照变化)以及不完整目标的检测方面仍存在明显不足.

基于额外信息辅助的方法则聚焦于如何利用额外信息来帮助模型更好地理解对象的 3D 信息. Ma 等[9]使用预训练模块获得深度信息和伪点云信息. Mono3D++[10]首先使用对象高度作为先验,并通过将深度与结构化多边形相结合来获得粗略的 3D 框. 然后,它用鸟瞰图纠正了粗略的 3D 框. OVM3D[11]自动将图像与标签 3D 对象组合在一起,以利用互联网规模的数据. MonoNeRD[12]使用神经辐射场模型来实现精确的 3D 感知,并采用体绘制来恢复 RGB 图像和深度图. AutoShape[13]采用关键点几何约束进行 2D/3D 回归. 他们提出了一种方法,可以自动将 3D 形状拟合到视觉观察中,然后为训练网络生成 2D/3D 关键点对的真实注释. YOLOBU[14]使用可变形 DETR 模型和交叉注意力机制来建立像素连接以进行检测. 虽然额外的信息有助于提高模型的检测性能,但推理速度较慢成本较高.

尽管基于单目图像的 3D 目标检测方法已取得很大进展,然而,其仍面临着如下挑战:(1)在深度信息估计方面,精确度仍有待提升;(2)现有方法通常采用 3D 检测分支与 2D 检测分支联合训练的策略,这种耦合方式限制了 2D 检测分支的性能优化.

针对上述挑战,本文提出了一种基于深度融合与噪声扰动增强的两阶段单目 3D 目标检测方法(a two-stage monocular 3D target detection method based on deep fusion and noise disturbance enhancement, DFND-M3D),该方法的创新性主要体现在两个方面:首先,设计了一种深度信息融合机制,通过评估不同深度估计结果的可靠性,采用自适应加权策略进行深度融合,显著提升了深度估计精度;其次,提出了一种解耦式训练策略,将 3D 检测分支与 2D 检测分支独立训练,并在 2D 检测分支中引入噪声扰动,通过数据增强的方式强化 2D 特征提取能力,从而为 3D 检测提供更可靠的 2D 信息支持.

1 模型总体框架

如图 1 所示,本文设计的是一个两阶段训练的模型,主要由主干特征提取网络 DLA(Deep Layer Aggregation)[15],3D 检测模块 3D Detection 和 2D 检测模块 2D Detection 组成. 在第一阶段,训练 3D

Detection 和 DLA 模块,其中,3D Detection 模块将多种方式获取的深度信息,基于可靠性加权策略进行融合. 在第二阶段,冻结在第一阶段训练好的 DAL 和 3D Detection 模块中的参数,仅训练 2D Detection 模块,同时,在原始特征上加入噪声扰动(Noise Perturbation)来增强特征,帮助 2D Detection 中的预测头(Predict Heads)和姿态检测器(Pose Detector)更好地学习目标特征.

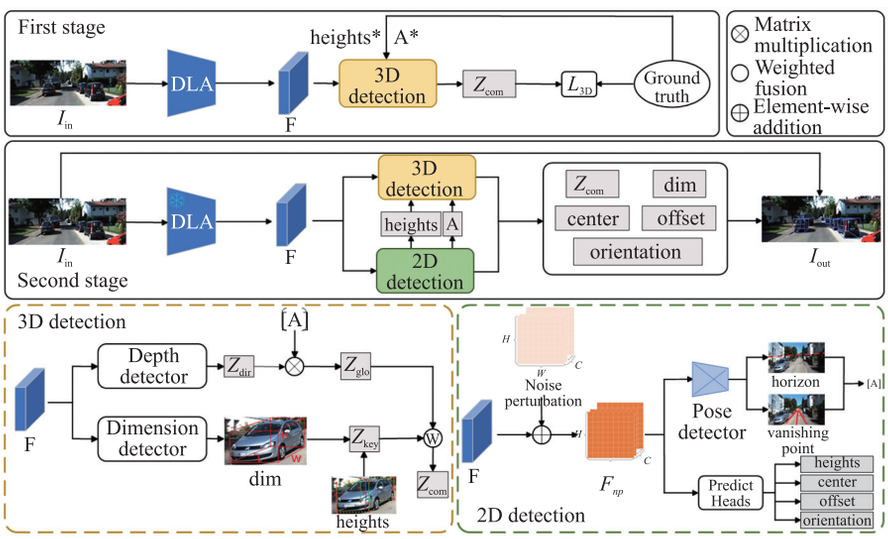

图 1　基于深度融合与噪声扰动增强的两阶段单目 3D 目标检测

Fig. 1　Two-stage monocular 3D object detection based on deep fusion and noise perturbation enhancement

1.1　基于可靠性加权融合的 3D 检测分支

在 3D Detection 模块中,使用两种基于图像的深度信息获取方法,深度检测器(Depth Detector)和维度检测器(Dimension Detector),前者直接估计得到深度,后者先直接估计物体的三维长、宽、高,再结合二维的高度信息,通过公式计算得到深度. 通过 Depth Detector 和 Dimension Detector 分别获得代表全局深度信息的深度图 Z_{dir}和代表局部深度信息的目标深度值 Z_{key}. 对于深度图 Z_{dir},使用变换矩阵 $\boldsymbol{A}$ 对其进行矫正,获得矫正过的深度图 Z_{glo}. 将 Z_{glo}与 Z_{key}基于可靠性加权策略进行融合,进而获得代表了融合全局深度信息和局部深度信息的综合深度 Z_{com}. 详细介绍如下:

1.1.1　基于 Depth Detector 的深度信息估计分支

首先,将主干网络 DLA 提取的特征 F 输入到深度检测器(Depth Detector)中,该检测器用来估计全局深度图与可靠性,从而获得代表全局深度信息的深度图 Z_{dir}和对应点的可靠性图 σ_{dir}.

其次,使用变换矩阵 $\boldsymbol{A}$ 来矫正 Z_{dir}. 在第一阶段训练时,$\boldsymbol{A}$ 来自 Ground Truth 中的 $\boldsymbol{A}^*$,在第二阶段训练和模型预测阶段,使用 2D Detection 中估计得到的变换矩阵 $\boldsymbol{A}$. 首先将深度图 Z_{dir}中坐标为(u,v)的点通过其对应的深度值 z_{uv}转换成三维空间的坐标(x,y,z_{uv}),如下所示:

$$x=\frac{(u-c_u)z_{uv}}{f_x},y=\frac{(v-c_v)z_{uv}}{f_y}, \tag{1}$$

式中,(c_u,c_v)是图像中心点的像素坐标,f_x 和 f_y 分别是 x 轴和 y 轴方向上的像机焦距.

然后,通过转换矩阵 $\boldsymbol{A}$ 将三维空间的点转换为校正后的坐标(x',y',z'_{uv}),公式如下所示:

$$(x',y',z'_{uv})=(x,y,z_{uv})\boldsymbol{A}, \tag{2}$$

再将校正过后的点根据式(1)映射回二维平面,就获得了矫正后的全局深度图 Z_{glo}.

1.1.2　基于 Dimension Detector 的深度信息估计分支

首先,使用 Dimension Detector 估计对象的三维尺寸信息,记为 dim,包含物体的长度 d_L、宽度 d_W 和高度 d_H.

再结合二维高度信息 heights 来计算局部深度 Z_{key}. 在第一阶段训练时,heights 来自 Ground Truth,记为 heights *. 在第二阶段训练和模型预测时,heights 来自于 2D Detection 模块中的估计值. heights 中包含 5 个高度信息(h_1,h_2,h_3,h_4,h_c),分别为 3D 边界框的 4 条支撑线高度和中心高度(如图 2 所示),具体信

息包含以像素为单位的高度值和对应的可靠性分数(当 $heights=heights*$ 时,可靠性分数为 1).

支撑线所在位置的深度可以通过 d_H 和 h_i ($i\in(1,2,3,4,c)$)来计算. 我们将 5 个高度 (h_1,h_2,h_3,h_4,h_c)分为 3 组,每组的两个高度可以计算出物体中心位置的深度,高度的分组方式如图 2 所示,相同颜色的高度分为一组.

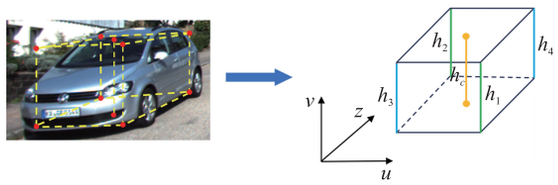

图 2　关键高度分组

Fig. 2　Key heights grouping

3 组中心深度值的计算公式如下:

$$z_i=\frac{f\times d_H}{h_i}, \tag{3}$$

$$z_{d1}=\frac{z_1+z_3}{2}, \tag{4}$$

$$z_{d2}=\frac{z_2+z_4}{2}, \tag{5}$$

式中,f 是像机的焦距.

对应地,3 组可靠性分数的计算公式如下:

$$\sigma_{d1}=\frac{\sigma_1+\sigma_3}{2}, \tag{6}$$

$$\sigma_{d2}=\frac{\sigma_2+\sigma_4}{2}. \tag{7}$$

由式(4),可以获得 z_c,由式(5)、(6)可以获得 z_{d1} 和 z_{d2},为通过不同计算方式获得的中心坐标 (x',y') 处的深度值,记为 $Z_{\text{key}}=\{z_{d1},z_{d2},z_c\}$.

同时,获取深度图 Z_{dir} 在 (x',y') 位置处对应的深度值 $z_{x'y'}$ 以及对应的可靠性分数. 然后根据其可靠性得分为 z_{d1}、z_{d2}、z_c 和 $z_{x'y'}$ 分配不同的权重,即基于可靠性加权策略进行融合,以获得更准确的深度 Z_{com},计算公式如下所示:

$$Z_{\text{com}}=\frac{\sum\limits_{k\in\{c,d1,d2,x'y'\}}\frac{z_k}{\sigma_k}}{\sum\limits_{k\in\{c,d1,d2,x'y'\}}\frac{1}{\sigma_k}}. \tag{8}$$

1.2　基于噪声扰动的 2D 检测分支

为使 3D Detection 和 2D Detection 模块可以共享特征提取主干网络,且 2D Detection 模块可以较好地使用原始共享特征,在第二阶段,使用已训练好的主干网络 DLA 提取出来的特征 F,为 F 添加扰动噪声来增强特征,以帮助 Predict Heads 和 Pose Detector 获得更好的检测效果.

单目图像的深度估计是通过主干网络提取图像中物体的特征,根据相对位置和大小来估计深度,因此主干网络提取出来的特征也是适用于 2D 检测任务的,但这些特征直接用于 2D 检测可能不是最优的. 为解决这两个任务之间的差异,以便于微调优化对于 2D 元素的检测,引入了噪声扰动,在 2D Detection 模块中对原始特征 F 添加噪声扰动,得到添加了噪声扰动后的增强特征 F_{np}. 噪声扰动增强了有目标区域的特征,帮助 2D Detection 和 Pose Detector 更好地关注到目标物体所在区域的特征,因此实现了较好地使用深度任务训练得到的特征 F.

具体来说,噪声扰动是一个可学习的张量,其大小与特征 F 一致,将其与特征 F 按对应元素位置相加,增强对应区域的特征. 噪声扰动学习的是数据集中目标出现区域的分布,目标出现多的区域噪声扰动值越大,能够更好地增强目标所在区域的特征,而背景区域的噪声扰动值较小,因为背景区域不是检测任务所关注的区域.

Pose Detector 获取地平线(horizon)与消失点(vanishing point),从而计算得到像机姿态的变化矩阵 $\boldsymbol{A}$. 地平线是图像中远端地面和天空相交形成的直线,由直线方程 $y=ax+b$ 表示在图像中的位置,消失点则

是由现实中平行的车道线经过透视变换在图像中相交于一点,坐标为(x_{vp},y_{vp}). 由地平线和消失点我们可以得到像机的翻滚角 θ_R 和俯仰角 θ_p,计算过程如式(10)和(11)所示,其中 x_c 是图像的中心坐标:

$$\theta_R=\arctan(a), \tag{9}$$

$$\theta_p=\arctan\frac{x_{vp}-x_c}{f}. \tag{10}$$

由翻滚角 θ_R 和俯仰角 θ_p 计算变换矩阵 $\boldsymbol{A}$ 的过程如下所示:

$$\boldsymbol{A}_R=\begin{bmatrix}\cos\theta_R & -\sin\theta_R & 0\\ \sin\theta_R & \cos\theta_R & 0\\ 0 & 0 & 1\end{bmatrix},\quad \boldsymbol{A}_P=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\theta_P & -\sin\theta_P\\ 0 & \sin\theta_P & \cos\theta_P\end{bmatrix} \tag{11}$$

$$\boldsymbol{A}=\boldsymbol{A}_R\boldsymbol{A}_P \tag{12}$$

最后,2D Detection 模块中的 Predict Heads 输出高度(heights)、2D 中心坐标(center)、偏移量(offset)、朝向角(orientation).

1.3　损失函数

3D 检测分支的损失函数 L_{3D} 由深度图对应的损失函数 L_{dep} 和维度的损失函数 L_{dim} 构成,如下所示:

$$L_{3D}=L_{dep}+L_{dim}=\frac{|Z_{dir}-Z^*|}{\sigma_{dir}}+\log(\sigma_{dir})+\sum_{k\in\{d_H,d_W,d_L\}}|k-k^*| \tag{13}$$

式中,Z^* 表示真实的深度图,σ_{dir}是对应的可靠性得分,k^* 表示物体尺寸的真值.

2D 检测分支的损失函数 L_{pos} 由 2D 中心损失 L_{cen}、偏移量损失 L_{off}、关键高度损失 L_{hei}、朝向角损失 L_{ori} 以及姿态损失 L_{pos} 组成,具体公式如下所示:

$$L_{2D}=L_{cen}+L_{off}+L_{hei}+L_{ori}+L_{pos}=\frac{1}{N}\sum_{i=1}^{N}|x_i-x_i^*|+|y_i-y_i^*|+\frac{1}{N}\sum_{i=1}^{N}|o_i^x-o_i^{x*}|+|o_i^y-o_i^{y*}|+$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|h_i-h_i^*|}{\sigma_i}+\log(\sigma_i)+\frac{1}{N}\sum_{i=1}^{N}|\theta_i-\theta_i^*|+\|\boldsymbol{A}-\boldsymbol{A}^*\|_F, \tag{14}$$

式中,h_i^* 是 2D 高度的对应 GT,σ_i 是检测头估计的对应可靠性得分;(x^*,y^*),(o_i^{x*},o_i^{y*})和 θ_i 分别代表来自 GT 的中心坐标、偏移量和朝向角.

2　实验数据及结果分析

2.1　数据集及评价指标

在基准数据集 KITTI[16] 上对 DFND-M3D 方法进行评价[17]. 该数据集包含 7 481 张图像及其对应的点云和标注数据,通常分为训练集(3 712 张图像)和验证集(3 769 张图像)两个部分. 根据目标的遮挡程度、截断程度和高度,将数据划分为简单(Easy)、中等(Moderate)和困难(Hard)三个难度等级,有助于评估算法在不同调整下的性能. KITTI 数据集为自动驾驶场景提供了多个广泛使用的基准任务,包括 3D 检测、鸟瞰(Bird's Eye View,BEV)检测. 论文对这些任务在三种难度等级下的平均精度(Average Precision,AP)进行性能评估.

在 KITTI 数据集中,标注数据包含汽车、自行车和行人三种类别. 本论文的研究专注于汽车类别,并按照 KITTI 的标准将汽车目标划分为三个难度等级. 汽车类别的评估默认采用 0.7 的交并比(IOU)阈值. 在测试集上进行了实验,以便与以往的研究结果进行公平比较.

2.2　实验细节

采用 DLA-34 作为骨干网络,输入图像分辨率为 1 280×384. 骨干网输出的特征图尺寸为 320×96×64. 模型中的 Depth Detector、Dimension detector 和 Pose Detector 都由一个 3×3×256 卷积层、归一化层、ReLU 激活函数和另一个 1×1×C 卷积层组成,其中 C 为输出通道数. 在训练阶段,使用 Adam 优化器,批处理大小为 8,共 100 个迭代周期. 初始学习率为 3×10^{-4},decay 权重为 1×10^{-5},硬件环境配置为单个 RTX 4090 GPU 的电脑.

2.3　实验结果分析

为验证论文方法的有效性,在 KITTI 的验证集上进行了定量实验. 如表 1 所示,第一的结果使用加粗

字体,第二好的结果使用下划线标出. 在 AP_{3D}上相比于同样不使用额外数据的纯单目图像方法,DFND-M3D 方法在 Easy 和 Hard 标准上获得了最好的效果,分别提升了 0.25%和 0.17%. 这是因为通过噪声扰动来增强目标所在区域的特征,使得检测头能更好地关注目标所在区域的特征. 然而,在 AP_{BEV}上 DFND-M3D 方法在 Mod.和 Hard 标准上获得了最好的效果,分别实现了 1.03%和 1.26%的提升. 这验证了方法中使用的融合方式在获得深度的同时,使用两阶段训练 3D Detection 模块优化了深度估计的性能,从而提升了对于目标检测的准确性.

表 1　在 KITTI 的测试集上与主流方法的 AP_{3D}和 AP_{BEV}结果对比

Table 1　Comparison of AP_{3D} and AP_{BEV} results with mainstream methods on the KITTI test set

Methods	Extra data	Test|AP_{3D}			Test|AP_{BEV}		
		Easy	Mod.	Hard	Easy	Mod.	Hard
DDMP—3D[18]	Depth	19.71	12.78	9.80	28.08	17.89	13.44
Kinematic3D[19]	Video	19.07	12.72	9.17	26.69	17.52	13.10
AutoShape[13]	CAD	22.47	14.17	11.36	30.66	20.08	13.10
MonoRUn[20]		19.65	12.30	10.58	27.94	17.34	15.24
CaDNN[21]	LiDAR	19.17	13.41	11.46	27.94	18.91	17.19
MonoDTR[22]		21.99	15.39	12.73	28.59	20.38	17.14
SMOKE[23]		14.03	9.76	7.84	20.83	14.49	12.75
MonoPair[24]		16.28	12.30	10.42	24.12	18.17	15.76
MonoDLE[25]		17.23	12.26	10.29	24.79	18.89	16.00
MonoFlex[26]	None	19.94	13.89	12.07	28.23	19.75	16.89
MonoGround[27]		21.37	14.36	12.62	30.07	20.47	17.74
MPMonoD[28]		20.08	13.72	11.34	—	—	—
GRAMO[29]		22.34	**15.67**	13.12	**32.44**	21.74	18.38
Ours	None	**22.59**	15.53	**13.29**	32.12	**22.77**	**19.64**

使用不同深度 Z_{key}、Z_{dir}、Z_{glo}和 Z_{com}分别作为 3D Detection 模块的输出,以验证深度融合的贡献,实验结果如表 2 所示. Z_{dir}和 Z_{key}是采用不同方式获取的深度值,其在三个难度级别上的检测性能各有参差. 与 Z_{dir}相比,Z_{glo}在三个难度级别上分别提升了 1.71%、0.39%和 0.62%. 这证明了使用变换矩阵矫正深度图操作的有效性. 与 Z_{glo}相比,Z_{com}有显著的提高,在三个难度级别上分别实现了 5.53%、3.19%和 3.16%的提升,证明了深度融合对于模型性能的贡献.

表 2　不同深度的贡献

Table 2　Contribution of different depths

深度	Val|AP_{3D}/%		
	Easy	Mod	Hard
Z_{key}	18.32	13.87	11.22
Z_{dir}	16.89	13.73	11.64
Z_{glo}	18.60	14.12	12.26
Z_{com}	**24.13**	**17.31**	**15.42**

表 3 显示了两阶段和噪声扰动的影响. 单阶段指在第二阶段同时训练 3D Detection 模块和 2D Detection 模块,而不需要第一阶段的训练. 从实验结果可以观察到,在所有难度水平上,单阶段的表现都低于两阶段. 在都不加入噪声扰动的情况下,两阶段相比单阶段提升了 1.28%、1.06%和 1.11%.这表明两阶段减少了训练过程中 3D Detection 模块与 2D Detection 模块之间的相互影响. 噪声扰动在 AP_{3D}上使模型性能提高了 3.51%、3.05%和 3.06%,这证明了噪声扰动对于特征增强的有效性.

表 3　两阶段和噪声扰动的贡献

Table 3　Contribution of the two-stage and noise perturbation

阶段	噪声扰动	*Val*|AP_{3D}/%		
		Easy	Mod.	Hard
单阶段	不加入	19.34	13.20	11.25
两阶段	不加入	20.62	14.26	12.36
	加入	24.13	17.31	15.42

从图 3 所示的检测结果来看,DFND-M3D 可以为检测周边环境中的汽车目标提供卓越的性能. 模型预测的 3D 包围框可以准确的覆盖大部分汽车目标,甚至一些较小的目标,这反映了噪声扰动对于突出有

目标区域的特征的有效性. 同时,模型在鸟瞰图上对于物体深度定位的准确性也十分明显,可见深度融合以及两阶段训练优化 3D Detection 模块的贡献.

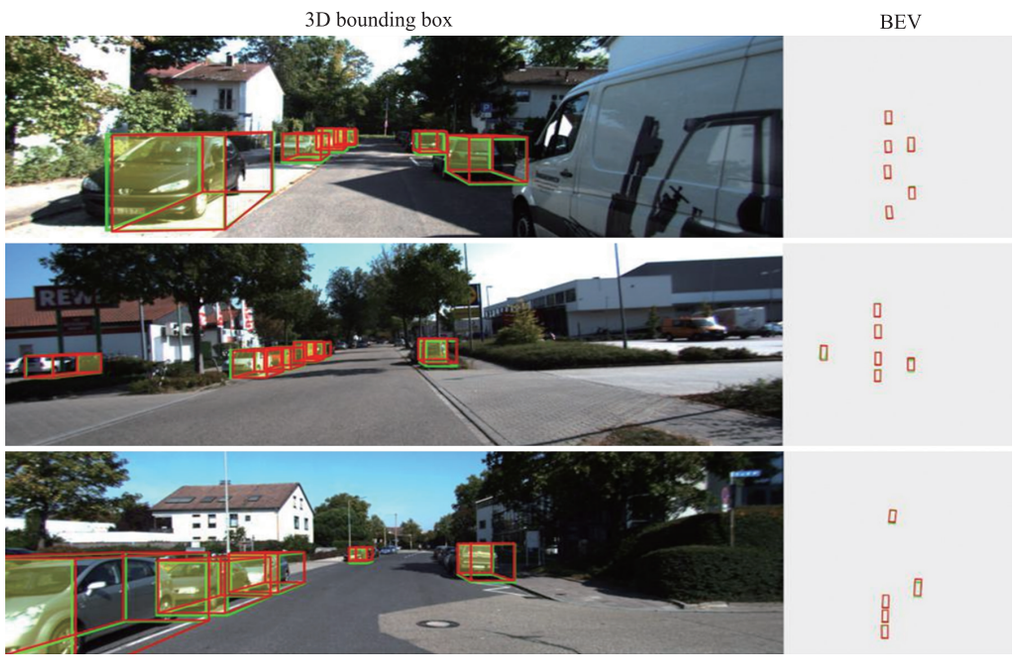

图 3　部分测试结果,左边是 3D 包围框,右边是鸟瞰图检测结果

Fig. 3　Part of the test results, the left side is the 3D bounding box, the right side is the bird's-eye view detection result

3　结论

本文提出了一种基于深度融合与噪声扰动增强的两阶段单目 3D 目标检测方法. 在第一阶段主要训练 3D Detection 和 DLA 模块,该模块以多种方式提取深度信息并通过基于可靠性加权策略进行融合,以提高深度的准确性. 在第二阶段,专注于通过噪声扰动增强特征来提升 2D Detection 模块的性能. 大量的实验结果表明,本文方法在 KITTI 数据集上取得了有竞争力的结果. 然而,检测器估计的可靠性得分可能存在偏差,这会影响融合深度图的准确性. 然而,当使用更加准确的可靠性评分准则时,本文方法性能将会继续得到提升,这也是本论文方法后续研究方向.

[参考文献]

[1]　黄哲,王永才,李德英. 3D 目标检测方法研究综述[J]. 智能科学与技术学报,2023,5(1):7-31.

[2]　刘永刚,于丰宁,章新杰,等. 基于激光点云与图像融合的 3D 目标检测研究[J]. 机械工程学报,2022,58(24):289-299.

[3]　DUAN K, BAI S, XIE L, et al. Centernet: Keypoint triplets for object detection[C]//International Conference on Computer Vision. Seoul: IEEE, 2019: 6569-6578.

[4]　MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]// Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017: 7074-7082.

[5]　JIANG X, JIN S, ZHANG X, et al. MonoMAE: enhancing monocular 3D detection through depth-aware masked autoencoders [C]//Neural Information Processing Systems. Vancouver: Morgan Kaufmann, 2024: 11392-11411.

[6]　TAO R, HAN W, QIU Z, et al. Weakly supervised monocular 3d object detection using multi-view projection and direction consistency[C]//Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 17482-17492.

[7]　ZHOU Y, ZHU H, LIU Q, et al. MonoATT: online monocular 3D object detection with adaptive token transformer[C]// Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 17493-17503.

[8]　WU Z, GAN Y, WANG L, et al. MonoPGC: Monocular 3d object detection with pixel geometry contexts[C]//International Conference on Robotics and Automation. London: IEEE, 2023: 4842-4849.

[9] MA X,WANG Z,LI H,et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]//International Conference on Computer Vision. Seoul:IEEE,2019:6851-6860.
[10] HE T,SOATTO S. Mono3D++:Monocular 3D vehicle detection with two-scale 3D hypotheses and task priors[C]//AAAI Conference on Artificial Intelligence. Hawaii:AAAI Press,2019,33(01):8409-8416.
[11] HUANG R,ZHENG H,WANG Y,et al. Training an open-vocabulary monocular 3d detection model without 3d data[C]//Neural Information Processing Systems. Vancouver:Morgan Kaufmann,2024,37:72145-72169.
[12] XU J,PENG L,CHENG H,et al. Mononerd:Nerf-like representations for monocular 3d object detection[C]//International Conference on Computer Vision. Los Angeles:IEEE,2023:6814-6824.
[13] LIU Z,ZHOU D,LU F,et al. Autoshape:real-time shape-aware monocular 3D object detection[C]//International Conference on Computer Vision. Montreal:IEEE,2021:15641-15650.
[14] XIONG K,ZHANG D,LIANG D,et al. You only look bottom-up for monocular 3d object detection[J]. Robotics and automation letters,2023,8(11):7464-7471.
[15] YU F,WANG D,SHELHAMER E,et al. Deep layer aggregation[C]//Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:2403-2412.
[16] GEIGER A,LENZ P,STILLER C,et al. Vision meets robotics:the kitti dataset[J]. The international journal of robotics research,2013,32(11):1231-1237.
[17] 朱文佳,张婷,程茹秋,等. 考虑背景失真的无参考视频质量评价方法[J]. 南京师大学报(自然科学版),2025,48(3):103-111.
[18] WANG L,DU L,YE X,et al. Depth-conditioned dynamic message propagation for monocular 3d object detection[C]//Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:454-463.
[19] BRAZIL G,PONS-MOLL G,LIU X,et al. Kinematic 3d object detection in monocular video[C]//European Conference on Computer Vision. Glasgow:Springer,2020:135-152.
[20] CHEN H,HUANG Y,TIAN W,et al. Monorun:monocular 3D object detection by reconstruction and uncertainty propagation[C]//Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:10379-10388.
[21] READING C,HARAKEH A,CHAE J,et al. Categorical depth distribution network for monocular 3D object detection[C]//Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:8555-8564.
[22] HUANG K C,WU T H,SU H T,et al. Monodtr:monocular 3D object detection with depth-aware transformer[C]//Conference on Computer Vision and Pattern Recognition. New Orleans:IEEE,2022:4012-4021.
[23] LIU Z,WU Z,TÓTH R. Smoke:single-stage monocular 3D object detection via keypoint estimation[C]//Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:996-997.
[24] CHEN Y,TAI L,SUN K,et al. Monopair:monocular 3D object detection using pairwise spatial relationships[C]//Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:12093-12102.
[25] MA X,ZHANG Y,XU D,et al. Delving into localization errors for monocular 3D object detection[C]//Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:4721-4730.
[26] ZHANG Y,LU J,ZHOU J. Objects are different:flexible monocular 3D object detection[C]//Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:3289-3298.
[27] QIN Z,LI X. Monoground:detecting monocular 3D objects from the ground[C]//Conference on Computer Vision and Pattern Recognition. New Orleans:IEEE,2022:3793-3802.
[28] SHI X,CHEN Z,KIM T K. Multivariate probabilistic monocular 3D object detection[C]//Winter Conference on Applications of Computer Vision(WACV). Hawaii:IEEE,2023:4270-4279.
[29] GUAN H,SONG C,ZHANG Z. GRAMO:geometric resampling augmentation for monocular 3D object detection[J]. Frontiers of computer science,2024,18(5):185706.

[责任编辑:陆炳新]