ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 5,Sep. 2025,pp. 1322－1332
DOI:10. 16337/j. 1004-9037. 2025. 05. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

MonoDI:基于融合深度实例的单目3D目标检测

赵　科，董浩然，业　宁

(南京林业大学信息科学与技术学院、人工智能学院,南京210037)

摘　要:单目3D目标检测旨在定位输入单个2D图像中物体的3D边界框,这在缺乏图像深度信息的情况下是一个极具困难的任务。针对2D图像在推理时的深度信息缺失以及深度图背景噪声干扰导致检测效果不佳的问题,提出一种融合深度实例的单目3D目标检测方法MonoDI。其关键思想在于利用有效的深度估计网络所生成的深度信息结合实例分割掩码得到深度实例,再与2D图像信息融合来帮助物体3D信息的回归。为了更好地利用深度实例信息,设计了一个迭代深度感知注意力融合模块(iterative Depth aware attention fusion module, iDAAFM),将深度实例特征与2D图像特征融合以得到含有物体清晰边界和深度信息的特征表示;另外,在训练和推理过程引入残差卷积结构代替一般的单一卷积结构,以保证网络在处理融合信息时的稳定与高效。同时,设计了一个3D边界框不确定性辅助任务,在训练中帮助任务学习边界框的生成,提高单目3D目标检测任务的精度。在KITTI数据集上对此方法进行验证,实验结果表明,MonoDI在3D目标检测任务中中等难度情况下的车辆类别的检测精度比基线提高了4.41个百分点,且优于MonoCon、MonoLSS等对比方法,同时在KITTI-nuScenes跨数据集实验中取得了较优的结果。

关键词:单目3D目标检测;实例分割;特征融合;残差卷积;辅助学习

中图分类号:TP391　　**文献标志码**:A

MonoDI:Monocular 3D Object Detection Based on Fusing Depth Instances

ZHAO Ke, DONG Haoran, YE Ning

(College of Information Science and Technology & Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China)

Abstract: Monocular 3D object detection aims to locate the 3D bounding boxes of objects in a single 2D input image, which is an extremely challenging task in the absence of image depth information. To address the issues of poor detection performance due to the absence of depth information during inference on 2D images and background noise interference in depth maps, this paper proposes a monocular 3D object detection method called MonoDI, which integrates depth instances. The key idea is to utilize depth information generated by an effective depth estimation network and combine it with instance segmentation masks to obtain depth instances, and then integrate the depth instances with 2D image information to aid in regressing 3D object information. To better use the depth instance information, this paper designs an iterative depth aware attention fusion module (iDAAFM), integrating depth instance feature with 2D image feature to obtain a feature representation with clear object boundaries and depth information.

基金项目:国家重点研发计划资助项目(2016YFD600101)。
收稿日期:2024-09-03;**修订日期**:2025-01-30

Subsequently, a residual convolutional structure is introduced during training and inference to replace the general single convolutional structure to ensure stability and efficiency of the network when processing fused information. Further, we design a 3D bounding box uncertainty auxiliary task to assist the main task in learning the generation of bounding boxes in training and improving the accuracy of monocular 3D object detection. Finally, the effectiveness of the method is validated on the KITTI dataset and experimental results show that the proposed method improves 3D object detection accuracy for the vehicle class at the moderate difficulty level by 4.41 percentage points compared with the baseline, and outperforms comparative methods such as MonoCon and MonoLSS. And it also achieves superior results on the KITTI-nuScenes cross-dataset evaluation.

Key words: monocular 3D object detection; instance segmentation; feature fusion; residual convolution; auxiliary learning

引　　言

3D目标检测是当前机器视觉研究领域中的一个重要课题,它在许多应用场景中都具有重要意义,例如自动驾驶、智能机器人、无人机等。许多早期研究通过使用激光雷达和立体相机等精确的传感器取得了显著的成果。例如,ROI-10D[1]结合深度特征图和估计的密集深度图来回归3D边界框;CMKD[2]使用跨模态知识蒸馏,将LiDAR模态知识转移到图像中。这些方法使用雷达信号提供的准确深度信息或利用立体匹配技术,在3D目标检测任务中展现出了卓越的性能。

然而,使用这些精确的传感器意味着高昂的成本,这对实际应用造成了一定的限制。为了降低成本并提升应用的普及性,一些研究者提出了仅基于图像的单目3D目标检测方法。例如,文献[3]提出了M3D-RPN,它建立了独立的3D区域建议网络,通过深度感知卷积预测目标的3D信息;SMOKE[4]是基于CenterNet[5]构建的端到端的3D检测网络,通过多步解缠提升模型收敛和检测效果;MonoFlex方法[6]使用边缘热图和融合模块改进了遮挡物预测;MonoDLE方法[7]采用从3D投影中心获取中心点的方法降低定位误差。这些方法在不依赖昂贵传感器的情况下,尝试通过单目相机进行3D目标检测。然而,由于单目图像天生缺乏深度提示信息,加上在单目视角下物体可能存在遮挡或角度变化等问题,这些方法往往难以达到理想的效果。对此,文献[8-9]提出了利用深度估计网络生成的深度图信息来生成伪雷达点云的方法,再利用基于3D点云的检测技术进行目标识别,这种方法在一定程度上弥补了单目图像缺乏深度信息的不足,但过于依赖深度生成网络的先进性,忽略了2D图像本身所蕴含的丰富语义信息。如何兼顾2D语义和深度语义是需要考虑的,而众多关于特征融合方法的研究提供了一定的指导。例如,Chen等[10]在RGB-D图像语义分割任务中提出SA-Gate在融合特征时通过分离、去噪、融合的方式减少深度噪声的干扰;Sun等[11]在显著目标检测中提出一个深度敏感的注意力融合模块来消除背景干扰;DAF_Net方法[12]在室内场景语义分割任务中使用深度加权交叉注意力融合方法,通过动态调整深度和RGB特征图上的融合权重来增强网络的3D感知能力;CMX方法[13]在图像分割任务中设计一种跨模态特征矫正方法,通过利用一种模态的特征来校正另一种模态的特征并校准双模态特征;马倩等[14]提出通道融合增强和非局部特征交互方法捕捉长距离特征依赖,提高了尺度变换和遮挡场景下的融合效果。同时,为了加强模型的鲁棒性,辅助学习的有效性得到了验证,MonoCon方法[15]在训练中设计辅助任务并在推理过程中丢弃它们以提高效率和模型鲁棒性。辅助学习与多任务学习的不同之处在于:在训练中设计与主任务相关的辅助任务帮助主任务回归,此过程被称为辅助学习,辅助任务在推理过程中被丢弃。辅助学习已经在计算机视觉等多个领域中展现了有效性。例如,Soomro

等[16]在中级低级表征中应用音素识别辅助监督任务以提高会话语音识别的性能;Liebel等[17]设计场景的全局描述简单的辅助任务,以提高模型对单个场景深度估计和语义分割的性能。通过选择学习任务对,还可以通过无监督的方式[18]执行没有真实标签的辅助学习。Flynn等[19]和Zhou等[20]提出了图像合成网络,通过预测多个相机的相对姿态作为辅助来执行无监督单目深度估计;Du等[21]建议使用余弦相似性作为自适应任务加权,以确定定义的辅助任务何时有效。通过选择适当的辅助任务,不仅可以进一步提高目标检测精度,还能够在不显著增加推理成本的前提下增强模型的鲁棒性。

鉴于现有研究的局限性,为了同时解决2D图像缺乏深度信息和深度图背景噪声干扰的问题,本文将2D图像信息与深度实例信息融合,并设计了迭代深度感知注意力融合模块(iterative Depth aware attention fusion module, iDAAFM),得到同时具有深度与清晰边界的深度实例信息,提高了模型在物体有遮挡和边界模糊情况下的检测效果;同时引入残差卷积检测头来保证模型在训练过程中的稳定性和效率;此外,设计了3D框不确定性辅助任务帮助模型在训练过程中的3D框回归,进一步提高了模型的检测精度;最后,在KITTI数据集[22]的单目3D目标检测和鸟瞰图任务上验证了所提出方法MonoDI的有效性。

1 本文方法

1.1 MonoDI主干

如图1所示,本文提出的MonoDI方法可以分为特征融合模块、3D框回归残差卷积头和3D预测框生成3个部分。

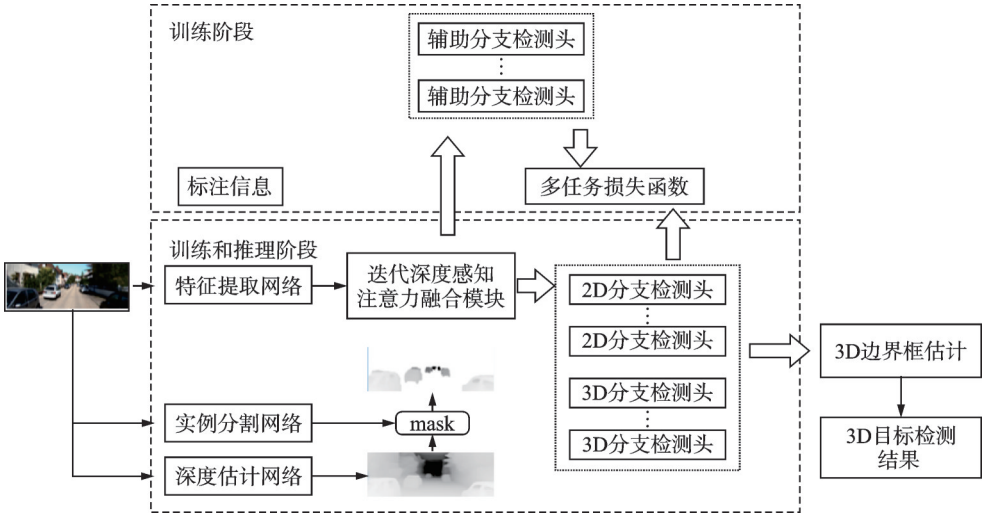


图1 网络结构图

Fig.1 Network structure diagram

1.1.1 特征融合模块

该模块由特征提取网络、实例分割网络、深度估计网络和迭代深度感知注意力融合模块组成。

(1)特征提取网络:假设输入2D图像$\boldsymbol{I}(3\times H\times W)$,通过特征提取网络$f(\cdot;\Theta)$将输入$\boldsymbol{I}$计算为输出特征图$\boldsymbol{F}(D\times h\times w)$,表达式为

$$F=f(\boldsymbol{I};\Theta) \tag{1}$$

式中:Θ表示收集2D图像中的可学习参数,D表示输出特征图的维度(例如,$D=256$),h、w由特征采样

时设置的步幅(采样率)决定。本文使用预训练的深度层聚合网络(Deep layer aggregation,DLA)[23](DLA-34),一个轻量级的、广泛应用于单目3D目标检测任务的特征提取网络。

(2)实例分割网络:输入$\boldsymbol{I}$,通过实例分割网络$q(\cdot;\Phi^{\mathrm{I}})$计算得到实例分割掩码$\boldsymbol{M}(D\times h\times w)$

$$\boldsymbol{M}=q(\boldsymbol{I}_{\Lambda};\Phi^{\mathrm{I}}) \tag{2}$$

式中:$q(\cdot;\Phi^{\mathrm{I}})$使用在CityScapes数据集[24]上预训练的掩码区域卷积神经网络(Mask region-based convolutional neural network,MaskRCNN),由于CityScapes和KITTI的数据集都专注于城市环境,预训练的模型能够更好地适应类似场景的特征,从而提高在KITTI上的表现。

(3)深度估计网络:同样输入$\boldsymbol{I}$,通过深度估计网络$g(\cdot;\Phi^{\mathrm{s}})$计算得到深度图$\boldsymbol{D}_{\mathrm{o}}(H\times W)$

$$\boldsymbol{D}_{\mathrm{o}}=g(\boldsymbol{I};\Phi^{\mathrm{s}}) \tag{3}$$

为了解决单目图像天生缺乏深度信息问题,同时减小模型训练过程中的计算压力,本文使用预训练的DepthAnything-L模型[25]作为深度图生成网络,一种先进的单目深度估计网络。它利用大量未标记的图像语料信息进行训练,在现阶段有着出色的零样本单目深度估计能力。

将$\boldsymbol{D}_{\mathrm{o}}$与$\boldsymbol{M}$相乘得到深度实例$\boldsymbol{D}_{\mathrm{I}}$,最后经过简单的卷积、激活和平均池化操作得到深度实例特征图$\boldsymbol{F}_{\mathrm{di}}$。

$$\boldsymbol{D}_{\mathrm{I}}=\boldsymbol{D}_{\mathrm{o}}\times\boldsymbol{M} \tag{4}$$

$$\boldsymbol{D}_{\mathrm{I}}\xrightarrow{\text{Conv}+\text{ReLU}+\text{Avgpool}}\boldsymbol{F}_{\mathrm{di}} \tag{5}$$

(4)迭代深度感知注意力融合模块:特征融合的目的是为了将深度特征与图像特征结合起来一同完成检测任务。通过融合特征,模型可以同时获得物体的外观和深度信息,这使模型在面对遮挡和角度变化问题时可以利用2D特征识别物体轮廓,利用深度特征识别物体位置与几何形状,从而提高模型在处理遮挡与角度变化问题上的检测效果与鲁棒性。

本文基于迭代注意力特征融合模块(iterative Attention feature fusion,iAFF)[26]设计了一个迭代深度感知的注意力特征融合模块iDAAFM,将iAFF初始的特征相加操作改变为特征拼接,同时保留2D和深度信息,考虑到单目3D目标检测任务对于空间信息的敏感性和输入特征尺度,使用空间注意力机制代替原有的多尺度通道注意力机制,并在最后阶段使用卷积注意力模块(Convolutional block attention module,CBAM)[27]以通过更精细的注意力机制改善特征融合过程,从而增强网络性能。其结构如图2所示。

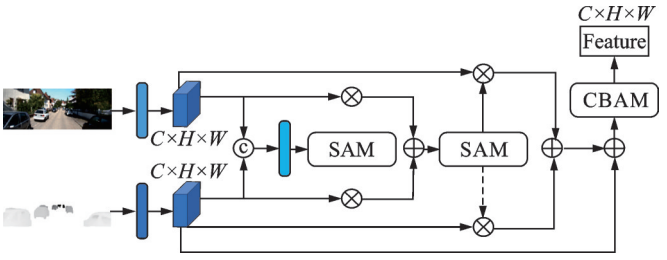


图2　迭代深度感知注意力融合模块

Fig.2　Iterative depth aware attention fusion module

通过特征融合模块iDAAFM($\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2;\Phi^{\mathrm{a}}$)计算得到最终融合后的特征表示$\boldsymbol{F}_{\mathrm{m}}(D\times h\times w)$

$$\boldsymbol{F}_{\mathrm{m}}=\mathrm{iDAAFM}(\boldsymbol{F},\boldsymbol{F}_{\mathrm{di}};\Phi^{\mathrm{a}}) \tag{6}$$

式中:$\boldsymbol{\Lambda}_1$和$\boldsymbol{\Lambda}_2$表示做融合操作的输入特征图,Φ^{a}表示融合模块中的可学习参数。

iDAAFM在开始阶段将$\boldsymbol{F}$与$\boldsymbol{F}_{\mathrm{di}}$通过拼接操作以保留原始拼接特征$\boldsymbol{F}_{\mathrm{c}}(2C\times h\times w)$,然后经过卷积操作得到$\boldsymbol{F}_{\mathrm{c}}'(C\times h\times w)$,再利用空间注意力(Spatial attention module, SAM)计算不同区域$\boldsymbol{F}_{\mathrm{c}}'$的特征权重并将2D图像特征与深度实例特征加权求和,在后续操作中迭代这一步骤以得到更精细的迭代特征表示$\boldsymbol{F}_{\mathrm{c}}^{\mathrm{i}}$。最后将$\boldsymbol{F}_{\mathrm{c}}^{\mathrm{i}}$与$\boldsymbol{F}_{\mathrm{di}}$相加再通过CBAM[27]计算得到最终融合特征$\boldsymbol{F}_{\mathrm{m}}$。

1.1.2 3D框回归残差卷积头

为了能够更好地利用融合深度信息，本文设计了一个具有残差结构的多任务通用检测头 $h(\boldsymbol{\Lambda};\phi)$，用于回归不同的检测任务，表达式为

$$\boldsymbol{H}=h(\boldsymbol{\Lambda};\phi) \tag{7}$$

不同的任务实体有着独立的可学习参数 ϕ，$\boldsymbol{H}$ 是有着任务目标特征的特征图（$c\times h\times w$），c 表示不同回归任务的输出类别（例如在回归2D边界框中心任务中，c 就代表目标类别数，在KITTI数据集中类别数为3，分别是汽车、行人和骑车的人）。

$$h(\boldsymbol{\Lambda};\phi)=m(\boldsymbol{\Lambda})+h_{\mathrm{r}}(\boldsymbol{\Lambda}) \tag{8}$$

式中：$m(\boldsymbol{\Lambda})$ 用来执行线性变换，确保残差连接的维度一致性，调整原始特征维度以允许网络学习输入数据的恒等映射，有助于缓解梯度消失的问题，使得网络可以更深入地学习。$h_{\mathrm{r}}(F_{\mathrm{m}})$ 由卷积层Conv、注意力归一化层AN和激活层ReLU组成。

在2D检测部分，融合特征 F_{m} 被送入尺寸检测头 $\mathcal{S}^{2\mathrm{D}}(2\times h\times w)$、偏移量检测头 $\mathcal{O}^{2\mathrm{D}}(2\times h\times w)$ 和类别热图检测头 $\mathcal{H}^{\mathrm{m}}(3\times h\times w)$，分别用于回归2D尺寸大小 $(h_{\mathrm{a}},w_{\mathrm{a}})$、2D偏移量 $(\Delta x_{\mathrm{a}},\Delta y_{\mathrm{a}})$ 和2D热图中心及其类别 $(x_{\mathrm{a}},y_{\mathrm{a}},C)$。

2D-3D检测部分包括关键点热图检测头 $\mathcal{K}^{\mathrm{m}}(\mathrm{kpts}\times h\times w)$，kpts为预设置的关键点个数，本文设置kpts=9，中心偏移检测头 $\mathcal{O}^{\mathrm{c}}(2{*}\mathrm{kpts}\times h\times w)$ 和关键点热图偏移检测头 $\mathcal{O}^{\mathrm{k}}(2\times h\times w)$，分别回归物体投影关键点的热图 $(x_{\mathrm{b}},y_{\mathrm{b}})$、物体中心到关键点的偏移量 $(\Delta x_{\mathrm{b}},\Delta y_{\mathrm{b}})$ 和投影关键点热图的偏移量 $(\Delta x_{\mathrm{k}},\Delta y_{\mathrm{k}})$。

3D检测部分，将融合特征 F_{m} 输入3D尺寸检测头 $\mathcal{S}^{3\mathrm{D}}(3\times h\times w)$、深度不确定性检测头 $\mathcal{D}^{\mathrm{u}}(2\times h\times w)$ 和方向检测头 $\mathcal{A}^{\mathrm{b}}(\mathrm{bins}\times h\times w)$，分别回归物体的3D尺寸 (h,w,l)、深度信息 (d,u) 和方向信息 θ，其中对于深度信息预测采用不确定性建模思想，两个输出通道分别表示深度和其不确定性。对于方向信息采用离散化角度仓库，bins为预设的仓库数量（例如，bins=12）。其中，在辅助分支中利用标注上下文信息回归辅助任务帮助预测，包括投影关键点热图、投影角点偏移向量、2D边界框尺寸、2D边界框中心和3D边界框不确定性。

3D边界框不确定性辅助任务：在原有3D框的基础上对其不确定性进行预测，假设原有3D框输出为 $\boldsymbol{B}=[x,y,z,l,w,h,\theta]$，其对应不确定性预测输出为 $\boldsymbol{\sigma}=[\sigma_x,\sigma_y,\sigma_z,\sigma_l,\sigma_w,\sigma_h,\sigma_\theta]$，其中 σ_i 对应3D框输出参数 i 的不确定性。对3D边界框进行不确定性预测可以帮助模型更好地理解和处理数据中的不确定性，提高模型对数据噪声的鲁棒性。

1.1.3 3D预测框生成

将融合特征作为多任务检测头的输入，通过类别热图检测头 $\mathcal{H}^{m}$ 计算结果结合非极大值抑制，为每个类别得到一组2D边界框中心。假设得到一个汽车的2D边界框中心 $(x_{2\mathrm{D}},y_{2\mathrm{D}})$，由 $\mathcal{O}^{2\mathrm{D}}$ 计算得出其偏移量 $(\Delta x_{2\mathrm{D}},\Delta y_{2\mathrm{D}})$，那么它的投影3D中心 $(x_{3\mathrm{D}},y_{3\mathrm{D}})=(x_{2\mathrm{D}}+\Delta x_{2\mathrm{D}},y_{2\mathrm{D}}+\Delta y_{2\mathrm{D}})$，对应的深度值 $z=\mathcal{D}^{\mathrm{u}}(x,y)$，根据原始相机参数信息能够计算出汽车的3D中心 (x,y,z)。同理通过其他任务得出汽车的尺寸 (h,w,l) 和观察角度 α。结合以上所有预测参数可以得到最终的3D预测框。

1.2 损失函数

总体损失由2D损失和3D损失组成，其值为各个分支损失加权之和，表达式为

$$L=\sum_k \omega_k L_k^{2\mathrm{D}}+\sum_j \omega_j L_j^{3\mathrm{D}} \tag{9}$$

式中：ω_k 与 ω_j 是预设置的损失权重（本文简单将 ω_j 设置为1，将2D尺寸损失对应 ω_k 设置为0.15），针对

2D和3D的不同任务目标采用不同的损失函数,这些损失函数被广泛使用在单目3D目标检测任务中。在2D检测任务中使用形状尺寸、偏移向量的标准L_1损失函数和用于热图的高斯核加权焦点损失函数[28];在3D检测任务中使用用于深度估计的拉普拉斯偶然不确定性损失函数[29-30]、观测角度中bin指数的标准交叉熵损失函数和3D尺寸的维度感知L_1损失函数;对于3D框不确定性使用高斯负对数似然损失函数[31]。本文简要讨论其中3个。

(1)拉普拉斯偶然不确定性损失:用$\boldsymbol{D}^{\mathrm{t}}(1\times h\times w)$表示真值深度图,其中带注释的3D边界框的真值深度被分配给$(h\times w)$格子中相应的真值2D边界框中心位置,即$\boldsymbol{D}^{\mathrm{t}}(x,y)$(应用与方程中相同的逆S形变换)。采用拉普拉斯分布建模深度不确定性$\boldsymbol{D}^{\mathrm{u}}$。$D_{\mathrm{u}}^{\mathrm{e}}=\boldsymbol{D}^{\mathrm{u}}(x,y)$表示在格点$(x,y)$处预测的不确定性值,$d=\boldsymbol{D}(x,y)$表示格点$(x,y)$处深度的预测值,$d^{\mathrm{t}}=\boldsymbol{D}^{\mathrm{t}}(x,y)$表示在格点$(x,y)$处深度的真值,对于预测深度$\boldsymbol{D}$,损失函数定义为

$$\mathcal{L}(\boldsymbol{D},\boldsymbol{D}^{\mathrm{t}})=\frac{1}{2|S|}\sum_{(x,y)\epsilon S}\frac{(d-d^{\mathrm{t}})2}{D_{\mathrm{e}}^{\mathrm{u}}}+\ln D_{\mathrm{e}}^{\mathrm{u}} \tag{10}$$

式中:S表示全部2D边界框中心点的集合,$|S|$为集合大小。

(2)高斯核加权焦点损失:在回归的热图为$\boldsymbol{\mathcal{H}}(1\times h\times w)$的情况下(例如物体2D边界框中心),假设真值热图$\boldsymbol{\mathcal{H}}^{\mathrm{t}}(1\times h\times w)$与$\mathcal{H}$分辨率相同,对于原始图片中的每一个中心点真值$(x_{\mathrm{c}}^{\mathrm{t}},y_{\mathrm{c}}^{\mathrm{t}})$,它在真值热图中的值为$\left(x_{\mathrm{c}}=\left\lfloor\frac{x_{\mathrm{c}}^{\mathrm{t}}}{s}\right\rfloor,\ y_{\mathrm{c}}=\left\lfloor\frac{y_{\mathrm{c}}^{\mathrm{t}}}{s}\right\rfloor\right)$(其中$s$是特征提取网络的总步长),符号$\lfloor\cdot\rfloor$表示向下取整运算,即取不大于给定数值的最大整数。高斯核$K(x,y)=\exp(-\gamma(\|x-x_{\mathrm{c}}\|^2+\|y-y_{\mathrm{c}}\|^2))$用于对中心点建模,其中$(x,y)$表示热图中坐标位置,$\gamma$使用预设的物体大小自适应方差的倒数。如果两个高斯核重叠,则保留元素最大值。原始图片中所有中心点对应的$K(\cdot,\cdot)$都被折叠到真值热图$\boldsymbol{\mathcal{H}}^{\mathrm{t}}$中。若$\mathcal{H}_{xy}$表示热图坐标$(x,y)$处的预测值,损失函数的定义为

$$L(\boldsymbol{\mathcal{H}},\boldsymbol{\mathcal{H}}^{\mathrm{t}})=-\lambda\sum_{x=1}^{h}\sum_{y=1}^{w}\begin{cases}(1-\mathcal{H}_{xy})^{\delta}\ln\mathcal{H}_{xy} & \mathcal{H}_{xy}^{\mathrm{t}}=1\\ (1-\mathcal{H}_{xy}^{\mathrm{t}})^{\beta}(\mathcal{H}_{xy})^{\delta}\ln(1-\mathcal{H}_{xy}) & \mathcal{H}_{xy}^{\mathrm{t}}\neq 1\end{cases} \tag{11}$$

式中:λ为真值点总数的倒数,δ和β为预定义的超参数(例如$\delta=0.2,\beta=0.4$)。

(3)高斯负对数似然损失:在3D框各个参数的预测值为δ_i的情况下,假设真实值为δ_i^t,损失函数定义为

$$\mathcal{L}(\boldsymbol{B},\boldsymbol{B}^{\mathrm{t}})=\sum_{i=1}^{P}\frac{(\delta_i^{\mathrm{t}}-\delta_i)^2}{\sigma_i^2}+\ln\sigma_i^2 \tag{12}$$

式中:$(\boldsymbol{B},\boldsymbol{B}^{\mathrm{t}})$表示3D边界框参数向量预测值和真值,$\sigma_i$为$i$处参数所对应的不确定性,$P$为3D边界框参数个数。

2　实验与结果分析

2.1　数据集

本文选用了公开的KITTI数据集[22]和nuScenes数据集[32]进行实验。两者都聚焦于自动驾驶车辆在真实道路环境中的感知任务,包括目标检测、跟踪、定位和场景理解,同时有着多模态的数据和高质量标注。KITTI数据集包含7 481张训练图像和7 518张测试图像。但由于其测试集缺少标注信息,本文采用Split1划分协议[33],将原训练数据集划分为3 712张图像组成的训练集和3 769张图像组成的验证集。nuScenes数据集提供了前置摄像头捕获的28 130张训练图像和6 019张验证图像,本文在其验

证集上进行跨数据集评估。

2.2 评估指标

本文遵循KITTI基准测试中的评估指标。对3D边界框和鸟瞰图的平均精度(Average precision, AP)进行评估检测($AP_{3D}|R40$和$AP_{BEV}|R40$),两者在3种难度设置(简单、中等和困难)下都使用了40个召回位置(R40),AP计算时对应预测值与真实值的交并比(Intersection over union, IoU)设置为0.7。

2.3 实验细节和参数设置

本文方法MonoDI在PyTorch框架下实现,其训练过程在单张GPU中进行,设置初始批量大小为8,最大训练周期为200,学习率优化器选择AdamW并设置初始学习率为2e−4,betas参数为(0.95,0.999),权重衰减系数为1e−5,使用单循环调度器将学习率先升高至1e−3,步长为0.4,然后学习率下降至2e−8,本文采用常用的数据增强方法,如随机裁减和水平垂直翻转[5,7]等。实验环境如表1所示。

表1 实验配置

Table 1 Experimental configuration

配置	值
CPU	Intel(R) CORE i9-14900K
内存/GB	64
GPU	Nvidia GeForce 4090*1
系统	Ubuntu22.04
CUDA	11.8

2.4 实验结果分析

本文聚焦模型在汽车类别检测任务中的表现,与现有主流研究方向一致。基于上述实验设置和评估方法,本文在KITTI数据集上测试了模型性能,并将实验结果与DD3D、MonoCon、MonoLSS等先进方法进行了对比分析,实验结果如表2所示。

表2 MonoDI与其他方法基于鸟瞰图和3D边界框的性能比较

Table 2 Performance comparison of MonoDI and other methods based on bird's eye view and 3D bounding boxes

方法	AP_{3D}(IoU=0.7|R40)/%			AP_{BEV}(IoU=0.7|R40)/%			运行时间/
	Easy	Mod	Hard	Easy	Mod	Har	ms
SMOKE[4]	14.76	12.85	11.50	19.99	15.61	15.28	30
D4LCN[34]	22.32	16.20	12.30	31.53	22.58	17.87	200
MonoDIS[35]	18.05	14.98	13.42	24.26	18.43	16.95	40
DD3D[36]	18.45	14.48	12.87	27.15	21.17	18.35	—
MonoDLE[7]	17.45	13.66	11.68	24.97	19.33	17.01	40
GUPNet[37]	22.76	16.46	13.72	31.07	22.94	19.75	35
MonoFlex[6]	23.64	17.51	14.83	—	—	—	35
MonoCon[15]	26.33	19.01	15.98	34.65	25.39	21.93	25
MonoLSS[38]	25.91	18.29	15.94	34.70	25.36	21.84	35
MonoDI(本文)	29.53	22.07	17.74	39.37	28.52	23.91	40

由表2可知,在严格条件(IoU≥0.7)下,本文提出的MonoDI在3D目标检测(AP_{3D})和鸟瞰图检测(AP_{BEV})两个评价指标上,在简单、中等与困难3个难度级别均优于SMOKE、GUPNet、MonoFlex等方法。与表2中的次优方法相比,MonoDI在3D目标检测任务的简单、中等、困难3个级别上分别较MonoCon提升了3.20、3.06和1.76个百分点;在鸟瞰图检测任务的简单级别上,较MonoLSS提升了4.67个百分点,在中等、困难级别上较MonoCon分别提升了3.17个百分点、1.98个百分点。实验结果显

示,本文方法MonoDI在单目3D目标检测任务中表现优异,不但在各个难度级别下维持了较高的检测精度,而且从运行时间来看,本文方法也具有较高的可行性,整体上相比其他先进方法具有显著优势。

2.5　消融实验

本文在KITTI数据集的Split1验证集上进行消融实验,以评估所提方法MonoDI各部分对模型性能的影响。本文提出融合2D图像与深度实例、迭代深度感知注意力融合模块、残差卷积和3D边界框不确定性辅助。

消融实验结果如表3所示,其中①表示残差卷积,②表示迭代深度感知注意力融合模块,③表示融合深度实例,④表示边界框辅助任务。本文所提方法MonoDI(基线①+②+③+④)在单目3D目标检测任务中明显提升了模型的性能。

表3　消融实验结果

Table 3　Ablation experiment results

方法	AP_{3D}(IoU=0.7|R40)/%			AP_{BEV}(IoU=0.7|R40)/%		
	Easy	Mod	Hard	Easy	Mod	Hard
基线	23.31	17.66	15.05	32.06	23.69	20.48
基线+①	24.61	17.92	15.19	33.76	24.19	20.78
基线+②	25.55	18.47	15.57	33.97	24.52	21.06
基线+①+②	26.72	19.13	15.97	34.43	25.42	21.95
基线+①+②+③	28.46	21.15	17.23	38.25	27.68	23.18
基线+①+②+④	27.09	19.36	16.28	37.02	26.16	22.69
基线+①+②+③+④	29.53	22.07	17.74	39.37	28.52	23.91

2.6　跨数据集验证

为了验证模型MonoDI在具有不同参数的相机所采集的数据集上的泛化能力,本文使用深度的平均绝对误差在KITTI验证集和nuScenes前置摄像验证集上对模型进行评估,在0～20 m、20～40 m、≥40 m和全局的距离范围上计算误差结果。具体结果如表4所示。

表4　跨数据集验证结果

Table 4　Cross-dataset validation results

方法	KITTI				nuScenes frontal			
	0～20 m	20～40 m	≥40 m	ALL	0～20 m	20～40 m	≥40 m	ALL
M3D-RPN[3]	0.56	1.33	2.73	1.26	0.94	3.06	10.36	2.67
MonoRCNN[39]	0.46	1.27	2.59	1.14	0.94	2.84	8.65	2.39
GUPNet[37]	0.45	1.10	1.85	0.89	0.82	1.70	6.20	1.45
MonoLSS[38]	0.35	0.89	1.77	0.82	0.59	2.01	5.40	1.42
MonoDI	0.33	0.85	1.62	0.77	0.68	1.72	4.35	1.27

从表4可以看出,本文方法在每个距离分段上的平均绝对误差值均有所降低,特别是在≥40 m距离分段上误差降低显著,这说明本文提出方法对模型在远距离目标的感知能力上有着较好的提升。

2.7　可视化结果分析

图3为针对基线方法与本文方法以及对比方法MonoCon在KITTI数据集上的可视化结果,在不同

难度的检测场景中预测并绘制物体的3D框，并对比各方法的预测结果，以全面评估本文方法MonoDI的有效性。

(a) Base results (b) MonoCon results (c) MonoDI results

图3 可视化结果

Fig.3 Visualization results

通过对比图像中的3D框结果可以看出，对于无遮挡、边界清晰且距离较近的物体，3种方法都能够较为准确地预测出对应3D边界框。但在物体存在较大部分遮挡、没有清晰边界和距离较远的场景中，本文方法具有明显优势。从图3(a)中可以看出基线方法对于距离较远和具有较大遮挡的物体检测效果较差，对没有清晰边界的物体出现了误检的情况；图3(b)中对比方法MonoCon虽然对于较远物体的检测效果有明显提升，但是在有遮挡的情况下仍有着明显不足；图3(c)中可以看出本文方法在有遮挡、没有清晰边界和距离较远这3种情况下均有着较好的检测效果。这也进一步验证了本文所提出的iDAAFM模块和融合深度实例对于单目3D目标检测任务的有效性。

3 结束语

虽然通过深度估计可以弥补2D图像对于深度信息的缺失，给予3D回归任务深度信息指导，但是通过深度估计所得的深度信息中存在大量的深度背景噪声干扰，模型并不能够准确识别没有清晰边界和存在大面积遮挡的物体。本文提出MonoDI，利用先进的深度估计网络结合实例分割掩码得到具有清晰边界和深度信息的物体特征表示，显著提高了模型在单目3D目标检测任务中的性能，特别是在没有清晰边界和存在大面积遮挡物体的情况下；同时使用残差卷积头回归最终结果，保证了网络的稳定性与效能；再结合3D检测框不确定性辅助任务，利用标注上下文信息回归3D检测框不确定性，进一步提高了网络的检测精度。最后，对比试验与消融实验的结果均表明本文所提MonoDI的有效性，同时跨数据集验证结果证明该方法也有着较好的泛化能力。

参考文献：

[1] MANHARDT F, KEHL W, GAIDON A. ROI-10D: Monocular lifting of 2D detection to 6D pose and metric shape[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, USA:

IEEE, 2019: 2069-2078.
[2] HONG Y, DAI H, DING Y. Cross-modality knowledge distillation network for monocular 3D object detection[M]//Computer Vision- ECCV 2022. Cham, Switzerland: Springer Nature Switzerland, 2022: 87-104.
[3] BRAZIL G, LIU X. M3D-RPN: Monocular 3D region proposal network for object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV). Seoul, South Korea: IEEE, 2019: 9287-9296.
[4] LIU Z, WU Z, TOTH R. SMOKE: Single-stage monocular 3D object detection via keypoint estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). Seattle, USA: IEEE, 2020: 996-997.
[5] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points[EB/OL]. (2019-04-16). https://doi. org/10.48550/arXiv.1904.07850.
[6] ZHANG Y, LU J, ZHOU J. Objects are different: Flexible monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, USA: IEEE, 2021: 3288-3297.
[7] MA X, ZHANG Y, XU D, et al. Delving into localization errors for monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, USA: IEEE, 2021: 4721-4730.
[8] WANG Y, CHAO W L, GARG D, et al. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, USA: IEEE, 2019: 8437-8445.
[9] SIMONELLI A, BULÒ S R, PORZI L, et al. Are we missing confidence in pseudo-LiDAR methods for monocular 3D object detection?[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal, Canada: IEEE, 2021: 3205-3213.
[10] CHEN X, LIN K Y, WANG J, et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation[M]//Computer Vision-ECCV 2020. Cham, Switzerland:: Springer International Publishing, 2020: 561-577.
[11] SUN P, ZHANG W, WANG H, et al. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, USA: IEEE, 2021: 1407-1417.
[12] ZOU W, PENG Y, ZHANG Z, et al. RGB-D Gate-guided edge distillation for indoor semantic segmentation[J]. Multimedia Tools and Applications, 2022, 81(25): 35815-35830.
[13] ZHANG J, LIU H, YANG K, et al. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14679-14694.
[14] 马倩, 曾凯, 吴家文, 等. 基于非局部融合的多尺度目标检测研究[J]. 数据采集与处理, 2023, 38(2): 364-374.
MA Qian, ZENG Kai, WU Jiawen, et al. Multi-scale object detection based on non-local feature fusion[J]. Journal of Data Acquisition and Processing, 2023, 38(2): 364-374.
[15] LIU X, XUE N, WU T. Learning auxiliary monocular contexts helps monocular 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1810-1818.
[16] SOOMRO K, ZAMIR A R, SHAH M. A dataset of 101 human action classes from videos in the wild[J]. Center for Research in Computer Vision, 2012, 2(11): 1-7.
[17] LIEBEL L, KÖRNER M. Auxiliary tasks in multi-task learning[EB/OL]. (2018-05-16). https://arxiv.org/abs/1805.06334.
[18] JADERBERG M, MNIH V, CZARNECKI W M, et al. Reinforcement learning with unsupervised auxiliary tasks[EB/OL]. (2016-11-16). https://arxiv.org/abs/1611.05397.
[19] FLYNN J, NEULANDER I, PHILBIN J, et al. Deep stereo: Learning to predict new views from the world's imagery[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, USA: IEEE, 2016: 5515-5524.
[20] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, USA: IEEE, 2017: 6612-6619.
[21] DU Y, CZARNECKI W M, JAYAKUMAR S M, et al. Adapting auxiliary losses using gradient similarity[EB/OL]. (2018-12-05). https://arxiv.org/abs/1812.02224.
[22] 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.

[23] YU F, WANG D, SHELHAMER E, et al. Deep layer aggregation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 2403-2412.

[24] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, USA: IEEE, 2016: 3213-3223.

[25] YANG L H, KANG B Y, HUANG Z L, et al. Depth anything: Unleashing the power of large-scale unlabeled data[EB/OL]. (2024-01-19). https://arxiv.org/abs/2401.10891.

[26] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision(WACV). Waikoloa, USA: IEEE, 2021: 3560-3569.

[27] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV). [S.l.]: Springer International Publishing, 2018: 3-19.

[28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision(ICCV). Venice, Italy: IEEE, 2017: 2999-3007.

[29] CHEN Y, TAI L, SUN K, et al. MonoPair: Monocular 3D object detection using pairwise spatial relationships[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, USA: IEEE, 2020: 12093-12102.

[30] KENDALL A, GAL Y, KENDALL A, et al. What uncertainties do we need in Bayesian deep learning for computer vision? [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA: ACM, 2017: 5580-5590.

[31] CIPOLLA R, GAL Y, KENDALL A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7482-7491.

[32] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: A multimodal dataset for autonomous driving[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, USA: IEEE, 2020: 11621-11631.

[33] CHEN X, KUNDU K, ZHU Y, et al. 3D object proposals for accurate object class detection[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems-Volume 1. Montreal, Canada: ACM, 2015: 424-432.

[34] DING M, HUO Y, YI H, et al. Learning depth-guided convolutions for monocular 3D object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). Seattle, USA: IEEE, 2020.

[35] CHONG Z Y, MA X Z, ZHANG H, et al. Monodistill: Learning spatial features for monocular 3D object detection[EB/OL]. (2022-01-26). https://arxiv.org/abs/2201.10830.

[36] PARK D, AMBRUŞ R, GUIZILINI V, et al. Is pseudo-lidar needed for monocular 3D object detection?[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal, Canada: IEEE, 2021: 3122-3132.

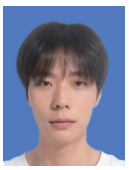
[37] LU Y, MA X, YANG L, et al. Geometry uncertainty projection network for monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal, Canada: IEEE, 2021: 3091-3101.

[38] LI Z, JIA J, SHI Y. MonoLSS: Learnable sample selection for monocular 3D detection[C]//Proceedings of the 2024 International Conference on 3D Vision (3DV). Davos, Switzerland: IEEE, 2024: 1125-1135.

[39] SHI X, YE Q, CHEN X, et al. Geometry-based distance decomposition for monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal, Canada: IEEE, 2021: 15152-15161.

作者简介:

赵科(1999-),男,硕士研究生,研究方向:单目3D目标检测,E-mail:1468039662@qq.com。



董浩然(1999-),男,硕士研究生,研究方向:异常检测,E-mail:1626835594@qq.com。

业宁(1967-),**通信作者**,男,教授,博士生导师,研究方向:生物信息学,数据挖掘和机器学习,E-mail:yening@njfu.edu.cn。

(编辑:张蓓,王婕)