计算机科学
COMPUTER SCIENCE

基于高深约束与边缘融合的单目3D目标检测

浦斌, 梁正友, 孙宇

引用本文

浦斌, 梁正友, 孙宇. 基于高深约束与边缘融合的单目3D目标检测[J]. 计算机科学, 2024, 51(8): 192-199.

PU Bin, LIANG Zhengyou, SUN Yu. Monocular 3D Object Detection Based on Height-Depth Constraint and Edge Fusion [J]. Computer Science, 2024, 51(8): 192-199.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500071

基于高深约束与边缘融合的单目 3D 目标检测

浦 斌[1] 梁正友[1,2] 孙 宇[1,2]
1 广西大学计算机与电子信息学院 南宁 530004
2 广西大学广西多媒体通信与网络技术重点实验室 南宁 530004
(pbbingo@foxmail.com)

摘 要 单目 3D 目标检测旨在通过单目图像完成 3D 目标检测,现有的单目 3D 目标检测算法大多基于经典的 2D 目标检测算法。针对单目 3D 目标检测算法中通过直接回归的实例深度估计不准,导致检测精度较差的问题,提出了一种基于高深约束与边缘特征融合的单目 3D 目标检测算法。在实例深度估计方法上采用几何投影关系下的实例 3D 高度与 2D 高度计算高深约束,将实例深度的预测转化为对目标的 2D 高度以及 3D 高度的预测;针对单目图像存在图像边缘截断目标,采用基于深度可分离卷积的边缘融合模块来加强对边缘目标的特征提取;对于图像中目标的远近造成的目标多尺度问题,设计了基于空洞卷积的多尺度混合注意力模块,增强了对最高层特征图的多尺度特征提取。实验结果表明,所提方法在 KITTI 数据集上的汽车类别检测精度相比基准模型提升了 7.11%,优于当前的方法。

关键词:单目 3D 目标检测;高深约束;边缘融合;多尺度特征;注意力机制
中图分类号 TP391

Monocular 3D Object Detection Based on Height-Depth Constraint and Edge Fusion

PU Bin[1],LIANG Zhengyou[1,2] and SUN Yu[1,2]
1 School of Computer and Electronics Information,Guangxi University,Nanning 530004,China
2 Guangxi Key Laboratory of Multimedia Communication and Network Technology,Guangxi University,Nanning 530004,China

Abstract Monocular 3D object detection aims to complete 3D object detection using monocular images,and most existing monocular 3D object detection algorithms are based on classical 2D object detection algorithms. To address the issue of inaccurate instance depth estimation through direct regression in monocular 3D object detection algorithms,which leads to poor detection accuracy,a monocular 3D object detection algorithm based on height-depth constraint and edge feature fusion is proposed. In the instance depth estimation method,the height-depth constraint is calculated by the instance 3D height and 2D height under the geometric projection relationship,mainly converting the prediction of instance depth into the prediction of 2D height and 3D height of the object. To address the issue of object truncation at image edges in monocular images,an edge fusion module based on depth separable convolution is used to enhance the feature extraction of edge objects. For the multi-scale problem caused by the proximity and distance of objects in the image,a multi-scale mix attention module based on dilated convolution is designed to enhance the multi-scale feature extraction of the highest layer feature map. Experimental results demonstrate the effectiveness of the proposed method,as it achieves a 7.11% improvement in car category detection accuracy compared to the baseline model on the KITTI dataset,outperforming the current methods.

Keywords Monocular 3D object detection,Height-Depth constraint,Edge fusion,Multi-scale feature,Attention mechanism

1 引言

在自动驾驶领域中,通常采用激光雷达、多目相机和单目相机 3 种传感器进行 3D 目标检测。相较于基于激光雷达和多目相机的方法,基于单目相机的检测方法在成本方面具有显著优势。单目 3D 目标检测算法依赖于单目图像以及单目相机自身参数完成 3D 检测。

在单目 3D 目标检测任务中,由于单目图像只能提供 2D 视角信息,使用直接回归深度的方法[1-4]难以实现精确的实例深度估计,限制了算法的检测精度。同时,车辆目标在单目图像上存在边缘截断的情况,造成车辆特征信息不完整。车辆距离的远近在单目图像上存在尺度上的差异,需要考虑图像

到稿日期:2023-05-10 返修日期:2023-08-10
基金项目:国家自然科学基金(62171145)
This work was supported by the National Natural Science Foundation of China(62171145).
通信作者:梁正友(zhyliang@gxu.edu.cn)

经过深度神经网络多次下采样后的特征图的多尺度问题。

为了解决这些问题,本文提出了一种基于几何投影下的高深约束(Height-Depth Constraint in Geometric Projection, HDCGP)与边缘特征融合的单目 3D 目标检测算法。首先,通过设计几何投影关系的 3D 高度与 2D 高度计算主要实例深度,将算法对实例深度的估计解耦为算法对 3D 高度与 2D 高度的估计,提高实例深度估计的可靠性。其次,针对图像边缘截断问题,提出了基于深度可分离卷积的边缘融合模块,有助于增强边缘目标特征提取,更好地解决目标截断问题。再次,针对由目标远近在图像上的尺寸差异引起的多尺度缺乏问题,本文设计 4 组不同空洞率的空洞卷积来完成对高语义低分辨率特征图的多尺度信息提取,并使用多尺度的混合注意力在空间与通道上逐尺度地对输出特征图进行加强特征提取。最后,在 KITTI 数据集上进行对比实验,结果验证了本文算法的有效性和优越性。

2　相关研究

随着深度学习技术的不断发展,单目 3D 目标检测取得了重大进展。目前,这些算法根据所依赖的额外数据类型可分为以下 4 种:

1)基于深度信息引导的方法。这类算法利用单目深度估计模型预先得到像素级深度图,将深度图与单目图像结合后输入检测器。D4LCN[5] 和 DDMP-3D[6] 是该类方法的经典算法。D4LCN 提出了一种局部卷积神经网络,通过自动学习基于深度图中的卷积核及其接受域,克服了传统二维卷积无法捕获物体多尺度信息的问题。DDMP-3D 提出了一种深度先验条件下的动态信息传递网络,并通过中心感知深度编码任务来减轻深度先验不准确的影响。这类方法依赖于预先得到的像素级深度图,受限于深度估计模型的准确性,其预测误差会被进一步引入单目 3D 目标检测模型中。

2)基于立体图引导的方法。此类方法通过双目立体 3D 目标检测算法引导单目 3D 目标检测算法的学习,代表性算法为 SGM3D[7]。SGM3D 采用了多粒度特征对齐机制,利用粗特征层次和细锚点层次的特征对齐来引导单目特征,并通过基于 IoU 匹配的特征对齐方法在立体和单目预测之间进行物体级别的特征对齐,来提升检测性能。SGM3D 利用立体图引导训练,会造成立体图信息与单目图像的特征不完全匹配,从而影响了所引导的单目 3D 目标检测模型的泛化能力。

3)基于雷达信息引导的方法。这类算法将稀疏的雷达点云信息作为辅助监督进行模型训练,在推理时只需输入图像和单目相机信息。MonoRUn[8],CaDDN[9] 和 MonoDTR[10] 是近年来的代表性算法。MonoRUn 可以在自监督情况下学习密集的 2D-3D 对应和几何信息,并利用网络的不确定性进行姿态估计和置信度计算。尽管 MonoRUn 能利用自监督重建在单目图像上进行 3D 检测,但它忽略了对实例深度信息的充分挖掘,在没有雷达信息引导的情况下,其检测性能会明显下降。CaDDN 通过将深度分类来生成视锥特征,并通过相机参数进一步转化为体素特征,并完成 BEV 特征生成和 3D 检测。由于 CaDDN 使用多个输入转换分支完成 3D 检测,因此其模型推理速度缓慢,不适用于实时场景。MonoDTR 则将 Transformer[11] 引入单目 3D 目标检测领域,通过深度感知特征增强模块和深度感知 Transformer 模块,实现全局上下文和深度感知特征的综合,使用深度位置编码向 Transformer 注入深度位置提示,可以更好地将 Transformer 应用于单目 3D 目标检测领域。但 MonoDTR 使用的自注意力机制难以处理多尺度目标,表现为对远端目标的检测能力下降。

4)基于直接回归的方法。代表性方法有 SMOKE[2],MonoPair[3],MonoFlex[12],GUPNet[13],MonoDLE[4],DEVIANT[14] 等。这些算法主要利用几何先验知识和深度估计的不确定性建模来提高算法性能。SMOKE 设计了基于关键点的 3D 检测分支并去除了 2D 检测分支。MonoPair 考虑了目标间的约束关系而设计了成对空间约束,计算出目标位置的不确定性预测与相邻目标对的 3D 距离。MonoFlex 设计了解耦截断目标和正常目标的预测方法,通过组合基于关键点的深度和直接回归深度进行精确的实例深度估计。GUPNet 利用几何不确定性投影模块解决几何投影过程的误差放大问题,并提出了分层任务学习来解决多任务下参数的学习问题。MonoDLE 在初期进行了一系列的实验,发现定位误差是影响单目 3D 目标检测模型性能的关键因素。因此,MonoDLE 改进了中心点的取法,采用了从 3D 投影中心而不是 2D 边界框中心获取中心点的方法,以提升模型性能。此外,在实例深度估计任务上,MonoDLE 采用了不确定性原理对实例深度进行估计。DEVIANT 提出了深度等变性网络来解决现有神经网络模块在处理 3D 空间中的任意平移时缺乏等变性的问题。这类基于直接回归的方法仅使用单目图像完成模型训练与推理,但是使用直接回归的实例深度估计方法[1-4] 不准确,使用几何投影的实例深度估计方法[12-14] 则缺乏准确的 2D 高度与 3D 高度预测。此外,这类方法都未能充分利用单目图像中的边缘截断特征以及目标的多尺度特征,限制了对截断目标以及远近尺度不一致目标的检测能力。

3　本文方法

单目图像下直接回归实例深度依赖于深度神经网络提取的特征。近处目标的下采样特征尺度大,故而实例深度估计准确;而对于过于接近检测器所在车辆的目标,可能出现截断在图像两端,进而导致漏检问题。远处目标的下采样特征尺度小,导致直接回归的实例深度估计不准,且远近目标的下采样特征尺度不一还会导致目标多尺度问题的出现。本文就以上问题提出解决方法,首先就直接回归的实例深度估计不准问题,提出了一种几何投影下的高深约束实例深度估计方法,将目标的实例深度估计主要转化为对目标的 2D 高度与 3D 高度预测,并对 3D 高度与 2D 高度设计新的目标损失函数。其次,对于图像边缘截断车辆的漏检问题,设计了基于深度可分离卷积的边缘融合模块(Edge Fusion Module based on Depth Separable Convolution,EF-DSC)。最后,对于不同距离目标在图像中造成的多尺度问题,设计了基于空洞卷积的

多尺度混合注意力模块(Multi-scale Mix Attention Module based on Dilated Convolution, MSMADC)。

随后 3.1 节介绍系统的总体结构,3.2 节介绍高深约束实例深度估计方法,3.3 节介绍基于深度可分离卷积的边缘融合模块,3.4 节介绍基于空洞卷积的多尺度混合注意力模块,3.5 节介绍本文算法使用的损失函数。

3.1 总体结构

在 2D 目标检测领域,Mask R-CNN[15] 使用二阶段的设计方法构造神经网络特征图的感兴趣区域,再对感兴趣区域进行分类与回归完成检测过程。受此启发,本文以 MonoDLE 网络为基础,设计了一个二阶段的单目 3D 目标检测网络,如图 1 所示。

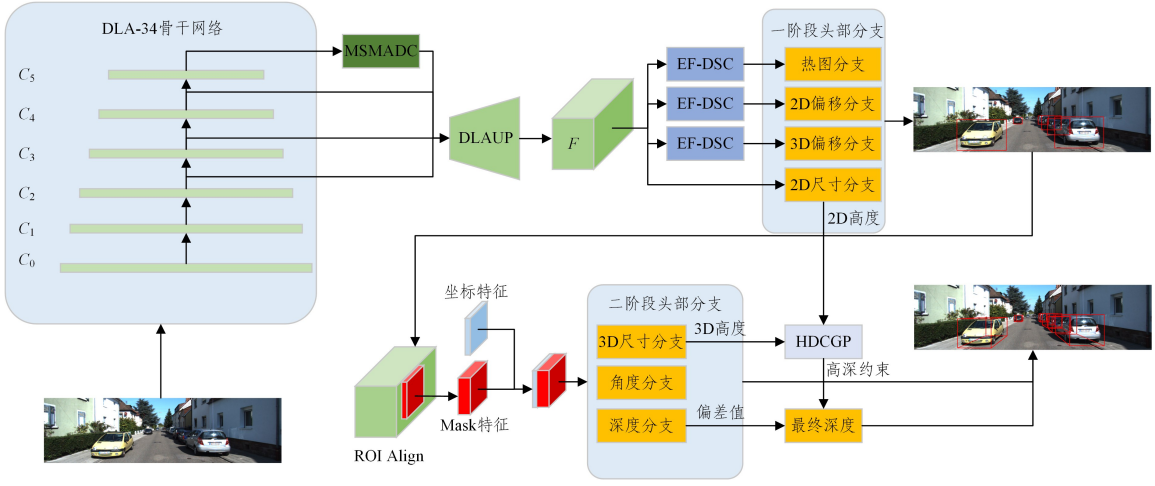

图 1 二阶段的单目 3D 目标检测网络结构

Fig. 1 Structure of two-stage monocular 3D object detection network

第一阶段为单目图像进行特征提取到完成 2D 检测的阶段。首先,输入图像经过 DLA-34[16] 骨干网络完成图像特征提取,输出不同下采样率的特征图,对 DLA-34 输出的 C_5 特征图采用 MSMADC 模块进行多尺度特征提取。其次,将 DLA-34 输出的 C_2、C_3、C_4 与完成多尺度特征提取的 C_5 使用 DLAUP 进行特征融合,最终输出 4 倍下采样特征图 F。最后,F 进入各个一阶段分支完成一阶段预测结果,同时输出 2D 预测结果,其中一阶段分支的热图分支与 2D 偏移分支以及 3D 偏移分支额外通过 EF-DSC 模块学习图像边缘特征。一阶段涉及的热图分支预测目标的分类以及 2D 边界框中心的粗坐标,2D 偏移分支预测 2D 粗坐标与 2D 边界框中心的偏移量,2D 尺寸分支预测 2D 边界框的宽高,3D 偏移分支预测 2D 粗坐标与 3D 边界框中心投影在图像上的偏移量。第二阶段的设计主要遵循 GUPNet 的设计方法,为对 2D 检测结果进行感兴趣特征提取到完成 3D 检测的阶段。根据第一阶段 2D 预测结果,先用 ROI Align(Region of Interest Align)进行感兴趣特征提取,再与坐标特征图在通道维度上结合获取带坐标信息的 Mask 特征图。带坐标信息的 Mask 特征图输入角度分支、3D 尺寸分支、深度分支完成第二阶段预测。后处理部分主要利用 HDCGP 方法求取高深约束,并将高深约束与网络直接回归的深度偏差相加作为最终输出深度。综合一阶段分支与二阶段分支的预测结果以及最终深度,获取实例目标的 3D 检测结果。

3.2 几何投影下的高深约束

本文设计的几何投影下的高深约束方法可以很好地建立目标边界框的 3D 高度与 2D 高度的关系,如图 2 所示。2D 高度 h 为目标在原始图像中的 2D 边界框高度,3D 高度 H 为目标在真实世界坐标系下的 3D 边界框高度。

根据相似三角形定理,实例目标在原始图像中的 2D 边界框的高度 h、相机焦距 f、3D 边界框高度 h 与实例深度 d 的关系如式(1)所示:

$$d = f * \frac{H}{h} \tag{1}$$

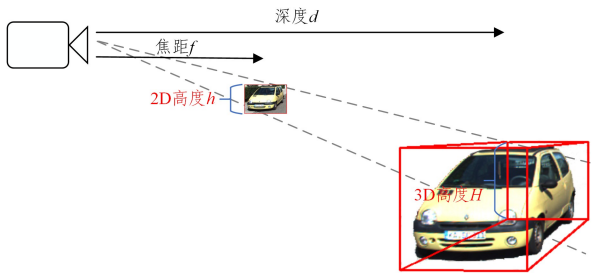

图 2 几何投影下的高深约束

Fig. 2 Height-Depth constraint in geometric projection

3.2.1 基于不确定性原理的高度求解

应用于计算机视觉的不确定性理论由 Kendall 等[17] 提出,他们认为深度学习模型的预测结果并非始终可靠,因此需要模型能够为错误的预测结果提供高度的不确定性,以判断深度学习模型的预测是否可靠。不确定性可分为认知不确定性与偶然不确定性,认知不确定性在理论上可以随着模型学习数据的增加而降低,偶然不确定性获取的是单目相机传感器本身的噪声和运动时造成的噪声。在许多先前的单目 3D 目标检测中,使用拉普拉斯分布建模深度的偶然不确定性可以显著减少实例深度的噪声[3-4]。然而,这些方法通过深度神经网络直接回归深度的不确定性,而忽略了高度和深度之间的关系。为此,本文设计的 HDCGP 方法利用不确定性理论在 2D 高度与 3D 高度上应用不确定性原理,以使深度神经网络输出更可靠的 2D 高度与 3D 高度及其不确定度分数。

假定本文模型预测的 2D 高度与 3D 高度都符合拉普拉斯分布，则可以对 2D 高度与 3D 高度应用拉普拉斯偶然不确定性损失函数[3]。其中，h^* 与 H^* 分别为 2D 高度的真实值与 3D 高度的真实值，h 与 H 分别为 2D 高度的预测值与 3D 高度的预测值，σ_h 与 σ_H 为 2D 高度和 3D 高度对应的不确定度。

$$\mathscr{L}_h(h,h^*,\sigma_h)=\frac{\sqrt{2}}{\sigma_h}|h-h^*|+\log(\sigma_h) \tag{2}$$

$$\mathscr{L}_H(H,H^*,\sigma_H)=\frac{\sqrt{2}}{\sigma_H}|H-H^*|+\log(\sigma_H) \tag{3}$$

3.2.2　深度偏差

为了获得更好的预测深度，在二阶段头部分支设计了深度分支，用来学习深度偏差 d_n，该偏差用于修正几何投影下的高深约束。和过去的工作一样[4,13-14]，将深度分支的输出张量 z 进行逆 sigmoid 变换得到深度偏差值 d_n，如式(4)所示：

$$d_n=\frac{1}{\mathrm{Sigmoid}(z)+\varepsilon}-1 \tag{4}$$

其中，ε 设为一个极小常数值，用于保证深度偏差值的稳定性。

假定深度偏差 d_n 与几何投影下的高深约束 d_h 都符合拉普拉斯分布，则由两个独立分布的深度相加计算的最终深度也符合拉普拉斯分布 $d=La(\mu,\sigma)$。根据最终深度与两个深度的关系为 $\mu=d_n+d_h$，则最终深度的不确定度 σ 与各自的不确定度大小的关系为 $\sigma^2=\sigma_h^2+\sigma_n^2$。最终深度及其不确定度的计算式如式(5)所示：

$$d=La(d_n+d_h,\sqrt{\sigma_n^2+\sigma_h^2}) \tag{5}$$

3.3　基于深度可分离卷积的边缘融合模块

本文针对截断特征学习问题所设计的 EF-DSC 模块如图 3 所示，该模块应用于一阶段头部分支的热图分支、2D 偏移分支、3D 偏移分支。首先，以特征融合层 DLAUP 输出的 4 倍下采样特征图作为输入，先经过一层卷积层和 ReLU 激活函数得到通道升维特征 F'。然后，将 F' 与边缘 Mask 进行网格采样生成图像 4 条边缘的一维特征。最后，使用深度可分离 1D 卷积对一维特征进行特征提取，并将边缘特征加在 F' 的对应边上得到最终输出特征。

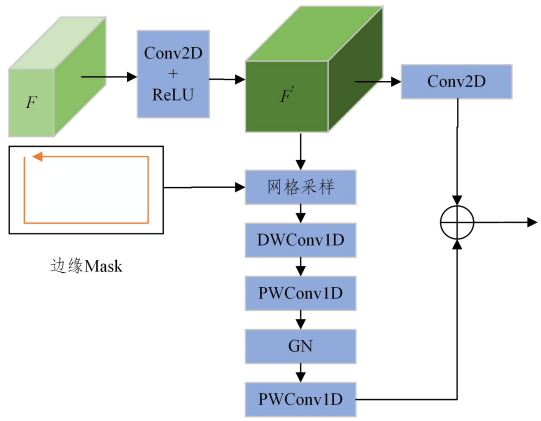

图 3　基于深度可分离卷积的边缘融合模块

Fig. 3　Edge fusion module based on depth separable convolution

深度可分离 1D 卷积由一层逐深度卷积、一层逐点卷积、群组归一化层(Group Normalization，GN)[18]和一层逐点卷积组成。逐深度卷积的卷积核长度为 3，分组数为 F' 通道数 256。第一层逐点卷积输入输出通道数均为 F' 的输入通道数 256，GN 层设置的分组数为 32，最后一层逐点卷积的输出通道为使用该边缘融合模块的分支的输出通道数，如在热图分支中 F' 经过 Conv2D 后输出通道为 3，则相应的最后一层逐点卷积的输出通道数为 3。逐深度卷积对一维特征的长区域进行特征提取，逐点卷积则负责混合逐深度卷积输出的通道信息，GN 将通道划分为组并在每组内计算归一化的均值和方差，这使得 GN 在不同 Batch 的训练下更稳定。最后的逐点卷积则将通道下降至对应头部分支的输出通道数，从而在图像边缘部分增强边缘获取的特征。

3.4　基于空洞卷积的多尺度混合注意力模块

本文设计的 MSMADC 模块用来捕获骨干网络输出的最高层特征图的多尺度特征，以解决单目图像中的目标多尺度问题。MSMADC 模块的整体结构如图 4 所示，其中 CBR 为卷积层＋ BN 层＋ReLU 激活函数，DilatedCBR 为空洞卷积层＋BN 层＋ReLU 激活函数。4 组空洞卷积的空洞率分别设置为 3/6/9/12，卷积核大小为 3＊3。

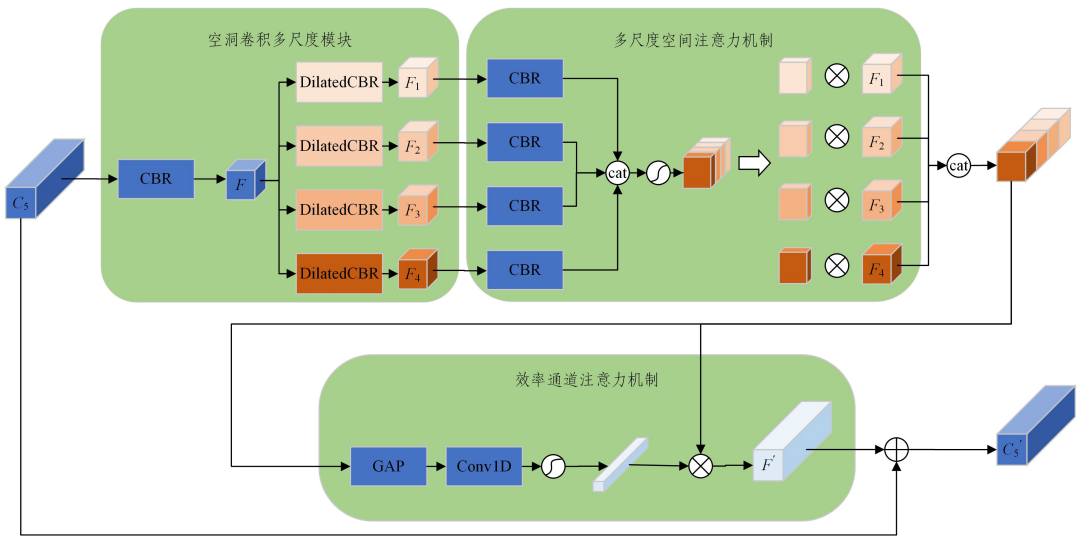

图 4　基于空洞卷积的多尺度混合注意力模块

Fig. 4　Multi-scale mix attention module based on dilated convolution

MSMADC 模块由空洞卷积多尺度模块、多尺度空间注意力机制、效率通道注意力机制以及残差连接组成。空洞卷积多尺度模块对 DLA-34 模型输出的高层特征 C_5 进行通道压缩，得到特征 F。压缩通道后的特征 F 可减少在 4 组空洞卷积中的训练和推理时长。特征 F 经过不同空洞率的空洞卷积后，得到不同尺度的特征 F_1-F_4。多尺度空间注意力机制主要完成对不同尺度特征的空间特征加权，进一步增强不同尺度特征的空间特征。该机制首先将 F_1-F_4 分别进行通道压缩至 1，然后通过 Sigmoid 函数处理得到不同尺度的空间注意力权值，最后将 F_1-F_4 特征分别与各自的权重值进行乘积得到多尺度空间特征。效率通道注意力机制[19]会在通道层面上对多尺度空间特征进行通道加权得到特征 F'。最终的多尺度特征 C_5' 由 F' 与 DLA-34 输出的高层特征 C_5 进行残差连接获得。残差连接[20]可以保证在深度神经网络进行反向传播过程中不会出现梯度弥散问题，使得模型可以学习到输出和输入之间的残差映射。

3.5 损失函数

本文算法中的各个头部分支一共使用了 4 类损失函数，包括：1）热图分支的高斯核加权的聚焦损失函数；2）深度分支、2D 高度、3D 高度的拉普拉斯偶然不确定性损失函数；3）角度分支中的分类子分支所使用的标准交叉熵损失函数；4）2D 尺寸分支、2D 偏移分支、3D 偏移分支、角度分支中的残差回归子分支所使用的标准 L1 损失函数。

3.5.1 热图损失函数

Lin 等[21]提出的聚焦损失函数，用于对 2D 目标检测中正负样本的权重以及难易样本分类的权重进行控制，使损失更侧重于难分类的样本。MonoDLE 和 CenterNet 都使用了聚焦损失函数作为热图分支的损失函数。聚焦损失函数如式(6)所示：

$$\mathcal{L}_{\text{seg}}(Y,Y^*)=\frac{-1}{N}\sum_{(x,y,c)}\begin{cases}(1-Y_{xyc})^{\alpha}\log(Y_{xyc}), & \text{if } Y^*_{xyc}=1\\ (1-Y^*_{xyc})^{\beta}(Y_{xyc})^{\alpha}\log(1-Y_{xyc}), & \text{otherwise}\end{cases} \tag{6}$$

其中，α 和 β 为超参数，N 是图像的目标数量。预设 α 和 β 分别为 2 和 4，Y_{xyc} 为热图分支的预测值，Y^*_{xyc} 为标注真实值，求和符号下的 xyc 表示所有热图上的坐标点（其中 c 表示目标类别）。

3.5.2 2D 偏移与 3D 偏移损失函数

2D 偏移与 3D 偏移损失使用标准的 L1 损失函数，2D 偏移损失计算 2D 偏移分支预测的 2D 粗坐标与 2D 边界框中心的偏移量 O_{2d} 与真实 2D 偏移量 O^*_{2d} 的平均绝对值误差，2D 偏移损失函数如式(7)所示：

$$\mathcal{L}_{\text{offset\_2d}}(O_{2d},O^*_{2d})=\mathcal{L}_1(O_{2d},O^*_{2d}) \tag{7}$$

3D 偏移损失计算 2D 粗坐标与 3D 边界框中心投影在图像上的偏移量 O_{3d} 与真实 3D 偏移量 O^*_{3d} 的平均绝对值误差，3D 偏移损失函数如式(8)所示：

$$\mathcal{L}_{\text{offset\_3d}}(O_{3d},O^*_{3d})=\mathcal{L}_1(O_{3d},O^*_{3d}) \tag{8}$$

3.5.3 2D 尺寸与 3D 尺寸损失函数

2D 尺寸分支预测目标 2D 边界框的宽度与高度，由式(2)给出 2D 边界框高度 h 的损失函数，2D 边界框宽度 w 损失则应用标准的 L1 损失函数计算宽 w 与真实值 w^* 的平均绝对值误差，最终 2D 尺寸损失函数如式(9)所示：

$$\mathcal{L}_{\text{size2d}}=\frac{1}{2}\mathcal{L}_h(h,h^*,\sigma_h)+\frac{1}{2}\mathcal{L}_1(w,w^*) \tag{9}$$

3D 尺寸分支预测目标 3D 边界框的长宽高，3D 边界框的长度与宽度损失应用标准的 L1 损失函数。L1 损失函数分别计算长 L 与真实值 L^*、宽 W 与真实值 W^* 的平均绝对值误差。3D 边界框的高度损失如式(3)所示，由此计算最终 3D 尺寸损失函数，如式(10)所示：

$$\mathcal{L}_{\text{size3d}}=\frac{1}{3}\mathcal{L}_1(L,L^*)+\frac{1}{3}\mathcal{L}_1(W,W^*)+\frac{1}{3}\mathcal{L}_H(H,H^*,\sigma_H) \tag{10}$$

3.5.4 深度拉普拉斯偶然不确定性损失函数

应用于式(5)求取的最终深度 d 的拉普拉斯偶然不确定性损失函数如式(11)所示：

$$\mathcal{L}_d(d,d^*,\sigma_d)=\frac{\sqrt{2}}{\sigma_d}|d-d^*|+\log(\sigma_d) \tag{11}$$

其中，d^* 是实例深度真实值。

3.5.5 角度损失函数

角度分支采用 Multi-Bin 方法[22]预测朝向角。朝向角的区间$[-\pi,\pi]$分为 12 个重叠的格子，通过角度分支预测出的置信度对该目标落于格子的类别进行分类，这一部分使用标准的交叉熵损失函数计算预测的角度分类类别 C 与真实值 C^* 的损失值。角度分支额外预测这 12 个格子各自的残差角度，所预测的残差角度 α 用于对所在的格子的局部朝向值进行修正，应用标准的 L1 损失函数计算残差角度 α 与真实值 α^* 的损失值。总的 Multi-Bin 损失函数如式(12)所示：

$$\mathcal{L}_{\text{heading}}=\mathcal{L}_{\text{cross\_entropy}}(C,C^*)+\mathcal{L}_1(\alpha,\alpha^*) \tag{12}$$

3.5.6 总损失函数

总损失函数由所有损失项的总和构成。本文中采用与 MonoDLE[4] 相同的损失权衡参数设置，即将每个损失项的权衡参数均设置为 1，得到的总损失函数如式(13)所示：

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{seg}}+\mathcal{L}_{\text{offset\_2d}}+\mathcal{L}_{\text{size2d}}+\mathcal{L}_{\text{offset\_3d}}+\mathcal{L}_{\text{size3d}}+\mathcal{L}_d+\mathcal{L}_{\text{heading}} \tag{13}$$

4 实验

4.1 KITTI 数据集

本文在 KITTI 数据集[23]上进行实验，训练集由 7481 张图像及其对应的标签信息与相机参数组成，测试集则仅由 7518 张图像及其对应的相机参数组成。由于 KITTI 测试集的标签由 KITTI 官方保密，为了能阶段性地验证模型的改进，将 KITTI 训练集划分为 3712 张图像的训练子集和 3769 张图像的验证子集。在模型验证阶段，训练子集参与模型的训练，验证子集仅用于阶段性验证模型的检测精度。在阶段性模型性能验证工作后，最终模型被确定，并重新训练 KITTI 全部训练集。全训练集训练的模型生成测试集标签，然后将生成的测试集标签交由 KITTI 官方服务器进行评估，从而得到最终的测试集精度。

4.2 实验环境与参数设置

本文实验的硬件环境为深度学习工作站，其中 CPU 为

i5-12490F，内存为 16 GB，GPU 为 RTX 3090。软件环境配置为 Windows 10 专业版操作系统，以 Python 3.6 语言编写算法，使用深度学习框架 PyTorch 1.9.1，对应的 CUDA 版本为 11.1。本文的实验参数遵循 MonoDLE 的设置，使用端到端的方式训练 140 个轮次。选取初始学习率为 0.001 25 和权重衰减为 0.000 01 的 Adam 优化器，在前 5 个轮次时采用预热策略，学习率分别在第 90 和 120 个轮次下降为原来的 1/10。Batch_Size 设置为 16，训练共计 9.5h。

4.3　评价指标

本文采用 KITTI 基准测试提供的建议，以 40 个召回位置（R40）和交并比（Intersection over Union，IoU）大于等于 0.7 的平均精度（Average Precision，AP）作为测试集评价标准，对 3D 边界框、鸟瞰图这两类不同评估方式采用上述评价标准。可将 3D 边界框评价指标记为 $AP_{3D|R40|IoU\geqslant 0.7}$，则鸟瞰图的评价指标记为 $AP_{BEV|R40|IoU\geqslant 0.7}$。在 3 种难度设置下（简单（Easy）、中等（Moderate）、困难（Hard））一共 6 个指标对两个评估方式进行汽车类别检测性能评估，所有指标的值越大代表算法的检测性能越好。

4.4　实验结果和分析

为了验证本文算法的有效性，将在 KITTI 数据集的测试集与验证集上进行对比实验。在测试集对比实验中以算法是否需要额外数据参与训练来界定不同算法的分类。KITTI 的测试集结果如表 1 所列，其中性能最好的指标用粗体突出表示。实验结果表明，在 KITTI 数据集中最关注的汽车类别检测性能上，相较于依赖额外深度数据训练的 DDMP-3D 和 D4LCN 算法，本文方法在 3D 检测任务和 BEV 任务上都表现出显著的性能提升；相较于依赖额外 Lidar 数据训练的 MonoDTR 算法，本文方法在 BEV 任务上明显领先，3D 检测任务则在简单难度下检测性能提升了 1.59。本文方法在 3D 检测任务上的性能与不依赖额外数据训练的 CenterNet（GeoAug）[24] 算法表现相当，在 BEV 任务 3 种难度（简单、中等、困难）中检测性能分别提升了 2.65，0.94，1.26。本文方法应用高深约束下的实例深度估计方法，进一步改进了对实例的深度估计，同时可以在不依赖额外数据的情况下对边缘特征以及多尺度特征进行增强提取。通过综合改进，本文所提模型在 KITTI 测试集上获得了先进的检测性能。

表 1　KITTI 测试集上汽车类别与不同方法的性能对比

Table 1　Performance comparison of car category with different methods on KITTI test set

Methods	Extra data	$AP_{3D\|R40\|IoU\geqslant 0.7}$			$AP_{BEV\|R40\|IoU\geqslant 0.7}$		
		Easy	Mod	Hard	Easy	Mod	Hard
D4LCN[5]	Depth	16.65	11.72	9.51	22.51	16.02	12.55
DDMP-3D[6]	Depth	19.71	12.78	9.80	28.08	17.89	13.44
MonoRUn[8]	Lidar	19.65	12.30	10.58	27.94	17.34	15.24
CaDDN[9]	Lidar	19.17	13.41	11.46	27.94	18.91	17.19
MonoDTR[10]	Lidar	21.99	15.39	12.73	28.59	20.38	17.14
SGM3D[7]	Stereo	22.46	14.65	**12.97**	31.49	21.37	18.43
SMOKE[2]	None	14.03	9.76	7.84	20.83	14.49	12.75
MonoPair[3]	None	13.04	9.99	8.65	19.28	14.83	12.89
MonoDLE[4]	None	17.23	12.26	10.29	24.79	18.89	16.00
MonoFlex[12]	None	19.94	13.89	12.07	28.23	19.75	16.89
GUPNet[13]	None	20.11	14.20	11.77	—	—	—
DEVIANT[14]	None	21.88	14.46	11.89	29.65	20.44	17.43
CenterNet(GeoAug)[24]	None	23.41	15.26	12.80	31.58	20.75	17.66
Ours	None	**23.58**	**15.49**	12.79	**34.23**	**21.69**	**18.92**

表 2 列出了本文方法在 KITTI 验证集上与其他方法的对比实验结果。使用额外训练数据的算法利用 DORN[25] 的预训练模型作为其深度估计器，而 DORN 模型的训练数据与 KITTI 数据集的验证集存在重叠，因此不将这些算法与仅使用 KITTI 数据集的算法进行比较，以确保比较结果的公正性和可靠性。

表 2　KITTI 验证集上汽车类别与不同方法的性能对比

Table 2　Performance comparison of car category with different methods on KITTI validation set

Methods	$3D@IoU=0.7$			$BEV@IoU=0.7$			$3D@IoU=0.5$			$BEV@IoU=0.5$		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
CenterNet[1]	0.60	0.66	0.77	3.46	3.31	3.21	20.00	17.50	15.57	34.36	27.91	24.65
MonoPair[3]	16.28	12.30	10.42	24.12	18.17	15.76	55.38	42.39	37.99	61.06	47.63	41.92
GUPNet[13]	22.76	16.46	13.72	31.07	22.94	19.75	57.62	42.33	37.59	61.78	47.06	40.88
DEVIANT[14]	**24.63**	16.54	14.52	32.60	23.04	19.99	61.00	46.00	40.18	65.28	49.63	43.50
CenterNet(GeoAug)[24]	24.53	17.23	14.32	—	—	—	—	—	—	—	—	—
MonoFlex[12]	23.64	17.51	**14.83**	—	—	—	—	—	—	—	—	—
MonoDLE[4]	17.45	13.66	11.68	24.97	19.33	17.01	55.41	43.42	37.81	60.73	46.87	41.89
Ours	24.56	**17.59**	14.67	**33.25**	**23.55**	**20.11**	**64.89**	**47.40**	**42.39**	**68.64**	**50.80**	**45.59**
Improvement	+7.11	+3.93	+2.99	+8.28	+4.22	+3.1	+9.48	+3.98	+4.58	+7.91	+3.93	+3.7

实验结果表明，相较于基准模型 MonoDLE，本文方法在 3D 检测任务与 BEV 任务上均取得了较大的提升，在设置严格条件下（$IoU=0.7$）的 3D 检测任务和 BEV 任务在 3 种难度上检测性能分别提升了 7.11，3.93，2.99 和 8.28，4.22，3.1。对于近期的工作 DEVIANT 和 CenterNet（AugGeo），本文方法在严格条件下的 3D 检测任务上取得了相似

的性能，在一般条件（$IoU=0.5$）的 3D 检测任务和 BEV 任务中相较于 DEVIANT 在 3 种难度下检测性能分别提升了 3.89，1.4，2.21 和 3.36，1.17，2.09。得益于高深约束的实例深度估计改进，显著提高的简单难度指标表明本文方法对近处简单实例的检测更接近真实值。对于中等以及困难指标，多尺度混合注意力模块可以融合远近尺寸大小不一的目标特征，边缘融合模块对图像边缘有截断属性的实例特征进一步增强，高深约束对较远目标的深度估计也更为鲁棒，三者的综合改进提升了基准模型的中等以及困难难度指标性能。可见本文方法在严格条件场景与一般条件场景中都表现出了优秀的检测性能。

4.5 可视化结果

为了更直观地展示本文方法与基准模型 MonoDLE 的检测效果差异，可通过在验证集上可视化点云和原始图像上的检测结果并进行对比，得到如图 5 所示的可视化结果图。

图 5 验证集可视化结果

Fig. 5 Visualization results of validation set

分析图像和点云中的检测结果可知，本文方法与基准模型在近处目标的检测框与真实 3D 边界框的重合度非常高，都表现出了良好的检测能力。而对于远处目标的检测，本文方法的实例深度估计更准确，所生成的 3D 边界框明显更接近于真实标签生成的 3D 边界框。此外，对于图像边缘存在的截断目标，本文方法也能够给出较为准确的检测结果。综上所述，本文方法不仅有效提升了基准单目 3D 目标检测算法的检测性能，而且还能够支持对图像边缘的截断目标进行检测。

4.6 消融实验

为了进一步验证本文提出的 EF-DSC 模块、MSMADC 模块和 HDCGP 方法的有效性，进行了消融实验。在基准模型 MonoDLE 的基础上，分别采用了不同组合的 EF-DSC 模块、MSMADC 模块以及 HDCGP 方法，共进行了 8 组实验。实验结果如表 3 所列，其中 MonoDLE* 表示基准模型在本文实验环境下的结果。分析实验结果可以发现，单独应用 EF-DSC 模块以及 HDCGP 方法的模型都能显著提升基准模型在 3D 检测任务和 BEV 任务中的性能，而单独应用 MSMADC 模块的模型性能提升幅度一般。由于 EF-DSC 模块和 HDCGP 方法的改进分别针对图像边缘特征提取和实例深度估计，对单目场景下的检测任务更具针对性，因此可显著提高检测精度。相比之下，MSMADC 模块只在特征层级上完成多尺度特征提取改进，因此其性能提升受限。尽管如此，该模块可以在任意位置添加，这是其优点之一。基于对模型复杂度的考虑，本文所提模型仅在最高层特征图上添加该模块。此外，任意两种方法或模块的组合都能显著提升检测性能。特别是在使用 HDCGP 方法后的改进模型在 3D 检测任务和 BEV 任务的简单难度指标上都有明显提升。这表明相对于网络直接输出的深度，使用高深约束求解的实例深度对近处无遮挡微截断的简单标注目标检测更鲁棒。最后，使用 EF-DSC 模块、MSMADC 模块和 HDCGP 方法集成的最终模型，在 3D 检测任务和 BEV 任务中性能均取得了大幅提升，充分证明本文方法综合考虑了边缘特征、多尺度信息和高深约束的实例深度估计的改进，相比基准模型 MonoDLE 性能表现得到提升。

表 3 KITTI 验证集上的消融实验

Table 3 Ablation studies on KITTI validation set

Methods	$3D@IoU=0.7$			$BEV@IoU=0.7$		
	Easy	Mod	Hard	Easy	Mod	Hard
MonoDLE*	16.97	13.86	11.79	24.10	19.87	17.44
MonoDLE+EF-DSC	20.20	15.55	13.91	28.66	22.07	19.33
MonoDLE+MSMADC	17.38	14.45	12.28	24.87	19.90	17.91
MonoDLE+HDCGP	21.54	15.57	13.83	29.98	22.13	19.05
MonoDLE+EF-DSC+MSMADC	21.00	16.82	14.38	30.41	22.89	19.79
MonoDLE+EF-DSC+HDCGP	22.91	16.83	14.25	31.33	22.71	19.43
MonoDLE+MSMADC+HDCGP	22.63	17.31	14.51	30.92	23.05	19.75
MonoDLE+EF-DSC+MSMADC+HDCGP	**24.56**	**17.59**	**14.67**	**33.25**	**23.55**	**20.11**

结束语 本文提出了一种基于高深约束与边缘特征融合的单目 3D 目标检测算法。通过改进的高深约束计算实例深度方法，并引入基于深度可分离卷积的边缘融合模块以及基于空洞卷积的多尺度混合注意力模块，构建了本文所提单目 3D 目标检测算法。消融实验结果表明，相较于基准模型，本文方法在汽车 3D 检测精度上提升了 7.11%。在测试集对比实验中，本文方法在汽车 BEV 检测任务的 3 种难度级别上分别取得了 34.23%，21.69%和 18.92%的检测精度，明显优于其他方法。

尽管本文方法能够利用更好的高度预测方法求解高深约束，但是单一实例深度求解没有考虑到多实例间的几何关系。今后的工作将探索利用更多几何方法求解实例深度，同时在训练上探究如何更好地利用雷达点云数据提升单目图像的 3D 空间表示能力。

参 考 文 献

[1] ZHOU X, WANG D, KRAHENBUHL P. Objects as points [EB/OL]. (2019-04-16)[2022-09-24]. https://arxiv.org/abs/

1904.07850.
[2] LIU Z，WU Z，TOTH R. Smoke：Single-stage monocular 3d object detection via keypoint estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020：996-997.
[3] CHEN Y，TAI L，SUN K，et al. Monopair：Monocular 3d object detection using pairwise spatial relationships[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：12093-12102.
[4] MA X，ZHANG Y，XU D，et al. Delving into localization errors for monocular 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：4721-4730.
[5] DING M，HUO Y，YI H，et al. Learning depth-guided convolutions for monocular 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020：1000-1001.
[6] WANG L，DU L，YE X，et al. Depth-conditioned dynamic message propagation for monocular 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：454-463.
[7] ZHOU Z，DU L，YE X，et al. SGM3D：stereo guided monocular 3d object detection[J]. IEEE Robotics and Automation Letters，2022，7(4)：10478-10485.
[8] CHEN H，HUANG Y，TIAN W，et al. Monorun：Monocular 3d object detection by reconstruction and uncertainty propagation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：10379-10388.
[9] READING C，HARAKEH A，CHAE J，et al. Categorical depth distribution network for monocular 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：8555-8564.
[10] HUANG K C，WU T H，SU H T，et al. Monodtr：Monocular 3d object detection with depth-aware transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：4012-4021.
[11] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. arXiv：1706.03762，2017.
[12] ZHANG Y，LU J，ZHOU J. Objects are different：Flexible monocular 3d object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：3289-3298.
[13] LU Y，MA X，YANG L，et al. Geometry uncertainty projection network for monocular 3d object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：3111-3121.
[14] KUMAR A，BRAZIL G，CORONA E，et al. Deviant：Depth equivariant network for monocular 3d object detection[C]// Computer Vision-ECCV 2022：17th European Conference，Tel Aviv，Israel，Part IX. Cham：Springer Nature Switzerland，2022：664-683.
[15] HE K，GKIOXARI G，DOLLAR P，et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2961-2969.
[16] YU F，WANG D，SHELHAMER E，et al. Deep layer aggregation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：2403-2412..
[17] KENDALL A，GAL Y，CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7482-7491.
[18] WU Y，HE K. Group normalization[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：3-19.
[19] WANG Q，WU B，ZHU P，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：11534-11542.
[20] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.
[21] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2980-2988.
[22] MOUSAVIAN A，ANGUELOV D，FLYNN J，et al. 3d bounding box estimation using deep learning and geometry[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：7074-7082.
[23] 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.
[24] LIAN Q，YE B，XU R，et al. Exploring Geometric Consistency for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：1685-1694.
[25] FU H，GONG M，WANG C，et al. Deep ordinal regression network for monocular depth estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：2002-2011.

PU Bin，born in 1997，postgraduate. His main research interests include monocular 3D object detection and image classification.

LIANG Zhengyou，born in 1968，Ph.D，professor，is a member of CCF（No. 16803M）. His main research interests include computer vision，artificial intelligence and parallel distributed computing.

（责任编辑：何杨）