激光与光电子学进展 先进成像

联合实例深度的多尺度单目3D目标检测算法

王凤随[1,2,3*]，熊磊[1,2,3]，钱亚萍[1,2,3]

[1]安徽工程大学电气工程学院，安徽 芜湖 241000；

[2]检测技术与节能装置安徽省重点实验室，安徽 芜湖 241000；

[3]高端装备先进感知与智能控制教育部重点实验室，安徽 芜湖 241000

摘要 针对单目3D目标检测算法中存在图像缺乏深度信息以及检测精度不佳的问题，提出一种联合实例深度的多尺度单目3D目标检测算法。首先，为了增强模型对不同尺度目标的处理能力，设计基于空洞卷积的多尺度感知模块，同时考虑到不同尺度特征图之间的不一致性，从空间和通道两个方向对包含多尺度信息的深度特征进行重新精炼。其次，为了使模型获得更好的3D感知，将实例深度信息作为辅助学习任务来增强3D目标的空间深度特征，并使用稀疏实例深度来监督该辅助任务。最后，在KITTI测试集以及评估集上对所提算法进行验证。实验结果表明，所提算法相较于基线算法在汽车类别的平均精度提升了5.27%，有效提升了单目3D目标检测算法的检测性能。

关键词 测量；单目3D目标检测；实例深度学习；多尺度；注意力机制；辅助学习

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP222627

Multiscale Monocular Three-Dimensional Object Detection Algorithm Incorporating Instance Depth

Wang Fengsui[1,2,3*], Xiong Lei[1,2,3], Qian Yaping[1,2,3]

[1]*School of Electrical Engineering, Anhui Polytechnic University, Wuhu* 241000, *Anhui, China*;

[2]*Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Wuhu* 241000, *Anhui, China*;

[3]*Key Laboratory of Advanced Perception and Intelligent Control of High-End Equipment, Ministry of Education, Wuhu* 241000, *Anhui, China*

Abstract To solve the problems of lack of depth information and poor detection accuracy in conventional monocular three-dimensional (3D) target detection algorithms, an algorithm for multiscale monocular 3D target detection incorporating instance depth is proposed. First, to enhance the processing ability of the model for targets with different scales, a multiscale sensing module based on hole convolution is designed. Then, the depth features containing multiscale information are refined from both spatial and channel directions to remove the inconsistencies among different scale feature maps. Further, the instance depth information is used as an auxiliary learning task to enhance the spatial depth characteristics of 3D objects, and the sparse instance depth is used to monitor the auxiliary task, thereby improving the model's 3D perception. Finally, the proposed algorithm is tested and validated on the KITTI dataset. The experimental results show that the average accuracy of the proposed algorithm in the vehicle category is 5.27% higher than that of the baseline algorithm, indicating that the proposed algorithm effectively improves the detection performance compared with the conventional monocular 3D target detection algorithms.

Key words measurement; monocular 3D object detection; instance depth estimation; multiscale; attention mechanism; auxiliary learning

收稿日期：2022-09-26；**修回日期**：2022-10-20；**录用日期**：2022-10-24；**网络首发日期**：2022-11-04

基金项目：安徽省自然科学基金(2108085MF197,1708085MF154)、安徽高校省级自然科学研究重点项目(KJ2019A0162)、检测技术与节能装置安徽省重点实验室开放基金(DTESD2020B02)、安徽工程大学国家自然科学基金预研项目(Xjky2022040)、安徽高校研究生科学研究项目(YJS20210448、YJS20210449)

通信作者：*fswang@ahpu.edu.cn

1　引　　言

3D目标检测在各种计算机视觉应用中发挥着至关重要的作用，例如自动驾驶[1-2]、无人驾驶飞机、机器人操作和增强现实。从单目图像中对目标进行3D检测时，由于图像中缺乏可直接计算的目标深度信息，因此相较于使用雷达信息[3]和多摄像头系统进行3D目标检测的方法更困难。但是单目3D目标检测如果可以达到可靠的检测性能，其在实际应用中将具有低成本、低功耗和灵活部署的优势。

单目图像缺乏目标的深度信息，导致从中获取3D目标包围框成为一个不适定问题。为了解决单目图像中缺乏深度信息的问题，PatchNet[4]和DDMP-3D[5]采用增加额外数据的方式，即使用卷积神经网络回归深度估计图，虽然深度估计有助于3D目标检测，但现有的单目深度估计算法能力有限，深度估计不准确，导致检测精度低。其次，由于使用额外的深度估计模块，这类方法的推理速度通常很慢。此外，最近的方法在网络中引入有效但复杂的几何先验方法，MonoPair[6]设计了一个额外的成对约束分支，计算对象位置的不确定性感知预测和相邻对象对的3D距离，随后通过非线性最小二乘法联合优化。MonoFlex[7]采用一种新的检测器来联合计算匹配对象对之间的对象位置和空间约束，成对空间约束被建模为位于两个相邻对象之间的几何中心的关键点，有效地编码所有必要的几何信息，捕获对象之间的几何上下文。GUPNet[8]提出一种包含几何不确定性投影模块和分层任务学习策略的几何不确定性投影网络来解决投影过程中的误差放大问题。这些方法利用几何约束来弥补图像中缺乏准确深度信息的问题。Ma等[9]发现定位误差是限制单目3D检测性能的原因，提出直接检测投影的3D中心、丢弃远处物体样本以及面向3D交并比损失的3种策略，该方法是在CenterNet[10]的基础上构建的网络模型，仅仅使用预测对象中心周围的局部特征不足以理解场景级几何线索，导致不能准确估计对象的深度。MonoGRnet[11]将单目3D目标检测任务分解为4个子任务，分别是2D目标检测、实例级深度估计、投影3D中心估计和局部角点回归。其中，实例级深度估计对于弥合2D到3D差距起到重要作用。同时，在相似的2D目标检测工作中，增加不同尺度的特征处理被证明是能够有效提高模型性能的方法[12-13]。YOLOv3-SPP[14]将空间金字塔(SPP)[15]模块引入YOLOv3[16]，SPP能够提取具有不同感受野的多尺度深度特征，并在特征图的通道维度中连接融合，进而提高检测精度，但是池化会造成部分信息的丢失。atrous spatial pyramid pooling(ASPP)[17]采用具有不同采样率的多个并行空洞卷积层在多个尺度上捕获对象和图像上下文。针对不同尺度的特征信息，这些方法通过拼接的方法进行简单融合，并不能充分利用不同尺度的特征。

本文基于MonoDLE方法在KITTI基准中对以上问题进行分析。首先，针对单目3D目标检测由于图像缺乏深度信息而造成的深度预测不准问题，在3D检测任务中增加实例深度辅助学习任务，通过该任务增强模型对3D感知特征的获取。其次，设计一个新的多尺度感知模块，通过该模块增强模型对不同尺度信息的获取能力，以此来提高对图像中不同尺度目标的检测能力。最后，针对实例深度辅助学习任务，采用稀疏实例深度表示和L1损失对其进行监督学习，同时该任务仅在训练阶段实施，在推理阶段丢弃。在KITTI数据集上的实验结果表明，所提方法优于基线方法，且与先进的单目3D目标检测方法相比具有竞争力。

2　所提算法

2.1　实例深度学习模块

单目3D目标检测因为缺乏准确的深度信息，检测精度受到深度不确定性的影响，导致性能下降。为了聚合对象全局的视觉特征作为深度估计的提示，提高3D定位的准确性，设计了一个实例深度学习模块(IDLM)来生成前景物体的深度信息特征，作为辅助学习任务，通过探索特征图中的大感受野来捕获粗略的实例深度，辅助网络学习到更好的3D特征表示。

实例深度学习模块的结构如图1所示，该模块的输入特征图为DLANet中最后一个分层深度聚合(HDA)结构输出的特征图$\boldsymbol{F}\in\mathbf{R}^{64\times96\times320}$，然后应用一个3×3卷积和一个1×1卷积来获得输入图像的实例深度图特征$\boldsymbol{F}_{\mathrm{d}}\in\mathbf{R}^{1\times96\times320}$。为了使网络能够得到图像中目标粗略的深度图特征$\boldsymbol{F}_{\mathrm{d}}$，通过输入图像中的对象深度标签对其进行监督训练，利用对象的2D包围框标签来生成分辨率为96×320的实例深度图标签和掩膜。实例深度图标签中的像素值由图像中对象的深度

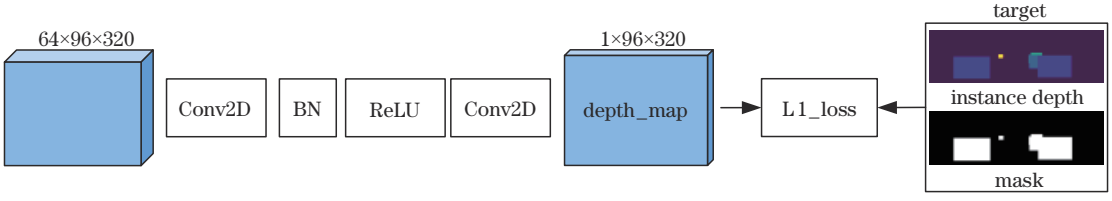

图1　实例深度学习模块

Fig. 1　Instance depth learning module

标签值来分配，若各个对象之间的包围框有重合，则重合位置的像素值为离相机最近的对象的深度标签值，这也符合图像的视觉外观。其次，将背景深度值赋值为 0。掩膜的像素值是由 0 和 1 组成的，由图像中对象的 2D 包围框标签生成，在包围框内的像素值赋值为 1，反之为 0，以此来对图像中的前景和背景进行分割。为了减少标注成本，不使用逐像素注释的像素级深度，而是使用稀疏监督直接预测目标 2D 边界框中的深度。

对于生成的实例深度图特征 $\boldsymbol{F}_{\mathrm{d}}$，使用 L1 损失函数对其进行约束学习，实例深度学习模块的表达式为

$$\boldsymbol{F}'=\mathrm{Conv}\left\{\mathrm{ReLU}\left\{\mathrm{BN}\left[\mathrm{Conv}(\boldsymbol{F})\right]\right\}\right\},\tag{1}$$

式中：Conv 表示二维卷积；ReLU 表示激活函数；BN 表示批量归一化层。

2.2　多尺度感知模块

对于不同尺度的目标用不同感受野去获取目标的特征图信息，能有效提高获取目标三维信息的能力。距离近的物体在图像中占据更多的像素，距离远的物体在图像中占据更少的像素，像素越少的图像携带的物体信息越少。因此，提出多尺度感知模块（MSS），自适应地学习不同尺度目标的特征信息。多尺度感知模块由空洞卷积模块、空间感知模块、通道感知模块和残差结构组成。空洞卷积模块负责提取不同感受野的特征图信息，空间感知模块和通道感知模块负责分别从空间和通道两个维度学习不同感受野的特征图之间的相关性，残差结构负责削弱空洞卷积带来的棋盘效应。

多尺度感知模块的结构如图 2 所示。该模块的输入特征图为 DLA34 的 level 5 层输出的深层特征图 $\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$。首先，特征图 $\boldsymbol{F}$ 通过空洞卷积模块得到不同感受野的特征图 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$，空洞卷积模块由 4 组 1×1 标准卷积块和 1 个 3×3 空洞卷积块并联组成，通过设置空洞卷积块中的空洞率得到不同大小感受野的特征图；其次，将特征图 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 分别输入空间感知模块和通道感知模块学习融合权重，空间感知模块先对特征图 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 分别使用 1×1 标准卷积降低通道维数，拼接后再使用 1×1 标准卷积进行重采样获得特征图 $\boldsymbol{Z}\in\mathbf{R}^{W\times H\times 4}$，然后将特征图 $\boldsymbol{Z}'$ 输入 Softmax 激活函数获得权重 $\boldsymbol{\alpha}\in\mathbf{R}^{H\times W\times 4}$，权重 $\boldsymbol{\alpha}$ 在通道维度上切片得到 $\boldsymbol{\alpha}_1\in\mathbf{R}^{H\times W\times 1}$、$\boldsymbol{\alpha}_2\in\mathbf{R}^{H\times W\times 1}$、$\boldsymbol{\alpha}_3\in\mathbf{R}^{H\times W\times 1}$、$\boldsymbol{\alpha}_4\in\mathbf{R}^{H\times W\times 1}$ 后，分别与特征图 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 相乘获得新的特征图后融合拼接输出特征图 $\boldsymbol{F}'\in\mathbf{R}^{H\times W\times C}$。空间感知模块的表达式为

$$\boldsymbol{F}'=\mathrm{cat}\left(\boldsymbol{\alpha}_1\cdot\boldsymbol{F}_1,\boldsymbol{\alpha}_2\cdot\boldsymbol{F}_2,\boldsymbol{\alpha}_3\cdot\boldsymbol{F}_3,\boldsymbol{\alpha}_4\cdot\boldsymbol{F}_4\right),\tag{2}$$

式中：cat(·)表示拼接；·表示按位相乘。

通道感知模块先对特征图 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 进行拼接融合，然后利用全局平均池化（GAP）产生全局上下文信息，通过一维卷积（$\mathrm{C1D}_k$）避免通道数据降维并捕捉跨维度联系，用 Sigmoid 激活函数输出通道感知注意力特征图 $\boldsymbol{\delta}\in\mathbf{R}^{1\times 1\times C}$。通道感知模块的表达式为

$$\boldsymbol{\delta}=\mathrm{Sigmoid}\left\{\mathrm{C1D}_k\left\{\mathrm{GAP}\left[\mathrm{cat}\left(\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3,\boldsymbol{F}_4\right)\right]\right\}\right\}。\tag{3}$$

将通道感知注意力特征图 $\boldsymbol{\delta}$ 映射乘以特征图 $\boldsymbol{F}'$，利用全局上下文信息优化不同感受野的特征图融合权重，进而得到特征图 $\boldsymbol{F}''\in\mathbf{R}^{W\times H\times C}$。为了削弱空洞卷积带来的棋盘效应，最后使用残差结构将原始输入深层特征图 $\boldsymbol{F}$ 与包含不同感受野的特征图 $\boldsymbol{F}''$ 相加输出特征图 $\boldsymbol{F}'''\in\mathbf{R}^{W\times H\times C}$。多尺度感知模块的表达式为

$$\boldsymbol{F}'''=\boldsymbol{F}+\boldsymbol{\delta}\times\boldsymbol{F}'。\tag{4}$$

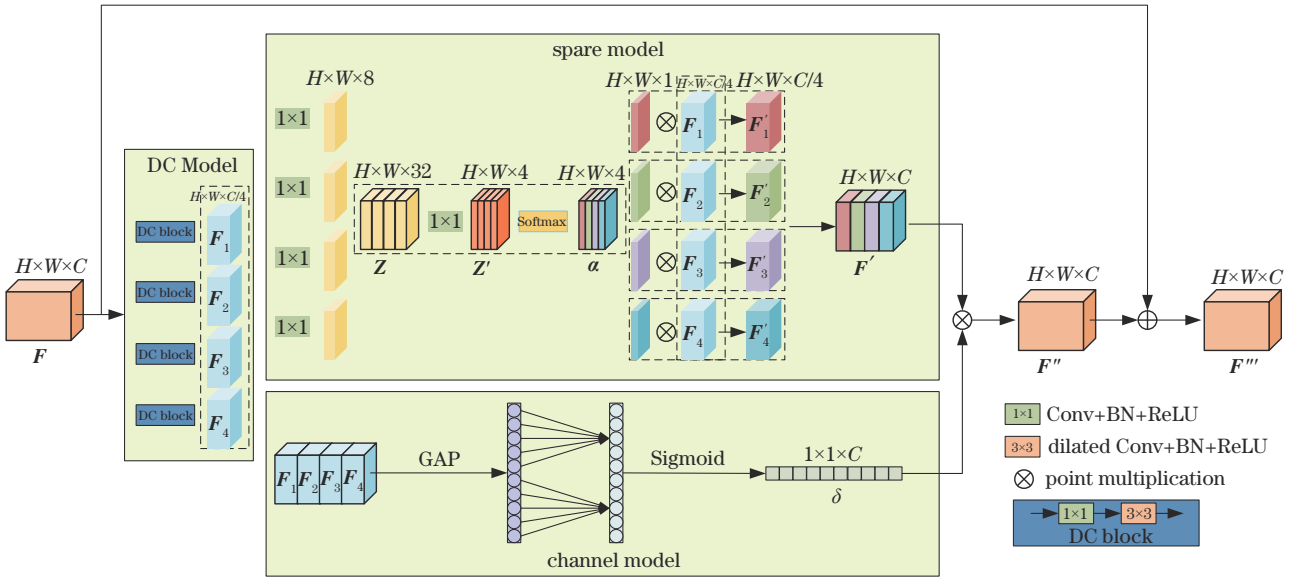

图 2　多尺度感知模块

Fig. 2　Multiscale sensing module

2.3　网络结构设计

网络结构如图 3 所示，所提基于 MonoDLE 的改进单目 3D 目标检测方法设计简单，由 4 部分组成：主干特征提取网络、多尺度感知模块、边界框回归头（DLAUp）、3D 边界框计算模块。

主干特征提取网络。给定分辨率为 3×384×1280 的输入图像 $\boldsymbol{f}$，使用 DLA 网络（DLA34）[18] 作为主干特征提取网络，计算维度为 512×12×40 的输出特征图 $\boldsymbol{f}_1$。

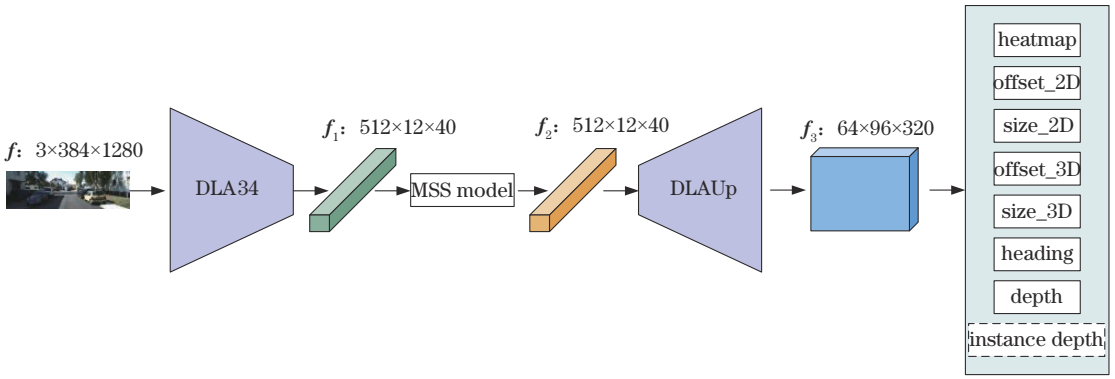

图 3　网络结构图

Fig. 3　Network structural diagram

多尺度感知模块。与之前工作不同的是，针对 3D 目标检测，为了获取不同感受野的特征信息，对于特征图 $\boldsymbol{f}_1$ 用多尺度感知模块对其进行重新采样得到大小为 512×12×40 的特征图 $\boldsymbol{f}_2$，使深度特征中包含更丰富的多尺度信息。

边界框回归头。热力图回归头使用投影的 3D 中心作为估计粗中心 C 的真值，预测粗中心。2D 偏移分支预测粗中心和 2D 边界框真实中心之间的偏移量。2D 尺寸分支预测 2D 边界框的大小。3D 偏移分支预测粗中心和投影 3D 边界框中心之间的偏移量。3D 尺寸分支预测 3D 边界框的大小。深度分支预测目标的真实深度。角度分支预测目标的观察角。新设计的最后一个分支，实例深度分支负责预测实例全局深度，仅用于训练阶段，在推理阶段丢弃，通过这种方式将实例深度学习作为辅助任务，有助于学习更好的 3D 感知特征。

3D 边界框计算模块。在推理阶段，基于最大池化和阈值（0.2）后的热力图峰值为每个目标的 2D 边界框中心，并索引出预测的 2D 和 3D 偏移量、2D 和 3D 边界框大小、深度和观察角度，继而确定目标的 3D 中心点以及包围框位置。

2.4　损失函数

MonoDLE 的损失函数由热力图损失函数 L_{k}、2D 偏移损失函数 $L_{\mathrm{o\_2d}}$、2D 尺寸损失函数 $L_{\mathrm{s\_2d}}$、3D 偏移损失函数 $L_{\mathrm{o\_3d}}$、3D 尺寸损失函数 $L_{\mathrm{s\_3d}}$、观察角损失函数 L_{head}，以及深度损失函数 L_{d} 组成，与 MonoDLE 不同的是，本研究在此基础上增加了由 L1 损失函数组成的实例深度损失函数 $L_{\mathrm{I\_d}}$，总损失函数可表示为

$$L_{\mathrm{total}}=L_{\mathrm{k}}+L_{\mathrm{o\_2d}}+L_{\mathrm{s\_2d}}+L_{\mathrm{o\_3d}}+L_{\mathrm{s\_3d}}+L_{\mathrm{head}}+L_{\mathrm{d}}+L_{\mathrm{I\_d}}。\tag{5}$$

热力图损失函数 L_{k} 采用 Focal 损失函数，解决样本不平衡问题，计算公式为

$$L_{\mathrm{k}}=\frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right), & Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\log\left(1-\hat{Y}_{xyc}\right), & \text{else}\end{cases},\tag{6}$$

式中：$\hat{Y}_{xyc}$ 为模型预测的热力图；Y_{xyc} 为热力图的标签值；α、β 为超参数，取 $\alpha=2$，$\beta=4$；N 是图像中目标的数量。

2D 偏移损失函数 $L_{\mathrm{o\_2d}}$、2D 尺寸损失函数 $L_{\mathrm{s\_2d}}$ 以及 3D 偏移损失函数 $L_{\mathrm{o\_3d}}$ 全部采用 L1 损失函数，用于约束模型的预测，计算公式为

$$L_{\mathrm{o\_2d}}=L_1\left(s_{\mathrm{o\_2d}},s^{*}_{\mathrm{o\_2d}}\right),\tag{7}$$

$$L_{\mathrm{s\_2d}}=L_1\left(s_{\mathrm{s\_2d}},s^{*}_{\mathrm{s\_2d}}\right),\tag{8}$$

$$L_{\mathrm{o\_3d}}=L_1\left(s_{\mathrm{o\_3d}},s^{*}_{\mathrm{o\_3d}}\right),\tag{9}$$

式中：$s_{\mathrm{o\_2d}}$ 表示模型预测的 2D 偏移量；$s^{*}_{\mathrm{o\_2d}}$ 表示 2D 偏移量的标签值；$s_{\mathrm{s\_2d}}$ 表示模型预测的 2D 包围框尺寸大小；$s^{*}_{\mathrm{s\_2d}}$ 表示 2D 包围框尺寸大小的标签值；$s_{\mathrm{o\_3d}}$ 表示模型预测的 3D 偏移量；$s^{*}_{\mathrm{o\_3d}}$ 表示 3D 偏移量的标签值；$L_1(\cdot)$ 表示 L1 损失函数。

3D 尺寸损失函数 $L_{\mathrm{s\_3d}}$ 采用维度感知 L1 损失函数，计算公式为

$$L_{\mathrm{s\_3d}}=L_1\left[\frac{\left(s_{\mathrm{s\_3d}}-s^{*}_{\mathrm{s\_3d}}\right)}{s_{\mathrm{s\_3d}}}\right],\tag{10}$$

式中：$s_{\mathrm{s\_3d}}$ 表示模型预测的 3D 包围框尺寸大小；$s^{*}_{\mathrm{s\_3d}}$ 表示 3D 包围框尺寸大小的标签值。

深度损失函数 L_{d} 采用拉普拉斯任意不确定性损失函数，计算公式为

$$L_{\mathrm{d}}=L_l\left(d,d^{*},\mathrm{d}m\right),\tag{11}$$

式中：d 表示模型预测的目标深度值；d^{*} 表示目标深度的标签值；$\mathrm{d}m$ 表示深度估计中的异方差任意不确定性；$L_l(\cdot)$ 表示拉普拉斯任意不确定性损失函数。

观察角损失函数L_{head}采用标准交叉熵损失函数，计算公式为

$$L_{head}=L_{cross\_entropy}(s_{head},s^{*}_{head}),\tag{12}$$

式中：s_{head}表示模型预测的观察角角度值；s^{*}_{head}表示目标的观察角标签值；$L_{cross_entropy}(\cdot)$表示标准交叉熵损失函数。

实例深度损失函数L_{I_d}采用L1损失函数，计算公式为

$$L_{I\_d}=L_1(V_{Id}\times V_{mask},V^{*}_{Id}),\tag{13}$$

式中：V_{Id}表示模型预测的实例深度值；V_{mask}表示掩膜值；V^{*}_{Id}表示实例深度标签值。

3　实验结果与分析

3.1　KITTI数据集介绍

在KITTI数据集[19-20]上评估所提方法的有效性，该数据集提供了7481张用于训练的图像和7518张用于测试的图像。由于缺少测试集的标签并且提交到官方服务器进行测试的访问受到限制，遵循先前工作的协议[21-22]将训练数据集分为训练集(3712张图像)和验证集(3769张图像)。基于此拆分进行消融研究并分析，最后在7481张图像上进行训练，并在KITTI官方服务器进行测试。

KITTI数据集为自动驾驶场景提供了许多广泛使用的评价指标，包括3D检测、鸟瞰(BEV)检测和平均方向相似度(AOS)。本研究报告了这些任务在3种难度设置(简单、中等和困难)下具有40个召回位置的平均精度(AP40)[23]，主要测试了汽车类别的表现，同时默认交并比(R_{IOU})阈值为0.7、0.5、0.5。

3.2　实验参数设置

实验中所使用的操作系统为Linux，GPU为两块RTX 2080Ti，处理器为24核Intel(R) Xeon(R) Platinum 8255C CPU @2.50 GHz，深度学习框架为PyTorch 1.1.0，为了公平起见，实验参数参照MonoDLE进行设置。以端到端的方式对网络进行140个epoch的训练，并将batch_size设置为16，同时使用初始学习率为1.25×10^{-3}的Adam优化器，并在90和120个epoch时将其衰减为原来的1/10。权重衰减设置为1×10^{-5}，并且前5个epoch使用预热策略。为了避免过度拟合，采用随机裁剪/缩放(仅用于2D检测)和随机水平翻转扩充数据。在此设置下，整个训练过程大约需要9 h。

3.3　定量评价

为了验证所提算法的有效性，在KITTI测试集上与其他算法进行对比，结果如表1所示。其中，Improvement中的数据代表所提方法相对于4种不同类型额外数据方法中表现最好的方法的提升量，例如，Improvement中Depth这一行的第一个数据+2.22表示所提方法的AP40相对于文献[25]算法提升2.22个

表1　汽车类别在KITTI测试集上的性能

Table 1　Performance of the Car category on the KITTI test set　　unit: %

Method	Extra data	AP40(3D@$R_{IOU}\geqslant0.7$)			AP40(BEV@$R_{IOU}\geqslant0.7$)		
		Easy	Moderate	Hard	Easy	Moderate	Hard
AM3D[24]	Depth	16.50	10.74	9.52	25.03	17.32	14.91
PatchNet[4]	Depth	15.68	11.12	10.17	22.97	16.86	14.97
DDMP-3D[5]	Depth	19.71	12.78	9.80	28.08	17.89	13.44
Reference [25]	Depth	20.28	13.12	9.56			
Kinematic3D[26]	Multi-frames	19.07	12.72	9.17	26.69	17.52	13.10
CaDDN[27]	LiDAR	19.17	13.41	11.46	27.94	18.91	17.19
MonoRUn[28]	LiDAR	19.65	12.30	10.58	27.94	17.34	15.24
MonoGRNet[11]	None	9.61	5.74	4.25	18.19	11.17	8.73
MonoDIS[23]	None	10.37	7.94	6.40	17.23	13.19	11.12
Reference [29]	None	20.89	14.49	12.19	29.57	20.77	17.88
MonoPair[6]	None	13.04	9.99	8.65	19.28	14.83	12.89
FADNet[30]	None	16.37	9.92	8.05	23.00	14.22	12.56
MonoDLE[9]	None	17.23	12.26	10.29	24.79	18.89	16.00
MonoGround[31]	None	19.48	14.36	12.62	30.07	20.47	17.74
MonoFlex[7]	None	19.94	13.89	12.07	28.23	19.75	16.89
MonoEF[32]	None	21.29	13.87	11.71	29.03	19.70	17.26
Reference [33]	None	21.65	13.25	9.91	29.81	17.98	13.08
GUPNet[8]	None	22.26	15.02	13.12	30.29	21.19	18.20
MonoCon[34]	None	22.50	**16.49**	**13.95**	31.12	22.10	19.00
Proposed method	None	**22.50**	16.19	13.49	**32.44**	**22.97**	**19.82**
Improvement	Depth	+2.22	+3.07	+3.32	+4.36	+5.08	+4.85
	Multi-frames	+3.43	+3.47	+4.78	+5.75	+5.45	+6.72
	LiDAR	+2.85	+2.78	+2.03	+4.5	+4.06	+2.63
	None	+0	−0.3	−0.46	+1.32	+0.87	+0.82

百分点。总体而言，所提算法相较于其他经典算法取得了更好的结果。例如，所提算法在3D检测任务的简单/中等/困难设置下相对于MonoDLE分别获得了5.27个百分点/3.93个百分点/3.2个百分点的改进。与具有额外数据的算法相比，例如DDMP-3D和PatchNet，所提算法仍然有一定的竞争力，这进一步证明了其有效性。其次，与没有额外数据的算法相比，在BEV任务下，所提算法的表现要高于MonoCon，而在3D任务下的MonoCon的检测能力要稍好。

表2展示了所提算法在KITTI验证集上的性能。由于DORN[35]的训练集与KITTI3D的验证集重叠，部分算法直接使用DORN提供的预训练模型作为其深度估计器，所以不与这些算法进行对比。从表2可以看出，所提算法无论是在3D任务还是BEV任务都表现得更好。与基线方法MonoDLE相比，所提算法在严格条件下（R_{IOU}为0.7）下检测目标的能力提升较大，在3D任务和BEV任务的简单/中等/困难设置下分别提升5.14个百分点/4.13个百分点/3.31个百分点和5.25个百分点/3.98个百分点/3.03个百分点。其次，所提算法在松散条件下（R_{IOU}为0.5）下检测目标的能力提升明显，基线方法MonoDLE相较于MonoPair在3D任务下优势不明显，且在BEV任务下表现略差。但是所提算法改进后，通过引入多尺度信息和实例深度辅助学习的方法，在3D任务的简单/中等/困难设置下相较于基线方法提升了8.81个百分点/4.07个百分点/4.91个百分点，在BEV任务的简单/中等/困难设置下相较于MonoPair提升了7.41/4.49/4.19。其次，在单个GTX 1080Ti GPU上测试所提模型，设置批量大小为1进行运行时间分析，相较于基线方法，其在推理阶段一张图像消耗时间增加5 ms。

表2 汽车类别在KITTI验证集上的性能

Table 2 Performance of the Car category on the KITTI validation set

Method	AP40 /%（3D@R_{IOU}=0.7）			AP40 /%（BEV@R_{IOU}=0.7）			AP40 /%（3D@R_{IOU}=0.5）			AP40 /%（BEV@R_{IOU}=0.5）			Runtime / ms
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	
CenterNet[10]	0.60	0.66	0.77	3.46	3.31	3.21	20.00	17.50	15.57	34.36	27.91	24.65	
MonoGRNet	11.90	7.56	5.76	19.72	12.81	10.15	47.59	32.28	25.50	48.53	35.94	28.59	60
MonoDIS	11.06	7.60	6.37	18.45	12.58	10.66							
M3D-RPN	14.53	11.07	8.65	20.85	15.62	11.88	48.53	35.94	28.59	53.35	39.60	31.76	161
MonoPair	16.28	12.30	10.42	24.12	18.17	15.76	55.38	42.39	37.99	61.06	47.63	41.92	57
MonoDLE	17.45	13.66	11.68	24.97	19.33	17.01	55.41	43.42	37.81	60.73	46.87	41.89	40
Proposed method	**22.59**	**17.79**	**14.99**	**30.22**	**23.31**	**20.04**	**64.22**	**47.49**	**42.72**	**68.47**	**52.12**	**46.11**	45
Improvement	+5.14	+4.13	+3.31	+5.25	+3.98	+3.03	+8.81	+4.07	+4.73	+7.41	+4.49	+4.19	

表3展示了所提算法分别添加多尺度感知模块和深度引导模块在KITTI验证集上的结果。从表3可以看出，分别在模型中添加多尺度感知模块和深度引导模块，都能提高模型的精度。首先，在模型中添加pyramid scene parsing（PSP）[36]和ASPP结构后，模型性能有所提升，这说明在网络中引入多尺度信息获取模块可以有效提升模型单目3D目标检测的性能。其次，在模型中添加所提多尺度感知模块后，性能提升相较于PSP和ASPP更为明显，在使用空洞卷积获取多尺度信息的同时考虑到不同尺度特征之间的不一致性，通过从空间和通道维度对不同尺度信息进行融合，利用残差结构削弱棋盘效应，进而获得丰富的多尺度信息来提高不同任务下的检测精度。

表3 添加不同模块在KITTI验证集上的性能对比

Table 3 Adding Performance Comparison of Different Modules on KITTI validation set unit: %

Method	AP40(3D)			AP40(BEV)		
	Easy	Moderate	Hard	Easy	Moderate	Hard
Baseline	17.45	13.66	11.68	24.97	19.33	17.01
+ASPP	18.04	14.72	12.47	25.60	20.84	18.20
+PSP	18.98	14.73	12.38	25.61	20.75	18.06
+MSS	20.48	15.89	14.06	28.58	22.41	19.46
+IDLM	21.76	16.19	14.24	28.71	22.44	19.43
+MSS+IDLM	**22.59**	**17.79**	**14.99**	**30.22**	**23.31**	**20.04**
Improvement	+5.14	+4.13	+3.31	+5.25	+3.98	+3.03

将实例深度学习模块作为训练过程中的辅助任务，以此来帮助单目3D目标检测，让模型在训练过程中学习实例深度信息，使模型中含有更多的3D感知特征，进而提高模型对目标的定位精度，在3D任务中简单设置下提升尤为明显。最后，同时将两个模块应用在算法中，在多尺度信息和深度信息两个方向同时增强网络的表达能力，进而提高模型的性能。

3.4 可视化结果分析

为了清楚地展示所提算法的有效性，在图4中可视化了在KITTI上的检测结果，将所提算法的预测框和基线算法的预测框以及真值框同时绘制在RGB图像和LiDAR信号中，图4中用绿色包围框、黄色包围框和红色包围框分别代表真值框、基线算法的预测框和所提算法的预测框。LiDAR信号仅用于可视化，且在LiDAR中标注了目标类别。

通过对比图像中的检测结果可以发现，无论是基线算法还是所提算法，针对合理距离内的目标都输出了较为准确的3D边界框。对比图4右侧图像可以发现，所提算法在位置预测上的表现要优于基线算法，预测框更贴近真值框。综上所述，所提算法可以有效改善单目3D目标检测模型的性能，提高检测精度。

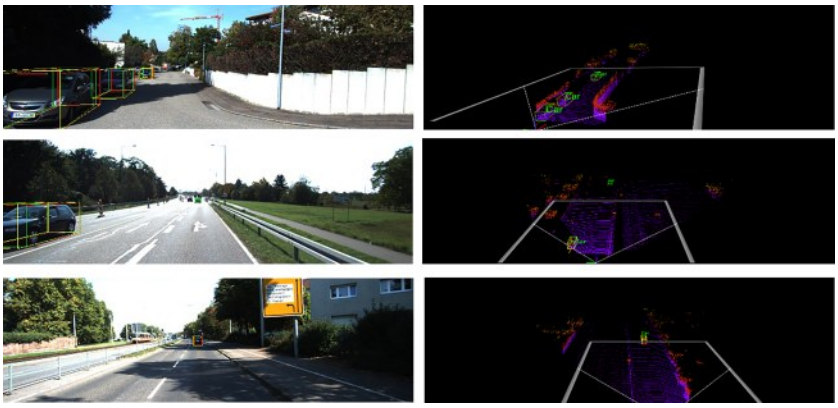

图4 KITTI的可视化结果

Fig. 4 Visualization results of KITTI

4 结 论

提出一种联合实例深度的多尺度单目3D目标检测算法。通过对深层特征使用多尺度感知模块，利用空洞卷积获取多尺度特征信息同时从空间和通道方向重新精炼特征，进而提高对不同尺度3D目标的检测能力。同时增加实例深度辅助学习任务，仅通过稀疏深度标签对其进行监督，有效提升了算法对3D特征的感知能力，进一步提高了算法处理3D目标的感知能力。在KITTI测试集上的实验结果验证了所提算法的有效性。

参 考 文 献

[1] 胡杰，刘汉，徐文才，等. 基于三维激光雷达的道路障碍物目标位姿检测算法[J]. 中国激光，2021，48(24)：2410001.
Hu J，Liu H，Xu W C，et al. Position detection algorithm of road obstacles based on 3D LiDAR[J]. Chinese Journal of Lasers，2021，48(24)：2410001.

[2] 赵亮，胡杰，刘汉，等. 基于语义分割的深度学习激光点云三维目标检测[J]. 中国激光，2021，48(17)：1710004.
Zhao L，Hu J，Liu H，et al. Deep learning based on semantic segmentation for three-dimensional object detection from point clouds[J]. Chinese Journal of Lasers，2021，48(17)：1710004.

[3] 龚威，史硕，陈博文，等. 机载高光谱激光雷达成像技术发展与应用[J]. 光学学报，2022，42(12)：1200002.
Gong W，Shi S，Chen B W，et al. Development and application of airborne hyperspectral LiDAR imaging technology[J]. Acta Optica Sinica，2022，42(12)：1200002.

[4] Ma X Z，Liu S N，Xia Z Y，et al. Rethinking pseudo-LiDAR representation[M]//Vedaldi A，Bischof H，Brox T，et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham：Springer，2020，12358：311-327.

[5] Wang L，Du L，Ye X Q，et al. Depth-conditioned dynamic message propagation for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 20-25，2021，Nashville，TN，USA. New York：IEEE Press，2021：454-463.

[6] Chen Y J，Tai L，Sun K，et al. MonoPair：monocular 3D object detection using pairwise spatial relationships[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 13-19，2020，Seattle，WA，USA. New York：IEEE Press，2020：12090-12099.

[7] Zhang Y P，Lu J W，Zhou J. Objects are different：flexible monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 20-25，2021，Nashville，TN，USA. New York：IEEE Press，2021：3288-3297.

[8] Lu Y, Ma X Z, Yang L, et al. Geometry uncertainty projection network for monocular 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 3091-3101.

[9] Ma X Z, Zhang Y M, Xu D, et al. Delving into localization errors for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 4719-4728.

[10] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points [EB/OL]. (2019-04-16)[2022-09-24]. https://arxiv.org/abs/1904.07850.

[11] Qin Z Y, Wang J L, Lu Y. MonoGRNet: a general framework for monocular 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5170-5184.

[12] 鞠默然, 罗江宁, 王仲博, 等. 融合注意力机制的多尺度目标检测算法[J]. 光学学报, 2020, 40(13): 1315002.
Ju M R, Luo J N, Wang Z B, et al. Multi-scale target detection algorithm based on attention mechanism[J]. Acta Optica Sinica, 2020, 40(13): 1315002.

[13] 刘芳, 吴志威, 杨安喆, 等. 基于多尺度特征融合的自适应无人机目标检测[J]. 光学学报, 2020, 40(10): 1015002.
Liu F, Wu Z W, Yang A Z, et al. Multi-scale feature fusion based adaptive object detection for UAV[J]. Acta Optica Sinica, 2020, 40(10): 1015002.

[14] Zhang P Y, Zhong Y X, Li X Q. SlimYOLOv3: narrower, faster and better for real-time UAV applications [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 37-45.

[15] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[16] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2022-09-24]. https://arxiv.org/abs/1804.02767.

[17] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[18] Yu F, Wang D Q, Shelhamer E, et al. Deep layer aggregation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2403-2412.

[19] 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.

[20] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[21] Chen X Z, Kundu K, Zhang Z Y, et al. Monocular 3D object detection for autonomous driving[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2147-2156.

[22] Chen X Z, Kundu K, Zhu Y K, et al. 3D object proposals for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 40(5): 1259-1272.

[23] Simonelli A, Bulò S R, Porzi L, et al. Disentangling monocular 3D object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 1991-1999.

[24] Ma X Z, Wang Z H, Li H J, et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, Seoul, Republic of Korea. New York: IEEE Press, 2019: 6850-6859.

[25] Liu H, Liu H P, Wang Y K, et al. Fine-grained multilevel fusion for anti-occlusion monocular 3D object detection[J]. IEEE Transactions on Image Processing, 2022, 31: 4050-4061.

[26] Brazil G, Pons-Moll G, Liu X M, et al. Kinematic 3D object detection in monocular video[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12368: 135-152.

[27] Reading C, Harakeh A, Chae J L, et al. Categorical depth distribution network for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8551-8560.

[28] Chen H S, Huang Y Y, Tian W, et al. MonoRUn: monocular 3D object detection by reconstruction and uncertainty propagation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 10374-10383.

[29] Zhou D F, Song X B, Fang J, et al. Context-aware 3D object detection from a single image in autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18568-18580.

[30] Gao T Z, Pan H H, Gao H J. Monocular 3D object detection with sequential feature association and depth hint augmentation[J]. IEEE Transactions on Intelligent Vehicles, 2022, 7(2): 240-250.

[31] Qin Z Q, Li X. MonoGround: detecting monocular 3D objects from the ground[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 3783-3792.

[32] Zhou Y S, He Y, Zhu H Z, et al. MonoEF: extrinsic

parameter free monocular 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10114-10128.

[33] Liu Y X, Yuan Y X, Liu M. Ground-aware monocular 3D object detection for autonomous driving[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 919-926.

[34] Liu X P, Xue N, Wu T F. Learning auxiliary monocular contexts helps monocular 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1810-1818.

[35] Fu H, Gong M M, Wang C H, et al. Deep ordinal regression network for monocular depth estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2002-2011. [LinkOut]

[36] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.