文章编号:1671-9352(2025)01-0063-11 DOI:10.6040/j.issn.1671-9352.4.2023.0213

结合深度信息引导和多尺度通道注意力机制的单目三维目标检测算法

刘青[1],李伟[1*],余少勇[2],宋宇萍[3],周启迪[1],邹伟林[1]
(1.厦门理工学院计算机与信息工程学院, 福建 厦门 361024; 2.龙岩学院数学与信息工程学院,福建 龙岩 364012; 3.厦门大学数学科学学院, 福建 厦门 361005)

摘要:针对三维边界框无法从缺少空间线索的单目图像中准确估计的问题,本文提出一种基于深度信息引导和多尺度通道注意力机制的单目三维目标检测算法。为了引入三维信息并有效地获取和利用不同尺度特征图的空间信息,在特征提取模块中利用多尺度分割注意力算法,分别从单目图像和深度图中提取多尺度预处理特征图,利用通道注意力算法进行权重标定,提高了特征图的表征能力。通过深度引导动态局部卷积网络,将包含多尺度信息的深度图特征作为单目图像特征的特定卷积核,引入三维信息作为指导,减少直接融合的误差累积,并解决单目视觉中近大远小的尺度敏感问题。选择不同的评估指标对模型的性能进行评价与比较。实验结果表明,同其他算法相比,本文算法的自动驾驶数据集中汽车、行人、骑自行车的人的三维目标检测平均精度均提高。

关键词:单目三维目标检测;深度引导;多尺度通道注意力机制;自动驾驶
中图分类号:TP391 **文献标志码**:A
引用格式:刘青,李伟,余少勇,等. 结合深度信息引导和多尺度通道注意力机制的单目三维目标检测算法[J]. 山东大学学报(理学版),2025,60(1):63-73,82.

Monocular 3D object detection algorithm combining depth guidance and multi-scale channel attention mechanism

LIU Qing[1], LI Wei[1*], YU Shaoyong[2], SONG Yuping[3], ZHOU Qidi[1], ZOU Weilin[1]
(1. School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China; 2. School of Mathematics and Information Engineering, Longyan University, Longyan 364012, Fujian, China; 3. School of Mathematical Sciences, Xiamen University, Xiamen 361005, Fujian, China)

Abstract: For issues where the absence of essential spatial structure signals makes it highly challenging to estimate 3D bounding boxes accurately from a single picture, a monocular 3D object detection algorithm is proposed based on a multi-scale channel attention mechanism plus depth guidance to conquer these challenges. To introduce 3D data and effectively capture spatial information from different scales of feature maps, the depth maps and monocular image feature maps are pre-processed in the feature extraction module using a pyramid split algorithm, respectively, and then on the basic of the weight using the channel-wise attention module to calibrate the corresponding feature vectors to generate a refined feature map which is richer in multi-scale feature information. A depth-guided dynamic local convolution network is suggested for applying depth maps as specific kernels that contain spatial structure signals to monocular image feature maps. This method mitigates error accumulation from direct fusion and addresses the scale sensitivity issue of objects looking larger or smaller with distance. The model's performance is assessed and also compared using various evaluation metrics. Experimental results demonstrate that the method proposed in this paper improves the 3D detection accuracy for cars, pedestrians and cyclists in the autonomous driving datasets when compared to other algorithms.

Key words: monocular 3D object detection; depth guidance; multi-scale channel-wise attention mechanism; autonomous driving

收稿日期:2023-06-02; **网络出版时间**:2024-02-28 19:20:44
基金项目:教育部人文社会科学研究规划基金资助项目(23YJAZH067);国家留学基金资助项目(202308350042);厦门市科学技术局产学研资助项目(2023CXY0409);厦门理工学院研究生教育教学改革研究资助项目(YJS20220617)
第一作者:刘青(1998—),女,硕士研究生,研究方向为深度学习、三维目标检测. E-mail: 2783687819@ qq.com
*** 通信作者**:李伟(1979—),男,副教授,博士,研究方向为机器学习、粒计算、图像处理、信息安全. E-mail: drweili@ hotmail.com

0 引言

近年来,基于深度学习的三维目标检测方法成为实现自动驾驶的关键技术之一,备受研究者青睐。现有的三维目标检测算法可分为基于视觉(图像)和基于激光雷达点云两类,其中基于激光雷达和立体相机的算法表现出较高性能,但设备昂贵,限制了它们的推广和应用。基于单目照相机的低成本、易于应用的三维目标检测算法以及基于伪激光雷达的高性能三维目标检测算法显示出巨大潜力,然而,前者仍然存在一些问题:(1)相机的透视投影导致不同距离的单目图像尺度发生显著变化;(2)相机无法提供三维目标检测所需的深度信息(即空间线索),很难区分背景和目标对象。后者尚未解决的问题包括:(1)该方法很大程度上依赖于从单目图像中进行深度估计的准确性;(2)获得空间信息的同时丢失了单目图像的高级语义信息,例如路障、电子箱或道路上的灰尘等区域可能导致错误的检测,但通过使用单目图像可以轻松识别它们。

考虑经济可代替性和高性能,本文提出一种基于深度信息引导和多尺度通道注意力机制的单目三维目标检测算法。在特征提取模块中引入多尺度分割注意力算法,利用该算法处理多尺度特征的空间信息,对单目图像和深度图同时进行高效的特征提取。在基于深度引导的动态局部卷积模块中,使用深度图引导操作代替标准的直接融合操作,被训练过的深度图特征成为单目图像特征的特定卷积核,能减少直接融合操作带来的误差累积,结合图像的语义信息,获取更好的 2D-3D 的表示(2D-3D 表示 2D 向 3D 转换)。

1 相关工作

基于图像的单目三维目标检测通常可分为以下 3 组:(1)标准单目三维目标检测[1-2]仅使用数据集提供的单目图像、注释和相机校准文件,引入新的几何约束来提高模型的表示能力。例如 Simonelli 等[3]提出设置十元组表示预测框的参数而不是直接回归 3D 框的 8 个点,将 2D、3D 损失函数直接解耦。Brazil 等[4]重新构造了独立的 3D 区域提议网络(region proposal network, RPN)。(2)添加额外数据。例如,文献[5-6]引入形状信息,尝试找到物体的关键点,将目标检测的问题转化为对关键点的回归问题。文献[7-8]采用了多尺度采样方式,不断加入辅助参数预测深度信息,减少了误差损失。(3)设置偏移量补偿 2D、3D 框的中心点错位。例如 Wang 等[9]提出了中心回归模块(center regression module, CRM),从点向特征中心的位置回归。一般来说,2D、3D 框的中心点并不一致,Chen 等[10]通过回归一个偏移量来补偿它们之间的差异,Yin 等[11]则通过关键点检测器检测对象的中心并回归三维大小、方向;然而,由于基于二维图像特征来表示三维结构可能导致误差累积,因此影响目标检测的准确性。为了获得更精确的坐标和结构信息,直接使用三维数据作为输入是一个可行的解决方案,然而,激光雷达传感器的高成本和有限的测距范围使得研究人员须要寻找低成本的替代方案来代替真实的雷达点云数据。

综合考虑以上 2 种方法的限制之后,研究者们提出并采用了基于伪激光雷达的单目三维目标检测方法。这种方法借鉴了深度估计、视差估计和基于真实激光雷达的三维目标检测方法的优势,实质上构建了图像与激光雷达(通常以点云表示)之间的桥梁。通过这种方法,可以在低成本的情况下获得相对准确的三维目标检测结果。在这个条件下,须要先从二维图像中估计出密集的深度图[12](或者视差图[13]并将其转换为深度图[14])。利用公式(1)推导出像素的三维位置(x,y,z),形成伪雷达信号数据:

$$\begin{cases} x=z(m-P_x)/f, \\ y=z(n-P_y)/f, \\ z=d, \end{cases} \tag{1}$$

式中,(P_x,P_y)表示主点,(m,n)表示二维位置,f表示焦距,d表示深度。最后,将其作为激光雷达检测器的输入,完成三维目标检测的任务。比如,文献[15]提出了基于图形的深度校正(graph-based depth correction, GDC)算法是把预测出来的伪雷达和实际稀疏的雷达数据进行融合修正。这一算法克服了基于图像与激光雷达方法之间的障碍,通过采用不同的深度估计器也证实了该方法的性能在很大程度上依赖深度图和视差图的准确性[16-17],视差中的小误差会导致遥远物体的深度上的大误差,在很大程度上影响了点云的坐标表示。基于伪激光雷达方法主要考虑如何产生可代替真实雷达点云数据的伪雷达数据;但考虑图像的语义信

息较少,可能导致目标检测的准确性下降,因此本文使用深度图作为引导,更好完成三维目标检测。

2　算法

本文提出一种提出一种基于深度信息引导和多尺度通道注意力机制的单目三维目标检测算法,该算法主要由 3 个关键模块组成:基于多尺度通道注意力的特征提取模块、基于深度引导的动态局部卷积模块、基于锚框的 2D-3D 检测模块。网络结构如图 1 所示(图中 C 为特征通道数,C_i 为特征通道分割后的各组通道数,分组序号 $i=1,2,\cdots,N$;H 和 W 分别是特征图的高和宽)。基于多尺度通道注意力机制[18]的特征提取模块分别提取单目图像 $\boldsymbol{R}$ 和深度图的特征 $\boldsymbol{D}$,处理多尺度特征图的空间信息;基于深度引导的动态局部卷积模块是将膨胀后移位[19]深度图特征 $\boldsymbol{G}(\boldsymbol{R})$ 作用在单目图像特征上得到融合特征 $\boldsymbol{R}'$,实现深度引导动态局部卷积;基于锚框的 2D-3D 检测模块使用基于先验的单级检测器完成 2D 数据在 3D 空间的表示,并采用 Confluence算法[20]生成三维预测框。

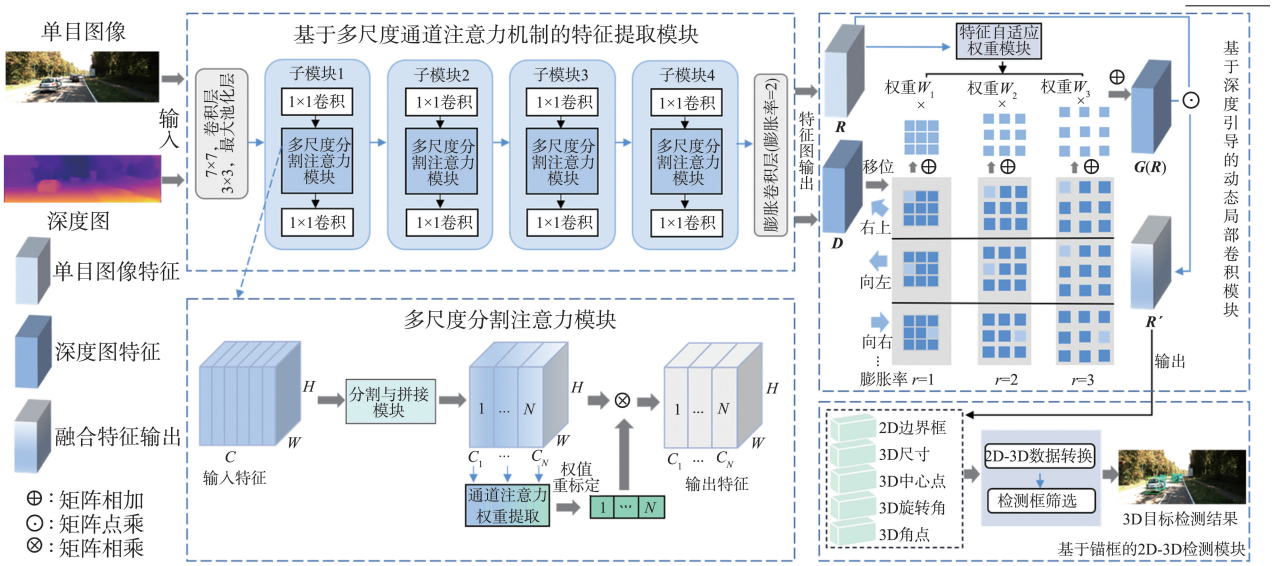


图 1　基于深度信息引导和多尺度通道注意力机制的单目三维目标检测算法网络结构

Fig.1　Illustration of a monocular 3D object detection algorithm combining depth guidance and multi-scale channel attention mechanism

2.1　基于多尺度通道注意力机制的特征提取模块

为了有效地获取和利用不同尺度的特征图的空间信息,丰富特征空间,本节采用多尺度通道注意力机制的网络[18]作为骨干网络完成特征提取。该模块包含单目图像和深度图 2 个输入数据[12]。2 个输入数据经过相同的特征提取网络,分别获得单目图像特征和深度图特征。

多尺度通道注意力网络 MSANet 整体结构的设计可以看作将残差网络[21](residual network, ResNet)中 3×3 卷积替换成多尺度分割注意力(multi-scale split attention, MSA)模块。MSA 模块包含以下 3 个步骤。

步骤 1　通过分割与拼接(split and concat, SC)模块实现多尺度特征提取,具体操作如下:先将输入特征定义为 $\boldsymbol{F}_X \in \mathbf{R}^{C\times H\times W}$,$C$、$H$ 和 W 分别为特征图的通道数、高和宽;再将其分割成 N 个部分,被分割后的每部分通道数 $C'=C/N$,分割后的特征图 $\boldsymbol{F}_{Xi}=\mathbf{R}^{C'\times H\times W}$,$i=1,2,\cdots,N$。接着,$\boldsymbol{F}_{Xi}$通过不同尺度的卷积核独立地学习多尺度空间信息。为了减少参数量,引入分组卷积,并设计了一种根据卷积核尺度大小 K 来自适应地调整分组大小 g 的策略得到多尺度预处理特征图为

$$\boldsymbol{F}'_{Xi}=G(K_i,g_i)\boldsymbol{F}_{Xi}, \tag{2}$$

式中 $g=2^{K-1/2}$。将分组卷积函数定义为 G,第 i 个卷积核大小和分组大小分别为 $K_i=2i+1$, $g_i=2^{K_i-1/2}$,利用拼接函数 T 得到的特征图 $\boldsymbol{F}''$为

$$\boldsymbol{F}''=T(\boldsymbol{F}'_{X1},\boldsymbol{F}'_{X2},\cdots,\boldsymbol{F}'_{XN})。 \tag{3}$$

步骤 2　对不同尺度提取的特征图 $\boldsymbol{F}'_{Xi}$ 通过通道注意力提取模块[22]进行权重提取,通过全局平均池化

将每个通道上的空间特征编码(压缩)为一个全局特征;再通过 2 个全连接层使高、低维度的通道之间进行交互;通过激活函数 sigmoid 得到不同尺度特征图 $\boldsymbol{F}'_{Xi}$ 的权重 $\boldsymbol{W}_i$;最后,将所有多尺度特征图的权重向量逐元素相加为

$$\boldsymbol{W}=\boldsymbol{W}_1\oplus\boldsymbol{W}_2\oplus\cdots\oplus\boldsymbol{W}_N。\tag{4}$$

步骤 3 为了实现局部和全局特征的注意力之间的交互,使用归一化指数函数 softmax(定义为 M)对权重 $\boldsymbol{W}_i$ 重标定,并与对应尺度的特征图逐元素相乘得到多尺度加权特征 $\boldsymbol{X}_i$,最后将其拼接获得更具表征能力的特征图 $\boldsymbol{X}$。$\boldsymbol{X}_i$、$\boldsymbol{X}$ 分别为

$$\boldsymbol{X}_i=M(\boldsymbol{W}_i)\odot\boldsymbol{F}'_{Xi},\quad i=1,2,\cdots,N,\tag{5}$$

$$\boldsymbol{X}=T(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N),\tag{6}$$

特征提取的整体网络结构是:先将图像输入到卷积核为 7×7 的卷积层,再经过卷积核为 3×3 的最大池化层;由 4 个子模块组成(数量分别为 3、4、6、3),每一个子模块都包含 2 个 1×1 卷积和一个 MSA 模块。为了不降低分辨率和产生过度的特征信号抽取,在网络的最后使用膨胀卷积代替最后的平均池化层和全连接层。

2.2 基于深度引导的动态局部卷积模块

为了解决单目视觉中近大远小的尺度敏感问题和深度缺失问题,本节提出基于深度引导的动态局部卷积模块。该模块的主要作用是将从特征提取网络中训练出的深度图特征作为单目图像对应的特定卷积核进行 2 个模态的信息融合,类似动态卷积网络[23]中生成的基于特定样本和特定位置的卷积核,因此可以捕获有意义的局部结构,以此来达到“动态地深度引导”的目的。

为了实现局部卷积操作,须要一个与单目图像特征映射大小相同的局部滤波器。为了减少空间卷积产生的参数量和避免由深度卷积导致的低算数强度,须要对深度图特征进行移位操作。具体过程如下:假设卷积核尺寸为 D_K,则存在 D_K^2 个可能的移位矩阵,每个移位矩阵对应一个移动的方向,并将每个移位矩阵中的一个值分配为 1,其余皆为 0。为了便于理解,本文将卷积核定义为一个 $D_K\times D_K$ 的移位网格$\{(\boldsymbol{G}_m,\boldsymbol{G}_n)\}$,其中 $\boldsymbol{G}\in(\mathrm{int})[1-D_K/2,\ D_K/2-1]$,通过移位卷积将深度图特征 $\boldsymbol{D}$ 向 $D_K\times D_K$ 个方向进行平移。将经过移位操作的深度图特征用 $\boldsymbol{D}'=\boldsymbol{D}(\boldsymbol{G}_m,\boldsymbol{G}_n)$ 表示。通过逐元素乘积操作将 $\boldsymbol{D}'$ 作用在单目图像特征图 $\boldsymbol{R}$ 上,得到融合特征为

$$\boldsymbol{R}'=\boldsymbol{R}\odot\frac{1}{D_K\cdot D_K}\sum\boldsymbol{D}'。\tag{7}$$

经过移位操作的特定卷积核的大小是固定的,即输出的感受野也是相同的;然而图像类别间与类别内尺度差异巨大,若模型能在将单目图像特征和深度图特征融合之前先学习出可缩放的卷积核,就可以解决单目视觉中近大远小的问题。为每个特定卷积核 $\boldsymbol{D}$ 分配不同的膨胀率 r[24],得到 $\boldsymbol{D}'=\boldsymbol{D}(\boldsymbol{G}_m r,\boldsymbol{G}_n r)$,并让 $\boldsymbol{R}$ 通过特征自适应权重模块学习卷积核的自适应权重,为不同膨胀率的移位卷积核进行加权计算,最后获得不同尺度的感受野。特征自适应权重模块的输入为特征 $\boldsymbol{R}$,包含:(1)一个自适应最大池化层,输出大小为$k\times k$,通道数为 c;(2)一个卷积核尺寸为 $k\times k$ 的二维卷积层,再经过一个维度重塑层;(3)一个归一化指数函数 softmax 层,生成 k 个和为 1 的权重 W_r,其中 $r\in(\mathrm{int})[1,k]$。本文采用的卷积核 $k=3$,卷积核的膨胀率为 1、2、3,分别对应的自适应权重为 W_{r1}、W_{r2}、W_{r3},最大膨胀率为 e。最终经过移位操作和自适应权重函数加权的卷积核为

$$\boldsymbol{G}(\boldsymbol{R})=\frac{1}{D_K\cdot D_K\cdot e}\sum W_r(\boldsymbol{R})\sum\boldsymbol{D}',\ r\in(\mathrm{int})[1,e]。\tag{8}$$

最终卷积核和单目图像特征经过逐元素相乘获得的融合结果被重新定义为 $\boldsymbol{R}'=\boldsymbol{G}(\boldsymbol{R})\odot\boldsymbol{R}$。深度引导动态局部卷积模块不仅捕捉了有意义的局部结构和尺度信息,而且充分利用了高级语义信息,进而实现了对三维空间信息的更好表示。

2.3 基于锚框的 2D-3D 检测模块

本节采用基于先验的 2D-3D 的单级检测头[25]作为基础检测器,它将边界框的输出空间离散化为每个特征图位置上不同长宽比和尺度的一组默认框。在预测时,网络为每个默认框中的每个物体类别生成存在

得分,并对框进行调整以更好地匹配物体形状。

(1) **输入**　将基于深度引导的动态局部卷积模块的输出 $\boldsymbol{R}'$作为输入,使用以下相机校准设置:假设相机内参数 $\boldsymbol{K}\in\mathbf{R}^{3\times3}$在训练和测试时都可用,3D 到 2D 的投影公式为

$$\begin{pmatrix}u\\v\\1\end{pmatrix}_{\mathrm{P}}\cdot\boldsymbol{Z}_{3\mathrm{D}}=\begin{pmatrix}f_x&0&c_x\\0&f_y&c_y\\0&0&1\end{pmatrix}_{\boldsymbol{K}}\cdot\begin{pmatrix}x\\y\\z\\1\end{pmatrix}_{3\mathrm{D}},\tag{9}$$

式中,f_x 和 f_y 分别表示 x 和 y 轴方向焦距的长度,(c_x,c_y)是主点实际位置(单位是像素)。$(x,y,z)_{3\mathrm{D}}$为 3D 点的坐标,$(u,v)_{\mathrm{P}}$ 表示 2D 点的投影坐标。

(2) **输出**　每个输入位置(m,n)包含以下输出:(a_x,a_y)、a_{w} 和 a_{h} 是预测的 2D 边界框的坐标宽、高;$(x,y)_{\mathrm{P}}$ 是三维角投影在 2D 平面上的位置;b_z 为深度;b_{w}、b_{h} 和 b_{l} 是 3D 检测框的宽、高、长,b_{α} 是旋转角;s 表示分类得分;a_{p}^{σ} 包含 8 个投影 3D 角,其中 $\sigma\in\mathrm{int}[1,8]$。再经过基于锚的转换得到实际输出。

(3) **预测**　使用 Confluence 算法对边界框进行预测。与传统的非极大抑制算法不同,Confluence 算法打破了基于交并比(intersection over union, IoU)的限制,使用曼哈顿距离作为边界框之前的重合度,根据置信度加权后的曼哈顿距离还作为最优边界框的选择依据。算法具体过程如下:

(ⅰ) 针对每个类别挑选 n 个边界框组成集合 B,每个边界框代表一个对象。

(ⅱ) 遍历 B 中的所有边界框并计算与其他边界框的曼哈顿距离 M_{dis}。因为每个框大小和位置不同,在没有标准度量的情况下直接将曼哈顿距离作为重合度是没有参考价值的,所以还须要对其坐标进行常规的 0-1 归一化处理,使其更具有辨识度。

(ⅲ) 归一化后的任意两框之间的 M_{dis}小于 2,可认为其是由高密度相似的并形成一个聚类,否则视为单独的对象。假设 C 为边界框的置信度,则加权曼哈顿距离为

$$M_{\mathrm{dis}}^{\mathrm{W}}=\frac{M_{\mathrm{dis}}}{C}。\tag{10}$$

(ⅳ) 在该聚类中进行排序选出 $M_{\mathrm{dis}}^{\mathrm{W}}$最小的边界框作为最优框加入最终结果集 R,且将其从 B 中删除;最后再将 B 中所有与最优框的曼哈顿距离小于预定义超参数的边界框去除。

(4) **损失函数**　使用 focal loss 平衡正负样本,总的损失函数为

$$L_{\mathrm{loss}}=(1-S)^{\lambda}(L^{\mathrm{cls}}+L_{\mathrm{reg}}^{\mathrm{2D}}+L_{\mathrm{reg}}^{\mathrm{3D}}+L^{\mathrm{cor}}),\tag{11}$$

式中,L^{cls}表示分类损失、$L_{\mathrm{reg}}^{\mathrm{2D}}$和 $L_{\mathrm{reg}}^{\mathrm{3D}}$表示 2D 和 3D 回归损失、$L^{\mathrm{cor}}$表示 2D-3D 角损失,聚焦参数 $\lambda=0.5$, S 表示目标类的分类得分;分类损失使用标准交叉熵(cross-entropy, CE)损失,即

$$\begin{aligned}
L^{\mathrm{cls}}&=-\log\,S_{t'},\\
L_{\mathrm{reg}}^{\mathrm{2D}}&=\sum_{i\in(a^x,a^y,a^w,a^h)}L_1(t_i',t_i),\\
L_{\mathrm{reg}}^{\mathrm{3D}}&=\sum_{j\in(b^w,b^h,b^l,b^z,b^{\alpha})}L_1(t_j',t_j)+\sum_{u\in(x,y)}L_1(t_u',t_u),\\
L^{\mathrm{cor}}&=\frac{1}{8}\sum L_1(t_{\alpha_{\mathrm{P}}^{\sigma}}',t_{\alpha_{\mathrm{P}}^{\sigma}})+\sum L_1(t_{b_z}',t_{b_z}),\quad\sigma=1,2,\cdots,8,
\end{aligned}$$

式中,$S_{t'}$为分数的预测值,t_i'和 t_i 分别为 2D 回归损失函数中二维检测框的预测值与真实值,t_j'和 t_j 分别为 3D 回归损失函数中三维检测框的预测值与真实值,t_u'和 t_u 分别为三维角投影到 2D 平面的位置的预测值与真实值,$t_{\alpha_{\mathrm{p}}^{\sigma}}'$和 $t_{\alpha_{\mathrm{p}}^{\sigma}}$分别为投影 3D 角的预测值与真实值,$t'_{b_z}$和 t_{b_z}分别为深度的预测值与真实值。$L_{\mathrm{reg}}^{\mathrm{2D}}$、$L_{\mathrm{reg}}^{\mathrm{3D}}$和 L^{cor}都使用函数 SmoothL1 回归损失,定义为 L_1。

3　实验过程及结果

本文将第 2 节所提出的算法与其他算法对比,采用 KITTI(Karlsruhe Institute of Technology and Toyota

Tchnological Institute)数据集[26],进行性能分析,给出三维预测框和鸟瞰图预测框的可视化效果,证明了本文方法的有效性。

3.1 数据集以及实验设置

(1) KITTI 数据集被广泛应用于基于单目和基于激光雷达的三维目标检测算法。该数据集包含 7 481 张训练图片、7 518 张测试图片,以及相应的点云数据和校准参数。检测目标根据对象类别主要分为汽车、行人和自行车。对于汽车,要求真实框和预测框的重叠度大于 70%,即交并比≥0.7;对于行人和自行车,要求真实框和预测框的重叠度要大于 50%,即交并比≥0.5。检测目标根据难度主要分为简单、中等和困难,中等难度最具代表性。该难度等级主要根据最小边界框高度、最大遮挡程度和最大截断程度来划分(具体标准如表 1 所示),本文将已分割的数据集称为"Split"。本文实现了鸟瞰图检测和三维目标检测任务,其中三维检测任务是本文的核心部分。

表 1 KITTI 数据集目标检测难度等级划分标准

Table 1 Settings of the difficulty level for object detection on the KITTI dataset

难度等级	最小检测框高度/像素	最大遮挡级别	最大截断/%
简单	40	完全可见	25
中等	25	部分遮挡	30
困难	25	难以看到	50

(2) **评估指标** 计算平均精度(average precision, AP)。2019 年以前使用 11 个点插值平均精度分别计算每个目标类和难度类,2019 后使用基于 40 个召回位置的平均精度。为了公平比较,实验给出 2 种指标下的检测结果,但在同一数据集下,基于后者可比较的文章较少。

(3) **训练** 在 GeForce RTX 3080 12GB 的图形处理器上以 batchsize 为 8、迭代次数为 40 000 来训练模型,学习率为 0.005。网络使用 paddle 框架实现,对图像先进行数据预处理,输入长×宽为 370 像素×1 224 像素,并采用水平翻转的方法实现数据增强,采用随机梯度下降(stochastic gradient descent, SGD)优化器对损失函数进行收敛。为了防止模型过拟合,在每个模块后使用丢弃率为 0.3 的正则化层,在输出主干网络后使用丢弃率为 0.5 的正则化层。在筛选预测框的 Confluence 算法中,曼哈顿距离重叠度的阈值 $p_{\text{thresh}}=0.6$。

3.2 比较结果

为了有效地评估算法,本文选取以下先进算法进行比较:①基于单目三维目标定位的几何推理网络(a geometric reasoning network for monocular 3D object localization, MonoGR)算法[2];②单目 3D 区域提取网络(monocular 3D region proposal network, M3DRPN)算法[4];③基于关键点的实时单目 3D 检测(real-time monocular 3D detection from object keypoints, RTM3D)算法[27];④基于立体影像精确类别检测的 3D 目标检测(3D object proposals using stereo imagery for accurate object class detection, 3DOP)算法[28];⑤三角化学习网络(triangulation learning network, TLNet)算法[29];⑥基于多级融合的单目三维目标检测(multi-level fusion based 3D object detection from monocular images, MF)算法[30];⑦基于距离归一化统一表示的 3D 目标检测(distance-normalized unified representation for monocular 3D object detection, UR3D)算法[31];⑧基于解耦结构化多边形估计和高度引导深度估计的单目 3D 目标检测(monocular 3D object detection with decoupled structured polygon estimation and height-guided depth estimation, DSP-HD)算法[32]。根据不同模态,将以上算法划分为 3 类:①—③属于单目彩色图像方法,④—⑤属于伪激光雷达方法,⑥—⑧和本文方法属于融合彩色图像和深度信息方法。

将以上算法基于 11 插值的平均精度分别定义为 $A^{\text{R11}}_{\text{MonoGR}}$、$A^{\text{R11}}_{\text{M3DRPN}}$、$A^{\text{R11}}_{\text{RTM3D}}$、$A^{\text{R11}}_{\text{3DOP}}$、$A^{\text{R11}}_{\text{TLNET}}$、$A^{\text{R11}}_{\text{MF}}$、$A^{\text{R11}}_{\text{UR3D}}$、$A^{\text{R11}}_{\text{DSP-HG}}$和 $A^{\text{R11}}_{\text{OURS}}$,基于 40 个召回位置的平均精度分别定义为 $A^{\text{R40}}_{\text{MonoGR}}$、$A^{\text{R40}}_{\text{M3DRPN}}$、$A^{\text{R40}}_{\text{RTM3D}}$、$A^{\text{R40}}_{\text{3DOP}}$、$A^{\text{R40}}_{\text{TLNET}}$、$A^{\text{R40}}_{\text{MF}}$、$A^{\text{R40}}_{\text{UR3D}}$、$A^{\text{R40}}_{\text{DSP-HG}}$和 $A^{\text{R40}}_{\text{OURS}}$。表 2—5 从 2 个视角展示了本文与 3 个模态下的算法在数据集 Split 中汽车类别的平均精度对比结果,加粗数据表示该视角及对应指标下某算法平均精度最优。从三维视角(表 2、3)可以看出:融合彩色图像和深度信息方法在 3 类模态算法中的平均精度最具优势,其中,在具有代表性的中等难度下,本文方法相较于 DSP-HD 算法,基于 11 插值的平均精度提升了 0.85%;相较于仅基于单目彩色图像的 M3D-RPN 算法,基于 40 个召回位置下的平均精度提升了 3.96%。从鸟瞰图视角(bird's eye view, BEV(表 4、5)来看:

本文方法的性能在 3 个模态中,在整体上有更好的表现(只有部分算法给出了鸟瞰视角下的结果);在中等难度下,相比 TLNet 算法,本文算法基于 11 插值的平均精度提升了 1.63%,相比 M3D-RPN 算法,本文算法基于 40 个召回位置的平均精度提升 3.78%。由于三维目标检测是直接在摄像机坐标系中进行目标检测,在目标定位上难度更大,因此相较于鸟瞰视角的检测结果有一定的差异。

表 2　汽车类别在三维视角下基于 11 插值的平均精度对比

Table 2　Comparison of the average precision based on 11-point interpolation for car 3D detection

难度等级	A^{R11}_{MonoGR}	A^{R11}_{M3DRPN}	A^{R11}_{RTM3D}	A^{R11}_{3DOP}	A^{R11}_{TLNET}	A^{R11}_{MF}	A^{R11}_{UR3D}	A^{R11}_{DSP-HG}	A^{R11}_{OURS}
简单	13.88	20.27	20.77	6.55	18.15	10.53	23.24	**26.95**	24.90
中等	10.19	17.06	16.86	5.07	14.26	5.69	13.35	18.68	**19.53**
困难	7.62	15.21	16.63	4.10	13.72	5.39	10.15	15.82	**17.29**

表 3　汽车类别在三维视角下基于 40 个召回位置的平均精度对比

Table 3　Comparison of the average precision based on 40 recall positions for car 3D detection

难度等级	A^{R40}_{MonoGR}	A^{R40}_{M3DRPN}	A^{R40}_{OURS}
简单	11.90	14.53	**22.30**
中等	7.56	11.07	**15.03**
困难	5.76	8.65	**11.93**

表 4　汽车类别在鸟瞰视角下基于 11 插值的平均精度对比

Table 4　Comparison of the average precision based on 11-point interpolation for car BEV detection

难度等级	A^{R11}_{M3DRPN}	A^{R11}_{TLNET}	A^{R11}_{MF}	A^{R11}_{OURS}
简单	25.94	29.22	22.03	**30.82**
中等	21.18	21.88	13.63	**23.51**
困难	17.90	18.83	11.60	**19.53**

3.3　多类别目标检测

表 6、7 分别在鸟瞰和三维 2 个视角展示了本文方法在 Split 数据集中汽车、行人和骑自行车的人 3 个类别上的平均精度检测结果。由于在 3 种不同模态下的单目三维检测算法中只有 M3D-RPN 算法存在相同类别、检测视角和基于 40 个召回位置的平均精度,因此将其与本文结果进行对比。

表 5　汽车类别在鸟瞰检测视角下基于 40 个召回位置的平均精度对比

Table 5　Comparison of the average precision based on 40 recall positions for car BEV detection

难度等级	A^{R40}_{M3DRPN}	A^{R40}_{OURS}
简单	20.85	**27.53**
中等	15.62	**19.40**
困难	11.88	**14.82**

M3D-RPN 算法和本文算法多类别目标平均精度分别定义为 A^{MC}_{M3DRPN} 和 A^{MC}_{OURS}。根据表 6、7 可以观察到:(1)汽车类的检测结果相对其他两类要好的多。因为相对形状较为统一且体积较大的汽车来说,人的姿态多变,难以捕捉,且深度信息难以估计,所以,对行人和骑自行车的人的检测十分困难;(2)本文算法在汽车和骑自行车的人类别上整体优于 M3D-RPN 算法,但在行人这一类别,与该算法存在一些比较。还须要注意的是,由于行人和骑自行车的人类别的训练样本数量相当少,因此性能可能存在一定程度的波动。

表 6　多类别目标在鸟瞰视角下的检测平均精度对比

Table 6　Comparison of the average precision for multi-class BEV detection

类别	难度等级	A^{MC}_{M3DRPN}	A^{MC}_{OURS}
行人	简单	**5.56**	4.35
	中等	4.05	**4.17**
	困难	3.29	**3.71**
骑自行车的人	简单	1.25	**3.22**
	中等	0.81	**3.15**
	困难	0.78	**3.00**
汽车	简单	20.85	**27.53**
	中等	15.62	**19.40**
	困难	11.88	**14.82**

表 7　多类别目标在三维视角下的检测平均精度对比

Table 7　Comparison of the average precision for multi-class 3D detection

类别	难度等级	A^{MC}_{M3DRPN}	A^{MC}_{OURS}
行人	简单	**4.92**	3.74
	中等	**3.48**	3.35
	困难	2.92	**3.22**
骑自行车的人	简单	0.94	**2.93**
	中等	0.65	**2.61**
	困难	0.47	**1.61**
汽车	简单	14.53	**22.30**
	中等	11.07	**15.03**
	困难	8.65	**11.93**

3.4　MSANet 性能评估

基于多尺度通道注意力的特征提取网络首先在 ImageNet[33] 数据集上进行预训练。每一个子模块中的 MSA 模块首先将通道分割为 4 个部分,每部分卷积核的大小分别为 3、5、7 和 9,根据自适应调整策略 $G=2^{K-1/2}$,分组卷积的分组大小分别为 1、4、8 和 16(卷积核尺寸为 3 所对应的分组大小默认为 1)。从表 8 可看

出,在参数量上,MSANet 相比 ResNet 少了 11.5%,计算代价(gigafloating-point operations per second,GFLOPs)减少了 12.1%,图像分类精确度提升了 3%。图 2 展示了目标检测任务在 KITTI 数据集上的分类精确度对比,MSANet 的精确度在整个测试过程中都比 ResNet 更高,在最后一轮中,ResNet 的精确度均为 0.81,MSANet 的精确度均为 0.86,相比较 ResNet 提升了 6.17%。

表 8 不同注意力方法在 ImageNet 上预训练的参数量(百万)、计算代价和精确度(%)对比

Table 8 Comparison of various attention methods on ImageNet (parameters (in millions)、GFLOPs and accuracy(%))

方法	骨干网络	参数量	计算代价	精确度
ResNet	ResNet50	25.56	4.12	75.20
MSANet		**22.61**	**3.62**	**77.49**

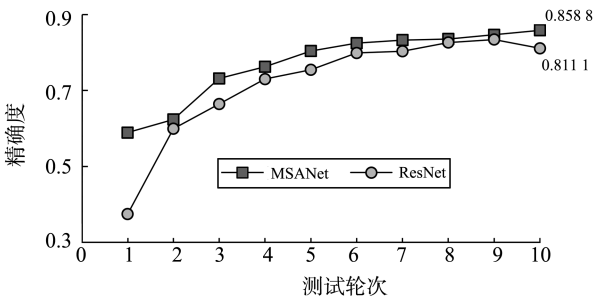


图 2 ResNet 和 MSANet 在 KITTI 数据集上的分类检测精确度对比

Fig.2 Accuracy comparison for the classification on the KITTI dataset between ResNet and MSANet

3.5 不同卷积模块的评估

为了证明本文的深度引导动态局部卷积模块对三维目标检测的有效性,本节将其与以下可代替方案进行比较:可变形卷积网络[34](deformable convolutional networks, DCN)和动态局部卷积网络[23](dynamic filter networks, DFN)。将以上算法基于 11 插值的平均精度分别定义为 A^{R11}_{DCN} 和 A^{R11}_{DFN},基于 40 个召回位置的平均精度分别定义为 A^{R40}_{DCN} 和 A^{R40}_{DFN};加粗表示对应指标下某卷积模块平均精度最优。可变形卷积网络通过偏移量得到非常规的卷积核,获得更适合目标的局部感受野,若卷积核太大,计算量增加,则会消耗掉巨大的空间。动态卷积网络通过注意力方法可自适应地调整卷积核权重,所以本文在此基础上增加膨胀卷积,生成不同感受野的自适应卷积核并局部应用在单目图像特征上。对于前一种方法,在图像和深度图分支上都应用可变形卷积,并按元素级乘积合并它们;对于后一种方法仅使用相同的深度图进行实验。由表 9、10 可以看出,本文的方法在具有代表性的中等难度上性能最优,能更好地从单目图像中捕获三维信息。

表 9 不同卷积模块在汽车类别上基于 11 插值的平均精度对比

Table 9 Comparisons of the average precision based on 11-point interpolation of different convolutional modules for car 3D detection

难度等级	A^{R11}_{DCN}	A^{R11}_{DFN}	A^{R11}_{OURS}
简单	23.95	**25.11**	24.90
中等	18.44	18.62	**19.53**
困难	16.11	16.57	**17.29**

表 10 不同卷积模块在汽车类别上基于 40 个召回位置的平均精度对比

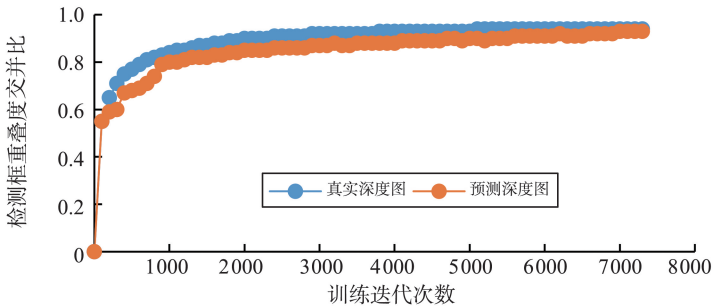
Table 10 Comparisons of the average precision based on 40 recall positions of different convolutional modules for car 3D detection

难度等级	A^{R40}_{DCN}	A^{R40}_{DFN}	A^{R40}_{OURS}
简单	19.05	21.78	**22.30**
中等	13.42	14.76	**15.03**
困难	10.07	11.31	**11.93**

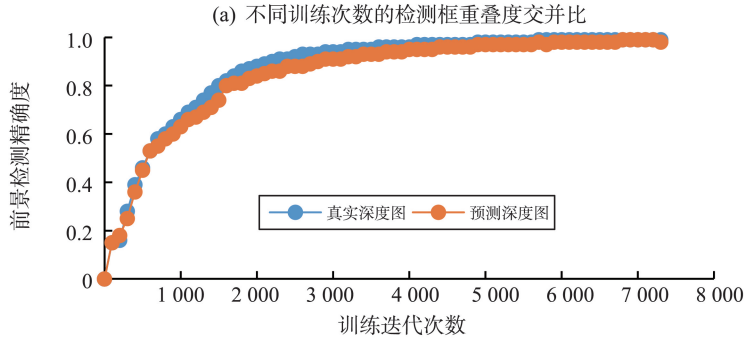
3.6 真实深度图与预测深度图对训练模型的影响

在真实深度图和通过序数回归预测的深度图在训练过程中,图 3(a)、(b)分别是在以上 2 种情况下,对预测的前景区域(即目标)的检测精确度和衡量三维真实框和三维预测框重叠程度(即位置准确性)的对比结果。由图 3(a)可知,在经过一定的训练迭代后,预测的深度图基本可以达到在真实深度图情况下的前景区域(即目标)检测精确度。由图 3(b)可知,在训练期间,2 种情况下的交并比逐渐提高,预测深度图呈现出明显向真实深度图靠拢并接近齐平的趋势,虽然还存在一定差距,但仍可以看出预测深度图对目标检测任务的有效性。如今,大部分单目三维检测算法[30]中融合的深度信息都使用直接从单目图像中预测后的深度图,避免了真实深度图针对特定场景须要额外且昂贵的数据采集和标注处理的局限性,降低了系统成本和复

杂性,也更具有更好的泛化性。



(a) 不同训练次数的检测框重叠度交并比



(b) 不同训练次数的前景检测精确

图 3　真实深度图和预测深度图在训练过程中的参数对比

Fig.3　Comparison of different parameters of the real depth map and the predicted depth map during training

3.7　定性研究

行人和骑自行车的人在数据集中的测试样本相当少,较为集中地出现在校园区,如图 4(a)所示。汽车类主要出现在数据集中的典型样本,即公路和住宅区,如图 4(b)、(c)所示。图 5、6 分别根据表 4(三维视角)、表 5(鸟瞰图视角),对汽车、行人和骑自行车的人 3 个类别进行了可视化比较(绿色表示汽车,浅蓝表示行人,深蓝表示骑自行车的人)。观察图 6 可以得出,本文方法在同一场景下能够检测出更多目标对象(橙色虚线圈出);从图 6(b)、(e)和(c)、(f)对比可看出,本文对远处小目标有更准确的定位和识别(蓝色虚线圈出);从图 6(c)、(f)对比可看出,重叠或被遮挡和截断的目标(红色虚线圈出)也被精确检测。这主要是因为多尺度的空间信息增强了特征表达能力,深度信息的引导捕捉到了局部对象并解决了单目图像近大远小的尺度敏感问题,使得二维图像在三维空间有了更好的表示。

(a) 校园　　(b) 公路　　(c) 住宅区

图 4　KITTI 数据集典型样本场景

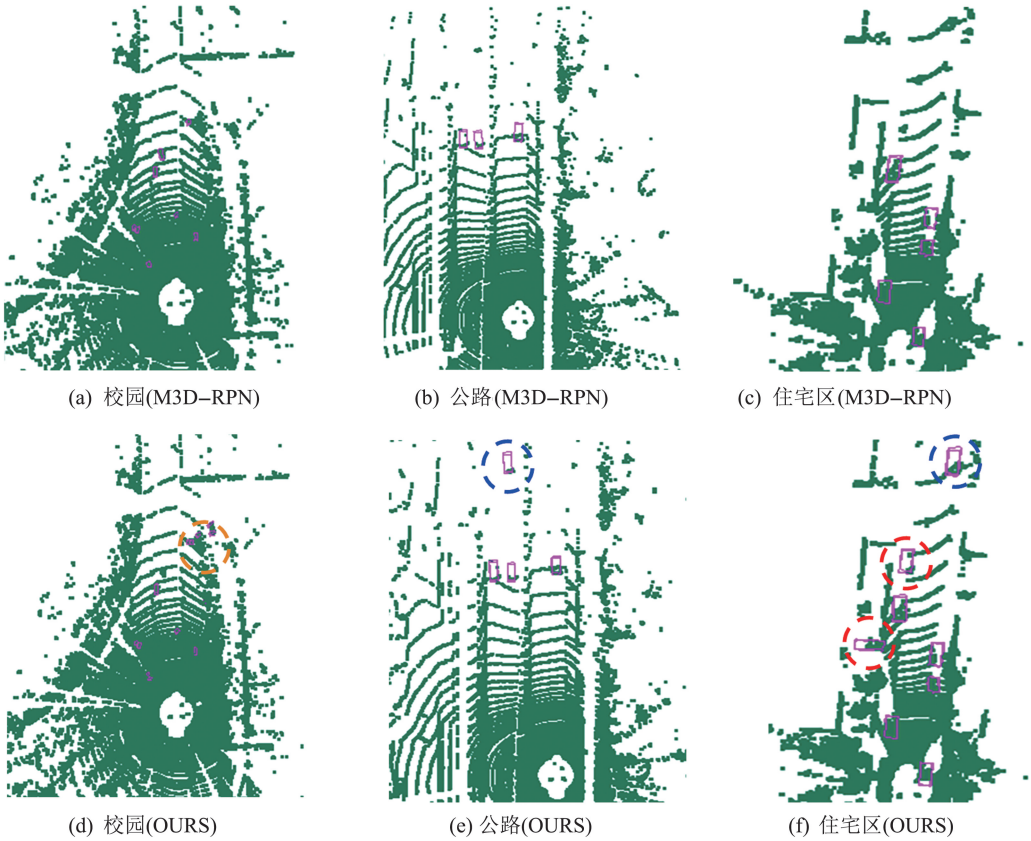
Fig.4　Typical sample scenarios in KITTI datasets

(a) 校园(M3D-RPN)　　(b) 公路(M3D-RPN)　　(c) 住宅区(M3D-RPN)

(d) 校园(OURS)　　(e) 公路(OURS)　　(f) 住宅区(OURS)

图 5　三维视角检测框的可视化结果比较

Fig.5　Comparison of visualization results for 3D detection boxes



(a) 校园(M3D-RPN) (b) 公路(M3D-RPN) (c) 住宅区(M3D-RPN)

(d) 校园(OURS) (e) 公路(OURS) (f) 住宅区(OURS)

图 6 鸟瞰视角检测框的可视化结果比较

Fig.6 Comparison of visualization results for bird's eye view detection boxes

4 结语

本文提出了一种基于深度信息引导和多尺度通道注意力机制的单目三维目标检测算法,基于多尺度通道注意力的特征提取网络输出的单目图像特征和深度图特征,同步输入到基于深度引导的动态局部卷积模块中进行融合,并将由深度图映射而来的特定卷积核应用在每个通道、像素上。这种方法弥补了二维、三维空间表示的差距,获得了更好的 2D-3D 的特征表示。最终实验表明,本文方法在三维目标检测公开数据集 KITTI 上的性能与其他相关算法相比,取得了较好的检测精度,但在不同检测类别中,检测精度有时不稳定,比如自行车和行人的检测精度不大理想,还须进一步优化算法,提高模型的精度和稳定性,这也是未来工作的重点。

参考文献:

[1] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017:5632-5640.

[2] QIN Zengyi, WANG Jinglu, LU Yan. Monogrnet: a geometric reasoning network for monocular 3D object localization[C]// 2019 33th AAAI Conference on Artificial Intelligence (AAAI-19). Hawaii: AAAI Press, 2019:8851-8858.

[3] SIMONELLI A, BULO S R, PORZI L, et al. Disentangling monocular 3D object detection: from single to multi-class recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3):1219-1231.

[4] BRAZIL G, LIU X M. M3D-RPN: monocular 3D region proposal network for object detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019:9286-9295.

[5] XIANG Y, CHOI W, LIN Y Q, et al. Data-driven 3D voxel patterns for object category recognition[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015:1903-1911.

[6] LIU Zongdai, ZHOU Dingfu, LU Feixiang, et al. Autoshape: real-time shape-aware monocular 3D object detection[C]// 2021

IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021:15621-15630.
[7] SONG Xibin, LI Wei, ZHOU Dingfu, et al. MLDA-Net: multi-level dual attention-based network for self-supervised monocular depth estimation[J]. IEEE Transactions on Image Processing, 2021, 30:4691-4705.
[8] GODARD C, AODHA O M, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019:3827-3837.
[9] WANG Qi, CHEN Jian, DENG Jiangqiang, et al. 3D-CenterNet: 3D object detection network for point clouds with center estimation priority[J]. Pattern Recognition, 2021, 115:107884.
[10] CHEN Yongjian, TAI Lei, SUN Kai, et al. Monopair: monocular 3D object detection using pairwise spatial relationships [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE Computer Society, 2020:12090-12099.
[11] YIN T W, ZHOU X Y, KRAHENBUHL P. Center-based 3D object detection and tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021:11779-11788.
[12] FU Huan, GONG Mingming, WANG Chaohui, et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018:2002-2011.
[13] MAYER N, ILG E, HAUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society, 2016:4040-4048.
[14] 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]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach : IEEE, 2019:8437-8445.
[15] PARK D, AMBRUS R, GUIZILINI V, et al. Is pseudo-lidar needed for monocular 3D object detection? [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021:3142-3152.
[16] MA Xinzhu, WANG Zhihui, LI Haojie, 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). Seoul: IEEE, 2019: 6850-6859.
[17] GARG D, WANG Y, HARIHARAN B, et al. Wasserstein distances for stereo disparity estimation[J]. Advances in Neural Information Processing Systems, 2020, 33:22517-22529.
[18] ZHANG Hu, ZU Keke, LU Jian, et al. EPSANet: an efficient pyramid squeeze attention block on convolutional neural network[C]//Asian Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022:541-557.
[19] WU Bichen, WAN Alvin, YUE Xiangyu, et al. Shift: a zero flop, zero parameter alternative to spatial convolutions[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018:9127-9135.
[20] SHEPLEY A J, FALZON G, KWAN P, et al. Confluence: a robust non-IoU alternative to non-maxima suppression in object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10):11561-11574.
[21] HE Kaiminh, ZHANG Xiangyu, REN Shaoqi, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society, 2016:770-778.
[22] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8):2011-2023.
[23] BRABANDERE D B, JIA X, TUYTELAARS T, et al. Dynamic filter networks[J]. Proceedings NIPS 2016, 2016, 29:1-9.
[24] WANG Xin, LV Rongrong, ZHAO Yang, et al. Multi-scale context aggregation network with attention-guided for crowd counting[C]//2020 15th IEEE International Conference on Signal Processing (ICSP). Beijing: IEEE, 2020:240-245.
[25] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//2016 14th European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016:21-37.
[26] 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 (CVPR). Providence: IEEE, 2012:3354-3361.
[27] LI Peixuan, ZHAO Huaici, LIU Pengfei, et al. Rtm3D: real-time monocular 3D detection from object keypoints for autonomous driving[C]//2020 16th European Conference on Computer Vision (ECCV). Beilin: Springer, 2020:644-660.
[28] CHEN X Z, KUNDU K, ZHU Y K, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017, 40(5):1259-1272.
[29] QIN Zengyi, WANG Jinglu, LU Yan. Triangulation learning network: from monocular to stereo 3D object detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019:7607-7615.

(下转第 82 页)