



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院遥感与数字地球研究所

中国图象图形学学会

北京应用物理与计算数学研究所

中国图象学报 图形学报

2020

06

VOL.25

ISSN1006-8961

CN11-3758/TB



三维点云分割 P1201



第25卷第6期 (总第290期)

2020年6月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿, 均视为同意在本刊网站及CNKI等全文数据库出版, 所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用, 所有内容, 未经许可, 不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院遥感与数字地球研究所

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 顾行发

编辑出版 《中国图象图形学报》编辑出版委员会

通信地址 北京市海淀区北四环西路19号

邮 编 100190

电子信箱 jig@radi.ac.cn

电 话 010-58887035

网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

(邮政信箱: 北京399信箱 邮编: 100048)

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Institute of Remote Sensing and Digital Earth, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa

Editor, Publisher Editorial and Publishing Board of Journal of
Image and Graphics

Address No. 19, North 4th Ring Road West, Haidian District,
Beijing, P. R. China

Zip code 100190

E-mail jig@radi.ac.cn

Telephone 010-58887035

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation

(P.O.Box 399, Beijing 100048, P.R.China)

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

ISSN 1006-8961

CODEN ZTTXFZ

国外发行代号 M1406

国内邮发代号 82-831

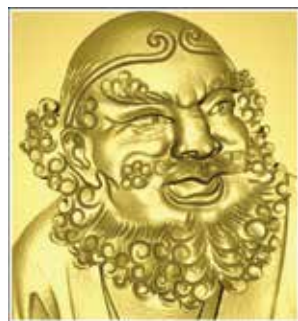
国内定价 60.00元



深度卷积神经网络图像语义分割研究进展(第1069页)



变分调整约束下的反向低秩稀疏学习目标跟踪(第1142页)



二次联合稀疏表示和低秩近似的浅浮雕优化(第1245页)

综述

深度卷积神经网络图像语义分割研究进展

青晨, 禹晶, 肖创柏, 段娟 1069

图像处理和编码

稳健局部特征非下采样小波域数字水印

牛盼盼, 杨思宇, 沈鑫, 杨红颖, 石齐良, 王向阳 1091

图像分析和识别

融合背景块再选取过程的显著性检测

陈星宇, 叶锋, 黄添强, 翁彬, 林文忠 1104

多尺度卷积神经网络显著物体检测

张晴, 左保川, 石艳娇, 戴蒙 1116

结合空间注意力多层特征融合显著性检测

陈凯, 王永雄 1130

变分调整约束下的反向低秩稀疏学习目标跟踪

田丹, 张国山, 张娟瑾 1142

自适应尺度突变目标跟踪

任俊丽, 郭浩, 董亚飞, 刘茹, 安居白, 王妍 1150

自适应多特征融合相关滤波目标跟踪

张艳琳, 钱小燕, 张淼, 葛红娟 1160

新型雷达辐射源识别

高欣宇, 张文博, 姬红兵, 欧阳成 1171

特征点聚类高精度视差图像拼接

谢从华, 张冰, 高蕴梅 1180

动态生成掩膜弱监督语义分割

陈辰, 唐胜, 李锦涛 1190

融合图卷积和差异性池化函数的点云数据分类分割模型

张新良, 付鹏飞, 赵运基, 谢恒, 王琬如 1201

图像理解和计算机视觉

外观表征分析下动态更新相关滤波跟踪

强壮, 石繁槐 1209

结合混合域注意力与空洞卷积的3维目标检测

严娟, 方志军, 高永彬 1221

计算机图形学

字典神经网络方法快速湍流烟雾合成

柏凯, 李伟, 刘晓培 1235

二次联合稀疏表示和低秩近似的浅浮雕优化

张倩倩, 计忠平, 付晓峰 1245

医学图像处理

双层水平集描述眼底图像视杯视盘分割

王莹, 于晓升, 迟剑宁, 雷晓亮, 吴成东 1260

遥感图像处理

结合深度学习和超像元的高分遥感影像变化检测

王艳恒, 高连如, 陈正超, 张兵 1271

CONTENTS

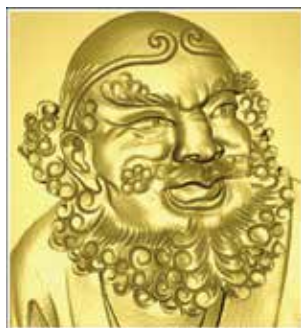
JOURNAL OF IMAGE AND GRAPHICS



Deep convolutional neural network for semantic image segmentation (P1069)



Object tracking via reverse low-rank sparse learning with variation regularization(P1142)



Bas-relief optimization based on twice-joint sparse representation and low-rank approximation (P1245)

Review

- Deep convolutional neural network for semantic image segmentation
Qing Chen, Yu Jing, Xiao Chuangbai, Duan Juan 1069

Image Processing and Coding

- Digital watermarking algorithm in the UDWT domain based on robust local features
Niu Panpan, Yang Siyu, Shen Xin, Yang Hongying, Shi Qiliang, Wang Xiangyang 1091

Image Analysis and Recognition

- Saliency detection based on the background block reselection method
Chen Xingyu, Ye Feng, Huang Tianqiang, Weng Bin, Lin Wenzhong 1104
- A multi-scale convolutional neural network for salient object detection
Zhang Qing, Zuo Baochuan, Shi Yanjiao, Dai Meng 1116
- Saliency detection based on multi-level features and spatial attention
Chen Kai, Wang Yongxiong 1130
- Object tracking via reverse low-rank sparse learning with variation regularization
Tian Dan, Zhang Guoshan, Zhang Juanjin 1142
- Adaptive scale sudden change object tracking
Ren Junli, Guo Hao, Dong Yafei, Liu Ru, An Jubai, Wang Yan 1150
- Correlation filter target tracking algorithm based on adaptive multifeature fusion
Zhang Yanlin, Qian Xiaoyan, Zhang Miao, Ge Hongjuan 1160
- New radar emitter identification method
Gao Xinyu, Zhang Wenbo, Ji Hongbing, Ouyang Cheng 1171
- High-precision parallax image stitching method using feature points clustering
Xie Conghua, Zhang Bing, Gao Yunmei 1180
- Weakly supervised semantic segmentation based on dynamic mask generation
Chen Chen, Tang Sheng, Li Jintao 1190
- Point cloud data classification and segmentation model using graph CNN and different pooling functions
Zhang Xinliang, Fu Pengfei, Zhao Yunji, Xie Heng, Wang Wanru 1201

Image Understanding and Computer Vision

- Dynamic update correlation filter tracking based on appearance representation analysis
Qiang Zhuang, Shi Fanhuai 1209
- 3D object detection based on domain attention and dilated convolution
Yan Juan, Fang Zhijun, Gao Yongbin 1221

Computer Graphics

- Fast turbulent smoke synthesis via a dictionary-based neural network
Bai Kai, Li Wei, Liu Xiaopei 1235
- Bas-relief optimization based on twice-joint sparse representation and low-rank approximation
Zhang Qiankan, Ji Zhongping, Fu Xiaofeng 1245

Medical Image Processing

- Segmentation of optic cup and disc based on two-layer level set describer in retinal fundus images
Wang Ying, Yu Xiaosheng, Chi Jianning, Lei Xiaoliang, Wu Chengdong 1260

Remote Sensing Image Processing

- Deep learning and superpixel-based method for high-resolution remote sensing image change detection
Wang Yanheng, Gao Lianru, Chen Zhengchao, Zhang Bing 1271

中图分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2020)06-1221-14

论文引用格式: Yan J, Fang Z J and Gao Y B. 2020. 3D object detection based on domain attention and dilated convolution. Journal of Image and Graphics, 25(06): 1221-1234(严娟, 方志军, 高永彬. 2020. 结合混合域注意力与空洞卷积的3维目标检测. 中国图象图形学报, 25(06): 1221-1234)
[DOI:10.11834/jig.190378]

结合混合域注意力与空洞卷积的3维目标检测

严娟, 方志军, 高永彬

上海工程技术大学电子电气工程学院, 上海 201620

摘要: **目的** 通过深度学习卷积神经网络进行3维目标检测的方法已取得巨大进展,但卷积神经网络提取的特征既缺乏不同区域特征的依赖关系,也缺乏不同通道特征的依赖关系,同时难以保证在无损空间分辨率的情况下扩大感受野。针对以上不足,提出了一种结合混合域注意力与空洞卷积的3维目标检测方法。**方法** 在输入层融入空间域注意力机制,变换输入信息的空间位置,保留需重点关注的区域特征;在网络中融入通道域注意力机制,提取特征的通道权重,获取关键通道特征;通过融合空间域与通道域注意力机制,对特征进行混合空间与通道的混合注意。在特征提取器的输出层融入结合空洞卷积与通道注意力机制的网络层,在不损失空间分辨率的情况下扩大感受野,根据不同感受野提取特征的通道权重后进行融合,得到全局感受野的关键通道特征;引入特征金字塔结构构建特征提取器,提取高分辨率的特征图,大幅提升网络的检测性能。运用基于二阶段的区域生成网络,回归定位更准确的3维目标框。**结果** KITTI(A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)数据集的实验结果表明,在物体被遮挡的程度由轻到高时,对测试集中的car类别,3维目标检测框的平均精度 AP_{3D} 值分别为83.45%、74.29%、67.92%,鸟瞰视角2维目标检测框的平均精度 AP_{BEV} 值分别为89.61%、87.05%、79.69%;对pedestrian和cyclist类别, AP_{3D} 和 AP_{BEV} 值同样比其他方法的检测结果有一定优势。**结论** 本文提出的3维目标检测网络,一定程度上解决了3维检测任务中卷积神经网络提取的特征缺乏视觉注意力的问题,从而使3维目标检测更有效地运用于室外自动驾驶。

关键词: 3维目标检测;注意力机制;空洞卷积;感受野;金字塔网络;卷积神经网络(CNN)

3D object detection based on domain attention and dilated convolution

Yan Juan, Fang Zhijun, Gao Yongbin

Department of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Abstract: Objective With the continuous development of convolutional neural network (CNN) used in deep learning in recent years, 3D object detection networks based on deep learning have also made outstanding development. 3D object detection aims to identify the class, location, orientation, and size of a target object in 3D space. It is widely used in the visual field, such as autonomous driving, intelligent monitoring, and medical analysis. The feature extracted by a deep learning network is important in detection accuracy. The detection task is similar to human vision; that is, it also needs to distinguish the difference between the background and the objects. In human vision, attention is given to target objects, while the background is disregarded. Therefore, paying more attention to the target area and less attention to the background area is better when performing object detection in an image. However, a CNN does not distinguish which areas and channels in an image should be given more and less attention. Thus, the features extracted by a CNN not only lack the

收稿日期: 2019-08-16; 修回日期: 2019-11-08; 预印本日期: 2019-11-15

基金项目: 国家自然科学基金项目(61802253, 61772328)

Supported by: National Natural Science Foundation of China (61802253, 61772328)

dependence relationship between different regions but also the dependence relationship between different channels. The current 3D object detection method based on a deep learning network uses a combination of pooling layers behind the multilayer convolution layer. These network structures generally use maximum or averaging pooling in feature maps. They aim to adjust the receptive field size of the extracted features. However, transforming the receptive field of the features of the pooling layers must be performed by removing some information, causing a considerable loss of feature information. Information loss may result in detected errors. Therefore, a CNN should expand the receptive field without losing information, obtaining good detection results. To address the shortcomings of the aforementioned 3D target detection methods, this study proposes a two-stage 3D object detection network that combines mixed domain attention and dilated convolution. **Method** In this study, a 3D object detection network based on a deep learning network is built. Integrating the spatial domain attention mechanism into the input layer of the network transforms the spatial position of the input information, preserving regional features that require more attention. Incorporating the channel domain attention mechanism into the network computes the channel weights of the extracted features, obtaining the key channel features. The features are mixed by combining the aforementioned spatial and channel domain attention mechanisms. Second, the output layer of the feature extractor integrates the network layer that is combined with the dilated convolution and the channel domain attention mechanism, and thus, our network can expand the receptive field of the extracted features without losing spatial resolution. In accordance with the different obtained receptive fields, the features can determine their channel weights and then fuse these feature weights through different schemes to obtain the channel weights of their global receptive fields and identify key channel features. In addition, the feature pyramid network structure is introduced to construct the feature extractor of our network, through which our network can extract high-resolution feature maps, considerably improving the detection performance of our network. Lastly, our network architecture is based on a two-stage region proposal network, which can regress to accurate 3D bounding boxes. **Result** A series of experiments has been conducted on the KITTI (A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago) dataset using the method proposed in this study. Cases wherein the object is slightly to severely occluded are denoted as “easy”, “moderate”, and “hard” in the tables. In the car class in the test set, the values of AP_{3D} that represents average accuracy of 3D object detection box obtained are 83.45%, 74.29%, and 67.92%; and the values of AP_{BEV} that represents average accuracy of 2D detection box from bird’s eye view obtained are 89.61%, 87.05%, and 79.69%. In the pedestrian class, the values of AP_{3D} obtained are 52.23%, 44.91%, and 41.64%; and the values of AP_{BEV} obtained are 59.73%, 53.97%, and 49.62%. In the cyclist class, the values of AP_{3D} obtained are 65.02%, 54.38%, and 47.97%; and the values of AP_{BEV} obtained are 69.13%, 59.69%, and 52.11%. We also perform ablation experiments on the test set. The experiment results show that in the car class and relative to the proposed method, the average value of AP_{3D} obtained after removing the pyramid structure is reduced by approximately 6.09%, the average value of AP_{3D} obtained after removing the mixed domain attention structure is reduced by approximately 0.99%, and the average value of AP_{3D} obtained after removing the dilated convolution structure is reduced by approximately 0.71%. **Conclusion** For the research on 3D object detection task, we propose a two-stage 3D object detection network that combines dilated convolution and mixed domain attention. The experiment results show that the proposed method outperforms several existing state-of-the-art 3D object detection methods and obtains accurate detection results, and it can be effectively applied to outdoor automatic driving.

Key words: 3D object detection; attention mechanism; dilated convolution; receptive field; feature pyramid network; convolutional neural network (CNN)

0 引言

随着计算机视觉的不断发展,从图像、视频等视觉信息中有效提取需求信息的方法越来越重要,从提取图像中提取基础信息的目标检测任务因此获得

高度关注。与2维目标检测任务不同,3维目标检测任务不仅需要对目标进行类别确定,而且要考虑目标的尺寸、3维位置与方向信息等。为了得到更加良好的检测结果,提取具有强泛化性、高代表性以及更具具体性的特征至关重要,基于此,一种良好的特征提取方法必不可少。

在过去的研究中,大都运用传统的基于机器学习的方法进行3维目标检测,运用的特征一般分为手工设计的特征和机器学习的特征。手工设计的特征主要有 Haar-like 特征(Papageorgiou 等,1998)、尺度不变特征变换(scale-invariant feature transform, SIFT)特征(Lowe, 1999)、方向梯度直方图(histogram of oriented gradient, HOG)特征(Dalal 和 Triggs, 2005);机器学习的特征主要有主成分分析特征、线性判别分析特征等。在传统的基于机器学习的3维目标检测方法中,一般先通过滑动窗口等方法提取候选框,之后对候选框中的图像信息提取特征,而后通过分类器判断候选框中是否为目标,最后只保留一定的目标候选框。基于机器学习的传统方法在目标检测研究中已取得了很多成果,但是这类方法的特征提取却存在一定缺陷,只会提取简单背景下的明显特征,而对于复杂背景下的多变目标,提取的抽象特征很难达到良好的目标检测效果。

2013年后,由于深度学习网络对特征表达的显著增强以及对分类精度的高速提升,开始用于3维目标检测任务,提取更复杂背景下的抽象特征。其中,类似于2维目标检测任务,在3维目标检测任务中,基于深度学习的网络模型也分为端到端网络与二阶段网络。端到端网络虽然相对二阶段网络更迅速,但由于检测准确度低、小目标检测效果差以及漏检率高等原因,二阶段网络用于3维目标检测任务更加普遍。二阶段网络的运行流程类似传统的机器学习方法,首先得到候选框,而后提取特征,之后对候选框对应的特征分类得到类别置信度,最后对置信度高的候选框回归尺寸与位置。与传统方法不同的是,该类网络一般运用多层卷积神经网络提取特征,比传统方法手工设计的特征更适用于复杂场景。基于此,二阶段网络对比传统的机器学习方法泛化能力更强,鲁棒性更好。由此可知,特征提取器的特征提取性能对于目标检测算法的重要性可见一斑。现有的二阶段网络的3维目标检测网络架构的特征提取器一般都是基于多层卷积神经网络,但一般的卷积神经网络都会在固定位置,用同等代价进行特征采样,得到的特征不仅缺乏旋转、缩放等几何变换性,还缺少对关键性信息的选择,这些特征会导致得到的结果检测框具有一定偏差。

对于图像中的物体而言,因为类别的不同、空间距离的差异以及环境场景的改变,物体显示的尺寸

可能大小不一,即在图像中所占的像素并非固定不变,基于此,需要网络具有良好的几何变换适应性,能在检测任务中对于图像空间域的局部区域的重要信息自适应地调整提取;现有的基于深度学习的3维目标检测方法是运用在多层卷积层中添加池化层的组合来调整特征的感受野大小,但此类多层网络结构一般都运用如最大值池化或平均值池化对特征图操作,即此类操作都是通过去除特征的某些信息变换特征的感受野,故此造成了特征信息的大量损失,这些特征的损失会造成检测的误差,故需要网络能无损信息地扩大感受野,从而得到良好的检测结果;此外,对于检测任务而言,图像中某些信息如目标是至关重要的,而有些信息如背景却是无关紧要的,因此需要特征提取器更集中提取对检测任务重要的特征,抑制次要特征,从而提升检测效果。现有的基于深度学习的3维目标检测方法对图像运用多层卷积神经网络提取特征时只是在固定位置以同等代价进行,提取的特征缺少关键性辨别、几何变换丢失,还损失了信息的完整性,对3维目标检测结果造成一定偏差。

针对上述问题,本文提出了一种结合混合域注意力与空洞卷积的二阶段3维目标检测方法。首先在网络的输入层融入空间域注意力机制,对输入信息进行几何变换,强化图像中的目标特征,突出图像中的目标;而后在网络中融入通道域注意力机制,运用学习的方式自动获取特征的每个通道的权重,使网络实现自适应校准特征通道权重以提取重要特征并抑制次要特征。融入空间域和通道域的注意力机制可以实现从两个方向对提取的特征进行分析。同时,在特征提取器的输出层融入结合空洞卷积与通道注意力机制的网络层,实现在不损失特征的空间分辨率的情况下扩大感受野,运用3种间隔数量的空洞卷积对输出特征进行操作,得到3种感受野的特征,根据3种特征通过通道域注意力机制提取各自的通道权重后进行融合,从而选择得到全局感受野的关键通道特征。受特征金字塔网络(feature pyramid networks, FPNs)(Lin 等,2017)对特征进行提取的启发,本文方法也运用金字塔结构构建网络的特征提取器,提取高分辨率的特征图,进行更精准的包括小目标的检测。最后,通过基于二阶段的区域生成级联网络的第2阶段检测网络,对第1阶段检测网络得到的3维目标候选框进行方向精调,以回归最终的3维目标检测框。

1 相关工作

1.1 基于深度学习网络的3维目标检测

随着深度学习网络的发展,越来越多的深度学习网络框架用于3维目标检测任务。这些网络框架可以被分为以下两类:第1类是端到端的网络框架,直接使用卷积神经网络(convolutional neural network, CNN)卷积特征,同时回归目标物体的类别,3维框的3维位置、尺寸与方向;第2类是二阶段网络,先提取3维候选框,根据候选框进行分类,得到类别置信度分数,最后根据类别回归3维位置、尺寸与方向。

在端到端类别的网络框架中,Li等人(2016)提出一种端到端的深度学习网络,输入是由雷达点云投影得到的2维前视图,将此图直接运用到2维检测网络的卷积神经网络,虽然这是运用于3维目标检测任务的网络,但检测流程与2维检测任务类似。Li等人(2017)同样提出一种端到端的深度学习网络,是一种直接运用3D卷积进行3维检测的网络架构,输入是一种由雷达点云构造的3D体素网格,而后运用3D卷积网络对输入进行特征提取,检测中没有运用候选框机制。Zhou和Tuzel(2018)提出的端到端的3维目标检测网络VoxelNet的网络架构扩展了Li等人(2016)提出的网络架构,输入的是运用逐点特征编码的3D体素网格。Simon等人(2019)提出的是扩展了端到端的2维目标检测网络(Redmon等,2016)的端到端的3维目标检测网络。上述端到端网络框架虽然可以应用于3维目标检测,但是相对于基于区域生成网络(region proposal network, RPN)的二阶段网络架构,检测精度有所不及。

在二阶段网络类别的框架中,MV3D(multi-view 3D object detection network)(Chen等,2017)的输入数据含有RGB图与雷达点云,其将雷达点云转换为雷达鸟瞰图和雷达前视图以作为网络的输入,网络架构是一种基于区域的融合网络,特征提取器对3种输入数据提取特征,并运用分类器对候选框进行类别判断,而后运用感兴趣区域池化对选出的候选框对应的所有输入数据的特征进行长度固定,以便于之后的特征融合操作,最后回归目标3维框的坐标值。F-PointNet(frustum pointnets)(Qi等,2018)与AVOD(an aggregate view object detection network)方法(Ku等,2018)的输入数据同为RGB

图与雷达点云,F-PointNet对RGB图先进行2维检测得到2维框,后将2维框扩展至3维空间得到3维信息并运用Charles等人(2017)提出的点云处理网络对3维信息进行处理,分割得到物体点云实例,最后基于物体点云实例回归3维目标检测框;与F-PointNet不同,AVOD的输入虽然含有雷达点云,但不是直接输入雷达点云,而是将雷达点云处理为雷达鸟瞰图后再进行输入,AVOD运用RPN对提取的特征进行候选框分类并得到3维候选框,而后运用二阶段检测网络进行精准3维目标检测框的回归;类似于F-PointNet,Zhou等人(2019)提出的方法同样是二阶段检测流程,但只运用了雷达点云数据。以上的二阶段3维目标网络相对于端到端3维目标检测网络检测精度高,运行效果好,但是对特征的要求高,需要更抽象更具有代表性的特征。运用一般的卷积神经网络对输入信息进行多层特征提取,提取的特征会缺少针对性、几何变换性和重要对比性。

Hu等人(2017)提出一种SeNet(squeeze-and-excitation networks)模型,网络根据loss去学习特征通道的权重,以此实现关键通道特征权重赋值大,不关键通道特征权重赋值小的效果;Vaswani等人(2017)提出一种注意力机制网络架构,免除了以往注意力机制中的循环与卷积结构,且让网络模型在训练中将更多的注意力关注特定目标的特征,以此达到校准每张特征图中主要与次要特征的目的;Yu和Koltun(2015)提出一种不同于普通的卷积操作的卷积结构,即空洞卷积,可以扩大卷积核尺寸时使核参数量不变,可以增加特征感受野时避免池化损失信息。受Hu等人(2017)、Yu和Koltun(2015)和Vaswani等人(2017)提出的方法的启发,本文提出一种结合混合域注意力与空洞卷积的二阶段3维目标检测方法。相比之前直接运用卷积神经网络进行特征提取的网络结构,本文提出的网络利用类似SeNet的通道域注意力机制与空间域注意力机制(Vaswani等,2017)进行特征选择,在通道与区域维度选择重要程度更高的特征,并抑制不重要的特征,此外,还利用了空洞卷积(Yu和Koltun,2015)对特征操作以得到多种扩大的感受野,以此映射更多更大范围的图像信息。

1.2 卷积神经网络

随着深度学习算法的快速发展,提出了越来越多的卷积神经网络模型。1994年提出的早期的卷

积神经模型是最早的卷积神经网络之一, 2012 年提出了 AlexNet 模型, 随后提出了 VGG (visual geometry group) 模型和 GoogleNet 模型, 之后又提出了 ResNet 模型。纵观以上 CNN 模型的发展历程可知, 其发展都围绕着两个主旨, 一是网络实现的效果与性能, 二是网络运行的计算复杂度, 对卷积网络的各种变形与组合都是围绕着平衡这两个主旨的。

本文网络的特征提取器中运用的网络结构即为 VGG-16 (visual geometry group 16-layer net) (Simonyan 和 Zisserman, 2014) 模型结构, 通过此结构提取有效的图像深层特征信息。

1.3 特征金字塔网络

Lin 等人 (2017) 提出了一种网络, 命名为特征金字塔网络, 可以将不同尺度的特征图都加以利用, 以改进对各种尺寸物体的检测效果。其针对底层特征与顶层特征融合的思想已被广泛运用于很多网络框架, 并取得了一定成效。FPNs 主要通过 3 部分的结合实现对不同尺度的特征图的运用, 包括自底向上的路径、自顶向下的路径以及中间的连接部分, 如图 1 所示。

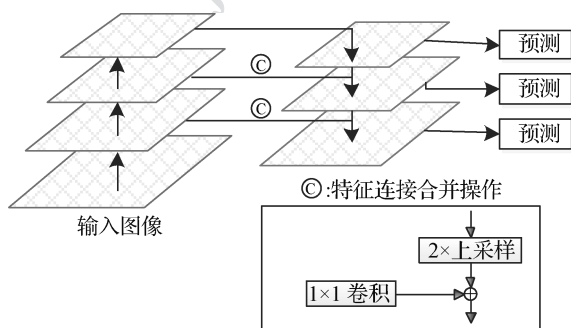


图1 特征金字塔网络模型结构图

Fig. 1 Structure of FPNs

自底向上的路径即为卷积网络的前馈计算, 其

中由不同比例的特征映射组成了不同的特征层级, 通常会运用多层卷积网络输出相同大小的特征映射, 这些层称为相同的网络阶段, 在 FPNs 中, 每个阶段为一个金字塔等级, 因为每个阶段的最深的层一般具有最强特征, 故将最后一层输出的特征映射作为 FPNs 的特征图参考。

自顶向下的路径即为上采样过程, 首先对在空间上更加抽象但是语义信息更强的高层特征图进行上采样以获取高分辨率的特征, 而后横向连接自底向上的路径中的特征图以增强高层特征。注意, 其中对高层特征图一般进行 2 倍上采样, 本文运用 2 倍最近邻上采样, 且横向连接中的自底向上的路径中的特征图与自顶向下的路径中的特征图尺寸相同。

中间的连接部分即将特征图连接合并的操作, 包括对自底向上的路径中的特征图进行 1×1 卷积操作, 以减少特征图的通道数, 并对自顶向下的路径中的特征图进行 2 倍上采样操作, 以扩张特征图的尺寸, 而后再将此两类特征图映射按元素相加进行合并。

本文为了提取得到高分辨率的特征图, 在网络中融入了以上特征金字塔思想, 通过对底层特征进行上采样操作, 最终得到了与原输入网络的图像相同分辨率的特征图输出。

2 结合混合域注意力与空洞卷积的第二阶段 3 维目标检测网络

本文提出的结合混合域注意力与空洞卷积的第二阶段 3 维目标检测方法的整体框架图如图 2 所示, 其网络框架主要由以下 3 部分组成。

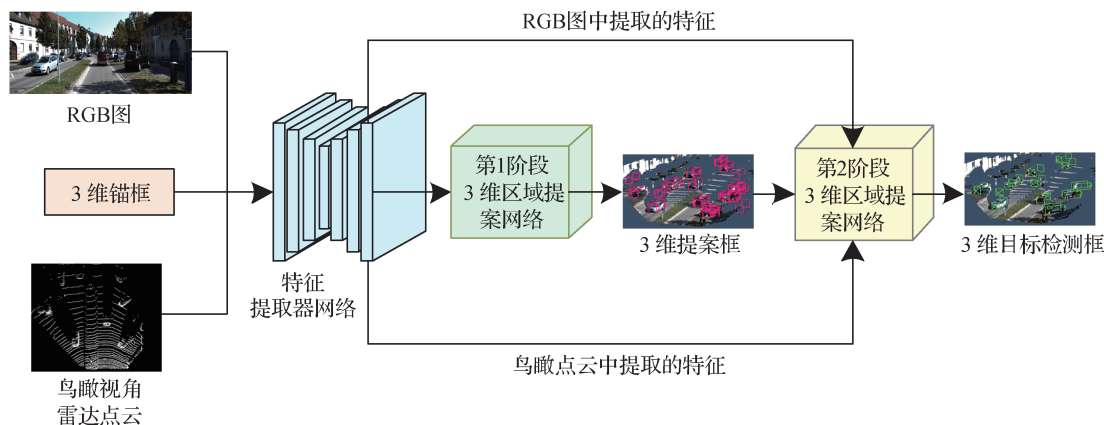


图2 网络整体架构图

Fig. 2 Architecture of our network

1) 特征提取器。本文网络的输入信息包括 RGB 图像与对应的雷达点云投影得到的鸟瞰图 (bird's eye view, BEV), 故特征提取器包括对以上两种输入信息的特征提取。

2) 区域生成网络 RPN。受 Ren 等人 (2017) 提出的网络的启发, 本文网络框架中第 1 阶段网络使用 RPN, 运用 RPN 回归目标的 3 维提案框。

3) 第 2 阶段检测网络。对 RPN 回归的 3 维目

标提案框执行更精准的定向与位置的精修。

网络的输入数据包括 RGB 图像与鸟瞰图, 最终的输出数据即为 3 维目标检测框。

2.1 特征提取器

本文网络运用的特征提取器结构类似特征金字塔, 如图 3 所示, 虽然网络中需要对 RGB 图像和鸟瞰图两种数据进行特征提取, 却使用了相同的特征提取器, 且都是进行了全分辨率的特征提取。

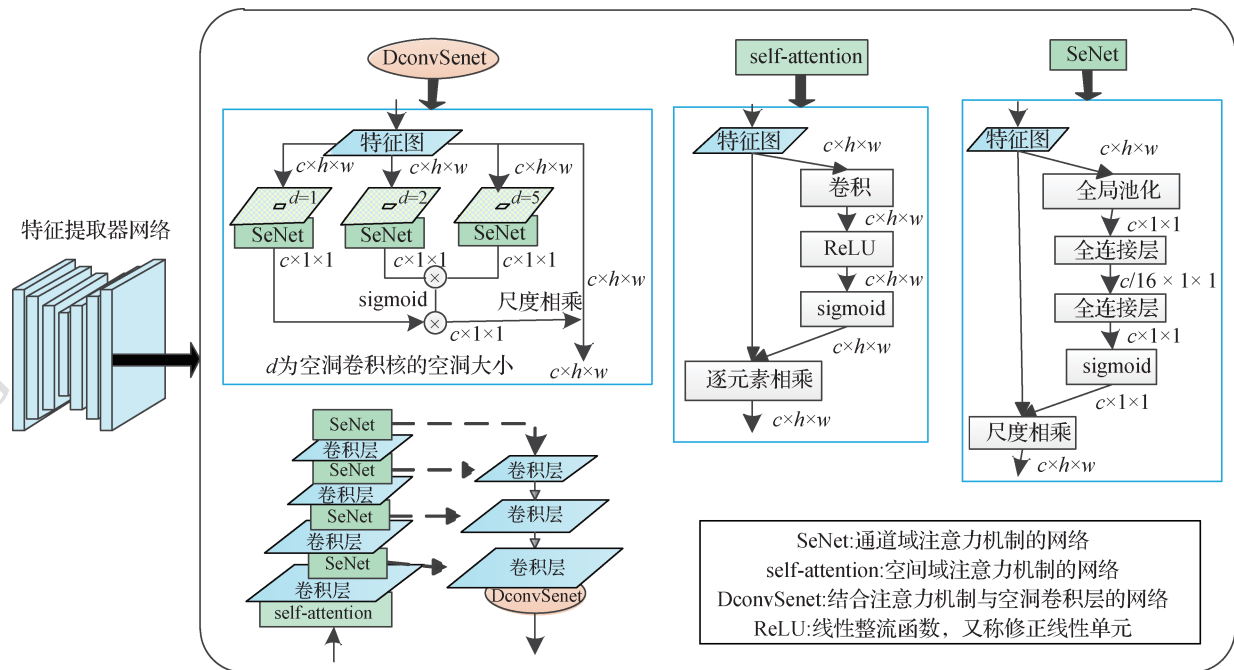


图 3 网络的特征提取器架构图

Fig. 3 Architecture of the feature extractor

2.1.1 特征提取器的网络结构

特征提取器的网络结构是受 Lin 等人 (2017) 的方法启发提出的, 由编码器和解码器两段组成。

1) 编码器的网络结构主要是以 VGG-16 为模型进行微小改动构造的。改动主要包括: (1) 各个卷积阶段的通道数减少一半, 通道数变为 $[32, 64, 128, 256]$; (2) 只运用了至 conv-4 阶段为止的网络, 故若输入图像的尺寸为 $W \times H \times C$, 则生成的特征图尺寸为 $\frac{W}{8} \times \frac{H}{8} \times C^*$, 即输出的特征图的尺寸变为原图的 $1/8$, 其层数 C^* 取决于通道数, 即卷积核的数目。至编码过程结束, 计算可知, 得到的特征图的图像分辨率降低到输入图像分辨率的 $1/8$, 而图像分辨率的降低会对小目标的检测有一定影响, 故需要提高特征图的分辨率以降低对小目标物体的检测难度。

2) 针对特征图的图像分辨率降低问题, 本文的特征提取器引入了解码器的概念, 解码器包括上采样和特征图融合步骤。上采样通过反卷积进行, 将特征图的尺寸 $W_1 \times H_1 \times C_1$ 变回卷积之前的输入尺寸 $2W_1 \times 2H_1 \times \frac{C_1}{2}$ (因为卷积操作的步长 $s=2$, 一个卷积阶段的卷积核数目以 2 的倍数递增)。上采样后, 分两步对特征图进行融合操作: (1) 在两个特征图的通道维度进行特征图拼接, 合并高分辨率的特征图与低分辨率的特征图; (2) 运用 3×3 的卷积核对拼接后的特征图进行卷积操作, 减少合并后的特征图的混叠效应。假设特征图 f_a 和 f_b 的尺寸分别为 $W_2 \times H_2 \times C_a$ 和 $W_2 \times H_2 \times C_b$, 则 f_a 和 f_b 拼接后的融合特征图 f_c 的尺寸为 $W_2 \times H_2 \times (C_a + C_b)$, 即融合后的特征图仅通道数目发生了改变, 其他尺寸

并没有改变。而后进行卷积操作后的特征图 f_d 的尺寸为 $W_2 \times H_2 \times C_d$, 其中 C_d 为卷积核的数目。

2.1.2 混合域注意力

为了选择对检测任务更加重要的特征, 在特征提取器中融入了两种注意力机制的网络结构, 分别是对特征的通道域和空间域进行关注的注意力机制, 这两种注意力机制构成了网络中进行关键特征提取的参数。本文方法运用的对特征的通道域进行关注的注意力机制为 SeNet 结构, 对特征的空间域进行关注的注意力机制则类似 Vaswani 等人(2017)提出的自注意力(self-attention), 通过在网络的特征提取器上结合以上两种注意力域, 实现在特征的通道与空间两个方面提取关键特征的效用。

对于目标检测任务, 一幅图像中不同空间区域的需求度是不同的, 有些区域需要更多的注意力, 而有些区域则可以忽略, 普通的卷积过程对所有特征都是等价的, 故并不能实现以上功能。受 Vaswani 等人(2017)的方法启发, 在特征提取器的输入中融入空间域的注意力机制可对输入图像进行空间变换, 目标区域实现了针对性地矫正转换。故在特征提取器的输入中融入一种应用在图像领域的空间域注意力机制, 通过在输入中融入自注意力机制可以使网络实现对图像进行空间变换的功能, 以强化输入的空间域中的目标特征, 使目标特征与背景特征的权重不一致, 从而突出图像感受野中的目标。

1) 在特征提取器的输入信息中融入自注意力机制, 如图3所示, 在输入图像输入编码器网络前融入空间域注意力机制。具体过程为: 首先运用 1×1 的卷积核输入机制的特征图 x 进行卷积, 从而实现在转换特征图通道域的同时保持特征图的空间域不变, 此时生成的新特征映射为机制的分数。输入 f_b 的维度为 $W_4 \times H_4 \times C_b$, 网络参数为 $W_\mu \in \mathbf{R}^{W_4 \times H_4 \times C_b}$, 则通过矩阵的点乘可求得新的特征映射 $\mu(x)$ 为

$$\mu(x) = W_\mu x \quad (1)$$

在经过特征变换得到新特征映射后, 使用 ReLU 激活函数对新特征映射执行非线性处理, 获得对应原始输入特征 f_b 的响应特征 $\mu(x_i)$ 为特征 i 对应的新特征映射, 特征 i 的响应特征 e_i 为

$$e_i = \max(\mu(x_i), 0) \quad (2)$$

根据获得的响应特征 e_i 可求得特征 i 的注意力得分 sc_i 为

$$sc_i = \frac{\exp(e_i)}{\sum_{i=1}^L \exp(e_i)} \quad (3)$$

式中, sc_i 表示 f_b 中特征 i 的注意力得分。

将注意力得分与输入 f_b 的新特征映射相乘可得最终输出的特征映射 f_{b_out} 为

$$f_{b_out} = \sum_{i=1}^L sc_i \mu(x_i) \quad (4)$$

通过在特征提取器的输入中添加空间域注意力机制, 可以使网络对图像中更值得注意的区域进行几何空间变换, 从而达到将更多的注意力集中于图像更有用的区域的目标, 对3维目标检测任务有一定的帮助。

运用卷积神经网络提取的特征图一般是全局感受野的信息, 是仅聚合局部感受野上的空间信息与通道信息后获取的, 故对获取全局特征通道间的依赖关系有些许欠缺。本文在特征提取器中融入了一种可对全局特征通道间的相互依赖关系建模的网络层, 此关系可以直接在特征提取器中被参数表示, 网络层为特征的通道域注意力机制 SeNet。SeNet 使网络可以自适应地校准特征通道间的关系, 表现为可以使网络学习到每个特征通道的重要程度并显示为权重, 最终增强权重重大即更重要的特征通道层并抑制权重小即更不重要的特征通道层。

2) 对特征提取器的编码器网络层的构造进行改动, 在编码器中卷积层的 conv1-4 阶段添加 SeNet 网络层, 如图3所示。主要过程为:

对输入到 SeNet 网络层的特征图进行 squeeze 操作, 即运用全局平均池化对输入到 SeNet 网络层的特征图按通道进行特征压缩, 生成代表每个特征通道的特征图的实数, 该实数在某种程度上代表每个特征通道的特征图的全局感受野, 表征在空间维度上的特征图响应的全局分布。输入到 SeNet 网络层的特征图 f_a 的维度为 $W_3 \times H_3 \times C_a$, 生成的代表所有特征通道的实数数列 z_a 的维度为 $1 \times 1 \times C_a$, 计算公式为

$$z_a = F_{sq}(f_a) = \frac{1}{W_3 \times H_3} \sum_{i=1}^{W_3} \sum_{j=1}^{H_3} f_a(i, j) \quad (5)$$

式中, $F_{sq}(f_a)$ 表示 squeeze 操作函数, $f_a(i, j)$ 表示第 a 通道的特征图第 i 行 j 列的特征值, 即局部描述子集合。

对实数数列进行 excitation 操作, 运用2个全连接层与1个 sigmoid 激活函数对实数数列 z_a 进行信

息融合,实现融合各个通道的特征图信息的目的。主要流程为:运用第1个全连接层 w_1 乘以实数数列

z_a 得到维度为 $1 \times 1 \times \frac{C_a}{16}$ 的数列, w_1 维度为 $\frac{C_a}{r} \times C_a$,

其中 r 是一个缩放参数,网络中设置 $r = 16$ 以减少特征通道个数,从而降低网络计算量,操作表示为 $d(w_1 z_a)$;接着继续与第2个全连接层 w_2 相乘, w_2

维度为 $C_a \times \frac{C_a}{r}$,此时得到的数列维度为 $1 \times 1 \times C_a$,

操作表示为 $q(z_a, w)$;最后,运用 sigmoid 激活函数对得到的数列进行处理,操作表示为 $s(q(z_a, w))$,得到刻画特征图中所有特征通道的特征图的权重实数数列 s_a ,其维度也为 $1 \times 1 \times C_a$ 。excitation 操作 $F_{ex}(z_a, w)$ 为

$$s_a = F_{ex}(z_a, w) = s(q(z_a, w)) = s(w_2 d(w_1 z_a)) \quad (6)$$

式中, w 表示全连接层的参数。

得到数列 s_a 之后,与输入到 SeNet 网络层的特征图 f_a 按通道进行相乘,即将每个特征通道的特征图作为一个 2 维矩阵,将矩阵中的每个值与 s_a 中对应通道的权重值相乘,操作表示为 $F_{sc}(f_a, s_a)$,得到最终维度为 $W_3 \times H_3 \times C_a$ 的输出特征 f_{a_out} 为

$$f_{a_out} = F_{sc}(f_a, s_a) = s_a \cdot f_a \quad (7)$$

通过运用 SeNet 对卷积输出的特征求通道域权重,本文网络可以使训练得到的模型相对只经过卷积的模型达到更好的检测效果。

2.1.3 特征空洞卷积

空洞卷积也称为扩张卷积,用于在不进行池化操作以损失特征信息的情况下,增大特征的感受野,让卷积输出的特征映射都包含较大范围的信息。

本文对特征提取器提取的特征图进行空洞卷积操作,以达到在不增加计算量的基础上得到不同感受野的特征信息,从而在全局与局部的层面对特征进行重要程度选择。主要流程如下:1)对特征提取器最后一层提取的维度为 $c \times h \times w$ 的特征图 f_c 进行 3 层空洞卷积, c 表示特征图通道数, w 与 h 表示特征图的长与宽,卷积核通道数目保持不变,都与输入到空洞卷积层的特征图的通道数目一样,但是空洞卷积的卷积核的间隔数量(dilation rate)呈现锯齿状结构,以保证对特征图的空间域中所有的特征映射都进行了卷积。本文的空洞卷积的卷积核的间隔数量为 $[1, 2, 5]$,尺寸保持不变,得到对应卷积的

输出为 $[f_{dc1}, f_{dc2}, f_{dc5}]$,其中每层输出的特征图维度则为 $c \times h \times w$;2)对每层空洞卷积的输出求通道域的注意力权重,运用 SeNet 求得对应权重为 $[c_{dc1}, c_{dc2}, c_{dc5}]$,每个对应权重的维度为 $c \times 1 \times 1$;3)对 $[2, 5]$ 两层空洞卷积的输出得到的注意力权重 c_{dc2} 与 c_{dc5} 逐元素相乘,并运用非线性函数 sigmoid 激活得到权重值 c_{dc25} ,维度为 $c \times 1 \times 1$;4)将权重 c_{dc1} 与 c_{dc25} 逐元素相乘得到特征图 f_{end} 的最终通道域的权重值,维度为 $c \times 1 \times 1$,将它们逐通道相乘得到最终特征映射 f_{ended} ,维度为 $c \times h \times w$ 。

2.2 区域生成网络

本文方法运用的 RPN 的结构与 2 维目标检测网络中运用的 RPN 类似,即在一系列的先验框与真实标签(ground truth)之间不断迭代得到网络模型参数,而这些先验框就是预先定义的锚框。2 维目标检测网络中的锚框是 2 维框,而本文网络的锚框是 3 维框,可以表示为 $[t_x, t_y, t_z, d_x, d_y, d_z]$,其中 $[t_x, t_y, t_z]$ 和 $[d_x, d_y, d_z]$ 分别代表锚框的质心坐标和轴对齐尺寸,是通过每个类别的训练样本进行聚类确定的。

得到网络的 3D 锚框后,将其中的每个锚框分别投影到 RGB 特征图与 BEV 特征图中。而后,通过 3 维感兴趣区域池化将锚框投影到 RGB 特征图与 BEV 特征图中的对应区域剪裁为感兴趣区域,并将 RGB 特征图与 BEV 特征图中的两种感兴趣区域的特征向量调整为等尺寸的特征向量,然后按元素取均值进行合并,得到合并后的特征向量。最后运用两个全连接层利用特征向量回归 3 维提案框,并输出一个评价提案框为“前景或背景”的分数。

本文网络 3 维框的回归是计算 3D 锚框与真实标签的质心坐标与轴对齐尺寸的差异。网络对 3 维框质心坐标与尺寸的回归运用了 smooth L1 函数作为代价函数计算,对“前景或背景”分数的回归运用了交叉熵函数作为代价函数计算。计算回归损失时,本文网络将背景锚框忽略,通过计算鸟瞰图中的 3D 锚框与真实标签 3 维框的交并比确定背景锚框。

2.3 第 2 阶段检测网络

本文网络运用的第 2 阶段网络为一个检测子网络,网络结构类似 RPN,网络输入是第 1 阶段的 RPN 网络得到的 3 维提案框,输出是最终 3 维目标检测框。具体流程为:1)将第 1 阶段得到的 3 维提案框作为输入分别投影到特征提取器提取的 RGB

特征图与 BEV 特征图,并在两种特征图上裁剪得到提案框的特征。因为提案框的数量相对锚框的数量少很多,故得到的提案框的特征相对第1阶段的锚框的特征数量少,此时得到的提案框的特征的深度信息与特征提取器提取的特征图的通道相同;2)将两种特征图上的提案框的特征尺寸调整为等尺寸的特征向量,并按元素取均值合并两种提案框的特征,得到合并后的提案框的特征向量;3)运用几组全连接层处理合并后的提案框的特征向量,回归每个3维提案框的坐标、尺寸、方向以及类别;4)运用非极大值抑制(non maximum suppression, NMS)算法(Neubeck 和 VanGool, 2006)删除重叠的3维提案框。本文网络中将检测子网络的 NMS 的阈值设置为 0.01。

第2阶段检测子网络使用的损失函数与 RPN 类似,对3维目标检测框的质心坐标与尺寸的回归及方向估计采用 smooth L1 函数作为代价函数计算,分类任务采用交叉熵函数作为代价函数计算。

2.4 模型训练

本文网络作为一个室外3维目标检测网络,主要用于检测室外的汽车、行人和骑自行车的人,故训练了两种模型,一种用于汽车,另一种用于行人和骑自行车的人。

在训练时,以端到端的方式联合训练第1阶段的 RPN 与第2阶段的检测子网络,训练优化器使用 Adam 优化器,对网络进行 120 000 次迭代,训练时,网络的初始学习率为 0.000 1,每进行 30 000 次迭代,将学习率以衰减系数为 0.8 的指数衰减。

3 实验与结果分析

3.1 实验数据

采用公开的3维目标检测数据集 KITTI(A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)(Geiger 等, 2013)中的3个类别对本文网络进行测试,训练集与验证集按照 MV3D 中对数据提供的训练帧的比例划分,即将 7 481 幅图像按约 1:1 的比例划分。用于3维目标检测任务的 KITTI 数据集包含 RGB 图和对应的激光雷达点云数据,每组数据对(RGB 图像与雷达点云)都含有对物体的3维框的标注信息,物体的遮挡程度按轻度(easy)、中度(moderate)与高度(hard)进行标识。

3.2 实验参数设置

实验中输入到网络的 RGB 图像是 3 通道的彩色图,剪裁为统一尺寸,激光雷达点云的处理按照 MV3D 中的处理,投影为等尺寸的鸟瞰图,鸟瞰图为 6 通道的体素网格。使用多层金字塔结构的卷积对输入的数据对进行特征提取,其中卷积核数量采用 VGG-16 结构前 4 层的卷积核数目,使用的详细参数如表 1 所示。

表 1 网络参数设置
Table 1 Parameter settings on network

超参数	值
RGB 图维度	[360, 1 200, 3]
鸟瞰图维度	[700, 800, 6]
编码器各层运用的核数目	32, 64, 128, 256
编码器+混合域注意力的核数目	32, 64, 128, 256
解码器各层运用的核数目	64, 32, 32
第1阶段网络 mini-batch	512
第2阶段网络 mini-batch	1 024

3.3 对比实验

3.3.1 其他方法

将本文提出的结合混合域注意力与空洞卷积的二阶段3维目标检测网络与其他方法在 KITTI 数据集上进行对比实验,以验证本文网络的先进性和有效性。

1)端到端的3维目标检测网络 VoxelNet。本文提出的网络为二阶段网络,使用了先验锚框并提取得到更加精确的提案框,而 VoxelNet 为端到端的网络结构,与本文提出的网络相比,在检测速度上没有优势,在提案框的生成精度上不及本文提出的网络。

2)二阶段的3维目标检测网络 MV3D。本文提出的网络的特征提取器采用了一种结合混合域注意力与空洞卷积的卷积神经网络作为特征提取器,可以生成高分辨率与更重要的特征图,有效识别小目标的物体,而 MV3D 仅运用普通 VGG-16 网络作为特征提取器,由于特征的缺陷,对目标的检测精度大大不及本文方法。

3)二阶段的3维目标检测网络 F-PointNet。本文提出的网络采用推广到3维的 RPN,先提出3维提案框,再运用第2阶段检测网络对3维提案框进行优化,从而得到3维目标检测框。而 F-PointNet 先运用2维目标检测框限制3维检测框的范围,再

利用点云处理网络 PointNet (Charles 等, 2017) 对 2 维检测框限定范围内的点云进行处理以回归 3 维目标检测框, 点云一般会因为距离、遮挡等原因或稀疏或不完整, 故 F-PointNet 对于远距离和互相遮挡的目标检测效果不如本文网络。

4) 二阶段的 3 维目标检测网络 AVOD。本文网络的特征提取器结合混合域注意力与空洞卷积的结构, 使得模型在训练过程中可以针对重要程度不同的特征进行选择以调整并优化模型的参数, 对比仅运用了金字塔结构的卷积神经网络作为特征提取器的 AVOD, 本文由于提取的特征对于检测任务更具针对性, 故检测效果更好。

3.3.2 消融实验

1) 本文网络 + 去除金字塔结构。在本文方法的特征提取器中去除金字塔结构。该网络不具备提取高分辨率特征图的能力, 对于小目标以及目标细节的检测效果较差。

2) 本文网络 + 去除混合域注意力。在本文方法的特征提取器中去除了混合域注意力。该网络不能自动学习获取特征的重要程度的参数, 既不能增强对当前检测任务有用的特征, 也不能抑制对当前任务用处不大的特征。

3) 本文网络 + 去除空洞卷积。在本文方法的特征提取器中去除了空洞卷积操作。该网络不能从多

个感受野的角度出发去分析通道特征的重要程度。

4) 基本网络。在本文方法的特征提取器中去除了金字塔结构、混合域注意力与空洞卷积层。该网络架构中运用的特征提取器与 MV3D 网络结构中运用的特征提取器一样, 都是利用 VGG-16 网络的前 4 个阶段的网络, 得到的最终特征图的尺寸上采样为原输入图像的 1/4。

3.4 实验结果与分析

基于 KITTI 数据集对 3.3 节所述的不同网络的模型精度进行验证, 并与本文提出的网络进行对比。采用目标检测任务中通用的评估基准——交并比 (intersection over union, IoU) 对网络模型的精度进行评价。本文网络学习训练时对 KITTI 数据集中的不同类别的数据设定了不同的 IoU 阈值。类别 car 的 IoU 阈值设置为 0.7, pedestrian 与 cyclist 这两个类别的 IoU 阈值都设置为 0.5。通过以上的对比实验可以有效验证本文提出的网络模型的先进性。不同网络在 KITTI 数据集中的测试集上对 car、pedestrian、cyclist 类别检测得到的 AP_{3D} 与 AP_{BEV} 值如表 2 和表 3 所示 (AP_{3D} 表示 3 维目标边界框的平均精度, AP_{BEV} 表示鸟瞰视角中目标 2 维边界框的平均精度, 也代表目标 3 维边框的坐标的平均精度)。表中 easy、moderate 与 hard 表示物体被遮挡的程度依次为轻度、中度与高度。

表 2 不同方法在 KITTI 数据集中的测试集上得到的 AP_{3D} 值
Table 2 Values of AP_{3D} obtained by different methods on test dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
MV3D	71.09	62.35	55.12	—	—	—	—	—	—
VoxelNet	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37
F-pointNet	81.2	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39
AVOD	81.94	71.88	66.38	50.8	42.81	40.88	64	52.18	46.61
本文	83.45	74.29	67.92	52.23	44.91	41.64	65.02	54.38	47.97

注: 加粗字体表示各列最优结果, “—”表示无数据。

由表 2 和表 3 的实验结果可知, 本文方法在 KITTI 数据集的测试集中, 无论物体的遮挡情况如何, 都比其他方法取得了相对更好的检测结果。其中, 对 car 类别检测时, 在中度遮挡情况下, 本文方法的 AP_{3D} 值比其他方法中表现最好的 AVOD 的 AP_{3D} 值高 2.41%, AP_{BEV} 值比其他方法中表现最好的

F-PointNet 的 AP_{3D} 值高 3.62%; 在所有的遮挡情况下, 本文方法的 AP_{3D} 和 AP_{BEV} 值比其他方法的 AP_{3D} 和 AP_{BEV} 值平均高 1.82% 和 2.29%。对 pedestrian 类别检测时, 在所有的遮挡情况下, 本文方法的 AP_{3D} 和 AP_{BEV} 值比其他方法的 AP_{3D} 和 AP_{BEV} 值平均高 0.60% 和 1.99%。对 cyclist 类别检测时, F-Point-

表3 不同方法在KITTI数据集中的测试集上得到的 AP_{BEV} 值
Table 3 Values of AP_{BEV} obtained by different methods on test dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
MV3D	86.02	76.9	68.49	—	—	—	—	—	—
VoxelNet	89.35	79.26	77.39	46.13	40.74	38.11	66.7	54.76	50.55
F-pointNet	88.7	84	75.33	58.09	50.22	47.2	75.78	61.96	54.68
AVOD	88.53	83.79	77.9	58.75	51.05	47.54	68.06	57.48	50.77
本文	89.61	87.05	79.69	59.73	53.97	49.62	69.13	59.69	52.11

注:加粗字体表示各列最优结果,“—”表示无数据。

Net 表现最好,主要因为 KITTI 数据集上的 cyclist 类别的实例较少,导致训练得到的模型参数不够准确。

本文方法在 KITTI 数据集中的测试数据集上对

car、pedestrian、cyclist 类别进行消融对比实验得到的 AP_{3D} 和 AP_{BEV} 如表4和表5所示。对各个网络同样基于得到的 AP_{3D} 与 AP_{BEV} 值进行验证对比。

表4 本文方法在KITTI数据集中的测试集上进行消融对比实验得到的 AP_{3D} 值
Table 4 Values of AP_{3D} obtained by our method on test dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
基本网络	73.59	65.78	58.38	38.28	31.51	26.98	60.11	44.9	38.8
去除混合域注意力	83.12	73.25	67.15	51.82	44.54	41.55	64.87	54.02	47.54
去除空洞卷积	82.84	72.88	66.97	51.63	44.36	41.53	64.74	53.95	47.45
去除金字塔	77.81	68.03	61.55	41.36	35.75	30.11	62.13	48.35	40.18
本文完整网络	83.45	74.29	67.92	52.23	44.91	41.64	65.02	54.38	47.97

注:加粗字体表示各列最优结果。

表5 本文方法在KITTI数据集中的测试集上进行消融对比实验得到的 AP_{BEV} 值
Table 5 Values of AP_{BEV} obtained by our method on test dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
基本网络	86.8	85.44	77.73	42.51	35.24	33.97	63.66	47.74	46.55
去除混合域注意力	88.81	86.73	78.36	59.54	53.18	49.42	68.77	59.02	51.24
去除空洞卷积	88.85	86.41	78.28	59.16	53.15	49.31	68.79	58.89	51.31
去除金字塔	87.92	87.01	78.13	45.13	40.12	37.68	65.34	51.67	48.97
本文完整网络	89.61	87.05	79.69	59.73	53.97	49.62	69.13	59.69	52.11

注:加粗字体表示各列最优结果。

由表4和表5的结果可知,对car类别的物体,相对本文提出的完整网络,采用基本网络、网络中去除金字塔结构、去除混合域注意力结构、去除空洞卷

积结构后的 AP_{3D} 值分别平均降低了约 9.30%、6.09%、0.99% 和 0.71%。对 pedestrian 与 cyclist 类别的物体,采用上述方法得到的结果与 car 类别

并无二致。

为了进一步分析本文提出的网络,在 KITTI 数据

集中的验证数据集上对比其他方法来验证网络模型的精度,不同网络模型的检测结果如表 6 和表 7 所示。

表 6 不同方法在 KITTI 数据集中的验证集上得到的 AP_{3D} 值
Table 6 Values of AP_{3D} obtained by different methods on validation dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
MV3D	71.29	62.68	56.56	—	—	—	—	—	—
VoxelNet	81.97	65.46	62.85	42.51	33.92	33.52	64.73	48.79	47.37
F-PointNet	83.76	70.92	63.65	53.65	45.73	41.73	74.38	57.29	51.79
本文	85.66	76.75	69.84	54.51	46.98	43.21	67.52	56.07	50.02

注:加粗字体表示各列最优结果,“—”表示无数据。

表 7 不同方法在 KITTI 数据集中的验证集上得到的 AP_{BEV} 值
Table 7 Values of AP_{BEV} obtained by different methods on validation dataset of KITTI

方法	car			pedestrian			cyclist		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
MV3D	86.55	78.1	76.67	—	—	—	—	—	—
VoxelNet	89.6	84.81	78.57	48.32	41.89	39.57	66.97	59.21	51.75
F-PointNet	88.16	84.02	76.44	58.11	50.2	48.41	75.86	61.99	55.77
本文	89.03	85.72	78.33	59.82	55.88	50.41	71.32	60.91	54.05

注:加粗字体表示各列最优结果,“—”表示无数据。

由表 6 和表 7 的结果可知,本文方法取得了相对更好的检测结果。其中,本文方法对 car 类别的物体进行检测后,在物体高度遮挡时, AP_{3D} 值相对其他方法中表现最好的 F-PointNet 方法的 AP_{3D} 值提高 6.19%,在所有的遮挡情况下,本文方法的 AP_{3D} 值相对所有其他方法的 AP_{3D} 值平均提高 4.64%。对比 VoxelNet 方法,本文方法的 AP_{BEV} 值仅在物体中度遮挡程度时占优,其他遮挡情况下均有所不如,但相对表中的其他方法均具有优势,与 MV3D 和 F-PointNet 相比,本文的 AP_{BEV} 值分别平均提高 3.92% 和 1.49%。

表 8 为本文方法在 KITTI 数据集中的验证数据集上检测 pedestrian 与 cyclist 两类物体的得到 AP_{3D} 与 AP_{BEV} 值。

除了采用 AP_{3D} 与 AP_{BEV} 对网络的性能进行分析,还使用 AP 与 AHS 对网络模型进行验证,其中, AP 表示目标边界框的平均精度, AHS 表示目标方向的平均精度。本文对 MV3D 和 AVOD 方法在 KITTI 数据集中的验证数据集上的 car 类别进行验

表 8 在 pedestrian 与 cyclist 类别得到的 AP_{3D} 与 AP_{BEV}
Table 8 Values of AP_{3D} and AP_{BEV} obtained on the pedestrian and cyclist class

类别	AP_{3D}			AP_{BEV}		
	easy	moderate	hard	easy	moderate	hard
pedestrian	54.41	46.92	44.41	61.87	55.28	51.21
cyclist	67.49	56.23	50.28	71.17	61.97	54.64

证,结果如表 9 所示。

由表 9 的数据可知,本文方法得到的 AHS 值相对 MV3D 方法具有明显优势,原因在于网络计算了矢量方向的损失值,从而使网络模型对方向有矢量估计。相对 AVOD 方法得到的 AHS 值,也有优势表明本文方法提取的特征的先进性。由此也进一步表明了本文方法提取的特征的先进性及具优势。

3.5 网络运行时间分析

基于相同的操作系统 ubuntu16.04 与硬件配置

表9 实验验证得到的AP与AHS值
Table 9 Values of AP and AHS obtained
with experimental validation

方法	easy		moderate		hard		/%
	AP	AHS	AP	AHS	AP	AHS	
MV3D	83.87	52.74	72.35	43.75	64.56	39.86	
AVOD	84.41	84.19	74.44	74.11	68.65	68.28	
本文	84.78	84.57	76.51	75.37	69.87	68.89	

NVIDIA GeForce GTX 1080 Ti 分析不同网络的训练时间以及测试时间。表10为不同网络训练完成一次迭代(epoch)的时间。

表10 不同方法训练时完成一次迭代的时间
Table 10 Runtime for accomplishing an epoch while
training with different methods

方法	时间	/s
基本网络	6.71	
去除混合域注意力	7.89	
去除空洞卷积	8.04	
去除金字塔	6.91	
本文完整网络	8.20	

由表10可知,本文完整网络在网络训练时完成一次迭代所需的时间为8.20 s,综合表4和表5可知,本文网络不仅有优秀的检测性能,而且网络的训练运行时间也耗费不大。

为了更进一步分析网络的运行性能,实验对不同网络对一幅图像进行检测所需的时间(即对一幅图像的测试时间)进行了对比,结果如表11所示。

表11 不同方法对一幅图像进行检测的运行时间
Table 11 Runtime for an image while testing
with different methods

方法	时间	/(s/帧)
基本网络	0.08	
去除混合域注意力	0.13	
去除空洞卷积	0.16	
去除金字塔	0.15	
本文完整网络	0.18	

由表11的数据可知,本文提出的完整网络对一幅图进行检测所需的耗时约0.2 s,约等于实时检测,表中数据也验证了本文网络在检测性能和运行性能方面都具有相对良好的表现。

4 结 论

针对现有的3维目标检测网络提取的特征缺乏区域与通道的依赖关系,且不能在无损分辨率的情况下扩大感受野,导致回归的3维框有一定误差的问题,提出一种结合混合域注意力与空洞卷积的二阶段3维目标检测方法,通过在网络中引入空间域与通道域注意力机制以获取特征映射的特征权重,解决了区分关键特征的问题;通过在网络的输出层中运用融入结合空洞卷积与通道域注意力机制的网络层,获取了全局感受野的关键通道特征;并且在特征提取器中融入了金字塔网络结构,提取了高分辨率的特征图。实验结果表明,本文算法因为提取的特征的优秀性,回归的检测结果相对很多优秀的3维目标检测方法具有更高的精确性。但是,本文方法针对网络添加网络层,增加了网络的参数量,也增加了网络的运行时间。未来的工作将针对优化网络的运行时间和模型参数量展开进一步研究。

参考文献 (References)

- Charles R Q, Su H, Kaichun M and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]
- Chen X Z, Ma H M, Wan J, Li B and Xia T. 2017. Multi-view 3D object detection network for autonomous driving//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE: 6526-6534 [DOI: 10.1109/CVPR.2017.691]
- Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]
- Geiger A, Lenz P, Stiller C and Urtasun R. 2013. Vision meets robotics: the KITTI dataset. International Journal of Robotics Research, 32(11): 1231-1237 [DOI: 10.1177/0278364913491297]

- Hu J, Shen L, Albanie S, Sun G and Wu E H. 2017. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*; 1-13 [DOI: 10. 1109/TPAMI. 2019. 2913372]
- Ku J, Mozifian M, Lee J, Harakeh A and Waslander S L. 2018. Joint 3D proposal generation and object detection from view aggregation// *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid: IEEE; 1-8 [DOI: 10. 1109/IROS.2018.8594049]
- Li B, Zhang T L and Xia T. 2016. Vehicle detection from 3d lidar using fully convolutional network. *Proceedings of Robotics: Science and Systems*; #42 [DOI: 10. 15607/RSS.2016. XII.042]
- Li B. 2017. 3D fully convolutional network for vehicle detection in point cloud//*Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE; 1513-1518 [DOI: 10. 1109/IROS.2017.8205955]
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI: IEEE; 936-944 [DOI: 10. 1109/CVPR.2017.106]
- Lowe D G. 1999. Object recognition from local scale-invariant features// *Proceedings of the 7th IEEE International Conference on Computer Vision*. Kerkyra, Greece: IEEE; 1150-1157 [DOI: 10. 1109/ICCV.1999.790410]
- Neubeck A and Van Gool L J. 2006. Efficient Non-Maximum Suppression//*Proceedings of the 18th International Conference on Pattern Recognition*, Hong Kong, China: IEEE; 850-855 [DOI: 10. 1109/ICPR.2006.479]
- Papageorgiou C P, Oren M and Poggio T. 1998. A general framework for object detection//*Proceedings of the 6th International Conference on Computer Vision*. Bombay, India: IEEE; 555-562 [DOI: 10. 1109/ICCV.1998.710772]
- Qi C R, Liu W, Wu C X, Su H and Guibas L J. 2018. Frustum Point-Nets for 3D object detection from RGB-D data//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE; 918-927 [DOI: 10. 1109/CVPR.2018.00102]
- Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE; 779-788 [DOI: 10. 1109/CVPR.2016.91]
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (6): 1137-1149 [DOI: 10. 1109/TPAMI.2016.2577031]
- Simon M, Amende K, Kraus A, Honer J, Sämann T, Kaulbersch H, Milz S and Gross H M. 2019. Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds [EB/OL]. [2019-10-24]. <https://arxiv.org/pdf/1904.07537.pdf>
- Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-10-24]. <https://arxiv.org/pdf/1409.1556.pdf>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need [EB/OL]. [2019-10-24]. <https://arxiv.org/pdf/1706.03762.pdf>
- Yu F and Koltun V. 2015. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2019-10-24]. <https://arxiv.org/pdf/1511.07122.pdf>
- Zhou J, Tan X, Shao Z W and Ma L Z. 2019. FVNet: 3D front-view proposal generation for real-time object detection from point clouds// *Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. Suzhou, China: IEEE; 1-8 [DOI: 10. 1109/CISP-BMEI48845.2019.8965844]
- Zhou Y and Tuzel O. 2018. Voxelnet: end-to-end learning for point cloud based 3d object detection//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE; 4490-4499 [DOI: 10. 1109/CVPR.2018.00472]

作者简介



严娟, 1994年生, 女, 硕士研究生, 主要研究方向为计算机视觉。

E-mail: juanyan_sues@163.com



方志军, 通信作者, 男, 教授, 主要研究方向为图像视频分析、计算机视觉、智能数据分析。

E-mail: zjfang@sues.edu.cn

高永彬, 男, 讲师, 主要研究方向为计算机视觉。

E-mail: gaoyongbin@sues.edu.cn