



主办: 中国科学院空天信息创新研究院
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

2022
08
VOL.27

ISSN1006-8961
CN11-3758/TB



工业低纹理目标检测 P2418



第27卷第8期（总第316期）
2022年8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 中国科学院空天信息创新研究院
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 吴一戎
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@aircas.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

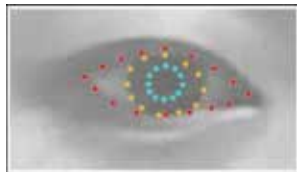
Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
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@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

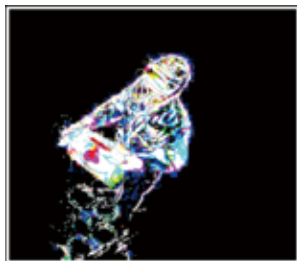
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

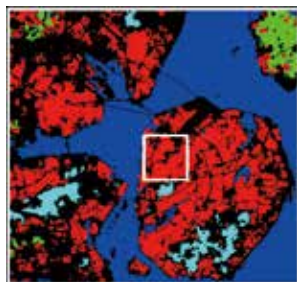
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



自然光普通摄像头的眼部分割及特征点定位数据集ESLD (第2329页)



双视图三维卷积网络的工业装箱行为识别(第2368页)



多源特征自适应融合网络的高分遥感影像语义分割(第2516页)

综述

虚拟现实图像客观质量评价研究进展

周玉, 汪一, 李雷达, 高陈强, 卢兆林 2313

数据集论文

自然光普通摄像头的眼部分割及特征点定位数据集ESLD

张俊杰, 孙光民, 郑鲲, 李煜, 付晓辉, 慈康怡, 申俊杰, 孟凡超, 孔江萍, 张玥 2329

图像处理和编码

双尺度顺序填充的深度图像修复

陈东岳, 朱晓明, 马腾, 宋园园, 贾同 2344

二维码位流长度最小化算法

袁泰凌, 徐昆 2356

图像分析和识别

双视图三维卷积网络的工业装箱行为识别

胡海洋, 潘健, 李忠金 2368

面向航拍图像中工程车辆检测与识别的改进胶囊网络

钟映春, 郑海阳, 张文祥, 王波, 罗志勇 2380

用于骨架行为识别的多维特征融合注意力机制

姜权晏, 吴小俊, 徐天阳 2391

高性能整数倍稀疏网络行为识别研究

臧影, 刘天娇, 赵曙光, 杨东升 2404

面向工业零件分拣系统的低纹理目标检测

闫明, 陶大鹏, 普园媛 2418

融合策略优选和双注意力的单阶段目标检测

戴坤, 许立波, 黄世旸, 李鉴铃 2430

融合视觉词与自注意力机制的视频目标分割

季传俊, 陈亚当, 车洵 2444

基于Gaussian-Hermite矩的旋转运动模糊不变量

郭锐, 贾丽, 郝宏翔, 墨瀚林, 李华 2458

图像理解和计算机视觉

RGB-D语义分割: 深度信息的选择使用

赵经阳, 余昌黔, 桑农 2473

自纠正噪声标签的人脸美丽预测

甘俊英, 吴必诚, 翟懿奎, 何国辉, 麦超云, 白振峰 2487

面向非对称特征注意力和特征融合的太赫兹图像检测

曾文健, 朱艳, 沈韬, 曾凯, 刘英莉 2496

面向纹理平滑的方向性滤波尺度预测模型

林俊彦, 刘春晓, 章金凯, 李泓易 2506

遥感图像处理

多源特征自适应融合网络的高分遥感影像语义分割

张文凯, 刘文杰, 孙显, 许光奎, 付琨 2516

结合上下文编码与特征融合的SAR图像分割

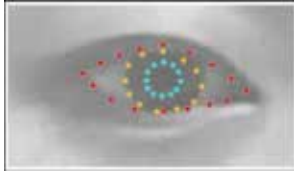
范艺华, 董张玉, 杨学志 2527

一阶全卷积遥感影像倾斜目标检测

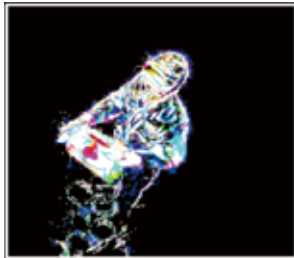
周院, 杨庆庆, 马强, 薛博维, 孔祥楠 2537

CONTENTS

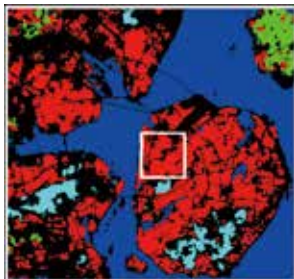
JOURNAL OF IMAGE AND GRAPHICS



ESLD: eyes segment and landmark detection in the wild (P2329)



Dual-view 3D ConvNets based industrial packing action recognition(P2368)



Multi-source features adaptation fusion network for semantic segmentation in high-resolution remote sensing images(P2516)

Review

Research progress in objective quality evaluation of virtual reality images

Zhou Yu, Wang Yi, Li Leida, Gao Chenqiang, Lu Zhaolin 2313

Dataset

ESLD: eyes segment and landmark detection in the wild

Zhang Junjie, Sun Guangmin, Zheng Kun, Li Yu, Fu Xiaohui, Ci Kangyi, Shen Junjie, Meng Fan-
chao, Kong Jiangping, Zhang Yue 2329

Image Processing and Coding

Depth image recovery based on dual-scale sequential optimized filling

Chen Dongyue, Zhu Xiaoming, Ma Teng, Song Yuanyuan, Jia Tong 2344

Minimization of bit stream length of QR codes

Yuan Tailing, Xu Kun 2356

Image Analysis and Recognition

Dual-view 3D ConvNets based industrial packing action recognition

Hu Haiyang, Pan Jian, Li Zhongjin 2368

Improved capsule network method for engineering vehicles detection and recognition in aerial ima-
ges

Zhong Yingchun, Zheng Haiyang, Zhang Wenxiang, Wang Bo, Luo Zhiyong 2380

M2FA: multi-dimensional feature fusion attention mechanism for skeleton-based action recognition

Jiang Quanyan, Wu Xiaojun, Xu Tianyang 2391

Action recognition analysis derived of integer sparse network

Zang Ying, Liu Tianjiao, Zhao Shuguang, Yang Dongsheng 2404

Texture-less object detection method for industrial components picking system

Yan Ming, Tao Dapeng, Pu Yuanyuan 2418

Single stage object detection algorithm based on fusing strategy optimization selection and dual
attention mechanism

Dai Kun, Xu Libo, Huang Shiyang, Li Yunling 2430

Visual words and self-attention mechanism fusion based video object segmentation method

Ji Chuanjun, Chen Yadang, Che Xun 2444

Rotational motion blur invariants based on Gaussian-Hermite moments

Guo Rui, Jia Li, Hao Hongxiang, Mo Hanlin, Li Hua 2458

Image Understanding and Computer Vision

RGB-D semantic segmentation: depth information selection

Zhao Jingyang, Yu Changqian, Sang Nong 2473

Self-correcting noise labels for facial beauty prediction

Gan Junying, Wu Bicheng, Zhai Yikui, He Guohui, Mai Chaoyun, Bai Zhenfeng 2487

Terahertz image detection combining asymmetric feature attention and feature fusion

Zeng Wenjian, Zhu Yan, Shen Tao, Zeng Kai, Liu Yingli 2496

Texture-smoothing-oriented directional filtering scales-predicting model

Lin Junyan, Liu Chunxiao, Zhang Jinkai, Li Hongyi 2506

Remote Sensing Image Processing

Multi-source features adaptation fusion network for semantic segmentation in high-resolution re-
mote sensing images

Zhang Wenkai, Liu Wenjie, Sun Xian, Xu Guangluan, Fu Kun 2516

The integrated contextual encoding and feature fusion SAR images segmentation method

Fan Yihua, Dong Zhangyu, Yang Xuezhi 2527

Improved one-stage fully convolutional network for oblique object detection in remote sensing ima-
gery

Zhou Yuan, Yang Qingqing, Ma Qiang, Xue Bowei, Kong Xiangnan 2537

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2022)08-2473-14

论文引用格式: Zhao J Y, Yu C Q and Sang N. 2022. RGB-D semantic segmentation: depth information selection. Journal of Image and Graphics, 27(08): 2473-2486 (赵经阳, 余昌黔, 桑农. 2022. RGB-D 语义分割: 深度信息的选择使用. 中国图象图形学报, 27(08): 2473-2486) [DOI: 10.11834/jig.210061]

RGB-D 语义分割: 深度信息的选择使用

赵经阳, 余昌黔, 桑农*

华中科技大学人工智能与自动化学院图像信息处理与智能控制教育部重点实验室, 武汉 430074

摘要: **目的** 在室内场景语义分割任务中, 深度信息会在一定程度上提高分割精度。但是如何更有效地利用深度信息仍是一个开放性问题。当前方法大都引入全部深度信息, 然而将全部深度信息和视觉特征组合在一起可能对模型产生干扰, 原因是仅依靠视觉特征网络模型就能区分不同物体, 在引入深度信息后可能产生错误判断。此外, 卷积核固有的几何结构限制了卷积神经网络的建模能力, 可变形卷积 (deformable convolution, DC) 在一定程度上缓解了这个问题。但是可变形卷积中产生位置偏移的视觉特征空间深度信息相对不足, 限制了进一步发展。基于上述问题, 本文提出一种深度信息引导的特征提取 (depth guided feature extraction, DFE) 模块。**方法** 深度信息引导的特征提取模块包括深度信息引导的特征选择模块 (depth guided feature selection, DFS) 和深度信息嵌入的可变形卷积模块 (depth embedded deformable convolution, DDC)。DFS 可以筛选出关键的深度信息, 自适应地调整深度信息引入视觉特征的比例, 在网络模型需要时将深度信息嵌入视觉特征。DDC 在额外深度信息的引入下, 增强了可变形卷积的特征提取能力, 可以根据物体形状提取更相关的特征。**结果** 为了验证方法的有效性, 在 NYUv2 (New York University Depth Dataset V2) 数据集上进行一系列消融实验并与当前最好的方法进行比较, 使用平均交并比 (mean intersection over union, mIoU) 和平均像素准确率 (pixel accuracy, PA) 作为度量标准。结果显示, 在 NYUv2 数据集上, 本文方法的 mIoU 和 PA 分别为 51.9% 和 77.6%, 实现了较好的分割效果。**结论** 本文提出的深度信息引导的特征提取模块, 可以自适应地调整深度信息嵌入视觉特征的程度, 更加合理地利用深度信息, 且在深度信息的作用下提高可变形卷积的特征提取能力。此外, 本文提出的深度信息引导的特征提取模块可以比较方便地嵌入当下流行的特征提取网络中, 提高网络的建模能力。

关键词: 语义分割; RGB-D; 深度信息引导的特征选择 (DFS); 深度信息嵌入的可变形卷积 (DDC); 深度信息引导的特征提取 (DFE)

RGB-D semantic segmentation: depth information selection

Zhao Jingyang, Yu Changqian, Sang Nong*

Key Laboratory of Ministry of Education for Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: **Objective** Semantic segmentation is essential to computer vision application. It assigns each pixel to its corresponding category in an image, which is a pixel leveled multi-classification task. It is of great significance in the fields of automatic driving, virtual reality and medical image processing. The emergence of convolutional neural network (CNN)

收稿日期: 2021-01-29; 修回日期: 2021-06-28; 预印本日期: 2021-07-03

* 通信作者: 桑农 nsang@hust.edu.cn

基金项目: 国家自然科学基金项目 (61433007)

Supported by: National Natural Science Foundation of China (61433007)

promotes the rapid development of neural network in various tasks of computer vision. The fully CNN has completely changed the pattern of semantic segmentation contexts. With the advent of depth camera, it is more convenient to obtain the depth images corresponding to color images. The depth image is single-channel, and each value should be ranged from the pixel to the camera plain in the image. Obviously, depth images contain spatial distance information, but color images are relatively insufficient. In the semantic segmentation task, it is difficult for the network to distinguish the adjacent objects with similar appearance in the plain image, but the application of depth image can be released to some extent. RGB-D semantic segmentation is focused on recently. The ways of depth information-embedded visual features can be roughly divided into the following three categories like one-stream, two-streams and multi-tasks. One-stream does not use depth images as additional input to extract features. It only has a backbone network to extract features from color images. In the process of feature extraction, the inherent spatial information of depth images is used to assist visual feature extraction for semantic segmentation improvement. Two-streams use the depth image as an additional input to extract features. There are mainly two backbone networks involved, each of which extracts features from color images and depth images each. In the encoding stage or decoding stage, the extracted visual features are fused with depth features to realize depth information application. Multi-task processes semantic segmentation, depth estimation and surface normal estimation at the same time. Such a method has one common backbone network only. In the process of feature extraction from color images, multi feature interaction can improve the performance of each task. Previous studies have challenged to effective depth information as well, and embedding all depth information into visual features may cause interference to the network. The inherent color and texture information can sometimes clearly distinguish two or more categories in a color image, where the addition of depth information is somewhat gilding the lily. For example, similar depth objects can be distinguished by visual features excluded different visual features, but the addition of depth information will make the network confused and even make wrong judgments. Moreover, the inherent structure of the convolution kernel limits its ability of feature extraction in CNN. To solve this problem, our proposed deformable convolution can learn the offset of the corresponding points according to the input, and extract more effective features in terms of the shape of the object, thus improving the modeling ability of the network. However, it is insufficient to learn the offset only by the input of visual features, because the spatial information of color images is very limited.

Method We develop a depth guided feature extraction module (DFE), which includes depth guided feature selection module (DFS) and depth embedded deformable convolution module (DDC). First, the proposed depth guided feature selection module concatenates the input of depth features and visual features in order to ignore the interference on the network derived from all depth information, and then selects the features with important influence from the fusion features through the channel attention method. Next, the weight matrix of the depth features is obtained through 1×1 convolution and sigmoid function. After multiplying the depth features and the corresponding weight matrix, the depth information to be embedded in the visual features is obtained. This depth information is then added to the visual features. Since the weight matrix corresponding to the depth features is obtained by learning, the network can adjust the number of depth information adequately, rather than accepting all the depth information. For instance, the proportion of depth information will be increased if the depth information is needed for classification. Otherwise, the proportion of depth information will be decreased. In order to promote the feature extraction capability of deformable convolution completely, the depth embedded deformable convolution module is proposed. Depth information-embedded visual features are taken as input to learn the offset of sampling points. The addition of depth features makes up for the deficiency of geometric information of visual features.

Result In order to verify the effectiveness of the method, a series of ablation experiments are carried out on the New York University Depth Dataset V2 (NYUv2) compared to the current methods. Mean intersection over union (mIoU) and mean pixel accuracy (mPA) are used as the measurement criteria. Our method achieved 51.9% of mIoU and 77.6% of PA on NYUv2, respectively. The visualization results of semantic segmentation are demonstrated to prove the effectiveness of the method.

Conclusion We facilitates the depth guided feature extraction module (DFE), which includes depth guided feature selection module (DFS) and depth embedded deformable convolution module (DDC). DFS can adaptively determine the proportion of depth information in terms of the input of visual features and depth features. DDC enhances the feature extraction capability of deformable convolution through embedding depth information, and can extract more effective features via the shape of objects. In addition, our designed module can be embedded into the current feature extraction net-

work. The depth information can be used to improve the modeling ability of the network effectively.

Key words: semantic segmentation; RGB-D; depth guided feature selection (DFS); depth embedded deformable convolution (DDC); depth guided feature extraction (DFE)

0 引言

语义分割是计算机视觉领域中的一项基本任务, 目标是将图像中的每一个像素分配给对应的类别, 是像素级别的多分类任务。在自动驾驶、虚拟现实和医学图像处理等领域具有重要意义。卷积神经网络促使神经网络迅速发展, 在计算机视觉各项任务中均取得最好效果 (Krizhevsky 等, 2012; Simonyan 和 Zisserman, 2015; He 等, 2016)。而全卷积神经网络 (Long 等, 2015) 彻底改变了语义分割等领域的格局, 依靠编码器—解码器的结构, 降低了对图像尺寸的依赖, 无论输入编码器图像尺寸如何, 都可在解码器输出部分得到相同大小的分割图像。一系列基于全卷积神经网络的语义分割研究 (Chen 等, 2017, 2018a; Lin 等, 2016; Liu 等, 2015) 极大地推动了语义分割的发展。

由于深度相机的出现 (Zhang, 2012), 可以比较方便地获取彩色图像对应的深度图像。深度图是单通道图像, 图像中的每个值对应该像素点到相机平面的距离。显而易见, 深度图含有彩色图像相对匮乏的空间深度信息。在语义分割任务中, 彩色图像中外观相似的相邻物体是网络难以区分的, 但深度图像的应用可以在一定程度上缓解这个问题。尽管外观特征相似, 但其在 3D 空间中相距甚远, 这可以帮助网络进行分类。Gupta 等人 (2014) 将深度图转化为三通道的 HHA (horizontal disparity, height above ground and angle with gravity) 图像, 丰富了深度图像信息。

近年有不少关于 RGB-D 语义分割的研究 (Jiang 等, 2018; Xing 等, 2020), 但如何更有效地提取深度信息并嵌入到视觉特征仍是一个值得探索的问题。彩色图像由于其明亮的颜色和清晰的纹理, 蕴含了良好的语义信息。而与彩色图像对齐的深度图像存在较为准确的深度数据, 含有区别于彩色图像的空间信息。两种图像既有差异性也有互补性, 如果简单地将全部深度信息和视觉特征组合在一起, 可能会对网络产生干扰。根据将深度信息嵌入视觉特征

的方式, 大部分研究方法可以大致分为以下 3 类: 单流 (one-stream) 方法 (Wang 和 Neumann, 2018; Xing 等, 2019)、双流 (two-stream) 方法 (Cheng 等, 2017; Lee 等, 2017) 以及多任务 (multi-task) 方法 (Xu 等, 2018; Zhang 等, 2019)。单流方法不会将深度图作为额外输入而对其提取特征, 只有一个对彩色图像进行特征提取的主干 (backbone) 网络, 特征提取过程中利用深度图固有空间信息辅助视觉特征提取, 以达到提高语义分割的效果; 或是采用多个卷积核沿深度方向对输入图像提取特征, 每个卷积核处理不同深度的像素点, 最后将多个卷积核的计算结果相加。双流方法将深度图作为额外输入而对其提取特征, 主要存在两个主干网络, 每个主干网络分别对彩色图像和深度图像提取特征, 在编码或解码阶段将提取的视觉特征与深度特征融合, 达到利用深度信息的目的。多任务方式与上述两种方式显著不同, 将语义分割、深度估计以及表面法线估计等多种任务一同处理, 这类方式往往只有一个共用的主干网络, 在对彩色图像进行特征提取过程中, 根据不同任务的监督可以得到多个任务相似的特征以及互补的特征, 不同任务间特征的交互会提高各个任务的性能。除此之外, Song 等人 (2017) 和 Zhong 等人 (2018) 利用深度信息将彩色图像映射到 3D 空间中, 使用 3D 卷积对图像进行特征提取, 然而这种方法得到的 3D 数据是比较稀疏的, 对于计算资源和存储空间要求较高, 从而限制了它的应用。Qi 等人 (2017) 在 3D 点云的基础上构建了 K-近邻图 (K-nearest neighbor graph), 通过循环迭代方式更新节点的特征表示。

以上研究都忽略了同一个问题, 即并非所有的深度信息都是必要的, 将全部深度信息嵌入视觉特征可能会对网络造成干扰。彩色图像固有的颜色和纹理信息有时完全可以清楚地区分两个或多个类别, 此时深度信息的加入多少有些画蛇添足的味道。例如, 深度特征相似但视觉特征不同的物体本可以由视觉特征区分, 此时加入深度信息反而会使网络模型做出错误判断。如图 1 所示, 图 1(a) 中的毛巾 (橙色框中物体) 仅通过视觉特征, 即仅将彩色图像

作为输入,网络便可清晰分辨,如图 1(b)橙色框中的分割结果。引入全部深度信息时,网络的分割结果反而变差,如图 1(c)橙色框中的分割结果。引入部分深度信息时,网络的分割结果又变好,如图 1(e)橙色框中的分割结果。此外,图 1(a)中的镜子(蓝色框中物体)和插座(绿色框中物体)仅使用彩色图像,分割结果如图 1(b)中的蓝色框与绿色框;或将全部深度信息嵌入视觉特征,分割结果如图 1(c)中的蓝色框与绿色框,都没有取得理想效果,仅使用视觉特征缺少空间深度信息,而引入全部深度信息可能会对网络产生干扰,但合理利用深度信息却取得了良好结果,如图 1(e)中蓝色框与绿色框中的分割结果。在网络不需要深度信息时不引入深度信息或者引入极少的深度信息,不会对模型造成额外干扰;在网络需要深度信息时,引入较多的深度信息。

而且,卷积神经网络中卷积核的固有结构限制了其特征提取的能力。Dai 等人(2017)和 Zhu 等人(2019)针对这个问题提出了可变形卷积,可以根据输入学习卷积点的位置偏移矩阵,扩大卷积的感受野,依据物体形状提取相关特征,从而提高模型的建模能力。但仅依靠视觉特征输入学习偏移矩阵略有

不足,因为彩色图像具有的深度信息是十分有限的。基于以上问题,本文提出了深度信息引导的特征提取模块(depth guided feature extraction, DFE),其中包括深度信息引导的特征选择模块(depth guided feature selection, DFS)和深度信息嵌入的可变形卷积模块(depth embedded deformable convolution, DDC)。为了避免引入全部深度信息对网络造成干扰,提出的 DFS 模块将视觉特征和深度特征连接,通过通道注意力模块(channel attention, CA)使网络关注图像中希望被关注的部分,筛选出对输出具有重要影响的视觉特征和深度特征。对于筛选出来的视觉特征和深度特征,采用深度信息选择模块(depth selection, DS)得到深度特征对应的权重矩阵,这样网络可以根据任务需要自适应地调整深度信息引入的程度,筛选出具有关键深度信息的多模特征。例如,需要深度信息辅助网络进行分类时,深度信息对应的比例变大。反之,深度信息占的比例减小。为了更好地提高模型的建模能力,提出的 DDC 对筛选出来含有关键深度信息的多模特征进行卷积运算,学习采样点的位置偏移矩阵,根据物体形状提取更相关的特征。

本文主要贡献为:1)提出深度信息引导的特征

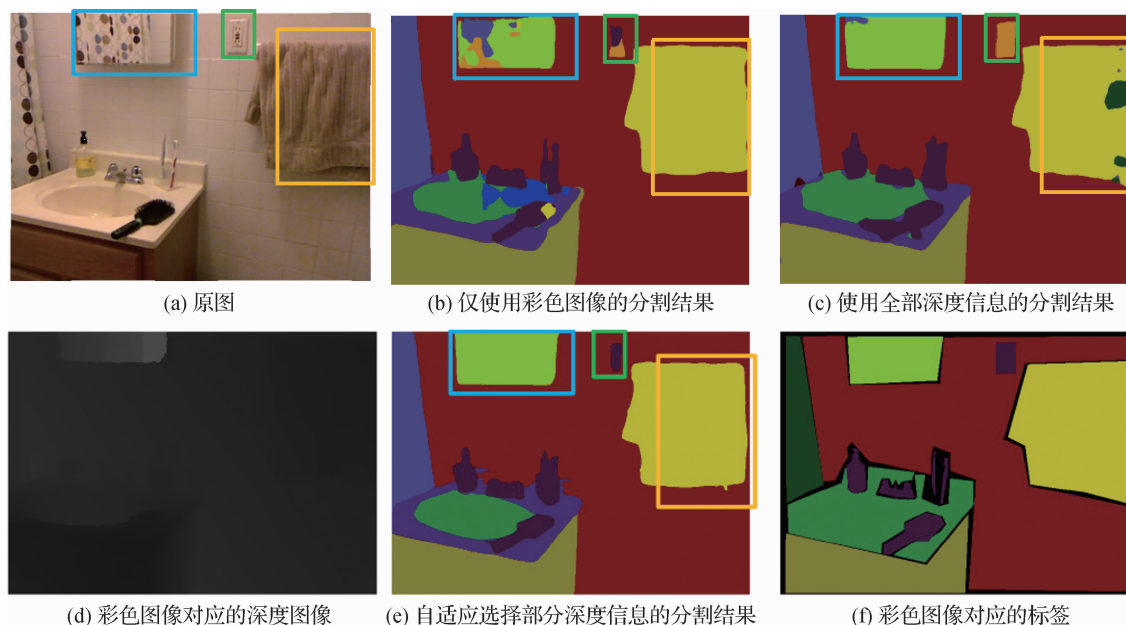


图 1 彩色图像、深度图像、标签以及不同方法的分割结果

Fig. 1 Color image, depth image, label and segmentation results of different methods

((a) original image; (b) segmentation result with color image only; (c) segmentation result with all depth information; (d) depth image; (e) segmentation result with partial depth information selected by network adaptively; (f) label image)

选择模块 DFS, 可根据视觉特征和深度特征的输入自适应地决定深度信息加入比例, 能够在一定程度上避免全部深度信息引入对网络造成的干扰; 2) 提出深度信息嵌入的可变形卷积模块 DDC, 深度图固有的空间深度信息可以帮助模型学习更准确的位置偏移, 从而根据物体形状提取更相关的特征; 3) 重新思考深度信息引入对 RGB-D 模型的影响, 并在 NYUv2 (New York University Depth Dataset V2) 数据集上达到最好的分割效果; 4) 设计的模块 DFS、DDC 以及 DFE 可以比较方便地嵌入目前流行的特征提取网络, 利用深度信息提高模型的建模能力。

1 相关研究

随着卷积神经网络的发展, 计算机视觉领域不断取得新的进展 (Simonyan 和 Zisserman, 2015; He 等, 2016; Szegedy 等, 2015)。对 2D 图像的语义分割, 基于全卷积神经网络 (fully convolutional networks, FCN) (Chen 等, 2018a; Zhao 等, 2017) 和编码器—解码器 (encoder-decoder) 模型的持续涌现 (Badrinarayanan 等, 2015; Ronneberger 等, 2015; Lin 等, 2017b), 极大地推动了语义分割技术进步。注意力机制 (Vaswani 等, 2017) 和可变形卷积 (deformable convolution, DC) 也进入科研人员视野。深度相机的出现使深度图获取相对容易, 人们很自然地将彩色图像的语义分割扩展到 RGB-D 图像的语义分割。研究表明, 深度图像的引入可以提高图像分割的效果, RGB-D 语义分割技术也在蓬勃发展。

1.1 RGB-D 语义分割

深度图像的加入使彩色图像的语义分割扩展到 RGB-D 图像的语义分割。早期 RGB-D 图像语义分割多是将深度图作为单独输入, 对深度图像和彩色图像分别提取特征, 在适当位置将两种模态的特征融合, 实现将深度信息嵌入视觉特征的目的。为了更好地将深度信息嵌入视觉特征, 自 RefineNet (Lin 等, 2017b) 扩展而来的 RDFNet (RGB-D fusion network) (Lee 等, 2017) 提出多模特征融合模块 (multi-modal feature fusion, MMF) 融合相同阶段的深度特征和视觉特征, 采用逐层优化的方法将不同阶段的多模特征融合, 显著提高了语义分割精度。Wang 等人 (2016) 在编码器和解码器之间提出一种特征转换网络, 分别挖掘视觉特征和深度特征之间的相似

特征以及独有特征, 通过相似特征之间的交互增强各自的特征表示。由反卷积网络扩展而来的 LSD-Net (locality sensitive deconvolution network) (Cheng 等, 2017) 在反卷积过程中利用深度信息改善物体边缘信息, 采用门控结构融合深度图像和彩色图像的分割结果。CFNet (cascaded feature network) (Lin 等, 2017a) 采用上下文感知的感受野模块提取不同深度的上下文信息, 采用级联方式将不同深度的上下文信息进行融合来改善语义分割效果。Wang 和 Neumann (2018) 基于同一物体具有相似深度的思想, 利用卷积点与周围像素点之间距离的差异, 显式地改变周围点对中心卷积点的贡献程度, 距离近的像素点贡献大, 距离远的像素点贡献小, 从而实现 RGB-D 的语义分割。2.5D 卷积网络 (Xing 等, 2019) 利用深度信息将彩色图像上的像素点按照深度划分至不同区域, 采用多个卷积核分别处理不同深度的像素点, 最后将多个卷积核处理的结果相加。3DGNN (3D graph neural network) (Qi 等, 2017) 采用 3D 图神经网络以点云为基础创建 K-近邻图, 通过循环迭代方式更新节点的特征表示。Zhang 等人 (2019) 将语义分割、深度估计以及表面法线估计等多种任务一同处理, 这些任务具有相似的特征以及互补的特征, 不同任务间特征的交互会提高各任务的性能。

以上 RGB-D 语义分割的研究均未考虑深度信息引入是否会对网络模型产生干扰。本文从一个新的角度思考深度信息的作用, 在合理利用深度信息的同时, 筛选关键的深度信息嵌入视觉特征, 从而改善语义分割效果。

1.2 注意力机制

源于人类视觉的注意力机制已广泛应用于计算机视觉领域的各种任务, 作为一种有效工具, 可以突出图像中含有重要信息部分, 即可以帮助网络关注到人类感兴趣的区域。SENet (squeeze-and-excitation network) (Hu 等, 2018) 中的通道注意力 (channel attention) 机制沿通道方向进行均值池化, 可以提取关键的全局信息, 这个全局信息可以作为权重对输入特征进行筛选, 即突出图像中人类感兴趣的部分。DFN (discriminative feature network) (Yu 等, 2018) 中的通道注意力模块 (channel attention block, CAB) 将高层特征与相邻的低层特征融合, 利用高层特征丰富的语义信息作为引导, 筛选出具有判别性的特征

对低层的特征进行逐层优化,提升了语义分割精度。本文采用 SENet 中的通道注意力机制,利用深度信息筛选关键特征。

1.3 可变形卷积

图像中物体形状和尺寸千差万别,卷积核只在固定位置采样,忽略了物体的空间位置信息,其固有的几何结构限制了卷积神经网络的建模能力。为了解决上述问题,Dai 等人(2017)提出可变形卷积,在标准 2D 卷积基础上增加一个新的卷积,用来学习采样点的位置偏移,此时可变形卷积可以通过输入视觉特征自适应地调整采样点位置,增强卷积神经网络提取特征的能力。Zhu 等人(2019)在 Dai 等人(2017)基础上进一步改进,由于采样点位置偏移是通过视觉特征卷积学习到的,所有新的采样点可能会超出感兴趣区域。为了解决这个问题,Zhu 等人(2019)采用调制机制,通过对输入视觉特征进行卷积,学习一个与采样位置相关的权重矩阵,超出感兴趣区域的采样位置会得到一个较小权重,可以在一定程度上缓解采样点超出相关区域问题。相对于彩色图像,深度图像含有丰富的空间深度信息,提出的深度信息嵌入的可变形卷积模块 DDC,在关键深度信息嵌入下,可以依据物体形状提取到更相关的特征。

SAG(separation-and-aggregation gate)(Chen 等, 2020)和 ACNet(attention complementary network)(Hu 等, 2019)都对特征进行了选择。但是,SAG 同时调整深度特征和视觉特征加入网络的程度,导致深度特征可能会占据主导地位,而深度特征缺少语义信息,仅依靠深度特征分辨不同物体对于网络是十分困难的。ACNet 仅依靠通道注意力的方法进行特征选择,略有不足,也没有考虑到深度特征语义信息不足的缺点。本文方法则以视觉特征为主,深度信息为辅,仅在模型需要时自适应地添加深度信息,即仅选择关键的深度特征。

2 方 法

彩色图像具有丰富的颜色和纹理信息,而与之对齐的深度图像蕴含较为丰富的空间深度信息,深度信息的引入可以在一定程度上提高彩色图像语义分割的效果。然而,如何将深度信息嵌入视觉特征

仍然是一个值得探索的问题。当前绝大多数的研究都在尝试找到一种最好的利用深度信息的方式来提高彩色图像语义分割的精度,但并没有考虑深度信息的引入是否会对网络产生干扰,有时仅依靠视觉特征网络就能区分不同的物体,在引入深度信息后反而会使网络模型作出错误判断。本文从这个角度出发,重新考虑深度信息的利用方式,提出了深度信息引导的特征提取模块 DFE,包括深度信息引导的特征选择模块 DFS 和深度信息嵌入的可变形卷积模块 DDC。DFS 首先将视觉特征与深度特征连接得到多模特征,然后通过注意力机制筛选对输出具有重要影响的多模特征,通过对这部分多模特征的学习,得到深度特征对应的权重矩阵,自适应地决定深度信息引入的程度,之后将输入视觉特征与筛选出的深度特征相加,得到含有关键深度信息的多模特征。DDC 对嵌入关键深度信息的多模特征进行学习,得到卷积采样点的位置偏移矩阵,从而确定新的采样点位置。由于深度信息的引入,可变形卷积可以在一定程度上克服彩色图像空间深度信息不足的缺点,根据物体的形状提取更相关的特征。

2.1 网络整体结构

图 2 为网络整体结构。本文采用双流方法分别对彩色图像和深度图像提取特征。提出的 DFE 模块包括 DFS 模块和 DDC 模块。DFS 可以筛选对输出具有重要影响的深度特征,并自适应地调整深度特征嵌入视觉特征的程度,得到含有关键深度信息的多模特征;DDC 结合含有关键深度信息的多模特征以及输入视觉特征,可以根据物体形状学习更相关的特征。网络编码器的输出通过解码器产生与原始图像尺寸相同的分割图像,本文未特别设计解码器结构,采用基准分割网络 deeplabV3+ 的解码器单元作为本文网络的解码器部分。

2.2 编码器

本文采用 ResNet-50(He 等, 2016)作为提取特征主干网络,在 ResNet 相邻模块间添加 DFE 结构,通过深度信息引导逐层提取重要的特征表示。

2.2.1 深度信息引导的特征提取模块 DFE

本文从一个新的辩证角度看待深度信息的作用,将全部深度信息引入神经网络有时会使网络做出错误判断,为了尽可能避免这种情况,合理利用深度信息,提出了 DFE 模块,如图 3 所示。深度信息引

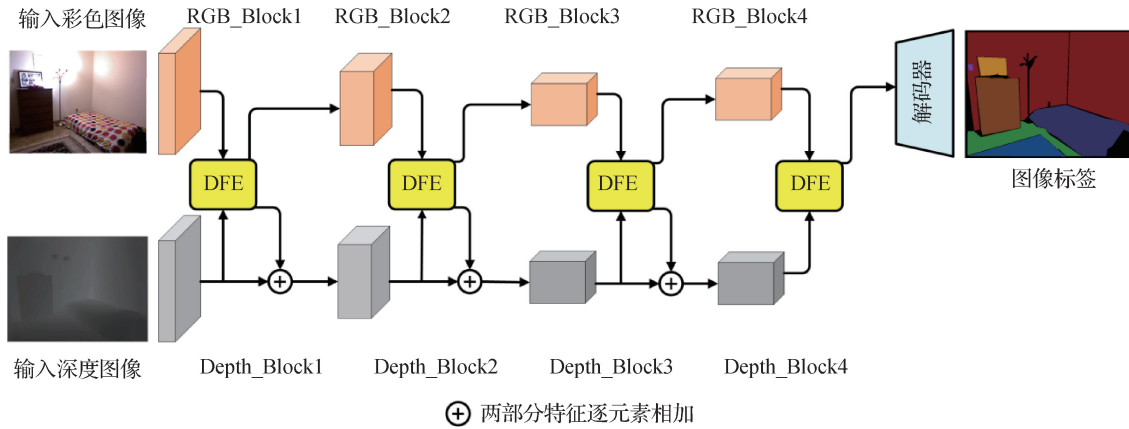


图 2 网络整体结构

Fig. 2 The overall structure of the network

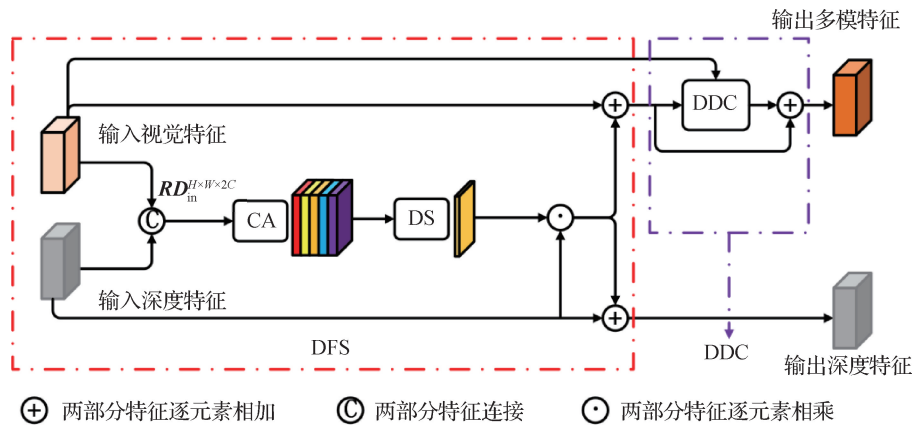


图 3 网络的 DFE 模块

Fig. 3 DFE module of the network

导的特征提取模块首先将深度特征与视觉特征连接,然后通过通道注意力方法从融合特征中筛选出具有重要影响的特征。得到的关键特征通过深度信息选择模块(DS)产生对应于深度特征的权重矩阵。深度特征与对应的权重矩阵相乘之后再与视觉特征相加,从而将深度信息嵌入视觉特征。由于深度特征对应的权重矩阵是通过学习得到的,网络可以根据需要自适应地调整深度信息引入的多少,而不是对深度信息照单全收。为了更好地发挥可变形卷积的特征提取能力,本文提出的深度信息嵌入的可变形卷积模块,将嵌入深度信息的多模特征作为输入,学习采样点的位置偏移,深度特征的加入弥补了视觉特征空间深度信息不足的缺点。

2.2.2 深度信息引导的特征选择模块 DFS

注意力机制可使网络关注需要的部分,筛选出

对输出有重要影响的特征。深度图像含有丰富的空间深度信息,然而需要从中挑选真正起作用的那部分信息,而不是全部引入。为了方便说明,本文用 $R_{in}^{H \times W \times C}$ 表示输入视觉特征, $D_{in}^{H \times W \times C}$ 表示输入深度特征。DFS 模块先将 $R_{in}^{H \times W \times C}$ 与 $D_{in}^{H \times W \times C}$ 连接在一起,得到 $RD_{in}^{H \times W \times 2C}$ 。即

$$RD_{in}^{H \times W \times 2C} = R_{in}^{H \times W \times C} \parallel D_{in}^{H \times W \times C} \quad (1)$$

式中, $RD_{in}^{H \times W \times 2C}$ 表示 RGB-D 特征,即多模特征, $\parallel$ 表示将 $R_{in}^{H \times W \times C}$ 与 $D_{in}^{H \times W \times C}$ 连接。

然后,使用通道注意力方法筛选出关键的 RGB-D 特征 $RD_{out}^{H \times W \times 2C}$ 。具体方法是,对连接后的 RGB-D 特征 $RD_{in}^{H \times W \times 2C}$ 沿通道方向进行全局均值池化,经过第 1 个全连接函数、ReLU 激活函数、第 2 个全连接函数和 sigmoid 函数,得到一个权重向量 $V^{1 \times 1 \times 2C}$ 。再将得到的权重向量 $V^{1 \times 1 \times 2C}$ 与多模特征 $RD_{in}^{H \times W \times 2C}$

点乘,相当于对多模特征进行了一个纠正操作,输出为 $RD_{out}^{H \times W \times 2C}$ 。即

$$RD_{out}^{H \times W \times 2C} = CA(RD_{in}^{H \times W \times 2C}) \quad (2)$$

式中, CA 表示通道注意力方法。

从多模特征中筛选出来的特征对输出有比较重要的影响,同时也影响着深度信息嵌入网络的程度。于是,通过深度信息选择模块 DS ,网络学习到了深度特征对应的权重矩阵 $\gamma^{H \times W}$ 。具体为

$$\gamma^{H \times W} = \sigma(f_2(R(f_1(f_{3 \times 3}(RD_{out}^{H \times W \times 2C})))) \quad (3)$$

式中, $f_{3 \times 3}$ 表示 3×3 的卷积函数, f_1 表示第 1 个 1×1 的卷积函数, f_2 表示第 2 个 1×1 的卷积函数, R 表示 ReLU 激活函数, σ 表示 sigmoid 函数。

将权重矩阵与相应深度特征点乘,得到筛选出的深度特征 $D_s^{H \times W \times C}$,即

$$D_s^{H \times W \times C} = D_{in}^{H \times W \times C} \cdot \gamma^{H \times W} \quad (4)$$

筛选出来的深度特征也是嵌入视觉特征的深度信息,将其再与输入视觉特征 $R_{in}^{H \times W \times C}$ 相加,得到嵌入关键深度信息的多模特征 $RD_s^{H \times W \times C}$,即

$$RD_s^{H \times W \times C} = D_s^{H \times W \times C} + R_{in}^{H \times W \times C} \quad (5)$$

DS 模块结构如图 4 所示。经过通道注意力方法的多模特征先进行 3×3 的卷积运算进一步提取特征,并对多模特征进行降维,减少运算参数量,之后经过 1×1 的卷积运算与 ReLU 激活函数得到通道数为 1 的目标矩阵,再经过 1×1 的卷积运算进行调整,调整后的目标矩阵通过 sigmoid 函数生成深度特征对应的权重矩阵,这个权重矩阵对深度特征进行了显式的筛选,可以根据任务需要自适应地确定引入网络的深度信息的多少。

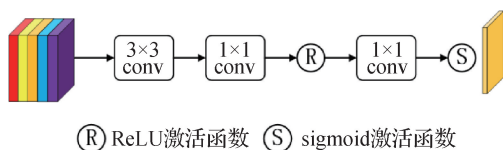


图 4 深度信息选择模块 DS

Fig. 4 Depth information selection module

2.2.3 深度信息嵌入的可变形卷积模块 DDC

为了改变卷积核固有的几何结构,提升卷积神经网络建模能力,可变形卷积通过对输入视觉特征进行卷积运算学习采样点的位置偏移,使采样点不再局限于固定尺寸的方格中,而是根据物体形状偏移。然而仅依靠视觉特征学习采样点的位置偏移略有不足,毕竟彩色图像含有的空间深度信息十分有

限。这时需要发挥深度图像的作用,将深度信息引入视觉特征,通过含有关键深度信息的多模特征来学习更加符合物体形状的位置偏移。

本文采用与可变形卷积(Dai 等,2017)类似的方式描述卷积运算,用 x 代表输入特征图, w 表示卷积核的权重, y 表示卷积的输出。卷积核采样点的感受野与卷积核的尺寸相关,假如采用 3×3 的卷积核,那么可以用集合 $A = \{(-1,1), (0,1), (1,1), (-1,0), (0,0), (0,1), (-1,-1), (0,-1), (1,-1)\}$ 表示卷积核的感受野,这也是卷积核采样点的位置坐标,中心坐标 $(0,0)$ 为卷积核采样的中心点 p_i 。标准卷积运算具体计算为

$$y(p_i) = \sum_{p_j \in A} w(p_j) \cdot x(p_j + p_i) \quad (6)$$

式中, $p_j \in A$ 为卷积核感受野中的点。

深度信息嵌入的可变形卷积同样包含两个卷积运算,第 1 个卷积运算对含有关键深度信息的输入多模特征进行卷积,得到第 2 个卷积运算的位置偏移矩阵。第 2 个卷积运算对输入视觉特征进行卷积,将初始采样点的坐标与第 1 个卷积运算得到的位置偏移坐标相加得到新的采样点的坐标,根据卷积核的权重对新的采样点完成加权求和。即

$$Off^{H \times W \times 2C} = \varphi_1(RD_s^{H \times W \times C}) \quad (7)$$

式中, φ_1 表示第 1 个卷积运算, $RD_s^{H \times W \times C}$ 表示含有关键深度信息的输入多模特征,由式(5)得到。 $Off^{H \times W \times 2C}$ 为对应第 2 个卷积运算的位置偏移矩阵,通道是 $2C$ 意味着存在水平方向和竖直方向的偏移。

第 2 个卷积运算的输出 $DDC_{out}^{H \times W \times C}$ 为

$$DDC_{out}^{H \times W \times C} = \varphi_2(R_{in}^{H \times W \times C}, Off^{H \times W \times 2C}) \quad (8)$$

式中, φ_2 表示第 2 个卷积运算, $R_{in}^{H \times W \times C}$ 表示输入视觉特征。第 2 个卷积运算的具体形式为

$$y(p_i) = \sum_{p_j \in A} w(p_j) \cdot x(p_j + p_i + \Delta p d_j) \quad (9)$$

式中, $p_j \in A$ 为感受野中其余采样点的坐标, $(p_j + p_i + \Delta p d_j)$ 为新的采样点的坐标; $\Delta p d_j$ 为采样点的位置偏移,来自 $Off^{H \times W \times 2C}$ 对应的栅格,由第 1 个卷积运算得到,即 $\Delta p d_j \in Off^{H \times W \times 2C}$,如图 5 所示。

DDC 与可变形卷积相比,主要区别在于产生采样点的位置偏移矩阵不同。对于第 1 个卷积运算,可变形卷积产生位置偏移矩阵的输入是视觉特征 $RGB_{in}^{H \times W \times C}$,而 DDC 产生位置偏移矩阵的输入是含

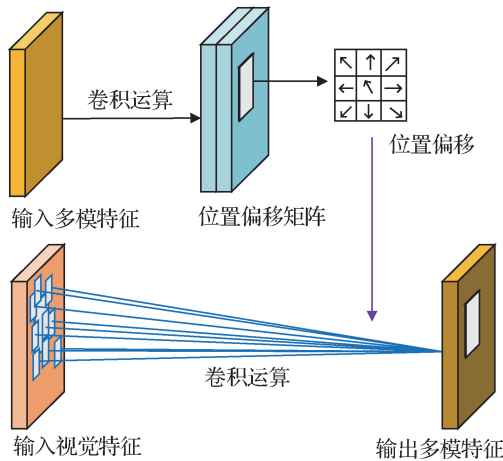


图 5 深度信息嵌入的可变形卷积 DDC

Fig. 5 Depth embedded deformable convolution

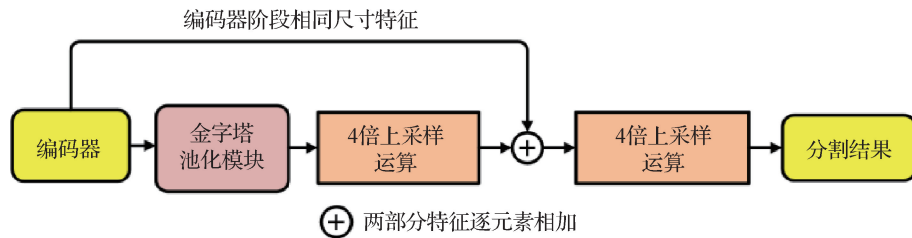


图 6 解码器网络结构

Fig. 6 Decoder network structure

解码器将编码器得到的语义特征通过金字塔池化模块得到多尺度的语义特征,经过 1×1 的卷积运算进行调整,接着通过 4 倍上采样运算得到中间特征,中间特征与编码器阶段 DFE 模块中相同尺寸的多模特征融合,融合后的特征经过 3×3 的卷积运算进一步提取特征,之后再行 4 倍上采样运算得到分割结果(Chen 等,2018b)。

3 实验

实验在 NYUv2 数据集(Silberman 等,2012)上进行,包括消融实验以及与其他方法的比较。

3.1 数据集

NYUv2 数据集是目前流行的 RGB-D 数据集之一,含有 1 449 幅像素级别标注的彩色图像以及与之对齐的深度图像,其中包括 795 幅训练集图像和 694 幅测试集图像。数据集图像均为常见的室内场景,如客厅、卧室和浴室等。本文使用的是适用于语义分割的 40 类标注的 NYUv2 数据集。

有关键深度信息的输入多模特征 $RD_s^{H \times W \times C}$;第 2 个卷积运算的输入相同,都是视觉特征 $R_{in}^{H \times W \times C}$ 。也就是说,与可变形卷积相比,DDC 具有两个输入,含有关键深度信息的多模特征 $RD_s^{H \times W \times C}$ 以及不含深度信息的视觉特征 $R_{in}^{H \times W \times C}$,如图 5 所示,而可变形卷积只有一个输入 $R_{in}^{H \times W \times C}$ 。

2.3 解码器

在语义分割任务中,解码器的作用主要是将分割图像的分辨率恢复到输入图像的尺寸。本文着重于利用深度信息的引导在解码器阶段提取更有效的语义特征,并没有设计专门的解码器,采用语义分割模型 DeeplabV3+(Chen 等,2018b)的解码器单元作为本文的解码器部分,结构如图 6 所示。

3.2 实验细节

实验基于 Pytorch 框架,在 4 块英伟达显卡(Titan X)上进行,采用随机梯度下降(stochastic gradient descent,SGD)的优化方法,动量(momenta)参数为 0.9,权重衰减系数(weight decay)为 0.000 1。选择 poly 学习策略,网络模型学习率随迭代次数衰减 $\left(1 - \frac{iter}{iter_{max}}\right)^{power}$, $iter$ 为当前迭代次数, $iter_{max}$ 为设置的网络最大迭代次数。输入图像尺寸为 480×640 像素,初始学习率为 0.02,运行了 400 个周期(epoch)。在数据增广方面,使用水平翻转、随机裁剪以及随机尺度变换,变换尺度因子为 $[0.5, 0.75, 1.0, 1.5, 1.75]$ 。为了衡量网络性能,采用平均像素准确率(pixel accuracy,PA)和平均交并比(mean pixel intersection over union,mIoU)作为评价指标。实验的主干网络是在 ImageNet 上预训练的 ResNet-50,为了比较的公平性,在 DeeplabV3+ 上验证本文方法的有效性。

3.3 消融实验

3.3.1 DDC 消融实验

一般意义上的可变形卷积对输入视觉特征先进

行第 1 个单独的卷积,得到对应采样点的位置偏移,使得中心采样点与周围采样点偏移至同一类别或相关类别,然后对位置偏移之后的新采样点进行第 2 个卷积,以打破卷积核固有的结构限制,提高卷积神经网络的建模能力。然而彩色图像中空间深度信息十分有限,仅依靠彩色图像学习到的位置偏移稍有不足。针对这个问题,提出了 DDC,对含有深度信息的多模特征进行卷积学习采样点的位置偏移,对新采样点完成卷积运算。

表 1 为 DDC 消融实验结果。实验 1 为实验基准,是将全部深度信息嵌入视觉特征(All)得到的分割结果。实验 2 在实验 1 基础上添加了一般意义上的可变形卷积(DC),图像分割结果为 46.2%,精度相对于基准提高了 0.2%;实验 3 在实验 1 基础上加入了 DDC,在参数数量相同情况下,分割结果为 46.6%,相对于基准提高了 0.6%,相对实验 2 提高了 0.4%,充分说明了 DDC 的有效性。在同时加入 DFS 情况下,DDC 的性能依旧高于一般意义上的可变形卷积,如实验 4 和实验 5,实验 5 相对于实验 4 提高了 0.4%。实验结果表明,在深度信息引导下,网络学习到了更好的位置偏移。

表 1 DDC 消融实验结果

Table 1 The results of DDC ablation study

实验	网络模型	mIoU/%
实验 1	DeeplabV3 + + All	46.0
实验 2	DeeplabV3 + + All + DC	46.2
实验 3	DeeplabV3 + + All + DDC	46.6
实验 4	DeeplabV3 + + DFS + DC	51.5
实验 5	DeeplabV3 + + DFS + DDC	51.9

注:加粗字体表示最优结果。

3.3.2 DFS 消融实验

深度图作为彩色图像的补充,提供了彩色图像相对匮乏的空间深度信息,在一定程度上提高了语义分割效果。然而不加考虑地将全部深度信息嵌入视觉特征也可能对网络产生干扰。在模型仅依靠视觉特征网络即可区分两类或多类物体时,加入的深度信息反而会使模型做出错误判断。对此,本文提出了 DFS,在深度信息的引导下,自适应地调整深度信息嵌入视觉特征的程度,在网络不需要深度信息

时,不引入或仅引入极少的深度信息,避免对网络产生不必要干扰;而在网络需要深度信息时,将关键的深度信息融入视觉特征。

为验证 DFS 的有效性,进行了一系列实验,结果如表 2 所示。实验 1 为实验基准,即将全部深度信息嵌入视觉特征(All)得到的分割结果。实验 2 为添加 DFS 的分割结果,相对于实验 1 提高了 5.4%。充分说明了 DFS 的有效性。在同时使用 DDC 情况下,实验 4 的分割结果为 51.9%,相对于实验 3 的 46.6%,提高了 5.3%。实验结果表明,合理利用深度信息可以在较好地提高网络性能的同时减少对网络的干扰。

表 2 DFS 消融实验结果

Table 2 The results of DFS ablation study

实验	网络模型	mIoU/%
实验 1	DeeplabV3 + + All	46.0
实验 2	DeeplabV3 + + DFS	51.4
实验 3	DeeplabV3 + + All + DDC	46.6
实验 4	DeeplabV3 + + DFS + DDC	51.9

注:加粗字体表示最优结果。

3.3.3 DFE 消融实验

表 3 为 DFE 消融实验结果。其中,DFE- n 表示第 n 个 DFE 模块。可以看出,在不使用 DFE 模块情况下分割结果为 46.0%,即对比基准。在仅使用第 1 个 DFE 模块情况下分割结果为 48.8%,相对于基准提高了 2.8%,在加入第 2 个 DFE 模块后分割结果提升至 49.3%,相对于只有 1 个 DFE 模块情况提高了 0.5%;使用 3 个和 4 个 DFE 模块的分割结

表 3 DFE 消融实验结果

Table 3 The results of DFE ablation study

网络模型	mIoU/%
DeeplabV3 +	46.0
DeeplabV3 + + DFE-1	48.8
DeeplabV3 + + DFE-1 + DFE-2	49.3
DeeplabV3 + + DFE-1 + DFE-2 + DFE-2	50.8
DeeplabV3 + + DFE-1 + DFE-2 + DFE-2 + DFE-2	51.9

注:加粗字体表示最优结果。

果分别为 50.8% 和 51.9%, 相对于基准分别提高了 4.8% 和 5.9%, 充分说明了 DFE 模块的有效性。

综合前面的实验, 形成表 4。通过表 4, 可以清楚地观察到本文方法的有效性。

表 4 DFE 综合实验结果

Table 4 The results of DFE comprehensive study

网络模型	mIoU/%
DeeplabV3 +	44.3
DeeplabV3 + + All	46.0
DeeplabV3 + + All + DDC	46.6
DeeplabV3 + + DFS	51.4
DeeplabV3 + + DFE	51.9

注: 加粗字体表示最优结果。

3.4 与其他方法的对比

为进一步验证本文方法有效性, 与其他方法进行对比。为了比较的公平性, 实验均采用 ResNet-50 作为主干网络。实验结果如表 5 所示。

表 5 不同方法的 mIoU 和 PA 对比

Table 5 Comparison of results of mIoU and PA among different methods

网络模型	/%	
	mIoU	PA
3DGNN(Qi 等, 2017)	43.1	—
CFN(Lin 等, 2017a)	44.5	—
RDF-101(Lee 等, 2017)	49.1	75.6
PADNet(Xu 等, 2018)	50.2	75.2
ACNet(Hu 等, 2019) [#]	48.3	—
PAP(Zhang 等, 2019)	50.4	76.2
2.5D(Xing 等, 2019)	49.1	75.9
Malleable 2.5D(Xing 等, 2020)	49.7	76.3
SAG(Chen 等, 2020)	51.3	—
VCD + RedNet(Xiong 等, 2020)	50.7	—
VCD + ACNet(Xiong 等, 2020)(3 个主干网络)	51.9	—
本文(ResNet-50)	51.9	77.6
本文(ResNet-101)	52.6	77.9

注: 加粗字体表示最优结果, “#”表示使用了 3 个主干网络, “—”表示原论文未提供该数据。

本文方法思路与 SAG 和 ACNet 两种方法相似, 与这两种方法相比, ACNet、SAG、DFS 和 DFE 方法的分割精度分别为 48.3%、51.3%、51.4% 和 51.9%, 本文方法取得了更优异的分割结果。且 ACNet、SAG、DFS 和 DFE 方法的参数量分别为 934.0 M、888.2 M、706.2 M 和 1 098.0 M, 相比于 ACNet 以及 SAG、DFS 方法使用较少参数实现了较好的分割效果。DFE 方法在结合 DDC 模块后参数量有一定增加, 但实现了最好的分割结果。VCD + ACNet 与本文分割结果相同, 然而 VCD + ACNet 使用了 3 个主干网络进行实验, 本文仅使用 2 个主干网络, 本文使用更少参数达到了相同的分割效果。使用 ResNet-101 作为主干网络时, 本文方法分割精度可以达到 52.6%。

3.5 图像语义分割结果的可视化

图 7 为图像语义分割结果的可视化。可以看出, 引入全部深度信息后可能会对分割结果产生负面影响(图 7(b)(c))。而在加入 DFS 模块和 DFE 模块后, 网络能够比较合理地利用深度信息, 在提升网络分割性能的同时, 尽量避免对网络造成不利影响(图 7(d)(e))。

4 结 论

本文重新思考了深度信息的利用方式。在视觉特征足够区分不同物体情况下, 深度信息的加入反而可能会使模型产生错误的分类结果。为了避免这种情况, 提出一种深度信息引导的特征提取模块 DFE。DFE 包括特征选择模块 DFS 和可变形卷积模块 DDC。DFS 模块在特征提取过程中可以自适应地调整深度信息引入的多少。DDC 模块通过深度信息的嵌入, 增强可变形卷积提取特征能力, 将嵌入关键深度信息的多模特征作为输入, 学习卷积核采样的位置偏移。由于深度图固有的空间深度信息, 弥补了彩色图像空间深度信息不足的缺点, 可以提取到较好的语义特征。与其他方法在 NYUv2 数据集上的比较中, 本文方法取得了最佳效果。此外, DFE 模块可以比较方便地嵌入当下流行的特征提取网络中。

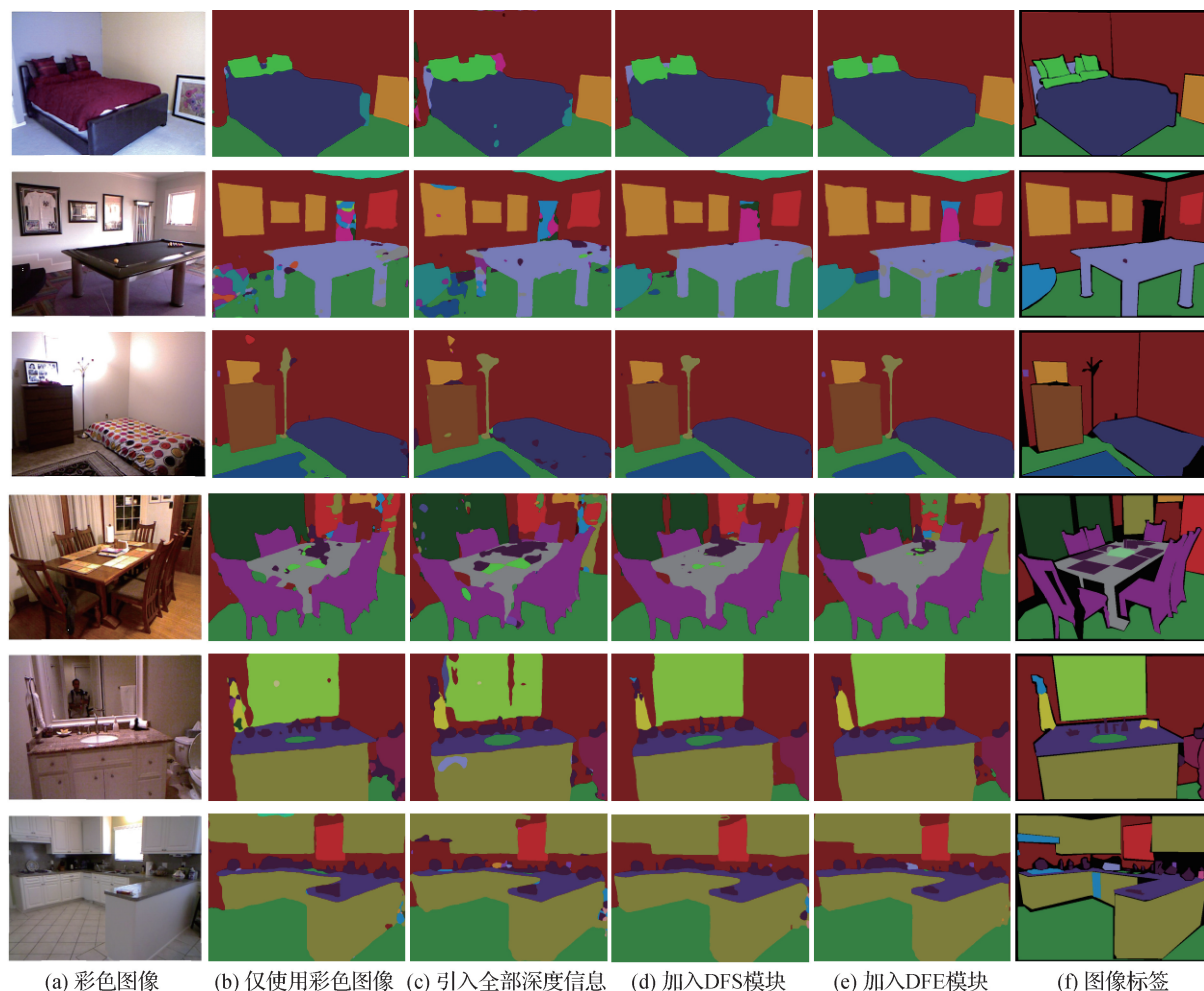


图7 图像分割结果的可视化

Fig. 7 Visualization of image segmentation results((a) color images;(b) results with color images only;
(c) results with all depth information;(d) results with DFS;(e) results with DFE;(f) label images)

参考文献 (References)

- Badrinarayanan V, Handa A and Cipolla R. 2015. SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labeling [EB/OL]. [2019-02-05]. <https://arxiv.org/pdf/1505.07293.pdf>
- Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018a. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]
- Chen L C, Papandreou G, Schroff F and Adam H. 2017. Rethinking atrous convolution for semantic image segmentation [EB/OL]. [2019-02-05]. <https://arxiv.org/pdf/1706.05587.pdf>
- Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018b. Encoder-decoder with atrous separable convolution for semantic

image segmentation [EB/OL]. [2021-01-27]. <https://arxiv.org/pdf/1802.02611v1.pdf>

- Chen X K, Lin K Y, Wang J B, Wu W, Qian C, Li H S and Zeng G. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 561-577 [DOI: 10.1007/978-3-030-58621-8_33]
- Cheng Y H, Cai R, Li Z W, Zhao X and Huang K Q. 2017. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1475-1483 [DOI: 10.1109/CVPR.2017.161]
- Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]
- Gupta S, Girshick R, Arbeláez P and Malik J. 2014. Learning rich fea-

- tures from RGB-D images for object detection and segmentation//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland; Springer; 345-360 [DOI: 10.1007/978-3-319-10584-0_23]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE; 770-778 [DOI: 10.1109/CVPR.2016.90]
- Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE; 7132-7141 [DOI: 10.1109/CVPR.2018.00745]
- Hu X X, Yang K L, Fei L and Wang K W. 2019. ACNET: attention based network to exploit complementary features for RGBD semantic segmentation//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China; IEEE; 1440-1444 [DOI: 10.1109/ICIP.2019.8803025]
- Jiang J D, Zheng L, Luo F and Zhang Z J. 2018. RedNet: residual encoder-decoder network for indoor RGB-D semantic segmentation [EB/OL]. [2021-01-27]. <https://arxiv.org/pdf/1806.01054v1.pdf>
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA; Curran Associates Inc.; 1097-1105
- Lee S, Park S J and Hong K S. 2017. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE; 4990-4999 [DOI: 10.1109/ICCV.2017.533]
- Lin D, Chen G Y, Cohen-Or D, Heng P A and Huang H. 2017a. Cascaded feature network for semantic segmentation of RGB-D images//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE; 1320-1328 [DOI: 10.1109/ICCV.2017.147]
- Lin G S, Milan A, Shen C H and Reid I. 2017b. RefineNet: multi-path refinement networks for high-resolution semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE; 5168-5177 [DOI: 10.1109/CVPR.2017.549]
- Lin G S, Shen C H, van den Hengel A and Reid I. 2016. Efficient piecewise training of deep structured models for semantic segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE; 3194-3203 [DOI: 10.1109/CVPR.2016.348]
- Liu Z W, Li X X, Luo P, Loy C C and Tang X O. 2015. Semantic image segmentation via deep parsing network//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile; IEEE; 1377-1385 [DOI: 10.1109/ICCV.2015.162]
- Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA; IEEE; 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Qi X J, Liao R J, Jia J Y, Fidler S and Urtasun R. 2017. 3D graph neural networks for RGBD semantic segmentation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE; 5209-5218 [DOI: 10.1109/ICCV.2017.556]
- Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany; Springer; 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Silberman N, Hoiem D, Kohli P and Fergus R. 2012. Indoor segmentation and support inference from RGBD images//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy; Springer; 746-760 [DOI: 10.1007/978-3-642-33715-4_54]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-02-05]. <https://arxiv.org/pdf/1409.1556.pdf>
- Song S R, Yu F, Zeng A, Chang A X, Savva M and Funkhouser T. 2017. Semantic scene completion from a single depth image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE; 190-198 [DOI: 10.1109/CVPR.2017.28]
- Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA; IEEE; 1-9 [DOI: 10.1109/CVPR.2015.7298594]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA; Curran Associates Inc.; 6000-6010
- Wang J H, Wang Z H, Tao D C, See S and Wang G. 2016. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer; 664-679 [DOI: 10.1007/978-3-319-46454-1_40]
- Wang W Y and Neumann U. 2018. Depth-aware CNN for RGB-D segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer International Publishing; 144-161 [DOI: 10.1007/978-3-030-01252-6_9]
- Xing Y J, Wang J B, Chen X K and Zeng G. 2019. 2.5D convolution for RGB-D semantic segmentation//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China, IEEE; 1410-1414 [DOI: 10.1109/ICIP.2019.8803757]
- Xing Y J, Wang J B and Zeng G. 2020. Malleable 2.5D convolution;

- learning receptive fields along the depth-axis for RGB-D scene parsing//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 555-571 [DOI: 10.1007/978-3-030-58529-7_33]
- Xiong Z T, Yuan Y, Guo N H and Wang Q. 2020. Variational context-deformable ConvNets for indoor scene parsing//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 3991-4001 [DOI: 10.1109/CVPR42600.2020.00405]
- Xu D, Ouyang W L, Wang X G and Sebe N. 2018. PAD-Net: multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 675-684 [DOI: 10.1109/CVPR.2018.00077]
- Yu C Q, Wang J B, Peng C, Gao C X, Yu G and Sang N. 2018. Learning a discriminative feature network for semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 1857-1866 [DOI: 10.1109/CVPR.2018.00199]
- Zhang Z Y. 2012. Microsoft Kinect sensor and its effect. IEEE Multimedia, 19(2): 4-10 [DOI: 10.1109/MMUL.2012.24]
- Zhang Z Y, Cui Z, Xu C Y, Yan Y, Sebe N and Yang J. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 4101-4110 [DOI: 10.1109/CVPR.2019.00423]
- Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA; IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]
- Zhong Y R, Dai Y C and Li H D. 2018. 3D geometry-aware semantic labeling of outdoor street scenes//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China; IEEE: 2343-2349 [DOI: 10.1109/ICPR.2018.8545378]
- Zhu X Z, Hu H, Lin S and Dai J F. 2019. Deformable ConvNets V2: more deformable, better results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 9300-9308 [DOI: 10.1109/CVPR.2019.00953]

作者简介



赵经阳, 1995年生, 男, 硕士研究生, 主要研究方向为室内图像的语义分割。

E-mail: jyzhao@hust.edu.cn



桑农, 通信作者, 男, 教授, 博士生导师, 主要研究方向为计算机视觉、模式识别。

E-mail: nsang@hust.edu.cn

余昌黔, 男, 博士研究生, 主要研究方向为场景语义分割。

E-mail: 201672484@hust.edu.cn