控制与决策
Control and Decision

基于双金字塔特征融合网络的RGB-D多类实例分割

张旭东， 王玉婷， 范之国， 付绪文

引用本文:
张旭东, 王玉婷, 范之国, 等. 基于双金字塔特征融合网络的RGB-D多类实例分割[J]. 控制与决策, 2020, 35(7): 1561-1568.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1520

文章编号: 1001-0920(2020)07-1561-08 DOI: 10.13195/j.kzyjc.2018.1520

基于双金字塔特征融合网络的RGB-D多类实例分割

张旭东†, 王玉婷, 范之国, 付绪文

(合肥工业大学 计算机与信息学院，合肥 230009)

摘 要: 针对RGB图像的实例分割任务在图像目标物体纹理相近但类别不同的区域可能出现分割错误的问题,引入Depth信息,结合RGB-D图像的三维几何结构特点,提出一种以双金字塔特征融合网络为框架的RGB-D实例分割方法. 所提出的方法通过构建两种不同复杂度的金字塔深度卷积神经网络分别提取不同梯度分辨率大小的RGB特征及Depth特征,将对应分辨率大小的两种特征相加输入区域候选网络,以此改变输入区域候选网络层的共享特征,共享特征再经过分类、回归与掩码网络分支输出定位与分类结果,从而实现RGB-D图像的实例分割. 实验结果表明,所提出的双金字塔特征融合网络模型能够完成RGB-D图像的实例分割任务,有效学习到深度图像与彩色图像之间的互补信息,与不包含Depth信息的Mask R-CNN相比,平均精度提高7.4 %.

关键词: 实例分割；RGB-D图像；金字塔网络；特征融合；区域候选

中图分类号: TP391 **文献标志码:** A

RGB-D multi-class instance segmentation based on double pyramid feature fusion model

ZHANG Xu-dong†, WANG Yu-ting, FAN Zhi-guo, FU Xu-wen

(School of Computer and Information，Hefei University of Technology，Hefei 230009，China)

Abstract: For RGB images instance segmentation, some segmentation errors may occur in areas with similar textures but different categories. This paper introduces depth information and makes use of three-dimensional geometric features of RGB-D images, proposing the double pyramid feature fusion model. The method constructs two pyramid depth networks with different complexity to extract RGB and Depth features of different resolutions, then add two features of corresponding resolution. In this way, we change the input features of region Proposal network, then the classification network, regression network and mask network output positioning and classification results to get RGB-D images instance segmentation results. The experimental results show that the proposed model can learn the complementary information between depth images and RGB images, and get satisfactory RGB-D instance segmentation results. Compared to the mask R-CNN model that does not contain depth information, the average precision of the proposed model is increased by 7.4 %.

Keywords: instance segmentation；RGB-D images；pyramid network；feature fusion；region proposal

0 引 言

实例分割[1]作为一项兼具物体检测[2-7]和语义分割[5,8-11]的任务,可以对场景中的每一个目标物体给出有效的检测结果,是计算机视觉领域中具有挑战性的研究任务之一.

现有实例分割大多是在RGB图像上展开研究的. Hariharan等[12]提出了SDS方法,首次将位置检测与分割结合在一起,利用MCG(modified conjugate gradient)算法提取候选区域,联合训练有两条路径的单一网络,分别提取边界特征和区域前景特征,基于CNN最后提取的特征训练SVM进行分类,对重复覆盖的区域进行非最大抑制. Hariharan等[13]对SDS方法进行改进,提出Hypercolumns方法,选择将高层特征与低层特征融合形成Hypercolumns,再用其训练SVM. 该方法在分类器中引入超列的概念,实现了对ROI的修正. Dai等[14]提出CFM(convolutional feature masking)算法,利用SSP实现从卷积特征中提取掩码而不是直接从原始图像提取掩码,将任意大小的区域生成一个固定大小的特征,用CFM代替矩形框,用不规则区域生成掩码,提取特征. Li等[15]提出的FCIS

收稿日期: 2018-11-05；**修回日期:** 2019-02-28.
基金项目: 国家自然科学基金项目(61876057, 61471154).
责任编委: 侯忠生.
†通讯作者. E-mail: xudong@hfut.edu.cn.

是首个全卷积、端到端的实例分割解决方案,使用位置敏感的特征融合方法进行特征提取,在同一时间对多张连在一起的图像进行分割和检测. He等[16]提出的Mask R-CNN在Faster R-CNN[4]的基础上,将RoIPooling层替换为RoIAlign层,引入双线性插值操作,解决仅通过Pooling直接采用带来的对齐问题,同时添加并列的FCN层,计算Mask损失,通过对每个类对应一个Mask有效避免了类间竞争,该模型在实例分割任务上取得了优秀的效果.

以上模型在RGB图像实例分割任务中都取得了阶段性成就,但真实场景往往比较复杂,当场景中的类别信息并不能只通过色彩特征进行区分时(如纹理相近而类别不同的区域),这类模型可能得不到理想的分割结果. RGB-D图像深度图中每个像素值是传感器距离物体的实际距离,其RGB图像与深度图的像素点也具有一对一的对应关系,因其具备可靠的Depth信息,能够较全面地反映场景中的三维几何关系,已广泛应用于图像分类[17-19]、图像分割[20-22]和三维重建[23-26]等领域中.

如何充分提取深度图像中的信息是RGB-D图像上分割任务研究的关键,Couprie等[27]将RGB图像和深度图像直接串联为四通道数据输入多尺度卷积网络,能够对彩色图像和深度图像进行实时处理,完成RGB-D图像语义分割. Gupta等[28]提出深度图像HHA编码方法,将原始深度图像转化为HHA特征图,然后利用两个CNN网络分别从RGB图像和深度图像提取特征,再连接两种特征放入最后的语义分类器中,对RGB-D图像进行语义分割和19类检测器的单实例分割. Park等[22]提出了RedNet模型,利用CNN分别提取RGB与Depth特征,通过4个名为MMFNet的模块融合不同阶段的多模态特征,再通过4个refineNet结构块从多个级别学习融合特征,采用编码解码器结构解决RGB-D图像语义分割. Long等[8]直接用两个独立的CNN模型进行特征提取,预测每种形态的评分图,然后按照等权重求和融合评分进行端到端的RGB-D语义分割. Cheng等[9]在RGB-D语义分割的研究中,也采用深度图像HHA编码方法,利用两个LS-DeconvNet网络分别训练彩色图片和HHA图像,再分别放入反卷积层预测每个像素的类别,实现语义分割.

从近几年的研究可以看出,RGB-D图像分割任务的研究侧重于语义分割. 而实例分割是针对场景中单独对象的分割任务,既能对目标物体进行定位又能对定位框内的物体进行分类,在具体类别基础上区别开不同的实例,具有更广阔的应用前景,因此本文的研究任务是对RGB-D图像进行实例分割.

本文基于RGB-D数据的特点,在Mask R-CNN模型基础上通过实验分析对比了彩色图像、彩色图像+原始深度图、彩色图像+编码后深度图的实例分割结果,进而提出双金字塔特征融合网络,利用卷积块结构搭建深度图像特征金字塔网络分支,采用残差网络提取彩色特征,改变RPN(region proposal network)层的共享特征图输入.

本文主要贡献归纳如下: 1) 基于卷积块设计了深度图的网络分支,提取不同分辨率的深度图像特征,构成深度图的特征金字塔,在Mask R-CNN的网络基础上,于RPN层之前融合对应分辨率大小的RGB特征与Depth特征; 2) 采用不同复杂度的网络分别提取彩色特征与深度图特征,由于彩色图像语义信息丰富而深度图像主要用于提高定位准确度,利用复杂度较高的残差网络提取彩色特征,复杂度较低的卷积网络提取深度图像特征; 3) 设计数据层融合模型和特征层融合的网络模型,通过对比实验分析了彩色信息及深度信息在实例分割中的作用,同时得出直接利用深度图像进行数据层融合能够提高检测种类的数量但检测精度较低,而双金字塔特征融合网络在提高分割丰富度的同时提高了分割平均精度.

1 本文方法

本文主要任务是解决RGB-D图像实例分割问题,由于Mask R-CNN在RGB图像实例分割任务中表现优异,为此首先改变Mask R-CNN网络的输入,对原始深度图进行HHA编码和表面法线编码,在该模型上进行RGB、RGB+原始深度图、RGB+编码后深度图的实例分割实验,将分割结果可视化输出,从定性和定量的角度进行对比. 实验表明,数据层融合网络能够提高分割丰富度但检测精度较低,基于此实验结果,进一步设计了一种在特征层融合的双金字塔特征融合网络,利用卷积网络学习深度图像特征,改变RPN的输入共享特征图,获得较高的RGB-D实例分割精度.

1.1 数据层融合模型

为了设计有效的网络结构以充分发挥RBG信息与Depth信息各自的优势,最直观的方法是将深度图和彩色图在网络底层进行融合,设计思路近似于将RGB-D图像作为四通道数据输入Mask R-CNN中. 该数据层融合模型利用两个单层卷积层分别对RGB图像与深度图像进行一次特征提取,如图1所示,首先采用的卷积核大小为7×7,步长为2,分别提取RGB与Depth特征,当前的权重比设置为4:1.

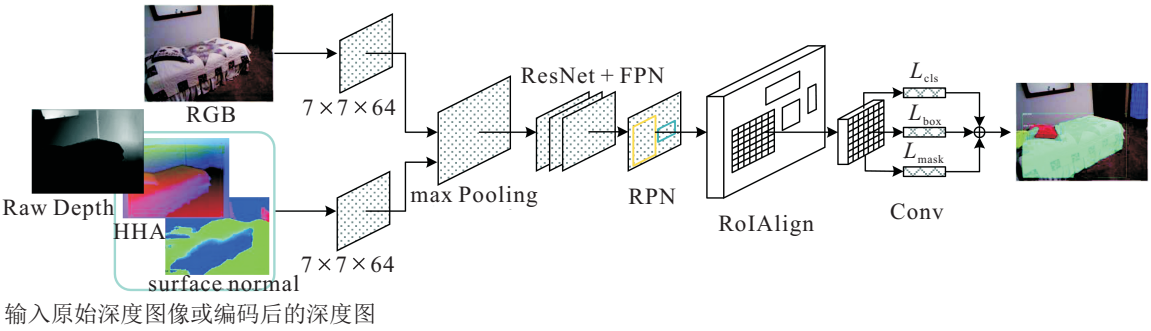


图1　数据层融合模型

图2为数据层融合模型在NYUD2数据集上的测试结果. 由图2可见,不管是原始深度图还是编码后的深度图,融合Depth信息之后,分割结果的丰富度都得到了提高. 在颜色纹理相近但深度存在差异的地方,如场景2中的盥洗台与浴缸,单纯利用RGB信息不足以识别盥洗台的存在,融合Depth信息之后可以检测到其位置. 又如场景1中位置分散的玩具,只学习RGB信息的模型无法检测其存在,融合Depth信息之后可以将前景(玩具)与背景(地板)区别开. 同时,编码后的深度图相比原始深度图得到的分割结果更丰富,且HHA编码比表面法线编码的位置信息更准确. 但是,这种融合方式准确度较差,引入Depth信息后对原本彩色特征的学习带来了干扰,导致很多分类错误的目标,所以最终的mAP反而小于只学习RGB特征的模型,编码后的检测结果更多,识别错误的结果也更多,其mAP比学习原始深度图的值更小. 与定性结果相对应,该模型定量的结果见表1.

表1　Mask R-CNN与数据层融合模型mAP比较

input	fusion	AP	AP_{50}	AP_{75}
RGB	1:0	33.9	51.7	35.2
RGB+Depth	64:64	7.4	25.7	8.7
RGB+Depth	64:16	26.1	47.5	27.8
RGB+HHA	64:16	15.3	33.6	16.8
RGB+SF	64:16	16.2	30.1	17.3

从利用数据层融合模型这种简单的融合方式可以看出,Depth信息的加入可以提高分割丰富度,尤其是在纹理相近、深度不同的区域,但是,采用一般的融合方式会导致模型学习到的冗余信息大于两种特征的互补信息,带来错误的分类结果,检测到的目标越多错误的结果越多,检测精度反而越低. 同时,实验结果也表明深度图编码方法在该融合方式的网络模型中难以起到正面影响,单纯依靠编码方法带来的改进并不足以有效解决RGB-D实例分割任务,有必要设计更有效的网络模型进一步将两种特征进行提取与融合.

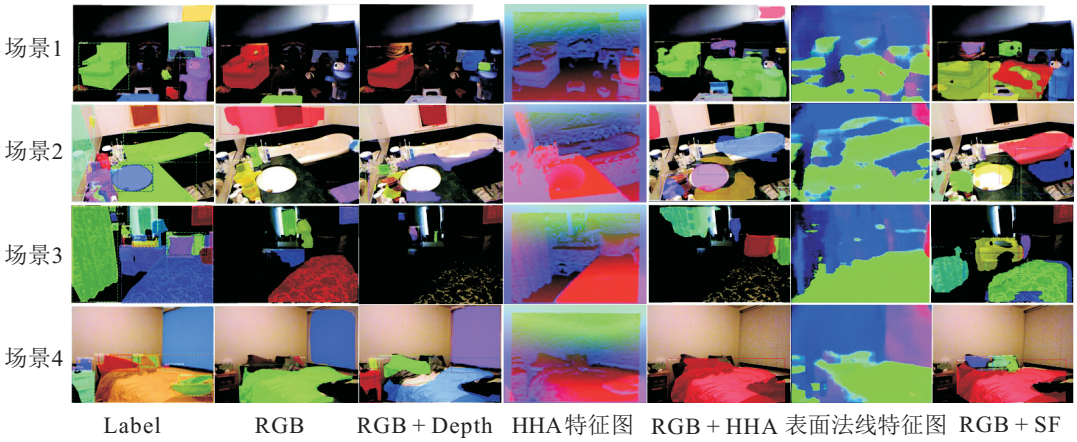


图2　数据层融合模型输出结果

1.2　金字塔网络提取深度特征

第1.1节的模型已经证明了Depth信息在实例分割任务中的有效性,但该模型最终的AP值较差,为此,分别设计不同复杂度的深度学习网络以提取彩色图像和深度图像特征,进而学习深度图像与彩色图像的互补信息. 选择用结构较深的残差网络提取RGB特征,用层数较少的卷积神经网络提取深度图像特征. 在提取特征过程中利用金字塔结构提取不同分辨率的RGB和Depth特征,并将对应大小的RGB特征与深度图像特征相加后输入RPN层,实现基于双金字塔特征融合网络的RGB-D多类实例分割,网络模型如图3所示.

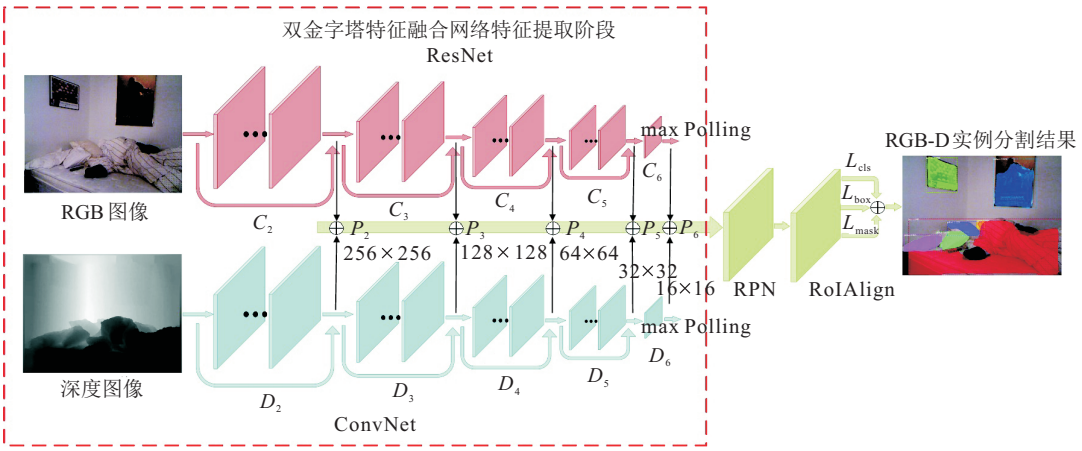


图3 双金字塔特征融合网络模型示意图

已有研究表明[7],高层的特征图包含的语义信息比较丰富,但是位置信息不够准确,低层的特征所包含的语义信息较少,但位置信息比较准确. 前文实验已经证明,Depth信息在定位任务中具有较好的表现,语义上的检测结果并不可靠,所以要获得两种特征有效信息互补的效果,输入RPN共享特征图中Depth特征应该比对应的RGB特征低阶. 因此,本文设计深度分支时并没有采用与RGB分支相同的残差网络,而是通过卷积块的结构构造网络. 卷积块的结构如图4所示,卷积块的主要通路由3个卷积层组成,shortcut连接部分由一层卷积层与一层BN层组成,卷积核大小为1 × 1,步长为2,每经过一个卷积块模型,特征图边长变为输入的一半. 残差网络的结构中有类似的网络结构,但是增加了深层的残差块模型,所以对于相同阶段的特征图而言,深度图像提取的特征比RGB图像提取的特征更低阶,符合期望的RGB特征引导语义分割,Depth特征补全位置检测的目标.

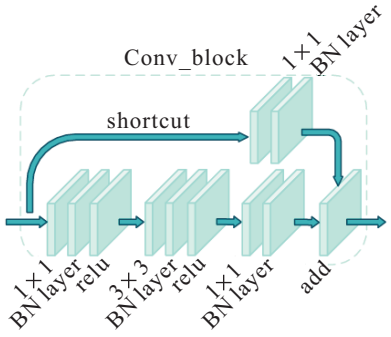


图4 卷积块结构

在特征提取的通路中,本文将深度金字塔分支网络提取Depth特征分别表示为D_2、D_3、D_4、D_5,对应于RGB通路中提取的特征C_2、C_3、C_4、C_5. 特征融合的通路中分别在P_2、P_3、P_4、P_5、P_6层融合相应的两种特征,其中P_6是P_5下采样的特征输出,每一层对应的特征图大小为256 × 256、128 × 128、64 × 64、32 × 32、16 × 16. 与Mask R-CNN不同的是,本文将最高层特征P_6也放入RPN进行训练.

双金字塔特征融合网络利用特征金字塔网络分支在产生目标框前对两种特征进行融合,通过改变输入共享特征图直接影响锚的生成. 不同于第1.1节模型在特征提取前融合两种特征,该网络采用金字塔网络单独提取深度图像的特征,省去了深度图像编码的步骤,通过卷积神经网络提取原始深度图像特征. 在特征提取的通路中,利用残差网络提取彩色图像的金字塔特征,再用卷积块提取对应大小的深度图像特征,并在特征融合通路中将不同分辨率的特征图放入对应的RPN网络中产生预选锚,经过RoIAlign层后输入Mask R-CNN的head与全连接层,输出分割结果.

2 实验结果及分析

2.1 实验设置

实验数据集选择NYUD2室内RGB-D数据集[29],参考Gupta等[30]之前的工作,将该数据集分为训练集、验证集和测试集(其中训练集包括381张,验证集包括414张,测试集654张). 此外,由于数据集数量少种类多,Gupta只列举出19种家具范畴的类别标签结果. 本文除了比较这19种类别外,还参考了COCO数据集的特点,从894类中选择出现次数较多的84类作为最终的目标类别. 为了保证实验的公平性与模型的通用性,按照上述样本挑选准则随机生成3份数据样本,取平均值作为最终分割精度.

实验模型基于TensorFlow与Keras搭建,首先利用COCO 2014数据集对网络RGB部分进行初始化训练,基于训练好的模型参数利用NYUD2数据集进一步训练,微调参数. COCO数据集输入图像的分辨率为1 024×1 024,NYUD2数据集的输入图像为480×640,因此需要通过zero padding保留横纵比将输入图像的大小进行统一. 本文模型中,通过调

整将最小边长设为800, 最大边长设为1 024. 每张图像输入分类器与Mask的RoI数目为128个, 并将mini-batch设置为1, RoI的比例设置为1:3 (positive比negative). 模型以0.002的学习率在2个GPU上进行训练, 每经过40 k次迭代, 学习率减少为原来的10倍. 每个epoch在NVIDIA TitanGPU上训练的时间约为15 min, 整个模型训练需要约16 h. 模型的权值衰减与动量分别设置为0.000 1与0.9. 本文模型在NMS (non-maximum suppression)之后保留了1 000个RoI, 然后选择前景分数最高的100个进行实例分割. 将每个RoI的第k个预测掩膜作为最终的预测类别, 并在设定的IoU阈值处计算平均精度.

2.2　实验结果

在第2.1节介绍的数据集选择标准之上, 本文模型完成84类室内NYUD2室内场景RGB-D数据的实例分割任务. 定性的实验结果如图5所示, 图中选择6个具有代表性的室内场景, 第1行显示了每个场景的标签, 第2行显示仅学习RGB特征的实例分割结果, 第3行显示加入深度网络分支学习Depth信息之后的实例分割结果. 由可视化结果可以得到, 双金字塔特征融合网络保留了数据层融合模型的优点, 学习Depth信息可以得到更丰富的分割结果, 如场景1中的灯与场景5中的瓶子. 对于颜色相近但距离不同的物体, 如场景2中的枕头与场景3中的电视, 仅通过颜色难以对这些目标物体进行区分, 双金字塔特征融合网络证明即使是相同颜色的不同物体也是可以被网络正确地区别开. 最重要的是, 基于卷积块构造的网络提取深度图像特征, 可以准确判断出场景中目标物体的类别, 相比于数据层融合模型的语义分割准确率得到明显的提高.

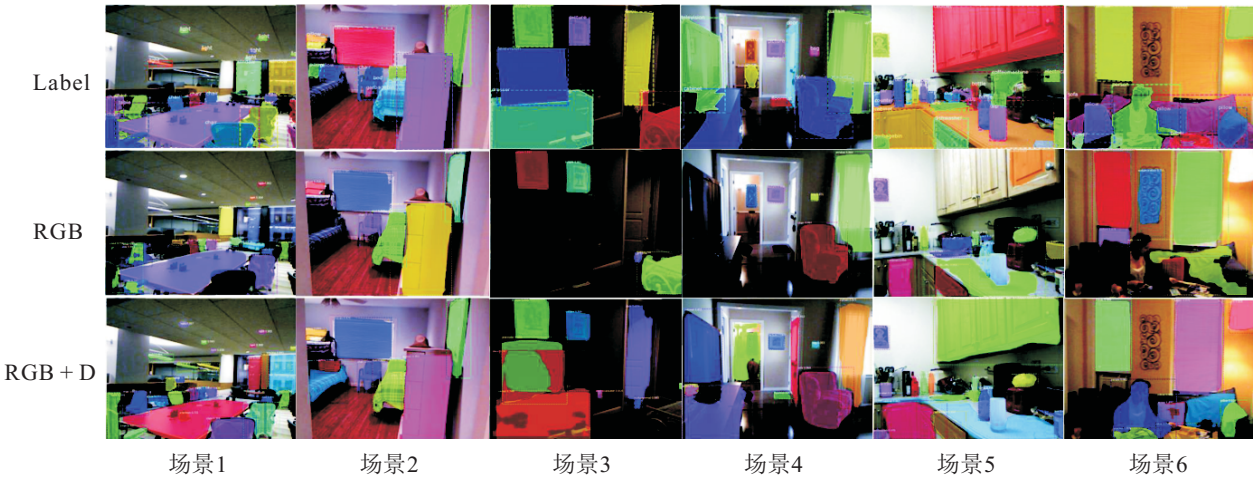


图5　双金字塔特征融合网络与Mask R-CNN的输出对比结果

为了评估实验的准确性, 需要定量的实验结果. 选择平均精度 (average precision, AP) 对分割结果进行定量的判定, 参考COCO数据集的评判指标, 分别计算AP、AP_{50}、AP_{75}, 以及场景中不同大小物体对应的检测指标AP_s、AP_m、AP_l. 本文模型定量的实验结果如表2所示, 表中给出了本文模型与当前最先进的几种RGB图像实例分割方法在NYUD2数据集上的对比结果. 可以看出, 本文模型相较于MNC[31]、FCIS[15]、Mask R-CNN, 所获得的AP值最高, 比Mask R-CNN AP值提升了7.4 %, 比AP_{75}提升了8.3 %. 且融合Depth特征后对中等大小的物体影响最大, AP_m提升了15.4 %. 定性与定量的实验结果表明, 相比于仅学习RGB特征的最新实例分割模型及数据层融合的模型, 本文所提出的双金字塔特征融合网络能够获得较高的平均精度 (由图5可视化RGB-D实例分割结果可以直观看出), 表明本文模型可以有效学习到RGB特征与Depth特征之间的互补信息. 一方面, 实验模型充分利用了Depth信息, 补偿了彩色信息在颜色相近区域无法正确区分的问题; 另一方面, 实验结果表明本文模型在NYUD2这样数量较少的样本上仍可以获得准确的实例分割结果.

表2　仅学习RGB图像模型与双金字塔特征融合网络的AP比较

input	model	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
RGB	MNC((ResNet-101)	20.6	38.1	21.5	1.8	8.2	28.1
RGB	FCIS(ResNet-101)	26.7	43.4	25.9	2.6	11.8	31.6
RGB	Mask R-CNN	33.9	51.7	35.2	3.0	15.4	36.0
RGB+Depth	本文方法	41.3	54.7	43.5	15.5	30.8	49.1

2.3 实例分割特征图对比

第1.2节指出,高层的特征图包含的语义信息比较丰富,但位置信息不够准确,低层的特征图所包含的语义信息较少,位置信息比较准确. 对于双金字塔特征融合网络RGB通路中提取的特征C_2、C_3、C_4、C_5、C_6以及对应分辨率大小的RGB特征和Depth特征融合之后的特征P_2、P_3、P_4、P_5、P_6,从定性及定量实验结果可以得到Depth信息对位置信息影响较明显,因此图6选择了测试集中具有代表性的一个场景图片,显示C_2、P_2层特征图,从特征阶段进一步证明模型的有效性. 图6所示的场景中,虚线框标注的两个目标物体与背景颜色相近,但深度图中两个物体的深度存在差异,C_2为双金字塔特征融合网络彩色图像对应的低层特征,P_2为融合RGB与Depth特征之后的低层特征图像. 从虚线框部分的特征图可以看出,融合Depth信息之后的特征图边缘信息更丰富,目标物体的识别度更高.

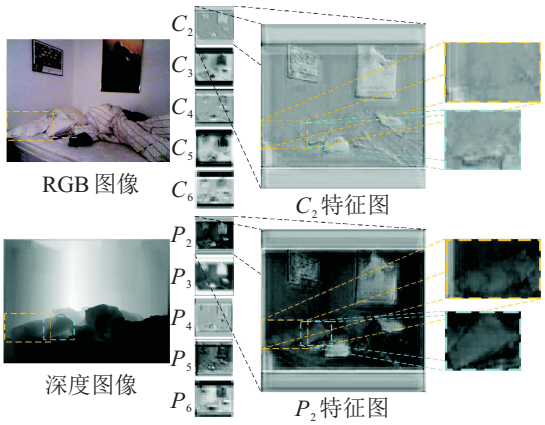


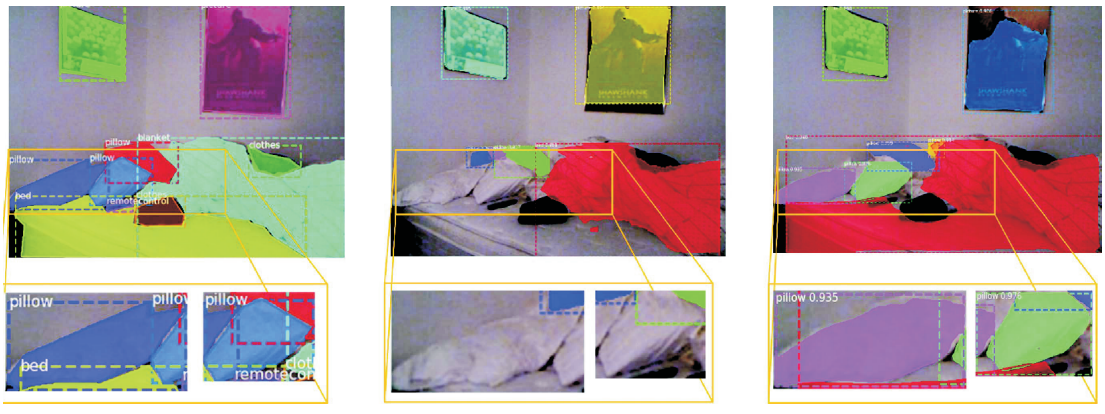
图6 彩色特征图与融合后特征图对比

图6特征图上目标区域的改进直接影响最终该区域的实例分割结果,图7展示了本场景在Mask R-CNN网络和本文双金字塔特征融合网络的实例分割结果. 对比标签图可以看出,依靠彩色信息的分割结果不能检测到这两个目标物体的位置及类别信息,但利用本文模型融合Depth信息后,可以正确分割出目标物体的位置及类别,并在边界框内完成物体的像素级分割,表明融合深度图像之后的分割结果与标签图吻合度更高,检测结果更丰富,且分类结果正确.

由实验结果可以看出,所提出的双金字塔特征融合网络可以从RGB特征与Depth特征的融合特征图中学习到实例分割的有效信息,不同于数据层融合模型的是,该模型用特征金字塔网络提取深度图像的特征,舍弃会增加计算开销的深度编码步骤,实验结果表明该设计能够高效地学习到深度图像的有效特征,即与彩色特征互补的特征信息. 从特征图对比实验中可以看出,融合后的特征图边缘特征更加明显,将融合后的特征图直接输入RPN层,能够改变输入共享特征图,直接影响锚的生成. 从最终的测试集可视化结果可以看出,该模型能够完成RGB-D图像的实例分割任务.

2.4 Gupta19类实例分割实验结果对比

之前的实验主要围绕网络模型与Depth信息进行讨论,通过大量自对比证明了双金字塔特征融合网络模型在RGB-D实例分割任务中的有效性. 为了证明本文模型的优越性,按照Gupta等[28]进行19类检测器的单实例分割的要求进行同样的实验,并计算对应的边界框平均精度AP_b与区域平均精度AP_r,区域平均检测精度相对于边界框精度,进一步计算像素的重叠区域,所以更能准确衡量实例分割的结果. 表3显示了19类AP_b与AP_r的分割结果,从结果可以看出,本文模型在19类分割任务中也获得了较好的分割结果,平均AP_b结果为46.2%,相较于Gupta等[28]的模型提高了8.9%,平均AP_r结果为41.4%,提高了9%,表明了本文模型在指定类的单实例分割任务中也十分有效.



Label图像 Mask R-CNN分割结果 双金字塔特征融合网络分割结果

图7 示例特征图实例分割结果对比

表 3　Gupta19类实例分割结果对比

	mean	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser
RGB DPM	9.0	0.9	27.6	9.0	0.1	7.8	7.3	0.7	2.5	1.4
RGBD DPM	23.9	19.3	56.0	17.5	0.6	23.5	24.0	6.2	9.5	16.4
RGB RCNN	22.5	16.9	45.3	28.5	0.7	25.9	30.4	9.7	16.3	18.9
文献[28]方法	37.3	44.4	71.0	32.9	1.4	43.3	44.0	15.1	24.5	30.4
本文方法(AP_b)	**46.2**	**50.0**	**83.8**	**54.4**	**3.6**	**64.4**	**59.5**	**20.0**	**28.1**	**38.5**

	garbage bin	lamp	monitor	nightstand	sink	pillow	sofa	table	television	toilet
RGB DPM	6.6	22.2	10.0	9.2	5.9	4.3	9.4	5.5	5.8	34.3
RGBD DPM	26.7	26.7	34.9	32.6	22.8	20.7	34.2	17.2	19.5	45.1
RGB RCNN	27.9	27.9	32.5	17.0	16.6	11.1	29.4	12.7	27.4	44.1
文献[28]方法	36.5	36.5	**52.6**	40.0	36.1	34.8	53.9	24.4	37.5	46.8
本文方法(AP_b)	**39.0**	**46.2**	50.7	**43.3**	**47.2**	**49.8**	**68.3**	**34.6**	**39.5**	**56.4**

	mean	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser
box	14.0	5.9	40.0	4.1	0.7	5.5	0.5	3.2	14.5	26.9
region	28.1	32.4	54.9	9.4	1.1	27.0	21.4	8.9	20.3	29.0
fg mask	28.0	14.7	59.9	8.9	1.3	29.2	5.4	7.2	22.6	33.2
文献[10]方法	32.1	18.9	66.1	10.2	1.5	35.5	32.8	10.2	22.8	33.7
本文方法(AP_r)	**41.1**	**38.4**	**78.8**	**30.6**	**3.0**	**57.4**	**48.8**	**13.6**	**27.7**	**35.0**

	garbage bin	lamp	monitor	nightstand	sink	pillow	sofa	table	television	toilet
box	32.9	1.2	40.2	11.1	9.4	6.1	13.6	2.6	35.1	11.9
region	37.1	26.3	48.3	38.6	30.9	33.1	30.5	10.2	33.7	39.9
fg mask	38.1	31.2	**54.8**	39.4	32.0	32.1	36.2	11.2	37.4	37.5
文献[10]方法	38.3	35.5	53.3	42.7	34.4	31.5	40.7	14.3	37.4	50.5
本文方法(AP_r)	**39.0**	**36.8**	50.6	**43.3**	**47.2**	**48.6**	**60.0**	**30.0**	**39.5**	**53.2**

3　结　论

本文针对RGB-D实例分割任务,通过实验分析了在Mask R-CNN框架下直接学习RGB+原始深度图像、RGB+编码后的深度图像的实例分割方法. 表明直接在数据层进行RGB与深度图像、深度图像编码的融合方式不能充分学习到深度图像的互补特征,检测到的目标物体数量虽然增加但准确度较差,以至于平均精度值较低. 因此进一步提出双金字塔特征融合网络,利用两种不同复杂度的网络分别提取特征,用复杂度较低的卷积神经网络提取深度图像特征,复杂度较高的残差网络提取RGB特征,并在RPN层之前融合对应分辨率的两种特征. 该设计既能保证从彩色图像学习到有效的语义信息,又能利用深度图像的信息提高定位精度. 实验结果表明,该网络模型可以有效学习两种特征的互补信息,带来更精确的实例分割结果. 下一步的任务是研究如何进一步改进框架结构,得到更容易训练与更高效的网络模型,提高实例分割的准确度.

参考文献(References)

[1] Romera-Paredes B, Torr P H S. Recurrent instance segmentation[C]. European Conference on Computer Vision. Amsterdam: Springer, 2016: 312-329.

[2] Gode C S, Khobragade A S. Object detection using color clue and shape feature[C]. International Conference on Wireless Communications, Signal Processing and Networking. Chennai: IEEE, 2016: 464-468.

[3] Girshick R. Fast R-CNN[C]. International Conference on Computer Vision. Santiago: IEEE, 2015: 1440-1448.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. International Conference on Neural Information Processing Systems. Montréal: MIT Press, 2015: 91-99.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference Computer Vision and Pattern Recognition. Washington, DC: IEEE, 2014: 580-587.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Rision. Piscataway: IEEE, 2017: 2980-2988.

[7] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2117-2125.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440.

[9] Cheng Y, Cai R, Li Z, et al. Locality-Sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation[C]. IEEE Conference on

Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1475-1483.

[10] Li Y, Zhang J, Cheng Y, et al. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation[C]. IEEE International Conference on Image Processing. Beijing: IEEE, 2017: 1262-1266.

[11] Su W, Wang Z. Regularized fully convolutional networks for RGB-D semantic segmentation[C]. Visual Communications and Image Processing. Chengdu: IEEE, 2016: 1-4.

[12] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation[C]. European Conference on Couputer Vision. Cham: Springer, 2014: 297-312.

[13] Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization[C]. IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 447-456.

[14] Dai J, He K, Sun J. Convolutional feature masking for joint object and stuff segmentation[C]. Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3992-4000.

[15] Li Y, Qi H, Dai J, et al. Fully convolutional instance-aware semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2359-2367.

[16] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C]. Proceedings of the IEEE International Conference on Computer Rision. Piscataway: IEEE, 2017: 2961-2969.

[17] Bo L, Ren X, Fox D. Depth kernel descriptors for object recognition[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. San Francisco: IEEE, 2011: 821-826.

[18] Xia Y, Shi X, Zhao N. Learning for classification of traffic-related object on RGB-D data[J]. Multimedia Systems, 2017, 23(1): 129-138.

[19] Eitel A, Springenberg J T, Spinello L, et al. Multimodal deep learning for robust RGB-D object recognition[C]. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg: IEEE, 2015: 681-687.

[20] Gupta S, Arbeláez P, Girshick R, et al. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation[J]. International Journal of Computer Vision, 2015, 112(2): 133-149.

[21] Qi X, Liao R, Jia J, et al. 3D graph neural networks for RGBD semantic segmentation[C]. IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5209-5218.

[22] Park S J, Hong K S, Lee S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation[C]. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 4980-4989.

[23] Zhou Q Y, Koltun V. Dense scene reconstruction with points of interest[J]. Acm Transactions on Graphics, 2013, 32(4): 1-8.

[24] 张辉, 王盼, 肖军浩, 等. 一种基于三维建图和虚拟现实的人机交互系统[J]. 控制与决策, 2018, 33(11): 1975-1982.
(Zhang H, Wang P, Xiao J H, et al. A human-robot interaction system based on 3D mapping and virtual reality[J]. Control and Decision, 2018, 33(11): 1975-1982.)

[25] Teng C H, Chuo K Y, Hsieh C Y. Reconstructing three-dimensional models of objects using a Kinect sensor[J]. Visual Computer, 2018, 34(11): 1507-1523.

[26] Zhao X, Chen W, Yan X, et al. Time stairs geometric parameters estimation for lower limb rehabilitation Exoskeleton[C]. 2018 Chinese Control And Decision Conference (CCDC). Shenyang: IEEE, 2018: 5018-5023.

[27] Couprie C, Farabet C, Najman L, et al. Indoor semantic segmentation using depth information[J]. arXiv preprint, arXiv: 1301.3572, 2013.

[28] Gupta S, Girshick R, Arbeláez P, et al. Learning rich features from RGB-D images for object detection and segmentation[C]. European Conference on Computer Vision. Zurich: Springer, 2014: 345-360.

[29] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[C]. European Conference on Computer Vision. Berlin: Springer-Heidelberg, 2012: 746-760.

[30] Gupta S, Arbelaez P, Malik J. Perceptual organization and recognition of indoor scenes from RGB-D images[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Oregon: IEEE, 2013: 564-571.

[31] Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2016: 3150-3158.

作者简介

张旭东 (1966–), 男, 教授, 博士, 从事机器视觉、传感器技术、智能信息处理机器相关应用系统等研究, E-mail: xudong@hfut.edu.cn;

王玉婷 (1994–), 女, 硕士生, 从事三维图像分割、智能信息处理的研究, E-mail: wyt0122@mail.hfut.edu.cn;

范之国 (1979–), 男, 副教授, 博士, 从事仿生偏振光导航、偏振光学探测及其智能信息处理等研究, E-mail: fzghfut@163.com;

付绪文 (1993–), 男, 硕士生, 从事深度图像超分辨率重建的研究, E-mail: fuxuwen@mail.hfut.edu.cn.

(责任编辑：郑晓蕾)