第 46 卷 第 10 期 计 算 机 工 程 2020 年 10 月
Vol. 46 No. 10 Computer Engineering October 2020

·图形图像处理· 文章编号：1000-3428(2020)10-0275-07 文献标志码：A 中图分类号：TP753

基于双目图像与跨级特征引导的语义分割模型

张 娣，陆建峰

（南京理工大学 计算机科学与工程学院，南京 210094）

摘 要： 为改善单目图像语义分割网络对图像深度变化区域的分割效果，提出一种结合双目图像的深度信息和跨层次特征进行互补应用的语义分割模型。在不改变已有单目孪生网络结构的前提下，利用该模型分别提取双目左、右输入图像的二维信息，并基于 ParallelNet 设计色彩深度融合模块，计算双目图像特征点的不同视差等级相似度提取深度信息，同时将其与二维信息进行融合获得深度特征。同时，在高层语义信息指导下使用跨级特征注意力模块得到准确的低层类别边界信息，以提高各尺度特征的利用率与边缘区域的准确率。实验结果表明，与传统 ParallelNet 双目基准模型相比，该模型分割得到图像的平均交并比与像素精度分别提高 3.67 和 3.32 个百分点，对栅栏和交通标志等相似区域的分割更细致准确。

关键词： 语义分割；双目图像；深度信息；跨级特征；注意力

开放科学(资源服务)标志码(OSID)：

中文引用格式： 张娣，陆建峰. 基于双目图像与跨级特征引导的语义分割模型[J]. 计算机工程，2020，46(10)：275-281，288.

英文引用格式： ZHANG Di，LU Jianfeng. Semantic segmentation model based on binocular images and guidance of cross-level features[J]. Computer Engineering，2020，46(10)：275-281，288.

Semantic Segmentation Model Based on Binocular Images and Guidance of Cross-Level Features

ZHANG Di，LU Jianfeng

（School of Computer Science and Engineering，Nanjing University of Science and Technology，Nanjing 210094，China）

【Abstract】 In order to improve the segmentation effect of semantic segmentation networks for monocular images on regions where image depth vary. To address the problem，this paper proposes a semantic segmentation model combining the depth information of binocular images and cross level features for complementary application. With no changes to its structure，the existing monocular twin network is used to extract two-dimensional information of input left and right binocular images，and to design color depth fusion module based on ParallelNet. On this basis，the similarity of different parallax levels of binocular image feature points is calculated to extract depth information，which is fused with the two-dimensional information to obtain depth features. At the same time，the cross-level feature attention module is used to get the accurate information of low-level category boundary under the guidance of high-level semantic information，so as to improve the utilization rate of each scale of features and the accuracy of edge regions. Experimental results show that compared with the traditional ParallelNet binocular benchmark model，the proposed model increases the mean Intersection over Union(mIoU) and the Pixel Accuracy(PA) by 3.67 and 3.32 percentage points respectively，and the segmentation of similar regions such as fences and traffic signs is more detailed and accurate.

【Key words】 semantic segmentation; binocular image; depth information; cross-level feature; attention

DOI： 10.19678/j.issn.1000-3428.0056292

0 概述

图像语义分割[1]是计算机视觉领域的热点问题之一，其任务是为图像中每个像素分配类别标签。语义分割技术对机器人和无人驾驶系统[2]的场景理解至关重要，如分割出道路与障碍物的位置等，为其安全行驶提供指导。

图形处理器(Graphics Processing Unit，GPU)具

基金项目： 国家重点研发计划(2017YFB1300205)。

作者简介： 张 娣(1994—)，女，硕士研究生，主研方向为双目视觉、语义分割；陆建峰，教授。

收稿日期： 2019-10-14 **修回日期：** 2019-11-23 **E-mail：** Lujf@njust.edu.cn

有强大的并行计算能力,在大规模像素级标注数据集出现后,基于深度学习的图像语义分割技术[1]得到进一步发展。2014 年 SHELHAMER 等人[3]提出的全卷积网络(Fully Convolutional Networks,FCN)首次将深度学习应用于语义分割。FCN 开创性地将目标分类网络中的全连接层替换为卷积层,并引入反卷积概念,实现了对任意尺寸图像的像素级语义分割。与传统非深度学习方法相比,FCN 的分割准确率更高且运行时间更短。但是从本质上来看,FCN 通过池化层逐渐缩小图像尺寸、扩大感受野,并利用卷积层提取不同层次的特征,然后采用反卷积将缩小后的特征图恢复至原始尺寸,图像在由大变小再变大的过程中,会丢失很多细节信息。因此,研究人员提出多种方法来提升语义分割对图像细节区域的处理能力。

文献[4]提出空洞卷积在不缩减特征图大小的情况下扩大感受野。部分研究者试图将不同尺度的特征进行融合。文献[5]设计了一种适合医学图像的 U 形对称网络(U-Net),采用跳跃连接的方法在通道维度上将不同特征图进行串联。文献[6]提出空间金字塔结构,通过聚合多尺度上下文特征获取全局信息。文献[7-9]将空洞卷积与空间金字塔相结合提出多孔金字塔池化,同时采用多个不同采样率的并行空洞卷积获取多尺度信息。文献[10]指出各尺度特征关注的信息层次不同,并采用多种方法加强高低层次特征之间的融合。由于透视成像过程丢失了深度信息[11],且单目图像缺乏足够的三维结构信息,因此大部分单目语义分割网络对三维结构特征显著的区域处理效果较差。

在 RGB-D 相机诞生后,研究者们利用额外的深度信息提升语义分割效果。早期的方法[12]是简单地将深度信息串联到 RGB 图像上,形成 1 个四通道数据并将其输入到神经网络中。文献[13]使用 2 个编码器分支分别提取 RGB 特征和深度特征,然后在特定节点将深度特征嵌入到 RGB 分支中,改变了原有特征提取网络的结构。此外,由于 RGB-D 相机测量范围太小,易受日光干扰,因此其仅适用于室内环境。为在更广泛的环境下利用深度信息,研究人员试图直接从成对的双目图像中提取深度信息。文献[14]提出的 3SP-Net 利用已有视差估计网络预测出深度信息,再将其与不同尺度的 RGB 特征融合。由于从双目图像得到深度信息计算量很大,这使整个网络不仅庞杂而且无法端到端地训练网络。文献[15]对已有的卷积神经网络进行微调,利用 L1 距离[16]匹配其左、右特征图之间的对应点,从而间接挖掘深度特征。该方法具有一定启发性,但是由于其在深度信息和二维图像信息融合上大量使用串联操作,因此结构不太合理且特征融合效率较低。

本文受文献[15]启发,利用已有的单目孪生网络提取双目图像二维信息,采用双目图像特征点在不同视差等级下的相似度间接表征深度信息,在不改变网络结构的前提下,通过少量计算提取双目图像的深度信息,以实现对环境三维特征的准确描述。

1 本文方法

1.1 网络整体结构

本文方法的网络结构包括编码器和解码器,如图 1 所示。其中:编码器的基础网络通过堆叠卷积层(Conv)和残差层(Res)构造 2 个完全相同的 ResNet50[17],以同步提取其左、右输入图像的二维信息。色彩深度融合模块(Color Depth Fusion Module,CDFM)用来提取不同尺度的深度特征,并将其与二维图像特征进行融合。解码器最顶层的融合特征应用注意力机制(Attention)[18-19]进行特征筛选以专注于更有用的信息,跨级特征注意力模块(Cross-level Feature Attention Module,CFAM)在高层语义信息的指导下,可获取更准确的低层边缘信息。将反卷积(Deconv)后的特征图与 CFAM 跨级融合后的特征图元素相加,并通过1×1卷积调整通道数可得到最终的分割图。

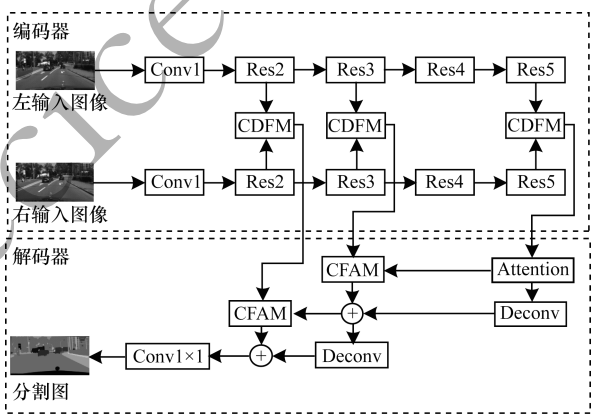

图 1 网络整体框架

Fig. 1 Overall network framework

1.2 色彩深度融合模块

为从二维图像特征中恢复深度特征,本文引入立体视觉的块匹配概念[20]来计算不同视差等级下对应点之间的相似度,并使用该间接相关的相似度表示深度特征。受 ParallelNet[15]启发,本文设计立体相似块(Stereo Similarity Block,SSB)提取更准确的深度信息。

具体地,令 $\boldsymbol{F}_{\mathrm{l}}$、$\boldsymbol{F}_{\mathrm{r}}$ 分别为双目视角下获得的左、右特征图,其维度均为 $h\times w\times c$,其中,h 为高度,w 为宽度,c 为通道数。以 $\boldsymbol{F}_{\mathrm{l}}$ 为例,该特征图可表示为:

$$\boldsymbol{F}_{\mathrm{l}}=\begin{bmatrix} l(1,1) & \cdots & l(1,d) & l(1,d+1) & \cdots & l(1,w) \\ \vdots & l(x,y) & \vdots & \vdots & & \vdots \\ l(h,1) & \cdots & l(h,d) & l(h,d+1) & \cdots & l(h,w) \end{bmatrix} \tag{1}$$

$$\boldsymbol{l}(x,y)=[\boldsymbol{l}(x,y)_1,\boldsymbol{l}(x,y)_2,\cdots,\boldsymbol{l}(x,y)_i,\cdots,\boldsymbol{l}(x,y)_c]_{1\times c} \tag{2}$$

其中,$\boldsymbol{l}(x,y)$为双目左特征图在(x,y)位置处的特征向量,其维度为 $1\times1\times c$,d 为视差偏移值。$\boldsymbol{F}_{\mathrm{r}}$ 的表达式与 $\boldsymbol{F}_{\mathrm{l}}$ 类似,其中,$\boldsymbol{r}(x,y)$为双目右特征图在(x,y)位置处的特征向量,其维度为 $1\times1\times c$。

SSB 的计算过程具体如下:

1)水平右移

固定左特征图 $\boldsymbol{F}_{\mathrm{l}}$,对右特征图 $\boldsymbol{F}_{\mathrm{r}}$ 中每个元素逐步水平右移 d_m 次,其中 d_m 为平移的最大值。$\tilde{\boldsymbol{F}}_{\mathrm{r}}$ 代表右移操作后的右特征图,其在位置(x,y)处的特征向量为 $\tilde{\boldsymbol{r}}(x,y)$,$\tilde{\boldsymbol{F}}_{\mathrm{r}}$ 具体表示为:

$$\tilde{\boldsymbol{F}}_{r}=\begin{bmatrix} 0 & \cdots & 0 & \boldsymbol{r}(1,1) & \cdots & \boldsymbol{r}(1,w-d) \\ \vdots & & \vdots & \vdots & \boldsymbol{r}(x,y) & \vdots \\ 0 & \cdots & 0 & \boldsymbol{r}(h,1) & \cdots & \boldsymbol{r}(h,w-d) \end{bmatrix} \tag{3}$$

当 $\boldsymbol{F}_r$ 向右移动 d 步时,$\tilde{\boldsymbol{r}}(x,y)$实际为原来位置$(x-d,y)$处的特征向量 $\boldsymbol{r}(x-d,y)$。与 ParallelNet[15] 中使用循环右移操作不同的是,本方法将最左边 d 列补零,以表明左图像最左边几列在右图像中无对应点,这比较符合在人眼视场中双目图像具有 1 个不重叠区域的特点。

2)相似度计算

计算左、右特征向量$\boldsymbol{l}(x,y)$和$\tilde{\boldsymbol{r}}(x,y)$之间的距离 L_2[16],计算公式为:

$$L_2(x,y,d)=\sum_{i=1}^{c}\sqrt{[\,l(x,y)_i-\tilde{r}\,(x,y)_i]^2}=\sum_{i=1}^{c}\sqrt{[\,l(x,y)_i-r(x-d,y)_i]^2} \tag{4}$$

距离 L_2越小表明特征之间的差异性越小,特征相关性越高,所有特征对之间的相似性构成相似度图。相较于 ParallelNet[15] 的距离 L_1,距离 L_2能更客观准确地描述 2 个特征向量之间的相似度。

3)串联

将 d_m 个相似度图串联可得到最终的深度特征。与 ParallelNet[15] 中设置固定 d_m 值不同的是,本文实验为了保证网络能够在给定的搜索范围内正确地找到匹配点,将 $d_{\max}$设置足够大,使其等于当前特征图的宽度。

值得注意的是,SSB 模块提取的是不同视差等级下左、右特征图之间的相似度,而深度信息实际上只与具有最高相似度的视差值有关。如果在实验中利用 argmin 操作[16] 手动选择可能性最大的视差值(即差异性最小时对应的视差值),实验结果(见 2.2.2 节)显示该操作无效果,推测这是因为 argmin 操作压缩过多维度,导致较多有用信息丢失。

色彩深度融合模块结构如图 2 所示。输入 1 对左、右特征图,先通过 SSB 模块获取深度特征,再对深度特征执行 1×1 卷积,使其通道数与二维图像特征通道数相等,然后分别对左特征图和深度特征图执行卷积、批量归一化(Batch Norm)和 ReLU 非线性化操作,然后将元素 D 与其相加以获得融合的 RGB-D 特征。

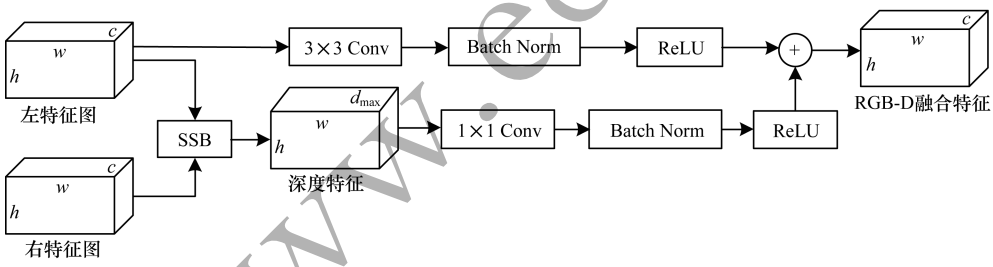

图 2　色彩深度融合模块结构

Fig. 2　Color depth fusion module structure

1.3　跨级特征注意力模块

语义分割网络通常由编码器和解码器组成。编码器直接使用 ResNet[17]、VGGNet[21] 等已有的卷积神经网络来获取分辨率逐渐降低、语义性逐渐增强的不同级别特征,解码器利用这些特征恢复不同类别像素的位置,从而预测出图像分割结果。

图像的高层特征和低层特征本质上是互补的。其中:高层特征用来指示图像中的语义信息,如道路、行人、汽车等类别信息;低层特征用来表征图像中的边缘、纹理、位置等信息。基于此,本文提出跨级特征注意力模块,以在高层语义信息指导下更准确地恢复低层的类别边界信息。

跨级特征注意力模块结构如图 3 所示。先对高层特征图执行窗口大小为(H_2,W_2)的全局池化(Global Pooling)操作以获得全局语义信息,再对全局语义特征执行 1×1 卷积、批量归一化和 ReLU 非线性化操作,使其通道数与低层特征图的通道数相等。同样对低层特征执行 3×3 卷积、批量归一化和

ReLU 非线性化操作,以获取更具表达力的低层特征。最后利用压缩后的全局语义信息指导低层特征在通道维度上的加权选择。该模块能够以高层特征为引导,选择性地保留低层特征中的有用信息,有助于融合跨级特征及提高语义边界定位准确率。

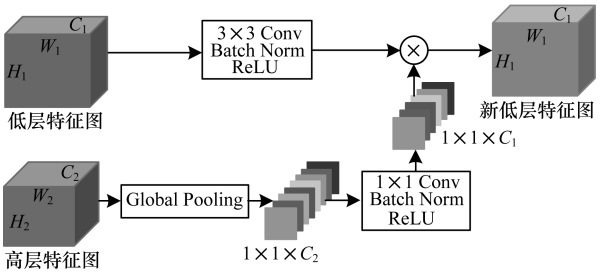

图 3 跨级特征注意力模块结构

Fig. 3 Cross-level feature attention module structure

2 实验与结果分析

2.1 实验设置

本文实验所用系统环境为 ubuntu 16.04、python 3.6.8 和 tensorflow 1.5.0[22],显卡为 NVIDIA TITAN Xp 12 GB,CPU 为 Intel® E5-2620 2.10 GHz。使用 Cityscapes 数据集[23],该数据集为目前少有的提供双目图像及语义标注的大型数据集。Cityscapes 数据集包含 5 000 张精确标注的图像和 20 000 张粗略标注的图像,这些图像是在不同季节和不同天气下从 50 个城市采集的街道场景。由于只有精确标注的图像提供了双目数据,因此本文使用 5 000 张精确标注的图像,并将这些图像分为训练集、验证集和测试集,数量分别为 2 975 张、500 张和 1 525 张。将平均交并比(mean Intersection over Union,mIoU)和像素精度(Pixel Accuracy,PA)作为语义分割的评价指标,计算公式如下:

$$\mathrm{mIoU}=\frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ij}}\times 100\% \tag{5}$$

$$\mathrm{PA}=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\times 100\% \tag{6}$$

其中,k 为类别数量,p_{ij} 为本属于类 i 但被预测为类 j 的像素数量。

本文对训练数据进行增强,通过图像归一化随机做高斯滤波使图像模糊。为保证正确学习双目特征点之间的匹配规则,未应用旋转、缩放、翻折等操作改变像素位置。图像随机裁剪为 512×512 大小。编码器部分的基础网络为 ResNet50[17],并加载在 ImageNet[24] 上预训练的参数。为更好地适配 ReLU 激活函数,网络中其他参数使用 He 初始化[25]方法,并使用 focal loss[26] 来减轻由于待测目标类别不平衡引起的分类困难问题。实验优化器为 Adam,使用多项式衰减的学习率策略,其中,基础学习率设置为 0.000 1,幂数为 0.9。受显卡容量限制,batch size 取 3,最大迭代次数设为 50 000。此外,采用早停策略以防止过拟合,每 60 次迭代后就在验证集上评估当前训练网络的性能,如果准确率在连续100 次的验证过程中没有得到提高,则提前结束训练。

2.2 对比实验

2.2.1 深度信息有效性评估

为评估深度信息的影响,在单目 FCN[3] 结构的基础上,将 CDFM 作用于原始特征图,并对融合深度后的特征图进行反卷积等操作以获取分割图。该网络称为 FCN + Depth,其具体结构和添加深度信息后不同方法的评价指标结果分别如图 4 与表 1 所示。由表 1 可知,添加深度信息后,语义分割性能得到明显提升。与基准模型 FCN 相比,采用本文提出的 FCN + Depth 方法得到的 mIoU 和 PA 分别提高 2.06 和 2.60 个百分点。

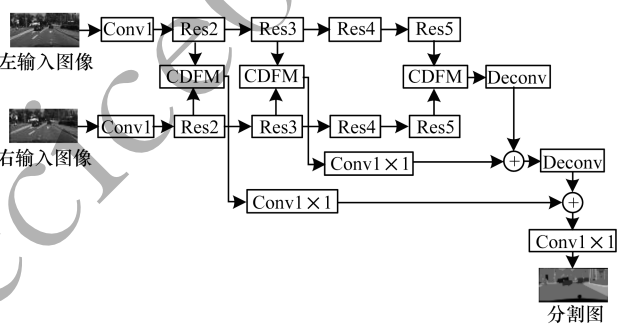

图 4 FCN + Depth 网络结构

Fig. 4 FCN + Depth network structure

表 1 添加深度信息后不同方法的评价指标结果

Table 1 Evaluation index results of different methods after adding depth information %

方法	mIoU	PA
FCN 方法	56.39	88.07
FCN + Depth 方法	58.45	90.67

2.2.2 特征筛选评估

由于 SSB 模块提取的是不同视差等级下左、右特征图之间的相似度,而深度信息只与具有最高相似度的视差值有关,因此本文实验尝试对 CDFM 模块采用不同方法进行 RGB-D 特征筛选并消除冗余信息,结果如表 2 所示。

表 2　不同特征筛选方法的评价指标结果

Table 2　Evaluation index results of different feature screening methods　%

方法	mIoU	PA
FCN + Depth 方法	58.45	90.67
FCN + Depth + argmin 方法	58.13	89.86
FCN + Depth + SE + RGB 方法	58.25	90.34
FCN + Depth + RGB + SE 方法	59.86	91.94
FCN + Depth + RGB + CBAM 方法	58.46	90.93

具体操作过程如下:

1)采用 FCN + Depth + argmin 方法,直接对 SSB 提取的深度特征实施 argmin 操作以选取可能性最大的视差值。由表 2 可知,采用 argmin 操作后评价指标均降低,这是因为在光照、视角、噪声等干扰因素下,匹配点之间的相似度不一定最高,而 argmin 操作将深度信息压缩至仅 1 个通道,所以会丢失很多有用信息。

2)采用 FCN + Depth + SE + RGB 方法,应用 SENet[18] 提出的 SE Attention 机制学习自动获取每个特征通道的重要程度,以实现深度特征的重标定,并将其与二维图像特征进行融合。由表 2 可知,该方法并未改善分割效果。

3)采用 FCN + Depth + RGB + SE 方法,先融合 RGB-D 特征,再对融合后的特征应用 SE Attention[18]。由表 2 可知,与未应用特征筛选的 FCN + Depth 方法相比,采用该方法得到的 mIoU 和 PA 分别提高1.41 和 1.27 个百分点。

4)采用 FCN + Depth + RGB + CBAM 方法,将 SE Attention 替换为在通道和空间 2 个维度上基于注意力机制的卷积块注意力模块(Convolutional Block Attention Module,CBAM)[19]。由表 2 可知,SE Attention 较 CBAM 分割效果更好。

2.2.3　跨级特征模块评估

为进一步评估跨级特征注意力模块 CFAM 的效果,先对最高层的 RGB-D 融合特征应用 SE Attention,再应用 CFAM 实现高层语义信息对低层边界信息的引导,网络框架如图 1 所示。将应用和未应用 CFAM 的方法分别记为 FCN + RGBD + SE + CFAM 和 FCN + RGBD + SE,得到的评价指标结果如表 3 所示。可以看出,引入 CFAM 后,mIoU 和 PA 分别提高0.80和 0.58 个百分点,有效提高了分割效果。

表 3　2 种方法的评价指标结果

Table 3　Evaluation index results of the two methods　%

方法	mIoU	PA
FCN + RGBD + SE 方法	59.86	91.94
FCN + RGBD + SE + CFAM 方法	60.66	92.52

2.3　综合评估

本文选取单目语义分割网络 FCN[3] 和双目语义分割网络 ParallelNet[15] 作为基准方法,在 ResNet50[17] 的基础上重新搭建 FCN 和 ParallelNet,并在 Cityscapes 数据集[22] 上将这 2 种基准方法与本文所提方法进行对比。

2.3.1　准确性评估

语义分割模型性能优劣主要通过其分割准确性来体现。优秀的分割模型对不同类别图像的辨识度更强,对语义边界刻画更细致。表 4 为采用 FCN 方法、ParallelNet 方法和本文方法得到的评价指标结果。可以看出,由于本文方法引入了间接深度信息,采用的双目语义分割网络比单目语义分割网络 FCN 效果更好。此外,由于本文方法考虑了特征筛选和跨级特征融合,与 ParallelNet[15] 相比,mIoU 和 PA 分别提高 3.67 和 3.32 个百分点。表 5 为 3 种语义分割方法对不同类别的像素精度对比,可以看出本文方法在交通标志、栅栏、行人、自行车上的分割准确率明显更高。

表 4　3 种语义分割方法的评价指标结果

Table 4　Evaluation index results of three semantic segmentation methods　%

方法	mIoU	PA
FCN 方法	56.39	88.07
ParallelNet 方法	56.99	89.20
本文方法	60.66	92.52

表 5　3 种语义分割方法对不同类别的 PA 对比

Table 5　Comparison of PA of three semantic segmentation methods for different categories　%

方法	行人	栅栏	摩托车	人行道	汽车	墙壁	骑车者	交通标志
FCN 方法	67.72	76.89	81.40	74.32	85.55	59.60	49.40	55.40
ParallelNet 方法	70.33	77.69	84.19	78.01	89.11	68.42	65.67	64.05
本文方法	71.74	79.13	87.23	82.53	91.02	69.93	69.74	70.74

图 5 是 Cityscapes 数据集原始图与不同方法在该数据集上的分割效果图,其中第 1 列、第 2 列分别为原始图与真值图,第 3 列 ~ 第 5 列分别为 FCN 方法、ParallelNet 方法和本文方法在 Cityscapes 数据集上的分割效果图。可以看出:FCN 方法对于相似类别图像的分辨力较差,如第 2 行示例场景中,其将属于交通标志类别的物体分类为栅栏;和 FCN 方法相比,ParallelNet 方法改善了深度特征与周围差别明显的部分区域分割效果,如树干、栏杆等边缘分割得更精细;本文方法由于采用深度信息和跨级

特征融合的方式,对图像细节及边缘的处理更准确细致。

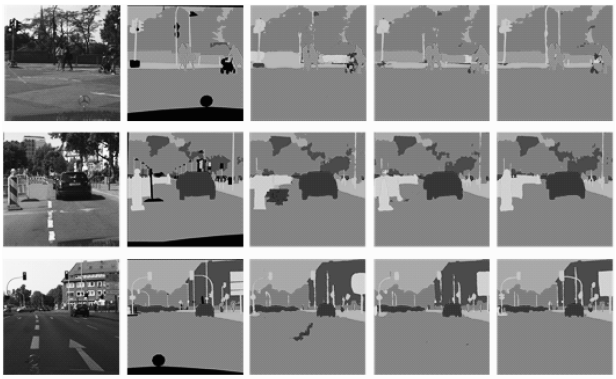

图 5 不同方法得到的分割效果图

Fig. 5 Segmentation effect images obtained by different methods

2.3.2 鲁棒性评估

为了评估模型的鲁棒性[27],本文对验证集中图像加入不同程度干扰项,观测并评估模型的分割效果。加入不同干扰项后,FCN 方法、ParallelNet 方法和本文方法在验证集上分割结果的 mIoU 如表 6 所示。

表 6 不同干扰项对 mIoU 的影响

Table 6 Influence of different interference terms on mIoU %

方法	未加干扰项	加椒盐噪声	图像调亮	图像调暗
FCN 方法	56.39	50.11	50.47	51.04
ParallelNet 方法	56.99	50.83	51.02	51.72
本文方法	60.66	56.85	57.65	58.35

首先对输入图像加入椒盐噪声[28],噪点数量占整幅图像像素点的 0.5% 。加入椒盐噪声后,FCN 方法、ParallelNet 方法和本文方法的 mIoU 与未加干扰项相比,分别降低 6.28、6.16 和 3.81 个百分点。然后通过伽马变换[29]调节输入图像亮度以模拟场景的照度变化:将验证集图像调亮后,FCN 方法、ParallelNet 方法和本文方法的 mIoU 与未加干扰项相比,分别降低 5.92、5.97 和 3.01 个百分点;将验证集图像调暗后,FCN 方法、ParallelNet 方法和本文方法的 mIoU 与未加干扰项相比,分别降低 5.35、5.27 和2.31 个百分点。由以上分析可知,对输入图像的数据加入干扰项后,模型性能在不同程度上均有所下降,但是本文方法较其他 2 种方法性能下降幅度更小,抗干扰能力更强。

图 6 为加入不同干扰项后不同方法在验证集部分场景下的分割结果鲁棒性对比情况。图 6(a) ~ 图 6(c)分别表示加入椒盐噪声、图像调亮和图像调暗 3 种干扰情况,从上至下分别为加入干扰的输入左图像、手工标注图、FCN 方法分割结果、ParallelNet 方法分割结果以及本文方法分割结果。由图 6(a)可以看出,当输入图像中存在大量随机出现的噪点时,由于 FCN 方法依赖局部区域内的颜色特征,因此其分割结果中会出现块状误判区域,而 ParallelNet 方法和本文方法由于考虑了双目图像的深度信息,因此均未出现明显的误判区域。在椒盐噪声干扰下,ParallelNet 方法在不同语义类别的边界处呈现毛躁的锯齿形态,而本文方法在语义边界区域分割更流畅。由图 6(b)可以看出,将输入图像调亮后,由于场景中栏杆与天空颜色接近,因此 FCN 方法未识别出栏杆,ParallelNet 方法分割出部分低矮的栏杆,而本文方法分割出大部分栏杆。由图 6(c)可以看出,将输入图像调暗后,FCN 方法将建筑物部分区域误判为天空,本文方法的分割结果更准确。3 种方法对右下角光线较暗骑行者的分割结果均不太理想,对行人和骑行者 2 种类别的分辨力有待加强。光线太暗也弱化了骑行者与自行车不同部位之间的辨识度,这也是可见光传感器在夜间性能较差的原因。

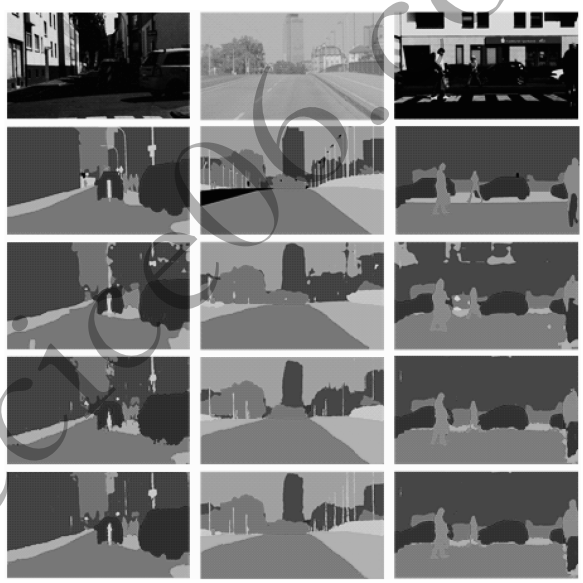

图 6 加入不同干扰项后不同方法的鲁棒性对比

Fig. 6 Robustness comparison of different methods after adding different interference terms

总体而言,由于 FCN 方法过分依赖图像的颜色特征,在加入干扰项后,会出现部分块状的误判区域。ParallelNet 方法考虑了深度信息,对图像颜色的依赖程度降低,但是对不同类别物体的边界识别不精细。本文方法由于不仅考虑了深度信息,还加强了对边界的关注,因此分割准确性更高且鲁棒性更强。

3 结束语

本文提出一种结合双目图像深度信息与跨级特征的语义分割模型。设计使用色彩深度融合模块计

算双目特征向量对的不同视差等级相似度以间接表征图像深度信息,并与原始特征图通过元素相加获得融合的深度特征。同时,通过跨级特征注意力模块利用富含语义信息的高层特征对低层特征进行加权选择,以更准确地恢复语义边缘。实验结果表明,该模型能更细致准确地分割图像边缘以及深度特征明显的区域。下一步将构建更多任务模型进行深度估计和语义分割,为三维场景建模提供更全面的信息。

参考文献

[1] YI Meng,SUI Lichun. Aerial image semantic classification method based on improved full convolution neural network[J]. Computer Engineering,2017,43(10):216-221. (in Chinese)
易盟,隋立春. 基于改进全卷积神经网络的航拍图像语义分类方法[J]. 计算机工程,2017,43(10):216-221.

[2] YANG Jingyu,TANG Zhenmin,ZHAO Chunxia,et al. Control method and implementation system in vehicle navigation based on monocular vision:CN200710019818.8[P]. 2008-08-06. (in Chinese)
杨静宇,唐振民,赵春霞,等. 基于单目视觉的汽车巡航控制方法及其实现系统:CN200710019818.8[P]. 2008-08-06.

[3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):640-651.

[4] YU F,KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1511.07122.

[5] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1505.04597.

[6] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017:6230-6239.

[7] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(4):834-848.

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

[9] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1706.05587.

[10] ZHANG Zhenlin, ZHANG Xiangyu, PENG Chao, et al. Exfuse:enhancing feature fusion for semantic segmentation[C]//Proceedings of 2018 European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 273-288.

[11] ZENG Zhihong, LI Jianyang, ZHENG Hanyuan. Depth information fused computational model of visual attention[J]. Computer Engineering,2010,36(10):200-202. (in Chinese)
曾志宏,李建洋,郑汉垣. 融合深度信息的视觉注意计算模型[J]. 计算机工程,2010,36(10):200-202.

[12] GUPTA S,GIRSHICK R,ARBELÁEZ P,et al. Learning rich features from RGB-D images for object detection and segmentation[C]//Proceedings of 2014 European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 345-360.

[13] HAZIRBAS C,MA L N,DOMOKOS C,et al. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture[C]//Proceedings of 2016 Asian Conference on Computer Vision. Berlin, Germany: Springer,2016:213-228.

[14] ZHOU Lingli, ZHANG Haofeng. 3SP-Net: semantic segmentation network with stereo image pairs for urban scene parsing[C]//Proceedings of Lecture Notes in Computer Science. Berlin,Germany:Springer,2018:503-517.

[15] LIU Shiyu, ZHANG Haofeng. ParallelNet: a depth-guided parallel convolutional network for scene segmentation[C]//Proceedings of PRICAI'18. Berlin,Germany:Springer,2018: 588-603.

[16] ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press,2016. (in Chinese)
周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[17] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA:IEEE Press,2016:770-778.

[18] HU J, SHEN L, ALBANIE S, et al. Squeeze and excitation networks[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1709.01507.

[19] WOO S,PARK J,LEE J Y,et al. CBAM:convolutional block attention module[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1807.06521.

[20] XIAO Jinsheng, TIAN Hong, ZOU Wentao, et al. Binocular stereo vision matching algorithm based on depth convolution neural network[J]. Acta Optica Sinica,2018,38(8):171-177. (in Chinese)
肖进胜,田红,邹文涛,等. 基于深度卷积神经网络的双目立体视觉匹配算法[J]. 光学学报,2018,38(8): 171-177.

[21] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1409.1556.

[22] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[EB/OL]. [2019-09-01]. https://arxiv.org/abs/1604.01685.

[23] ABADI M,BARHAM P,CHEN J,et al. TensorFlow:a system for large-scale machine learning[C]//Proceedings of Operating Systems Design and Implementation. New York, USA:ACM Press,2016:266-283.

(下转第 288 页)

（上接第 281 页）

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]// Proceedings of International Conference on Neural Information Processing Systems. Washington D. C., USA: IEEE Press, 2012: 1097-1105.

[25] HE Kingming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2015: 1026-1034.

[26] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[27] DU Wei, CAI Meng, DU Haifeng. Study on indices of network structure robustness and their application[J]. Journal of Xi'an Jiaotong University, 2010, 44(4): 93-97. (in Chinese)
杜巍, 蔡萌, 杜海峰. 网络结构鲁棒性指标及应用研究[J]. 西安交通大学学报, 2010, 44(4): 93-97.

[28] ZHANG Hao, CHEN Mingliang. Median filtering for eliminating high density salt and pepper noise by adjacent moving window[J]. Journal of electronic measurement and instrument, 2018, 32(9): 169-175. (in Chinese)
张皓, 陈明亮. 邻近移动窗消除高密度椒盐噪声的中值滤波[J]. 电子测量与仪器学报, 2018, 32(9): 169-175.

[29] CHEN Xiaonan, ZHANG Shufang, LEI Zhichun. High dynamic range image generation method by fusing multi-level gamma-transformed images[J]. Laser and Optoelectronics Progress, 2018, 55(4): 191-196. (in Chinese)
陈小楠, 张淑芳, 雷志春. 一种基于多层伽马变换融合的高动态范围图像生成方法[J]. 激光与光电子学进展, 2018, 55(4): 191-196.

编辑 宋 圆