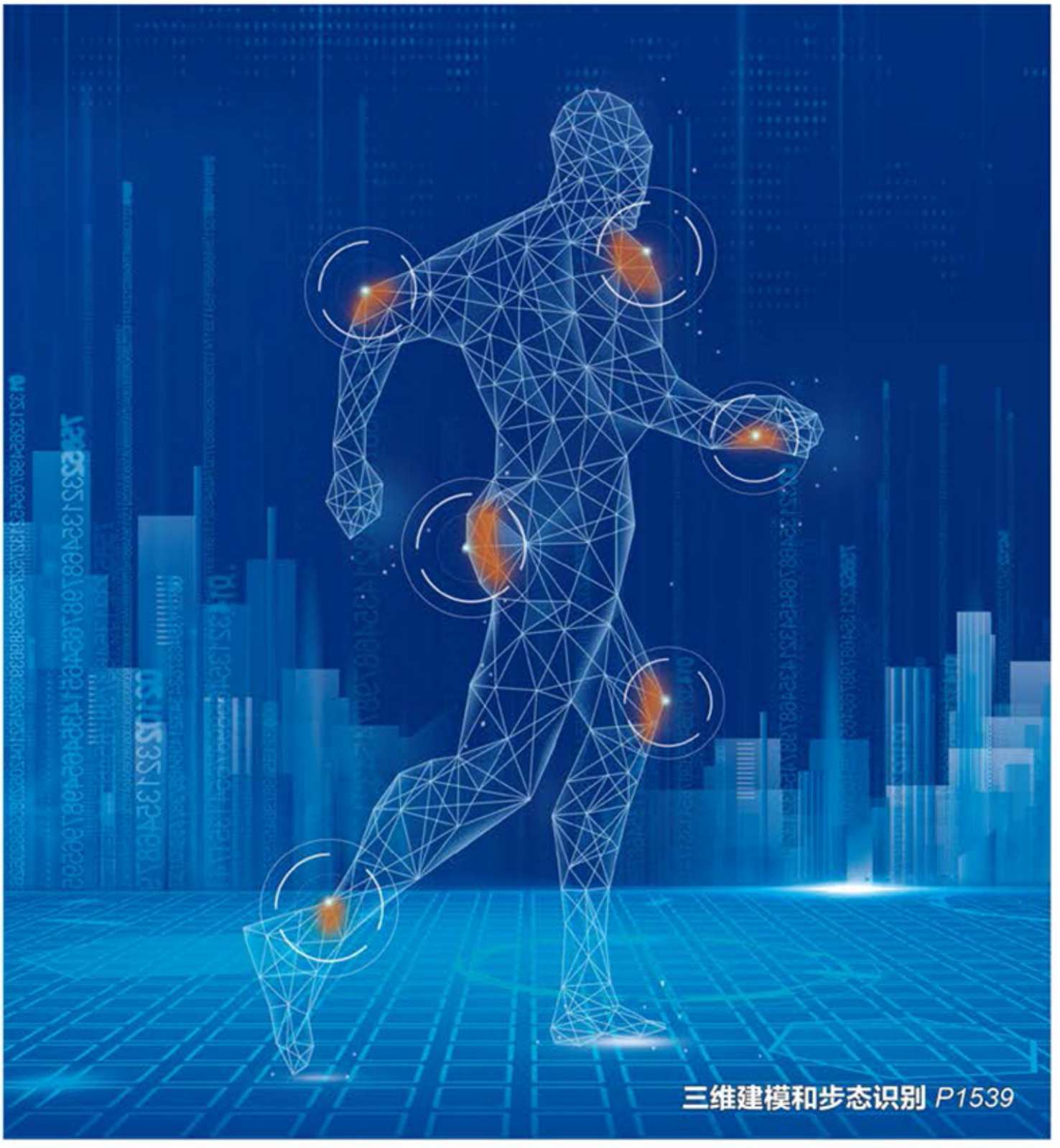
JOURNAL OF IMAGE AND GRAPHICS

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

中国图象图形学报

2020
08
VOL.25

ISSN1006-8961
CN11-3758/TB



三维建模和步态识别 *P1539*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第8期（总第292期）
2020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

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

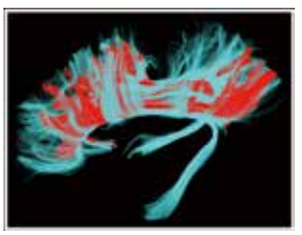
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

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

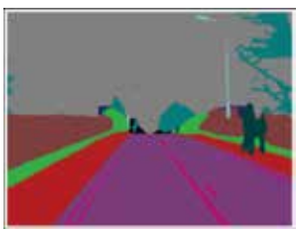
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年8月 第25卷 第8期（总第292期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

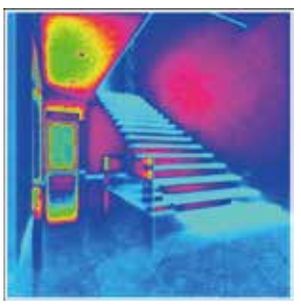
目次



神经纤维跟踪算法研究进展(第1513页)



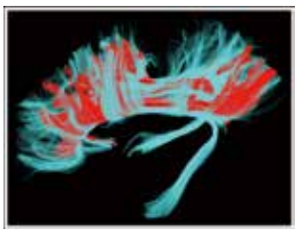
结合注意力机制的双路径语义分割(第1627页)



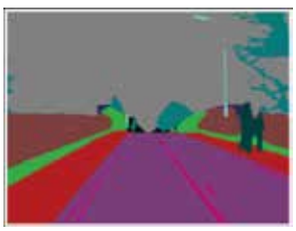
突变策略下多通路Metropolis光照传播(第1658页)

Volume 25, Number 8
Published August 16, 2020

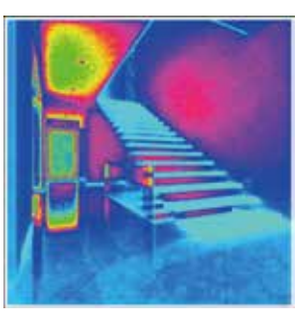
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Research progress of neural fiber tracking(P1513)



Two-path semantic segmentation algorithm combining attention mechanism(P1627)



Improved multiplexed Metropolis light transport based on fusion mutation strategy(P1658)

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2020)08-1627-10

论文引用格式:Zhai P B, Yang H, Song T T, Yu K, Ma L X and Huang X S. 2020. Two-path semantic segmentation algorithm combining attention mechanism. Journal of Image and Graphics,25(08):1627-1636(翟鹏博,杨浩,宋婷婷,余亢,马龙祥,黄向生. 2020. 结合注意力机制的双路径语义分割. 中国图象图形学报,25(08):1627-1636)[DOI:10.11834/jig.190533]

结合注意力机制的双路径语义分割

翟鹏博[1,2],杨浩[1],宋婷婷[1],余亢[1],马龙祥[1,2],黄向生[3]

1. 中国科学院微电子研究所,北京　100029;2. 中国科学院大学,北京　100049;3. 中国科学院自动化研究所,北京　100190

摘　要:**目的**　针对现有语义分割算法存在的因池化操作造成分辨率降低导致的分割结果变差、忽视特征图不同通道和位置特征的区别以及特征图融合时方法简单,没有考虑到不同感受视野特征区别等问题,设计了一种基于膨胀卷积和注意力机制的语义分割算法。**方法**　主要包括两条路径:空间信息路径使用膨胀卷积,采用较小的下采样倍数以保持图像的分辨率,获得图像的细节信息;语义信息路径使用 ResNet(residual network)采集特征以获得较大的感受视野,引入注意力机制模块为特征图的不同部分分配权重,使得精度损失降低。设计特征融合模块为两条路径获得的不同感受视野的特征图分配权重,并将其融合到一起,得到最后的分割结果。**结果**　为证实结果的有效性,在 Camvid 和 Cityscapes 数据集上进行验证,使用平均交并比(mean intersection over union,MIoU)和精确度(precision)作为度量标准。结果显示,在 Camvid 数据集上,MIoU 和精确度分别为 69.47% 和 92.32%,比性能第 2 的模型分别提高了 1.3% 和 3.09%。在 Cityscapes 数据集上,MIoU 和精确度分别为 78.48% 和 93.83%,比性能第 2 的模型分别提高了 1.16% 和 3.60%。**结论**　本文采用膨胀卷积和注意力机制模块,在保证感受视野并且提高分辨率的同时,弥补了下采样带来的精度损失,能够更好地指导模型学习,且提出的特征融合模块可以更好地融合不同感受视野的特征。

关键词:语义分割;卷积神经网络;感受视野;膨胀卷积;注意力机制;特征融合

Two-path semantic segmentation algorithm combining attention mechanism

Zhai Pengbo[1,2], Yang Hao[1], Song Tingting[1], Yu Kang[1], Ma Longxiang[1,2], Huang Xiangsheng[3]

1. Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: **Objective** Semantic segmentation is a fundamental problem in the field of computer vision where the category of each pixel in the image needs to be labeled. Traditional semantic segmentation uses a manual design to extract image features or edge information and uses the extracted features to obtain the final segmentation map through machine learning algorithms. With the rise of deep convolutional networks, some scholars have applied convolutional neural networks to semantic segmentation to improve its accuracy. However, the existing semantic segmentation algorithms face some problems. First, a contradiction is observed between the perceptual field of view and the resolution. When obtaining a large perceptual field of

收稿日期:2019-10-17;**修回日期**:2020-01-20;**预印本日期**:2020-01-27

基金项目:国家科技重大专项(2018ZX01031201);国家自然科学基金项目(61573356)

Supported by: National Science and Technology Major Project of China (2018ZX01031201); National Natural Science Foundation of China (61573356)

view, the resolution is often reduced, thereby generating poor segmentation results. Second, each channel of the feature map represents a feature. Objects at different positions should pay attention to different features, but the existing algorithms often ignore such difference. Third, the existing algorithms often use simple cascading or addition when fusing feature maps of different perception fields. For objects of different sizes, the various characteristics of the visual field are not given the same amount of importance. **Method** To address the aforementioned problems, this paper designs a semantic segmentation algorithm based on the dilated convolution and attention mechanism. First, two paths are used to collect the features. One of these paths uses dilated convolution to collect spatial information. This path initially utilizes a convolution kernel with a step size of 2 for fast downsampling and then experimentally selects a feature map with a downsampling factor of 4. Afterward, an expansion convolution with expansion ratios of 1, 2, and 4 is selected, and the obtained results are cascaded as the final feature map output while maintaining the resolution to obtain high-resolution feature maps. Meanwhile, the second path uses ResNet to collect features and obtain an expanded field of view. The ResNet network uses a residual structure that can well integrate shallow features with deep ones while simultaneously avoiding the difficulty of convergence during training. Feature maps are extracted with 16 and 32 times downsampling size from the ResNet network to expand the perception field. The attention mechanism module is then applied to assign weights to each part of feature maps to reflect their specificity. The attention mechanism is divided into the spatial attention, channel attention, and pyramid attention mechanisms. The spatial attention mechanism obtains the weight of each position on the feature map by extracting the information of different channels. Global maximization and average pooling are initially utilized in the spatial dimension. In this way, the characteristic information of different channels at each position is synthesized and used as the initial attention weight. Afterward, the initial spatial attention weights are processed by convolutional layers and batch normalization operations to further learn the feature information and are outputted as the final spatial attention weights. The channel attention mechanism obtains the weight of different channels by extracting the information of each channel. Maximum and average global pooling are initially applied in the channel dimension. Through these two pooling operations, the feature information on each channel is extracted, and the number of parameters is reduced. The output is treated as the initial attention weight, and a 3×3 convolution layer is used to further learn the attention weight, enhance the salient features, and suppress the irrelevant features. The pyramid attention mechanism uses a convolution of different perception fields and obtains the weights for each position in different channels. The initial feature map is operated by using convolution kernels with sizes 1×1, 3×3, and 5×5 to fully extract features at different levels of the feature map, and the output is treated as the initial attention weight. The feature maps are then multiplied to obtain the final state. Weights are assigned to the channel of each position on the feature map to effectively make the salient features get more attention and to suppress the irrelevant features. The three weights obtained are then merged to obtain the final attention weight. A feature fusion module is designed to fuse the features of different perception fields of the two paths in order to ascribe varying degrees of importance to the features of different perception fields. In the feature fusion module, the feature maps obtained from the two paths are upsampled to the same size by using a quadratic linear difference. Given that the features of different perception fields have varying degrees of importance, the obtained feature maps are initially cascaded, global pooling is applied to obtain the feature weights of each channel, and weights are adaptively assigned to each feature map. These maps are then upsampled to the original image size to obtain the final segmentation result. **Result** The results were validated on the Camvid and Cityscapes datasets with mean intersection over union (MIoU) and precision as metrics. The proposed model achieved 69.47% MIoU and 92.32% precision on Camvid, which were 1.3% and 3.09% higher than those of the model with the second-highest performance. Meanwhile, on Cityscapes, the proposed model achieved 78.48% MIoU and 93.83% precision, which were 1.16% and 3.60% higher than those of the model with the second-highest performance. The proposed algorithm also outperforms the other algorithms in terms of visual effects. **Conclusion** This paper designs a new semantic segmentation algorithm that uses dilated convolution and ResNet to collect image features and obtains different features of the image in large and small perception fields. Using dilated convolution improves the resolution while ensuring the perceived field of view. An attention mechanism module is also designed to assign weights to each feature map and to facilitate model learning. A feature fusion module is then designed to fuse the features of different perception fields and reflect their specificity. Experiments show that the proposed algorithm outperforms the other algorithms in terms of accuracy and shows a certain application value.

Key words: semantic segmentation; convolutional neural network; perception field; dilated convolution; attention mechanism; feature fusion

0 引 言

语义分割是计算机视觉领域中的一项基本任务,要求根据上下文信息以及空间结构信息对图像中的每一个像素标明类别(姜枫 等,2017)。图像的语义分割在各领域都有重要应用。在健康医疗领域,语义分割技术可以从患者的 X 光或 B 超图像中精准地提取病变部位,提高诊断效率;在自动驾驶领域,通过语义分割技术可以获得障碍物的具体信息,帮助汽车躲避障碍;在安防领域,语义分割技术可以从安检图像中准确分割危险物品,实现危险品的自动检测,提高安检效率。随着深度学习技术的发展,以及 GPU(graphics processing unit)运算能力的提高,训练大规模、深层次的神经网络已经成为语义分割领域的主流模式,并取得了极大进展。

AlexNet(Krizhevsky 等,2012)是第一个被提出的深度卷积神经网络模型,在 2012 年的 ILSVRC(ImageNet large scale visual recognition challenge)竞赛上,AlexNet 凭借 84.6% 的 top-5 准确率取得第 1 名的成绩,而第 2 名的准确率仅为 73.8% 。AlexNet 的出现引起了众多学者的关注,揭开了深度学习的大幕。Long 等人(2015)提出全卷积神经网络(fully convolutional networks,FCN),通过将原来分类网络 VGG16(visual geometry group 16-layer net)(Simonyan 和 Zisserman,2014)最后部分的全连接层改为卷积层,实现了将深层次粗糙的网络层的语义信息与浅层精细的网络层的表层信息结合的目的,从而实现精确分割。Badrinarayanan 等人(2015)提出语义分割网络 SegNet 在解码器中使用反池化的方式对产生的特征图进行上采样,使得一些细微边缘信息保存完整。同时在特征提取部分,SegNet 使用与 FCN 相同的全卷积网络,大幅减少了参数数量。Ronneberger 等人(2015)提出的 U-Net 采用 U 型结构,将编码器阶段的特征图拼接到解码器阶段对应大小的特征图,将空间语义信息与像素信息进行融合,允许解码器学习在下采样中丢失的信息,提升了分割效果。Chen 等人(2014)提出的 DeepLabV1 采用空洞卷积,在减小特征图尺寸的同时保持分辨率不变,并且在网络的最后使用条件随机场(conditional random field,CRF)模型,恢复边缘的信息,实现了准确定位。Paszke 等人(2016)针对低延迟高速度的任务提出 ENet,大幅提高了模型速度,比原来的网络模型快了 18 倍,比 FLOPs(floating-point operations per second)少了 79 倍,参数数量少了 79 倍,准确率与原来模型接近。Zhao 等人(2017)提出了 PSPNet(pyramid scene parsing network),采用膨胀卷积(Yu 和 Koltun,2015)修改基础的残差网络 ResNet(residual network)(He 等,2016)架构,经过最初的池化层,在整个编码器网络中采用相同的分辨率。另外,PSPNet 将额外的损失函数加入整体学习,取得了最优的结果。马震环等人(2019)提出了一种新型的解码器结构,首先将深层特征与浅层特征级联起来,然后通过自注意力机制,得到最后的解码特征,减少了解码上采样带来的精度损失。程晓悦等人(2019)提出了一种新型的语义分割算法,将密集层结构添加到网络中,并且采用分组卷积加快计算速度,同时添加注意力机制改善分割效果。

尽管语义分割领域取得了很多成就,但是仍然存在以下问题:1)采用池化、下采样等操作扩大感受视野时会带来分辨率降低问题,导致分割精度降低;2)浅层特征与深层特征直接级联或相加的融合方法过于简单,忽视了不同特征的特异性,带来分割精度损失;3)在下采样—上采样过程中造成信息丢失,导致分割准确率降低。

针对以上问题,本文设计了一种语义分割算法,采用膨胀卷积减少下采样过程中分辨率降低的问题,设计注意力机制减轻下采样过程中的信息丢失,并指导模型更好地学习,同时采用专门的特征融合模块更好地将浅层特征与深层特征进行融合。

1 网络模块设计

1.1 整体网络设计

本文网络的整体架构没有采用传统的 U 型结构,而是采用两条路径下采样(Yu 等,2018),如图 1 所示。空间信息路径采用膨胀卷积,获得较多的细节信息。语义信息路径采用ResNet提取特征,获得

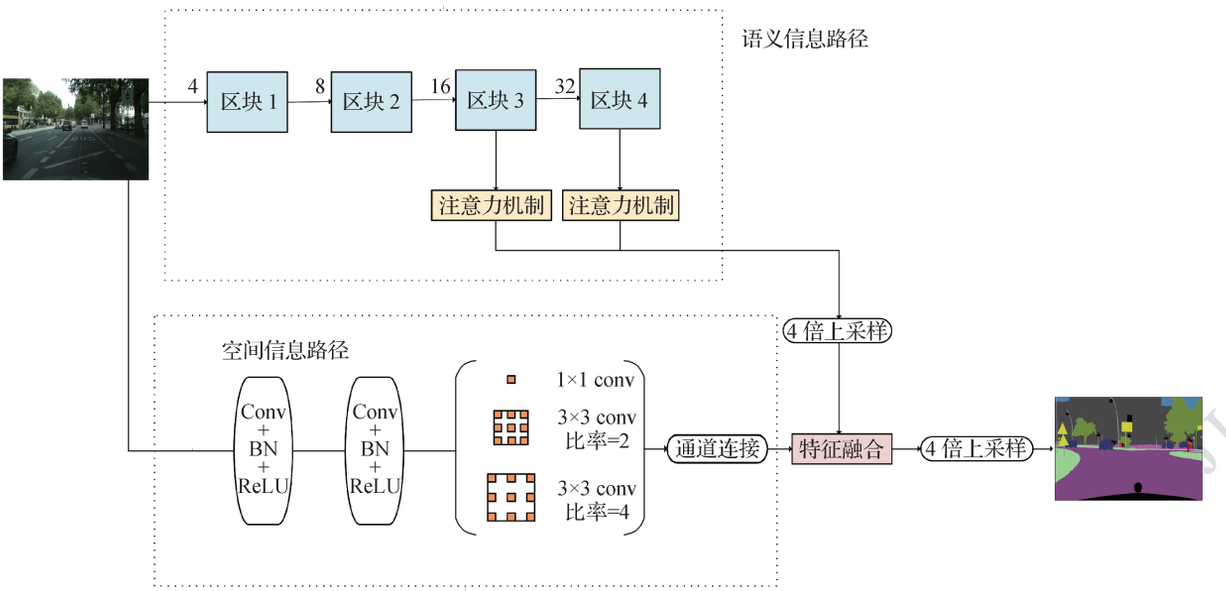


图 1　整体架构示意图

Fig. 1　Structure of model

较大的感受视野,同时仿照人类视觉机制设计注意力模块(Vaswani 等,2017)用以提高准确率。最后设计了特征融合模块,将采集到的不同感受视野的特征融合到一起,获得更佳的结果。

1.2　膨胀卷积

为了在保持一定感受视野的同时提高分辨率,本文算法在空间信息路径使用了膨胀卷积。普通卷积的表达式为

$$O(x,y)\cdot H(x,y)=\sum_{i=0}^{w}\sum_{j=0}^{h}H(i,j)\cdot O(x-i,y-j) \tag{1}$$

式中,$O(x,y)$ 是原始图像在点 (x,y) 处的像素值,$H(x,y)$ 是与其相乘的卷积核,大小为 $w\times h$。

膨胀卷积计算为

$$O(x,y)\cdot H'(x,y)=\sum_{i=0}^{w}\sum_{j=0}^{h}H'(i,j)\cdot O(x-l\times i,y-l\times j) \tag{2}$$

式中,l 为膨胀因子,$H'(x,y)$ 为膨胀卷积核。

从式(1)和式(2)可以看出,膨胀卷积实质上就是对卷积核进行了 0 填充,这样做可以在增加卷积核感受视野的同时保留原始的像素信息,增大了分辨率。若卷积核的尺寸为 k,膨胀率为 l,则膨胀卷积的实际有效尺寸为 $k+(k-1)\times(l-1)$。与相同大小的普通卷积相比,膨胀卷积不仅扩大了感受视野,还保持了与普通卷积相同的分辨率,示意图如图 2 所示。

1.3　注意力机制

为了弥补下采样造成的细节丢失,更好地指导模型训练,本文采用了一种注意力机制,通过对特征图进行加权处理达到增强目标特征并且抑制背景的目的(Gilra 和 Gerstner,2017)。注意力机制主要由空间注意力机制、通道注意力机制和金字塔注意力机制组成,如图 3 所示。

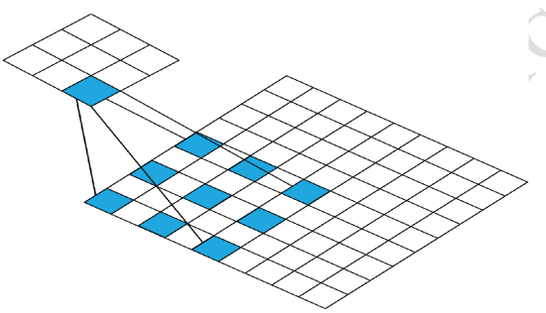


图 2　膨胀卷积示意图

Fig. 2　Dilated convolution

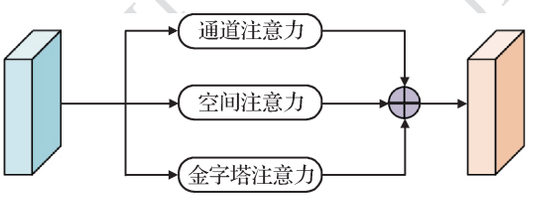


图 3　注意力机制架构

Fig. 3　Structure of attention

1.3.1　空间注意力机制

空间注意力机制主要是仿照人的视觉机制提出

的。人眼在看到一幅图像时，会自动给予关键位置更大的注意力，例如看到兔子时，会更多地关注兔子的耳朵。因此，图像特征图的不同部分应该有不同的权重（Zhang 等，2018）。本文提出的空间注意力机制如图 4 所示。

得到的特征图在通道维度上使用最大值全局池化和平均值全局池化，将特征图上不同位置的特征信息进行比较提取，得到每一位置上的特征权重。

最大值全局池化的表达式为

$$O(i,j) = \max(I_m(i,j)),\ m \in (1,c) \tag{3}$$

式中，$I_m(i,j)$ 为第 m 个通道上的特征图在 (i,j) 位置的特征值，$O(i,j)$ 位置为经过最大值池化后的特征图，c 为通道数。

平均值全局池化的表达式为

$$P(i,j) = \frac{1}{d}\sum_{n=1}^{d} T_n(i,j) \tag{4}$$

式中，$T_n(i,j)$ 为第 n 个通道特征图在 (i,j) 位置的特征值，$P(i,j)$ 为经过平均值池化后的特征图，d 为通道数。

采用通道连接的方式将这两幅特征图在通道维度连接起来。为了使采用两种方式提取的信息得到更好融合，采用大小为 1 × 1 卷积核进行学习，将得到的特征图通过 sigmoid 函数计算出最后的注意力权重，以避免得到的权重系数过大，出现错误。

sigmoid 函数的公式为

$$S(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

式中，$S(x)$ 为输出的响应，x 为输入。

空间注意力机制可以有效提取特征图上每个位置的显著性信息，并据此为每个位置的特征值分配一个权重，有效提取目标的主要特征，并抑制背景部分。

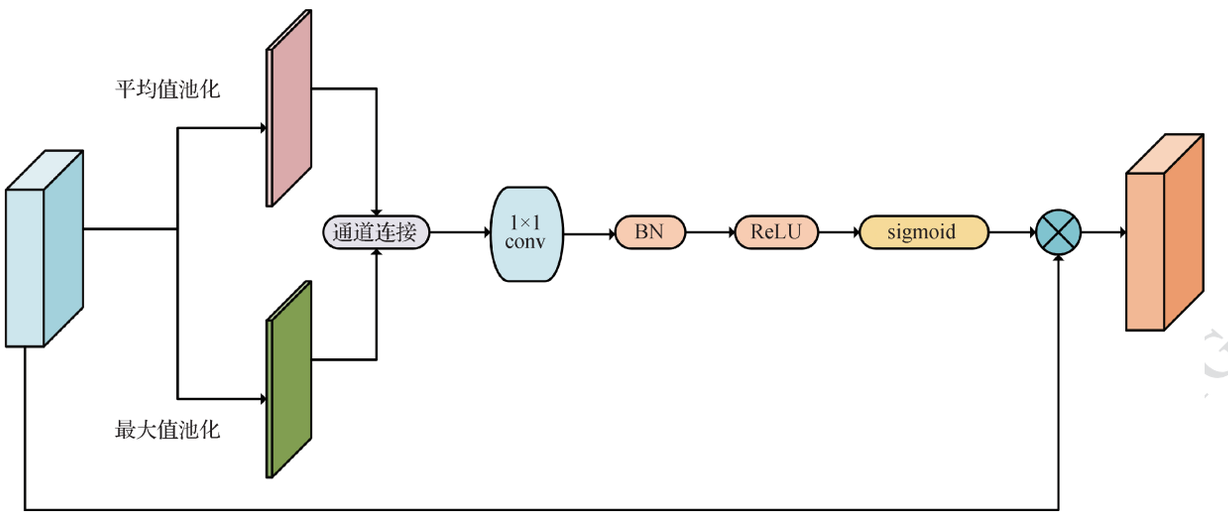


图 4　空间注意力示意图

Fig. 4　Spatial attention

1.3.2　通道注意力机制

卷积神经网络提取的特征图的每个通道都代表一种图像特征，例如纹理，形状等。在图像中，每种特征蕴含的信息不同，对图像分割的贡献程度也不同。因此应当对每种不同特征给予不同关注度（Woo 等，2018），分配不同权重。为此，设计通道注意力机制为特征分配权重，使得网络集中关注重要特征，如图 5 所示。

将特征图通过全局最大池化和全局平均池化，在空间维度上对通道信息进行建模，得到每个通道的特征信息。

最大池化的计算过程为

$$O_c = \max(I_c(i,j)) \tag{6}$$

式中，$i \in (1,h)$，$j \in (1,w)$，c 代表第 c 个特征图，O_c 代表特征图经过最大池化后的输出。

平均池化的计算过程为

$$P_m = \frac{1}{w}\frac{1}{h}\sum_{i=1}^{h}\sum_{j=1}^{w} I_m(i,j) \tag{7}$$

式中，$i \in (1,h)$，$j \in (1,w)$，m 代表第 m 个特征图，P_m 代表特征图经过平均池化后的输出。

为了将通过全局池化得到的特征图进行融合且仅增加少量计算，将两个特征图分别经过大小为

1 × 1 的卷积核,然后经过 BN 和 ReLU 层增添非线性成分,使得模型更好地进行拟合,并且能够在一定程度上防止产生过拟合现象。最后将得到的两个特征图相加,进行融合,并通过 sigmoid 函数得到最终权重。

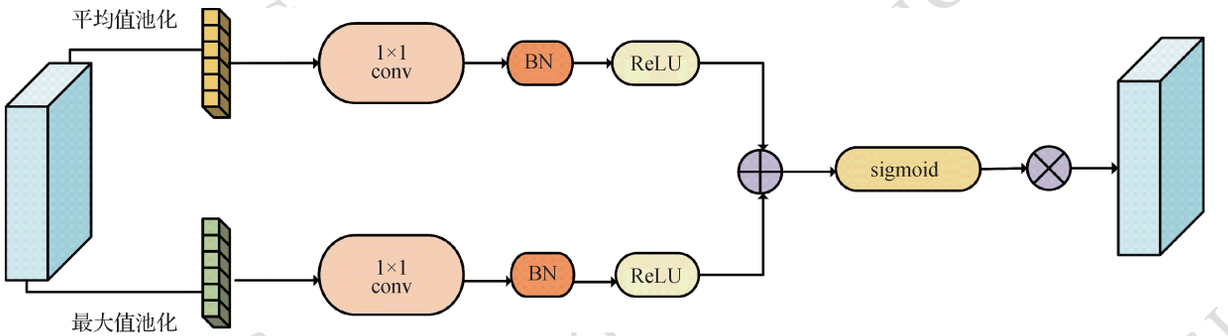


图 5　通道注意力示意图

Fig. 5　Channel attention

1.3.3　金字塔注意力机制

人类视觉在辨别物体时往往综合多种信息。例如,区别兔子和猫时更关注耳朵的形状特征,区别熊猫与狗熊时,更关注颜色特征。由此可以看出,特征图上不同位置的不同特征获得的关注度也应当不同。为了能够更好地获得图像不同位置的不同信息,提出一种金字塔式的注意力模型,通过提取图像不同感受视野的特征图,获取不同感受野下的图像信息,将这些信息融合,获得最后的权重系数,如图 6 所示。

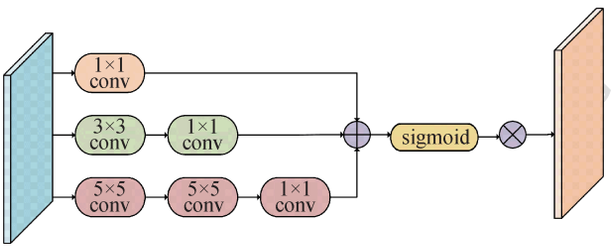


图 6　金字塔注意力示意图

Fig. 6　Pyramid attention

将特征图分别通过 3 × 3 、1 × 1 、5 × 5 大小的卷积核,由于在分辨率较小的高层次特征图使用这些卷积核,所以并不会带来太多的计算负担。然后将特征图通过不同大小的卷积核,采集不同尺度下的特征信息,这样可以更好地集成上下文的信息,获取不同尺度下的特征信息。通过这些不同大小的卷积核,获得了不同感受视野下的特征信息,基于这些特征信息,通过 1 × 1 的卷积核获得不同位置的特征权重。最后将得到的特征图相加,进行融合,通过 sigmoid 函数得到最终权重。

1.4　特征融合模块

不同类别的物体,感受视野大小的重要性不同。对较大的物体,大感受视野获取的特征比较重要,而对较小的物体,较大感受视野的特征会采集过多的周边信息,引来误差。传统的特征融合方法一般是级联或相加,这样的简单做法没有考虑不同特征图的感受视野不同,忽略了特征之间的特异性。对此,本文设计的特征融合模块为不同感受视野的特征图分配不同权重,实现了更好的特征融合,如图 7 所示。

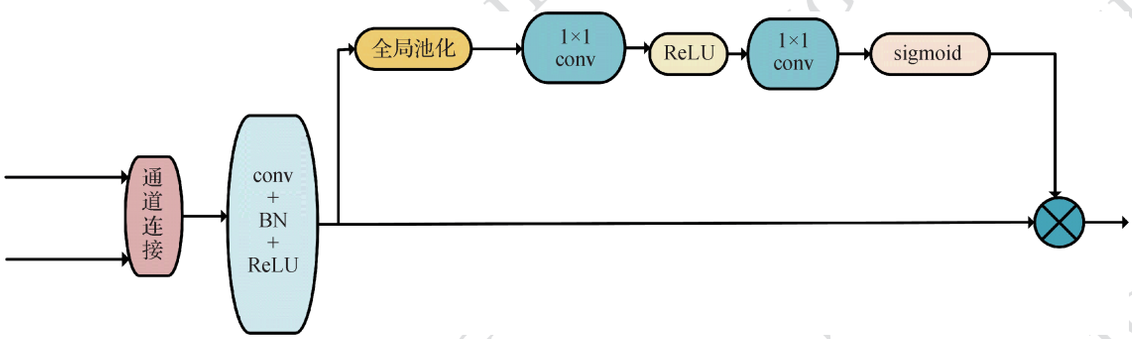


图 7　特征融合模块示意图

Fig. 7　Feature fusion

首先，将输入的两幅特征图在通道维度上级联起来。其次，将级联后的特征图通过大小为 3 × 3 的卷积核，实现特征图信息的初步融合，并将得到的特征图进行全局池化操作，提取每幅特征图的信息。然后，将得到的特征图通过大小为 1 × 1 的卷积核，让网络根据每幅特征图的整体信息学习权重。最后通过 sigmoid 函数得到最后的权重，并与原始特征图相乘。通过特征融合模块，为不同感受视野下的特征图分配了权重，使得不同感受视野下的特征特异性得到了体现，让特征更好地进行融合。

2 实验结果与分析

实验数据采用语义分割领域常见的 Camvid（Brostow 等，2009）和 Cityscapes（Cordts 等，2016）数据集。Camvid 数据集是剑桥大学主推的数据库，是第 1 个具有对象类语义标签的数据库，包含 32 类别，800 幅图像。Cityscapes 数据集是奔驰公司发布的数据库，包括 50 个城市的 30 类物品的标注，其中训练集图像 2 795 幅，测试集图像 500 幅。两个数据集的示例和本文算法在两个数据集上的结果展示如图 8 和图 9 所示。

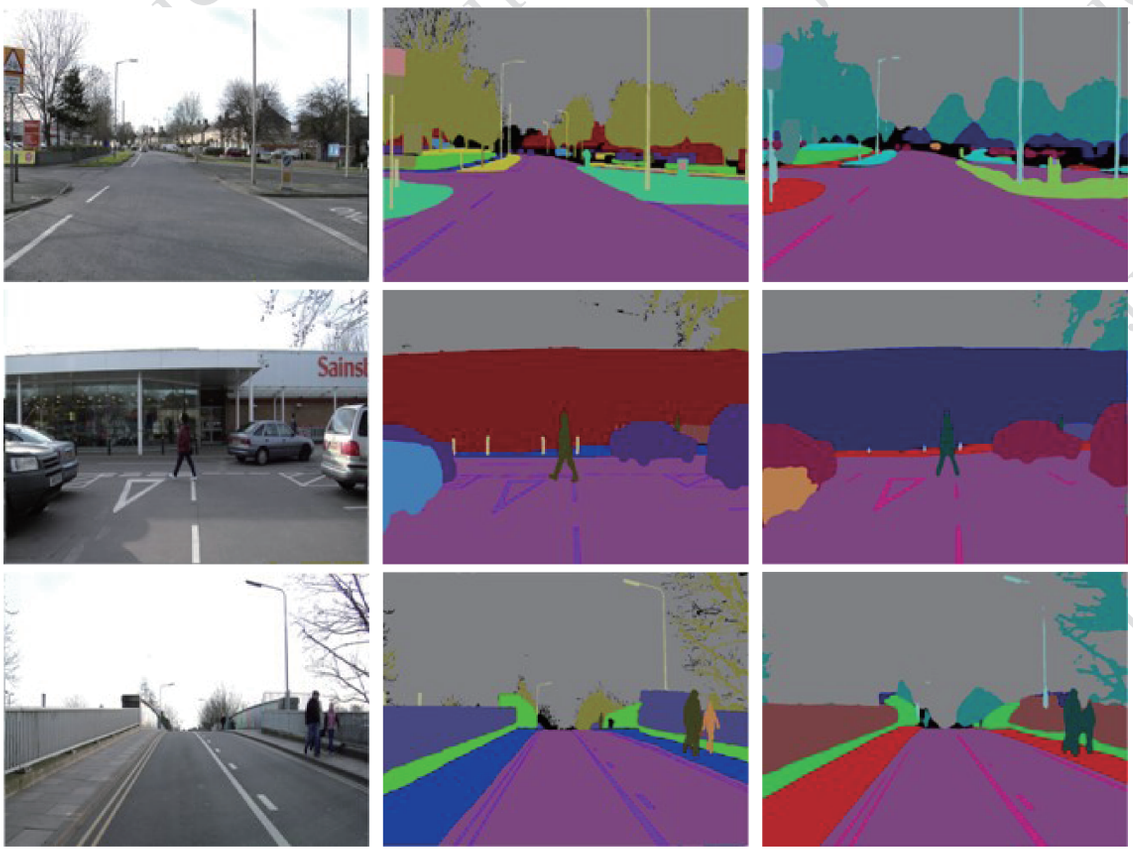
2.1 评价标准

使用平均交并比（mean intersection over union，MIoU）和精确度（precision）作为度量标准。

MIoU 是语义分割领域常用的度量标准，主要是计算两个集合的交集与并集之比。使用时需在每个类别内部分别计算 MIoU，然后再计算整体的平均值。MIoU 越大，模型的效果越好，具体计算为

$$M = \frac{1}{m+1}\sum_{i=0}^{m}\frac{p_{ii}}{\sum_{j=0}^{m}p_{ij}+\sum_{j=0}^{m}p_{ji}-p_{ii}} \tag{8}$$

式中，m 代表类别的数目，p_{ii} 代表某个像素实际为 i 并且结果预测为 i 的数目，p_{ij} 代表某个像素的实际类别为 j 但是结果预测为 i 的数目，p_{ji} 代表某个像素的实际类别为 i 但是结果预测为 j 的数目。

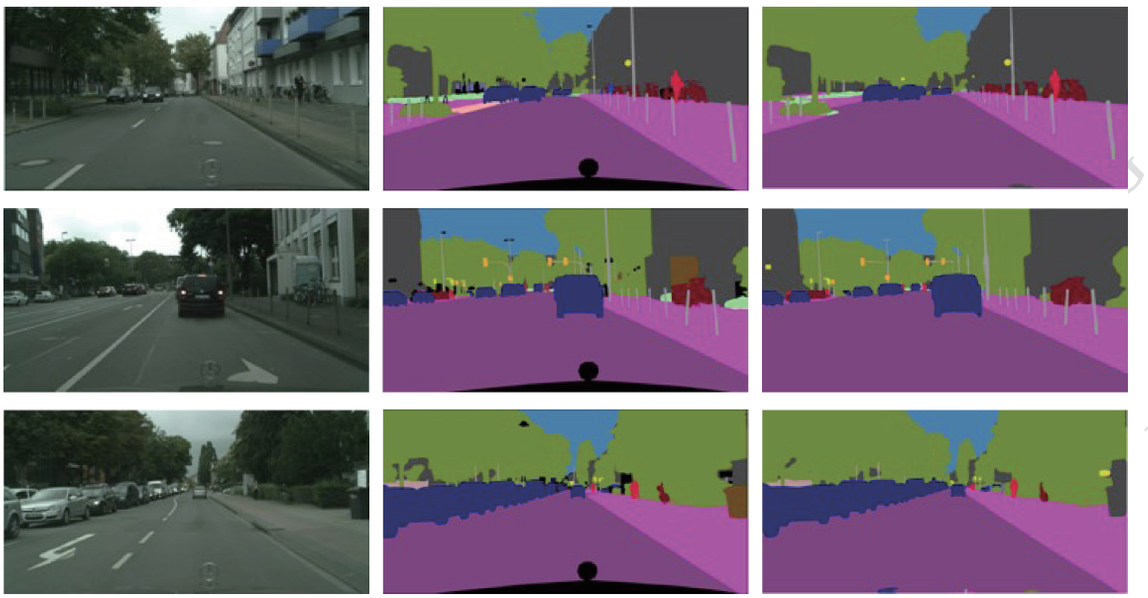


(a) 数据原图　(b) 标注图像　(c) 预测图像

图 8　Camvid 数据集示例及本文算法在该数据集上的测试结果

Fig. 8　Camvid dataset example and the test results of our method on this dataset

((a) original images; (b) label images; (c) predict images)



(a) 数据原图　(b) 标注图像　(c) 预测图像

图 9　Cityscapes 示例及本文算法在该数据集上的测试结果

Fig. 9　Cityacapes dataset example and the test results of our method on this dataset

((a) original images; (b) label images; (c) predict images)

精确度是计算标记正确的像素占所有像素的比例,即

$$P = \frac{i}{s} \tag{9}$$

式中,i 为分割结果中标记正确的像素总数,s 是像素总数目。精确度 P 的结果越大,模型的精度越高,越准确。

2.2　结果分析

实验采用的框架为 Pytorch1.0,CUDA 版本为 10.1,CUDNN 版本为 7.4,工作站配置了 48 个 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20 GHz,内存 128 GB,4 张 GTX 1080TI 显卡。

对输入数据进行增强处理,将输入图像按一定比例做水平翻转、随机裁剪和随机明暗亮度变化,最后得到的图像大小为 1 024×1 024 像素。使用 ResNet-18 作为基础网络,加载在 ImageNet 上预训练的权重。网络采用带动量的随机梯度下降为优化器,动量为 0.9,初始学习率为 0.01,衰减率为 0.999 5,训练 80 000 次。

本文网络架构首先在空间路径上选取 2、4、8、16、32 倍等不同的下采样倍数进行测试,实验结果表明,4 倍下采样的 MIoU 和精确度分别为 77.28% 和 92.28%,优于其他下采样倍数。然后选取不同比率的膨胀卷积进行测试,结果显示,比率为 2 和 4 时的 MIoU 和精确度分别为 77.83% 和 92.86%,优于其他比率的实验。以上实验均在 Cityscapes 数据集上完成,具体实验结果见表 1 和表 2。同时,为了证明注意力机制模块的有效性,在 Cityscapes 数据集上对网络添加注意力模块和未添加注意力模块进行测试。实验证明,添加注意力模块后,网络性能得到了提高。实验结果如表 3 所示。

表 1　不同下采样倍数的实验结果

Table 1　Experimental results of different downsampling times

下采样倍数	MIoU/%	精确度/%
2	76.43	90.12
4	**77.28**	**92.28**
8	74.65	86.75
16	71.20	83.65
32	69.32	82.48

注:加粗字体表示最优结果。

为了证明设计的网络对比其他算法的优势,在相同硬件设备与软件环境下,在 Camvid 和 Cityscapes 数据集上分别进行实验。训练时对输入图像均按一定比例做水平翻转、随机裁剪、随机明暗亮度变化,各算法的最优结果如表4所示。可以看出,

表 2　不同比率的膨胀卷积的实验结果

Table 2　Experimental results of different ratio for expansion convolution

膨胀卷积比率	MIoU/%	精确度/%
2，4	**77.83**	**92.86**
4，8	77.23	90.28
8，16	76.65	88.42
16，24	74.21	86.63
24，32	73.14	85.25

注：加粗字体表示最优结果。

表 3　添加及未添加注意力机制的实验结果

Table 3　Experimental results of adding attention mechanism or not

注意力机制	MIoU/%	精确度/%
未添加	77.83	91.83
添加	**78.51**	**93.91**

注：加粗字体表示最优结果。

基于本文算法模型的 MIoU 和精确度在 Camvid 数据集上分别为 69.47% 和 92.32%，在 Cityscapes 数据集上分别为 78.48% 和 93.83%，均优于其他算法。

表 4　不同算法实验结果比较

Table 4　Comparison of experimental results for different algorithms

方法	Camvid 数据集		Cityscapes 数据集	
	MIoU/%	精确度/%	MIoU/%	精确度/%
BiseNet	68.17	89.23	77.32	90.23
DeepLabV3(Chen 等,2017)	60.35	84.32	72.32	84.21
PSPNet	60.57	85.32	74.52	86.75
AdapNet(Valada 等,2017)	58.87	83.32	69.34	82.54
ENet	51.62	80.28	64.43	80.32
SegNet	55.48	82.34	66.25	81.24
DeepLabV1	58.45	83.12	70.4	83.29
本文	**69.47**	**92.32**	**78.48**	**93.83**

注：加粗字体表示最优结果。

3　结　论

本文提出了一种基于膨胀卷积和注意力机制采用双路径结构的语义分割算法，通过膨胀卷积，在下采样获得较大感受视野的同时，保证了一定的分辨率。同时网络中添加了注意力机制，使得网络可以自适应学习权重，提高了分割精度。最后设计了特征融合模块，更好地将不同感受视野的特征进行融合。将本文算法在 Camvid 和 Cityscapes 数据集上进行测试，并将最后结果与多种算法比较，本文算法取得了更优秀的结果。但是由于小目标物体本身的像素数目较少，从中获取的特征信息有限，无法很好地对目标整体进行建模，存在小目标物体分割精度不高、边缘分割不够精细的问题。接下来将进一步改进注意力机制模块，探究不同位置和通道特征之间的特异性与共同点，提取更加有效的特征信息，提升算法在小目标物体上的分割准确率。

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 labelling[EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1505.07293.pdf

Brostow G J, Fauqueur J, and Cipolla R. 2009. Semantic object classes in video: a high-definition ground truth database. Pattern Recognition Letters, 30(2): 88-97 [DOI: 10.1016/j.patrec.2008.04.005]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A. 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFS[EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1412.7062.pdf

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

Cheng X Y, Zhao L Z, Hu Q and Shi J P. 2019. Fast semantic segmen-

tation based on dense layer and attention mechanism [EB/OL]. Computer Engineering: 1-7 [2020-02-05]. (程晓悦,赵龙章,胡穹,史家鹏. 2019. 基于密集层和注意力机制的快速语义分割[EB/OL]. 计算机工程:1-7[2020-02-05]. https://doi-org-443.webvpn.las.ac.cn/10.19678/j.issn.1000-3428.0054245.)

Cordts M, Omran M, Ramos S and Ramos S. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 3213-3223 [DOI:10.1109/cvpr.2016.350]

Gilra A and Gerstner W. 2017. Non-linear motor control by local learning in spiking neural networks[EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1712.10158.pdf

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 770-778 [DOI: 10.1109/CVPR.2016.90]

Jiang F, Gu Q, Hao H Z, Li N, Guo Y W and Chen D X. 2017. Summary of content-based image segmentation methods. Journal of Software, 28 (1): 160-183 (姜枫, 顾庆, 郝慧珍, 李娜, 郭延文, 陈道蓄. 2017 基于内容的图像分割方法综述. 软件学报, 28(1): 160-183) [DOI: 10.13328/j.cnki.jos.005136]

Krizhevsky A, Sutskever I and Hinton G E. 2012. Imagenet classification with deep convolutional neural networks//Advances in Neural Information Processing Systems. Nevada, USA: NIPS, 1097-1105 [DOI: 10.1145/3065386]

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]

Ma Z H, Gao H J and Lei T. 2019. Semantic segmentation algorithm based on enhanced feature fusion decoder [EB/OL]. Computer Engineering: 1-6 [2020-02-05] (马震环,高洪举,雷涛. 2019. 基于增强特征融合解码器的语义分割算法[EB/OL]. 计算机工程:1-6[2020-02-05]. https://doi-org-443.webvpn.las.ac.cn/10.19678/j.issn.1000-3428.0054964.)

Paszke A, Chaurasia A, Kim S, Kim S and Culurciello E. 2016. Enet: A deep neural network architecture for real-time semantic segmentation [EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1606.02147.pdf

Ronneberger O, Fischer P and Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation//Proceedings of 2015 International Conference on Medical image computing and computer-assisted intervention. Munich, Germany: MICCAI, 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1409.1556.pdf

Valada A, Vertens J, Dhall A, and Burgard W. 2017. Adapnet: Adaptive semantic segmentation in adverse environmental conditions//Proceedings of 2017 IEEE International Conference on Robotics and Automation. Marina Bay Sands Singapore: IEEE: 4644-4651 [DOI:10.1109/icra.2017.7989540]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J and Jones L. 2017. Attention is all you need//Advances in Neural Information Processing Systems. Nevada, USA: NIPS, 5998-6008

Woo S, Park J, Lee J Y and Kweon I S. 2018. Cbam: convolutional block attention module//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: ECCV, 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Yu C, Wang J, Peng C, Gao C, Yu G and Sang N. 2018. Bisenet: bilateral segmentation network for real-time semantic segmentation//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: ECCV: 325-341 [DOI:10.1007/978-3-030-01261-8_20]

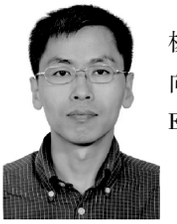
Yu F, and Koltun V. 2015. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1511.07122.pdf

Zhang H, Goodfellow I, Metaxas D and Odena A. 2018. Self-attention generative adversarial networks[EB/OL]. [2019-02-05]. https://arxiv.org/pdf/1805.08318.pdf

Zhao H, Shi J, Qi X, Wang X and Jia J. 2017. Pyramid scene parsing network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2881-2890 [DOI: 10.1109/CVPR.2017.660]

作者简介

翟鹏博,1994 年生,男,硕士研究生,主要研究方向为机器学习、计算机视觉。
E-mail:zhaipengbo@ime.ac.cn



杨浩,通信作者,男,副研究员,主要研究方向为图像处理、中医设备现代化。
E-mail:yanghao@ime.ac.cn

宋婷婷,女,硕士研究生,主要研究方向为图像处理。
E-mail:songtingting@ime.ac.cn

余亢,男,硕士研究生,主要研究方向为图像处理。
E-mail:yukang@ime.ac.cn

马龙祥,男,硕士研究生,主要研究方向为计算机视觉。
E-mail:malongxiang@ime.ac.cn

黄向生,男,副研究员,主要研究方向为计算机视觉。
E-mail:huangxiangsheng@ia.ac.cn