文章编号：1672-8785(2018)04-0033-06

基于深度学习的语义分割网络

代具亭[1,2,3]　汤心溢[1,3]　刘　鹏[1,3]

（1. 中国科学院上海技术物理研究所，上海　200083；
2. 中国科学院大学，北京　100084；
3. 中国科学院红外探测与成像技术重点实验室，上海　200083）

摘　要：提出了一种基于深度学习的语义分割网络。该网络通过多孔卷积设计了一个能提取图像多尺度信息的空间金字塔模块，并通过大量实验探索了空间金字塔模块中多孔采样率和多尺度分支对于网络场景解析能力的影响。讨论了网络训练中不同超参数对于网络性能的影响。在 SUN RGB-D 数据集上的测试结果显示，与其它 state-of-the-art 的语义分割网络相比，本文设计的网络性能突出。最后，还对基于红外图像的语义分割进行了初步探索。

关键词：卷积神经网络；语义分割；多尺度；红外图像

中图分类号：TP391.41　**文献标志码**：A　**DOI**：10.3969/j.issn.1672-8785.2018.04.007

Semantic Segmentation Network Based on Deep Learning

DAI Ju-ting[1,2,3], TANG Xin-yi[1,3], LIU Peng[1,3]

(1. *Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China*;
2. *University of Chinese Academy of Sciences, Beijing 100084, China*;
3. *Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China*)

Abstract: A semantic segmentation network based on deep learning is proposed. The network designs a spatial pyramid module which can extract multi-scale information from images through Atrous convolution. It also explores the influence of Atrous convolution sampling rate and multi-scale branches on the performance of network through extensive experiment. the impact of hyperparameters on network performance during training is discussed. The test results on the SUN RGB-D dataset show that compared with other state-of-the-art semantic segmentation networks, the performance of the network we proposed is outstanding. Finally, the semantic segmentation based on infrared images is explored preliminarily.

Key words: convolutional neural network; semantic segmentation; multi-scale; infrared image

收稿日期：2018-03-06

基金项目：国家“十三五”国防预研项目(Jzx2016-0404/Y72-2)；中国科学院青年创新促进会项目(2014216)；上海市现场物证重点实验室基金项目(2017xcwzk08)

作者简介：代具亭(1989–)，男，河南安阳人，博士生，主要从事基于深度学习的语义分割、语义 SLAM 系统等方面的研究。E-mail：djting@mail.ustc.edu.cn

0 引言

图像语义分割是计算机视觉领域最关键且最具挑战性的难点之一，其可被广泛应用到无人机、无人驾驶和增强现实等领域。图像语义分割技术是指利用计算机自动区分和识别图像中的不同物体，对图像中每一个像素都进行类别标注，并在像素级别上对图像包含的场景进行理解。图像语义分割技术并不是一个孤立的技术，它是图像场景理解技术发展过程中的重要一步。

深度学习指的是通过设计具有很多隐含层的深度学习模型和创建海量的训练数据，来更好地学习数据的特征，从而提高分类或者识别任务的准确率。而卷积神经网络则是一种深度前馈的深度学习网络，其主要应用于图像分类和图像识别等任务。Long 等人[1]提出的 FCN 网络成功地将卷积神经网络应用到图像语义分割任务中。该网络使用 VGG-16[2]作为基准网络，去掉最后的分类层，并将最后的全连接层替换为卷积层，最后将预测结果上采样到输入图像的尺度。随后，更多文献[3-5]研究了将用于图像分类任务的卷积神经网络应用到图像语义分割任务的方法。

在图像语义分割任务中，一个重要的挑战在于现实场景中存在的物体具有多尺度形式，以及对于同一物体，从不同距离和角度拍摄时，该物体在图像中的尺度也不尽相同。本文在设计网络时，主要从提取图像中多尺度信息的角度出发，通过多孔卷积方法设计了一个用于提取图像多尺度信息的具有捷径恒等连接的空间金字塔模块，并通过大量实验验证了本文设计的网络结构的有效性。

1 具有空间金字塔结构的语义分割网络

图 1 为本文提出的网络结构示意图。该网络以 ResNet50[6]网络为基准网络设计而成，并被应用到了图像语义分割任务中。该网络通过多孔卷积设计了空间金字塔模块(Atrous Spatial Pyramid with Shortcut，ASPS)，用于提取图像的多尺度信息，同时在 ResNet50 中也多次使用了多孔卷积，使得在不进行下采样操作的情况下也可以使卷积核的感受域保持不变，从而增加了最终得到的特征图尺寸，使得最后得到的语义分割结果更加精细。在卷积神经网络中，低层特征感受局部信息，高层特征感受全局信息，因此，将空间金字塔模块设计安排在 ResNet50 网络的 Block5 网络模块(res5c)之后，并将 ResNet50 网络最后的池化层和全连接层替换为卷积层，最后通过逐像素的分类层得到图像的语义分割结果。我们将本文设计的网络记为 ASPSNet 网络。

1.1 多孔卷积

在正常的卷积过程中，卷积核的作用区域是连续的。当卷积核的尺寸变小或者前面某一层的步长减小时，为了保证该层感受域不变，可以将卷积核的作用区域变成有间隔的连接。如图 2 所示，(a)为正常的卷积操作得到的稀疏特征提取，(b)为通过在卷积层滤波器间插入 0 值，使得卷积核的作用像素区域相互之间间隔 1，从而实现了多孔卷积。

以一维信号为例，设输入为 $x[i]$，滤波函数 $w[k]$的长度为 K，则多孔卷积的输出 $y[i]$为

$$y[i]=\sum_{k=1}^{K}x[i+r.k]w[k] \qquad (1)$$

式中，rate 参数 r 是对输入信息采样的步长，在标准卷积中 $r=1$，而在图 2 所示的多孔卷积中 $r=2$。

1.2 空间金字塔模块

虽然对于 ResNet 网络来说，其理论上的

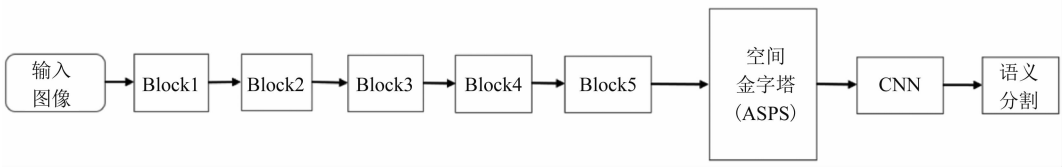
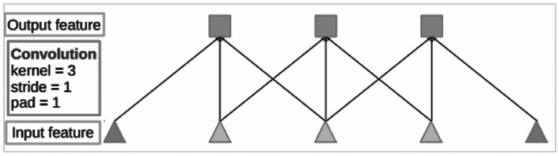
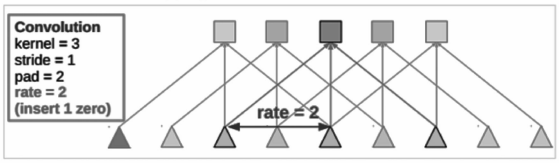


图 1 网络结构框图



(a) Sparse feature extraction



(b) Dense feature extraction

图 2　多孔卷积

感受域已经大于输入图像的尺寸。然而，在卷积神经网络中，卷积核(特别是高层次卷积核)的实际感受域要远低于其理论感受域[7]。因此本文通过设计空间金字塔模块来增强网络对于图像多尺度信息的提取能力，并将空间金字塔模块设计安排在高层网络模块(Block5)之后，用于提取高层次特征的多尺度信息。本文设计的空间金字塔模块结构如图 3 所示，该模块由多个并行的网络分支组成，其中每个网络分支通过设置多孔采样率大小的不同提取图像特征特定尺度的信息。将多个分支输出的结果融合，可以得到输入特征多尺度的信息。同时受 ResNet 网络残差结构的启发，如图 3 所示，在空间金字塔结构设计中还加入了捷径恒等连接网络分支。

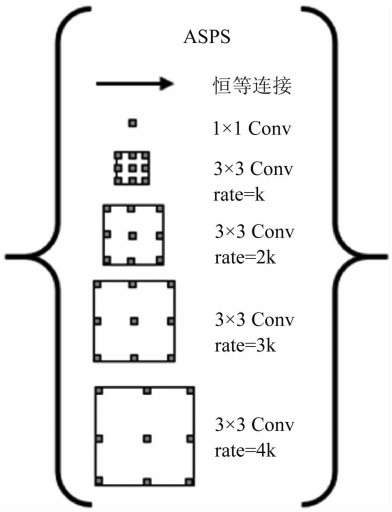


图 3　ASPS

2　实验结果分析

本文在 SUN RGB-D[8]数据集上对设计的网络结构性能进行了评估，该数据集包含有 10335 张具有语义标注信息的 RGB-D 图像(本文只用到 RGB 图像)。训练数据集和测试数据集分别包含 5285 张和 5050 张图片。同时，为了保证在网络改进过程中测试数据集不受影响，在训练数据集中随机选择其中的 1000 张图片作为验证数据集，将剩余的 4285 张图片作为训练数据集。测试数据集只对最终训练好的网络进行测试。为了防止网络训练过程中的过拟合现象，对训练数据集还进行了数据扩充。通过将训练数据集中的每张图片进行随机镜像、随机旋转、随机高斯模糊以及在(0.7～1.5)尺度间进行随机缩放等操作，训练数据集被扩充至原数量的 8 倍。在网络改进的过程中，深度学习网络均在 SUN RGB-D 数据集的 13 个类别[9]上进行训练和测试。为了与其它网络进行性能对比，最终设计好的网络还在 SUN RGB-D 数据集的 37 个类别上进行了训练和测试。

2.1　训练方法

在实验过程中，我们使用开源深度学习训练库 Caffe[10]对网络进行了训练。学习速率采用“poly”速率更新方法，将初始学习速率设为 0.0005，将 power 设为 0.9，并使用随机梯度下降法对网络参数进行更新，网络训练的迭代次数设为 10 K。由于 SUN RGB-D 数据集中各类之间的分布非常不均匀，在训练过程中还根据各个类的分布对每个类的损失函数值进行了相应的加权。

本文使用平均交除并(Intersection Over Union，IOU)作为网络性能的评价标准，其中 IOU 表示预测结果与标注信息的交集区域比上其并集区域，而平均 IOU 则为所有类 IOU 值的平均值。

2.2　实验结果分析

2.2.1　训练超参数的选择

为了选择最优的超参数对网络进行训练，本节主要对训练超参数初始学习速率和参数更

新一次所需要的迭代次数进行了探索。该部分使用的网络为基准网络 Base，即 ASPSNet 网络去掉空间金字塔模块。表 1 为使用不同的初始学习速率对网络进行训练时，训练得到的网络在 SUN RGB-D 验证数据集上的精度，其中每组评价指标最高值加粗显示。从中可以看出，当学习速率为 0.0005 时，训练所得到的网络性能最优。

表 1 初始学习速率测试

学习速率	Mean IOU(%)
0.001	61.02
0.00075	61.10
0.0005	**61.38**
0.00025	60.26

参数 iter_size 与小批量大小 batch_size 的乘积表示对前向网络迭代多少次，对网络参数更新一次。该参数可以防止网络震荡过大，且通过多次迭代求平均值，可以更好地反映数据集的整体表征，从而使网络参数尽可能地向总体误差更小的方向进行更新。表 2 为将 batch_size 设置为 3、并将 iter_size 设置为不同值时的测试结果。从中可以看出，当 iter_size 设置为 2，即每迭代 6 次更新一次网络参数时，训练得到的网络性能最优。

表 2 参数更新所需的迭代次数

Iter_size	Mean IOU(%)
1	60.04
2	**61.38**
4	60.96
6	60.50
8	60.53
10	60.38

通过以上的实验探索可以得出，当将初始学习速率设置为 0.0005，且每迭代 6 次对网络参数进行一次更新时，最终训练得到的网络性能最好。因此，在以后的实验中均以此超参数对设计的网络进行训练。

2.2.2 ASPS 评估

本文还对空间金字塔模块中多尺度并行网络分支的数量以及各个分支多孔采样率 rate 的大小对网络性能的影响进行了探索。将网络分支数分别设计为 4 个、5 个和 6 个分支，而多孔采样率则按照图 3 所示的(1,k,2k,3k…)的方式设置，对于 k∈(4,5,6,7,8)的情况进行了探索。表 3 为将多尺度分支数和采样率 k 设置为不同值时网络在平均 IOU 评价指标上的测试结果。从中可以看出，当多尺度分支数为 5 且 k=5 时，网络在平均 IOU 上的测试结果最好。因此，本文选择多尺度网络分支数为 5 且 k=5 作为空间金字塔模块的设置参数，即空间金字塔模块中各分支的多孔采样率分别为 1、5、10、15、20。

为了更深入地探索空间金字塔模块对于网络性能的影响，本文将 ASPSNet 网络和基准网络在 SUN RGB-D 验证数据集的 13 个类别的 IOU 结果以及平均 IOU 结果进行了比较。从表 4 中可以看出，除了 Books 类别的测试结果略微下降外，ASPSNet 网络对于其它 12 个类别的语义分割结果均有不同程度的提升，特别是对于 Bed、Furniture 和 Sofa 三个类别的语义分割精度提升特别明显，这三个类别属于大尺度的物体，即通过空间金字塔模块，ASPSNet 网络能够提取图像多尺度的信息，从而提高了网络对于各个尺度物体的语义分割精度，同时平均 IOU 测试结果也从 61.38%提高到了

表 3 空间金字塔模块的平均 IOU(%)测试结果

	Rate(k)=4	Rate(k)=5	Rate(k)=6	Rate(k)=7	Rate(k)=8
多尺度分支数=4	62.23	63.81	63.53	63.10	63.58
多尺度分支数=5	62.63	**64.20**	63.92	64.07	64.18
多尺度分支数=6	63.76	63.91	63.71	63.67	63.57

64.20%，从而验证了本文提出的空间金字塔模块的有效性。

图 4 为 ASPSNet 网络和基准网络语义分割结果的直观比较。图 4 中左上和右上两张图像分别为验证数据集中的一张图像及其语义标注结果，左下和右下的两张图像分别为 ASPSNet 网络和基准网络的语义分割结果。从图 4 中可以看出，相比于基准网络，ASPSNet 网络对于大尺度目标 Bed 的预测更加准确。

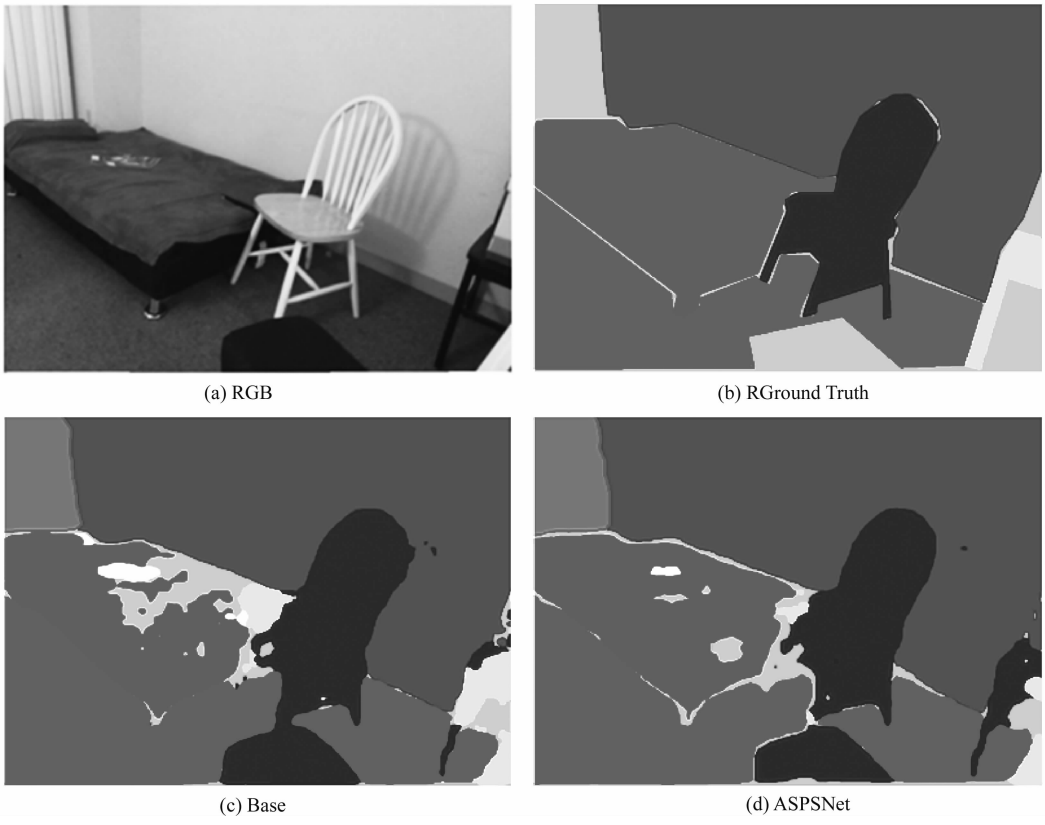
本文还将 ASPSNet 网络在 SUN RGB-D 的训练数据集和验证数据集上训练并在测试数据集上进行了测试，最终的平均 IOU 测试结果为 62.24%(13 类)。

2.2.3　与其它网络的比较

以上的实验均在 SUN RGB-D 数据集的 13 个类别上进行。为了更好地评估本文所设计的网络结构性能，本文在 SUN RGB-D 数据集的 37 个类别上对 ASPSNet 网络进行了训练并测试。表 5 为本文所设计的网络 ASPSNet 和其它 state-of-the-art 网络的测试结果对比。从表 5 中可以看出，本文所设计的网络在平均 IOU 评价指标上的测试结果明显优于其它网络的测

表 4 ASPSNet 网络与基准网络在各个类别 IOU(%)上的比较

类别	Base	ASPSNet
Bed	69.35	**77.00**
Books	**41.89**	40.13
Ceiling	71.02	**72.04**
Chair	69.12	**71.12**
Floor	87.68	**88.08**
Furniture	51.83	**57.23**
Objects	48.79	**51.83**
Picture	52.75	**52.84**
Sofa	53.03	**61.93**
Table	63.02	**66.02**
TV	48.60	**52.94**
Wall	78.42	**79.24**
Window	62.44	**64.21**
Mean IOU	61.38	**64.20**



(a) RGB　(b) RGround Truth　(c) Base　(d) ASPSNet

图 4　ASPSNet 与基准网络比较

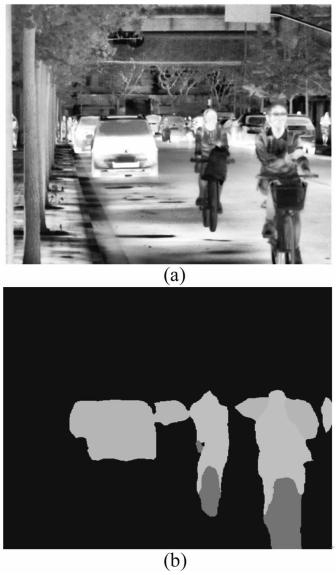
试结果，从而验证了本文所设计的网络的有效性。

表 5 与其它网络在 SUN RGB-D 测试数据集上的结果比较(37 类)

网络结构	Mean IOU(%)
Bayesian SegNet [11]	30.7
FuseNet-SF5 [12]	37.29
Context-CRF [13]	42.3
ASPSNet	**44.38**

2.3 基于红外图像的语义分割

可见光虽然包含了比较丰富的场景信息，可以为语义分割提供更多的判据。然而，当光线不足或者处于黑暗条件下时，可见光的成像质量会大大降低甚至无法成像。而红外图像因具有全天候的成像能力，因此当可见光无法成像时，探索基于红外图像的语义分割算法则具有重要意义。但大部分的语义分割数据集都是基于可见光图像的(如 SUN RGB-D 数据集)。如果创建与可见光图像数据集相同尺度的红外图像语义分割数据集将耗费大量的精力。由于红外图像和可见光图像的一些共性，本文探索了将在可见光图像上训练的语义分割网络应用于红外图像的可行性。图 5(a)是红外图像，将该图像归一化到 0～255 区间，并将单通道的红外图像拼接成 3 通道的图像，然后将其看做 RGB 图像进行语义分割。图 5(b)为语义分割的结果。从图 5 中可以看出，虽然该深度学习网络是基于可见光语义分割数据集训练的，但输出结果仍然较准确地将红外图像中的人、车和自行车检测了出来，这对于基于深度学习的红外图像处理研究具有重要意义。然而，由于红外图像和可见光图像的特性有很多不同，因此语义分割结果并不精细。因此，下一步工作准备创建一定尺度的红外图像语义分割数据集来对深度学习网络进行再训练，使得深度学习网络能更好地提取红外图像的特征信息，并优化最后的预测结果。



(a)

(b)

图 5 红外图像语义分割

3 结论

本文所设计的网络在 SUN RGB-D 数据集上的语义分割性能突出。详细介绍了训练过程中超参数的选择方法，同时通过数据扩充、带权重的损失函数等方法有效防止了网络训练过程中的过拟合现象，从而保证了网络的性能。还设计了一个空间金字塔模块，用于提取图像的多尺度信息，从而使得设计的网络能有效提取图像中的多尺度信息。本文所设计的网络并没有使用条件随机场(Conditional Random Field，CRF)作为后处理模块，使得网络在测试时有较好的实时性，从而可以将该网络应用于无人驾驶、增强现实等领域。最后，还对基于红外图像的语义分割进行了初步探索，并研究了下一步工作的开展方向。

参考文献

[1] Long J，Shelhamer E，Darrell T. Fully Convolutional Networks for Semantic Segmentation [C]. Bostn：IEEE Conference on Computer Vision and Pattern Recognition，2015:3431-3440.

[2] Simonyan K，Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. *arXiv*：1409.1556，2014.

(下转第 48 页)