计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190100119

基于深度特征融合的图像语义分割

周鹏程[1] 龚声蓉[1,2] 钟 珊[1,2] 包宗铭[1] 戴兴华[1]
1 苏州大学计算机科学与技术学院 江苏 苏州 215006
2 常熟理工学院计算机科学与工程学院 江苏 苏州 215500
(20165227048@stu.suda.edu.cn)

摘 要 在图像语义分割中使用卷积网络进行特征提取时,由于最大池化和下采样操作的重复组合引起了特征分辨率降低,从而导致上下文信息丢失,使得分割结果失去对目标位置的敏感性。虽然基于编码器-解码器架构的网络通过跳跃连接在恢复分辨率的过程中逐渐细化了输出精度,但其将相邻特征简单求和的操作忽略了特征之间的差异性,容易导致目标局部误识别等问题。为此,文中提出了基于深度特征融合的图像语义分割方法。该方法采用多组全卷积 VGG16 模型并联组合的网络结构,结合空洞卷积并行高效地处理金字塔中的多尺度图像,提取了多个层级的上下文特征,并通过自顶向下的方法逐层融合,最大限度地捕获上下文信息;同时,以改进损失函数而得到的逐层标签监督策略为辅助支撑,联合后端像素建模的全连接条件随机场,无论是在模型训练的难易程度还是预测输出的精度方面都有一定的优化。实验数据表明,通过对表征不同尺度上下文信息的各层深度特征进行逐层融合,图像语义分割算法在目标对象的分类和空间细节的定位方面都有所提升。在 PASCAL VOC 2012 和 PASCAL CONTEXT 两个数据集上获得的实验结果显示,所提方法分别取得了 80.5%和 45.93%的 mIoU 准确率。实验数据充分说明,并联框架中的深度特征提取、特征逐层融合和逐层标签监督策略能够联合优化算法架构。特征对比表明,该模型能够捕获丰富的上下文信息,得到更加精细的图像语义特征,较同类方法具有明显的优势。

关键词:图像语义分割;深度特征;空洞卷积;特征融合;上下文信息;条件随机场
中图法分类号 TP391

Image Semantic Segmentation Based on Deep Feature Fusion

ZHOU Peng-cheng[1], GONG Sheng-rong[1,2], ZHONG Shan[1,2], BAO Zong-ming[1] and DAI Xing-hua[1]
1 School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China
2 School of Computer Science and Engineering, Changshu Institute of Technology, Suzhou, Jiangsu 215500, China

Abstract When feature extraction is performed by using convolutional networks in image semantic segmentation, the context information is lost due to the reduced resolution of features by the repeated combination of maximum pooling and downsampling operations, so that the segmentation result loses the sensitivity to the object location. Although the network based on the encoder-decoder architecture gradually refines the output precision through the jump connection in the process of restoring the resolution, the operation of simply summing the adjacent features ignores the difference between the features and easily leads to local misidentification of objects and other issues. To this end, an image semantic segmentation method based on deep feature fusion was proposed. It adopts a network structure in which multiple sets of fully convolutional VGG16 models are combined in parallel, processes multi-scale images in the pyramid in parallel efficiently with atrous convolutions, extracts multi-level context feature, and fuses layer by layer through a top-down method to capture the context information as far as possible. At the same time, the layer-by-layer label supervision strategy based on the improved loss function is an auxiliary support with a dense conditional random field of pixels modeling in the backend, which has certain optimization in terms of the difficulty of model training and the accuracy of predictive output. Experimental data show that the image semantic segmentation algorithm improves the classification of target objects and the location of spatial details by layer-by-layer fusion of deep features that characterize different scale context information. The experimental results obtained on PASCAL VOC 2012 and PASCAL CONTEXT datasets show that the pro-

到稿日期:2019-01-15 返修日期:2019-04-07 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61272005;61702055);江苏省自然科学基金项目(BK20151254,BK20151260);江苏省六大高峰人才项目(DZXX-027);教育部科技发展中心"云数融合科教创新"基金(2017B03112)
This work was supported by the National Natural Science Foundation of China(61272005, 61702055), Natural Science Foundation of Jiangsu Province(BK20151254, BK20151260), Six Peak Talents Projecof Jiangsu Province(DZXX-027) and Cloud Integration Science and Education Innovation Foundation of Ministry of Education Science and Technology Development Center(2017B03112).
通信作者:龚声蓉(shrgong@suda.edu.cn)

posed method achieves mIoU accuracy of 80.5% and 45.93%, respectively. The experimental data fully demonstrate that deep feature extraction, feature layer-by-layer fusion and layer-by-layer label supervision strategy in the parallel framework can jointly optimize the algorithm architecture. The feature comparison shows that the model can capture rich context information and obtain more detailed image semantic features. Compared with similar methods, it has obvious advantages.

Keywords Image semantic segmentation, Deep feature, Atrous convolution, Feature fusion, Context information, Conditional random field

1 引言

语义分割是计算机视觉领域的重要基石之一，其不仅对图像中的每个像素进行分类，还标注该像素在图像中所属的对象类别，这样不但能分割出区域，还能对区域的内容进行标注。

语义分割通常可以分为几类不同的任务，如图 1 所示。图 1(a)可以描述为：给定一张图像，区分图像中属于人的所有像素和属于马的所有像素，将每种类别的像素分别标注为不同的颜色，以实现像素级别的图像分割。图 1(b)中的场景可以描述为：沙发的前面是 1 张桌子，或者桌子旁边围绕着 3 张椅子。其关键在于将整个场景分解成几个单独的实体，以推理目标的不同行为。虽然目标检测有助于绘制某些实体的大致边框，但还无法以像素级别的精细程度对每个实体标记精确的边界。对自动驾驶和智能机器人等的研究都需要对周围的环境进行深入的理解，其背后的实际需求均是精确且高效的分割技术。在图 1(c)的分类问题中，仅需回答有摩托车或有山羊，与分类问题不同的是，进行语义分割时需要分割模型对给定图像进行密集的像素级预测，在逐个分类像素点的同时把 where 和 what 两个问题结合在一起解决。与分类或检测相比，语义分割对图像进行了更加细致的了解，这种了解在增强现实及图像搜索引擎等领域都非常重要。

(a)像素级别的分割 (b)场景解析 (c)定位和分类的结合

图 1 语义分割示意图

Fig. 1 Examples of semantic segmentation

早期语义分割一般使用基于像素自身低阶视觉信息的无监督方法，或依赖于手工提取特征并与分类器相结合的传统机器学习方法。Long 等[1]基于深度卷积神经网络(Deep Convolutional Neural Network，DCNN)提出了全卷积网络(Fully Convolutional Network，FCN)方法，以卷积层代替全连接层构造全卷积网络，并将其应用到语义分割上，获得了更高的分割精度。全卷积网络由于不需要全连接层，因此能够对任意分辨率的图像进行语义分割，其端到端特性在一定程度上突破了传统机器学习方法中手工提取特征困难且提取的特征表达能力受限等问题，从而成为现代语义分割方法的开山之作。当前语义分割领域中几乎所有的先进方法都是基于该模型进行扩展的。

FCN 模型的强大和灵活在于其拥有充分学习分层特征的能力，然而每层标准卷积模块后紧随的连续最大池化和下采样操作逐渐降低了特征的分辨率，导致上下文信息丢失，使得分割结果失去对目标空间位置细节的敏感性。Yu 等[2]和 Wang 等[3]使用空洞卷积支持 FCN 中感受野的指数级扩展，以有效地聚合图像的全局信息而不丢失分辨率，一定程度地缓解了该问题。Liu 等[4]和 Nguyen 等[5]结合图模型将空间信息合并到 FCN 中，希望更好地控制空间约束，从而提高基于 DCNN 的语义分割方法的性能。此外，还有基于多分辨率重建的方法[6]重建对象的分段边界，而 Bertasius 等[7]则引入了一个简单且有效的卷积随机游走网络，来解决边界定位不良和空间碎片预测的问题。这些方法试图借助传统机器学习清晰的数学原理，在卷积网络中添加可解释的额外信息以提高语义分割的性能。文献[8-9]认为基于强监督的语义分割需要大规模令人信服的像素精确标记数据的存在，这种昂贵的像素注释限制了可训练数据集的大小，间接影响了深度网络的性能，因此提出了仅利用容易获得的边界框线索和少量注释便可得到具有竞争力的准确性的弱监督方法。最新的研究[10-13]利用基于对抗生成网络的方法进行语义分割的场景自适应，利用所学习的源和目标表示欺骗域鉴别器，在无监督方面取得了优异的成果。Kendall 等[14]采用最大池化索引代替 FCN 中的编码器特征来达到占用更少内存空间的目的，但却以牺牲最终的分割性能为代价。Bulo 等[15]引入了损失最大池化概念来处理训练数据分布不平衡的问题，使得训练过程更加平滑和容易。Lin 等[16-17]通过捕获相邻图像区域之间的语义相关性，并结合条件随机场(Conditional Random Field，CRF)构建深度结构化模型，并证明了其对于提高性能是有效的。然而，条件随机场原理复杂且训练不易，需要小心地选择合适的超参数，其本身也只能作为后端优化模块，并没有对卷积网络内部结构做出改进。目前流行的基于编码器-解码器架构的模型，如 Chen 等[18]提出的将能够进行多尺度空间采样的 FCN 模型与 CRF 相结合的方法聚合了上述众多方法的优点，在定性和定量两方面均提高了语义分割的性能。但其对含有多尺度上下文信息的邻接特征的融合操作只是将相邻特征裁剪采样至相同维度，再简单求和，这显然忽略了不同层级特征之间的差异性，容易引发对象局部区域的误识别。

针对基于编码器-解码器的 FCN 方法在平衡高层抽象的分类问题和低层精确的定位问题方面能力受限，从而导致分割结果中目标边界细节不清晰的问题，本文提出了一种基于深度特征融合的 FCN 架构，通过多组全卷积 VGG16 模型并联组合的框架，结合空洞卷积高效并行地提取多种尺度的图像深度特征，并通过自顶向下的逐层融合方法尽可能地捕获

全局上下文信息，有效缓解了算法在分类与定位之间平衡能力受限的问题。该方法在训练阶段改进损失函数，利用逐层标签进行深度监督，增强模型各个层级的学习能力；最后，通过全连接条件随机场进行像素建模，联合优化最终的分割效果。

本文第 2 节介绍了模型架构中关于空洞卷积与条件随机场的背景知识；第 3 节介绍了本文提出的深度特征融合架构，包括多层初始深度特征提取的模块细节、融合特征的策略及生成方法和改进的新型损失函数；第 4 节在 PASCAL VOC 2012 和 PASCAL CONTEXT 两个数据集上进行了实验，并与 DeepLab，DPN 以及 Piecewise 等几种方法进行了对比分析；最后总结全文并展望未来。

2 背景知识

2.1 空洞卷积

语义分割的输出分辨率应与输入图像一致。基于全卷积网络的语义分割方法虽然能够接受任意尺寸的输入图像，但连续的池化操作在增大感受野的同时也减小了特征的分辨率。虽然通过上采样可以将缩小的特征图还原到图像的原始尺寸，但这个过程必然造成丢失的信息无法还原，上采样恢复的特征图将失去对图像细节的敏感性，并且频繁的上采样操作也需要额外的内存和时间。采用最初应用于信号处理领域的小波变换分析中的空洞卷积方法可以解决这一问题[2]。

首先考虑一维信号。一维输入信号 $x[i]\in R$ 与长度为 K 的滤波器 $w[k]\in R$ 的空洞卷积输出 $y[i]$ 的定义如下：

$$y[i]=\sum_{k=1}^{K}x[i+r\cdot k]w[k] \tag{1}$$

其中，扩张率参数 r 是对输入信号进行采样的步幅，相当于将输入 x 与在两个连续滤波值之间插入 $r-1$ 个 0 而得到的滤波器进行卷积，因此称为空洞卷积。标准卷积相当于扩张率 $r=1$ 的特殊情形。

对于二维信号，将全分辨率图像与空洞滤波器进行卷积操作。首先将原始滤波器上采样 2 倍，并在滤波器值之间插入零值。虽然有效滤波器的尺寸有所增加，但无须考虑中间插入的零值，即空洞，因此滤波器参数的数量和每个位置的操作数量保持不变。通过改变扩张率参数 r 来自适应地修改感受野的大小，可以有效地控制卷积网络中特征的分辨率而无须学习额外的参数。

输入图像在经过连续 3 次 3×3 的标准卷积后，感受野尺寸分别为 3×3，5×5 和 7×7。若连续卷积操作的核尺寸为 $(2d+1)\times(2d+1)$ 且不变，则第 n 层感受野的尺寸为：

$$f_n=2dn+1,d\in\mathrm{N}^{+} \tag{2}$$

即标准卷积下感受野大小呈线性增长。而若为空洞卷积选取适当的扩张率参数，则可以使卷积网络在不增加额外计算量的情况下使感受野呈指数型增长，从而获取图像的密集特征。

2.2 全连接条件随机场

对分割架构的输出进行调优并强化其捕捉细粒度信息的一个通用方法是引入条件随机场作为全卷积网络的后端处理模块，将卷积网络的识别能力和全连接条件随机场的定位精度优化能力耦合在一起[10-12]，这可以成功地应对分割模型的定位挑战，在对象的轮廓和边界上恢复更多的细节。

对于图像 $\boldsymbol{I}$ 以及标签 $\boldsymbol{X}$，定义 $\boldsymbol{I}$ 为变量 $\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_N\}$ 上的一个随机场，$\boldsymbol{X}$ 为变量 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 上的一个随机场，其中 $\boldsymbol{I}_j$ 是像素 j 的颜色向量，$\boldsymbol{X}_j$ 是分配给像素 j 且取值自 $\{l_1,l_2,\cdots,l_M\}$ 的标签，M 是图像中可能的对象类别个数，包括背景。条件随机场 $(\boldsymbol{I},\boldsymbol{X})$ 用吉布斯分布可以描述为：

$$\mathrm{P}(\boldsymbol{X}|\boldsymbol{I})=\frac{1}{Z(\boldsymbol{I})}\exp\Big(-\sum_{c\in C_g}\phi_c(\boldsymbol{X}_c|\boldsymbol{I})\Big) \tag{3}$$

其中，$Z(\boldsymbol{I})$ 是归一化因子，$g=(V,E)$ 是 $\boldsymbol{X}$ 上的无向图模型，C_g 中的每一项 c 都包含一个势函数 ϕ_c。标签 $x\in L^N$ 的吉布斯能量为 $E(x|\boldsymbol{I})=\sum_{c\in C_g}\phi_c(x_c|\boldsymbol{I})$。对应的最大后验概率标签为 $x=\arg\max_{x\in L^N}P(\boldsymbol{X}|\boldsymbol{I})$。

全连接条件随机场模型中，g 是 $\boldsymbol{X}$ 上的无向完全图，而 C_g 是所有一元项和二元项的集合。相应的吉布斯能量可以表示为：

$$E(x)=\sum_i\psi_u(x_i)+\sum_{i<j}\psi_p(x_i,x_j) \tag{4}$$

一元势函数 $\psi_u(x_i)$ 可以利用图像中的形状、结构、位置、颜色、纹理、梯度等信息，二元势函数 $\psi_p(x_i,x_j)$ 的表达式如下：

$$\psi_p(x_i,x_j)=\mu(x_i,x_j)\underbrace{\sum_{m=1}^{K}w^{(m)}k^{(m)}(\boldsymbol{f}_i,\boldsymbol{f}_j)}_{k(\boldsymbol{f}_i,\boldsymbol{f}_j)} \tag{5}$$

$$k^{(m)}(\boldsymbol{f}_i,\boldsymbol{f}_j)=\exp\Big(-\frac{1}{2}(\boldsymbol{f}_i-\boldsymbol{f}_j)^{\mathrm{T}}\Lambda^{(m)}(\boldsymbol{f}_i-\boldsymbol{f}_j)\Big) \tag{6}$$

其中，每个 $k^{(m)}$ 都是从像素 i 和 j 提取的特征 $\boldsymbol{f}$ 决定的高斯核函数并由参数 $w^{(m)}$ 加权，向量 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 分别是像素 i 和 j 位置上任意维度的特征向量，$w^{(m)}$ 是线性组合项的权重。因为模型的因子图是全连接的，即图像中任意两个像素 i 和 j 无论彼此相距多远，它们之间都存在一个二元项。误差惩罚项：

$$\mu(x_i,x_j)=\begin{cases}1, & x_i\neq x_j\\ 0, & x_i=x_j\end{cases} \tag{7}$$

式(7)表明只有在像素 i 和 j 分配的标签不一样时上述二元势函数才有值，如果两个像素分配的标签相同，则二元势函数的值为零。

全连接条件随机场模型适用于高效的近似概率推理算法[19]，在完全可分解的平均场近似下传递更新消息可以表示为特征空间中带有高斯核的卷积，使用高维滤波算法[20]可以加快这一计算的速度。

3 深度特征融合架构

3.1 基于 VGG16 的初始特征提取

本文提出的模型以 VGG16 为基础架构。网络中的每一层数据都是尺寸为 $h\times w\times d$ 的三维数组，其中空间维度 h 和 w 为特征的高和宽，而 d 是特征图的通道数。第一层数据是分辨率为 $h\times w$ 且彩色通道数 $d=3$ 的原始图像。标准卷积网络最初被设计用于图像分类，采用固定尺寸的输入产生非空间的输出，这些网络的全连接层输出固定长度的一维向量而丢弃空间信息。语义分割这种密集分类问题在结构上与图像分类不同，需要修改网络的最后 3 层全连接层为卷积层来构造全卷积网络，以适应语义分割问题。

基于 VGG16 的 FCN 模型经过每次池化操作后输出的数据体变为原来的 1/2，第 5 次池化操作后输出尺寸为 $\frac{h}{32}\times\frac{w}{32}\times 512$，将其后 3 层全连接层转换为卷积层，卷积核的大小分别为 $1\times1\times4096$，$1\times1\times4096$ 和 $1\times1\times N$，最终输出 $\frac{h}{32}\times\frac{w}{32}\times N$ 的数据体。其中 N 是目标对象类别的个数，例如 PASCAL VOC 2012 数据集中共有 20 个对象类别和 1 个背景类别，即 $N=21$。如图 2 所示，将其中两组卷积模块的标准卷积改为空洞卷积即可将其用于提取图像的初始深度特征。

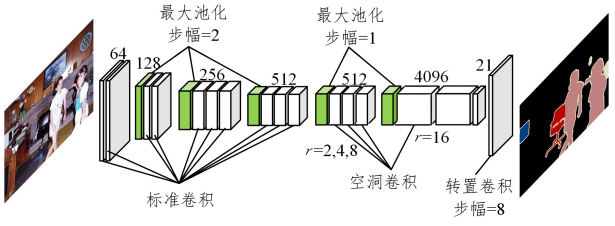

图 2 初始深度特征提取

Fig. 2 Extraction of initial deep feature

为了维持部分卷积和池化操作的输出数据体的尺寸与输入数据体一致，利用 padding 对输入数据体进行零值填充，同时设置采样步幅 $stride=1$。padding 的大小通常取 $p=(f-1)/2$，其中滤波器尺寸 f 通常为奇数。保持 VGG16 的前 4 个卷积模块不动，其间 3 个最大池化层步幅 $stride=2$，输出通道依次为 64，128，256，512。第 4 个最大池化层初始步幅 $stride=1$，在第 5 个卷积模块中分别引入扩张率 $r=2,4,8$ 的空洞卷积，保持特征的分辨率不变并在不增加计算量的情况下扩大感受野。第 5 个模块的最大池化层步幅也设为 $stride=1$，并紧随通道为 4096 且 $r=16$ 的空洞卷积。由于前 3 次最大池化步幅均为 $stride=2$，模型最后输出的特征尺寸变为原始图像的 1/8，因此需要在模型的最后添加步幅 $stride=8$ 的转置卷积进行上采样，将其恢复成原始图像的分辨率大小，从而在对每个像素产生预测的同时保留原始图像的空间信息。

3.2 基于上下文信息融合的特征生成方法

本文的模型需要在全卷积网络的基础上结合不同层级的特征输出，即对多种尺度的信息予以整合。一方面，细粒度或者相对局部的信息对于提高像素级别标注的正确性来说非常关键，另一方面，整合图像的全局上下文信息对于解决局部模糊性问题来说也十分重要。在高层抽象的语义信息与低层精确的细节信息之间取得平衡，能够最大限度地提高输出空间的精度。根据上述分析，本文采用直观的特征融合架构，包括缩放生成不同尺度的图像、提取不同尺度的层级特征、使用先行融合逐步调优的策略等。

3.2.1 特征融合网络架构设计

首先通过原始图像构建 4 层的图像金字塔，用于提取不同层级的特征。金字塔是一组图层，图层越高尺寸越小，每个层由上至下编号，第 $i+1$ 层图像 G_{i+1} 的分辨率小于第 i 层图像 G_i。在原始图像 G_0 上迭代生成 4 层的图像金字塔，将图像金字塔的各个图层输入多组并行的全卷积 VGG16 模型网络架构，相当于提取原始图像不同层级的特征。

特征融合架构并不直接融合所有的层级特征。如图 3 所示，尺寸邻近的特征作为相对局部与相对全局的信息先行融合可以更好地恢复由于分辨率降低而丢失的空间精度。从顶层 1/8 大小的图像开始提取得到 1/64 分辨率的特征，为了获得高质量的分割，从下一层开始将上一层得到的特征与当前层经过 5 次卷积池化模块后得到的特征一起作为输入进行特征融合操作。经过中间两层分支的融合有助于恢复和细化原本粗略的预测，即自顶向下地逐步调优，最终将不同尺度的上下文特征嵌入到网络架构的最后一层分支。该逐步调优过程在得到很好细节的基础上获得了尽可能强的语义信息，更加有效地集成了不同区域的上下文，进行了逐步细化。为了从特征图中得到最终的分割效果，需要进行步幅 $stride=8$ 的上采样操作，并加上 softmax 对各像素归属不同类别的概率进行评估。

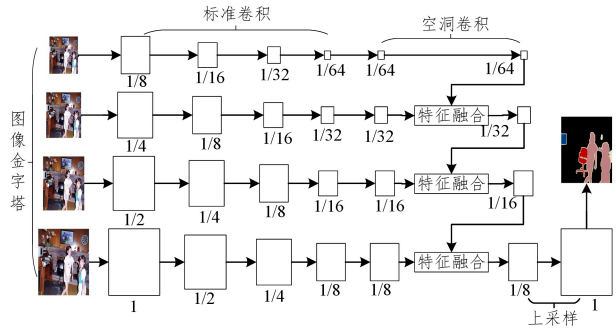

图 3 特征融合网络架构

Fig. 3 Network architecture of feature fusion

3.2.2 相邻特征的融合方法

如图 4 所示，模型使用特征融合策略合并不同分辨率的层级特征，结合不同层级的上下文先验学习更好的特征表示。该融合方法包含两个输入，分别是尺寸为 $\frac{H}{2}\times\frac{W}{2}\times C_1$ 的特征 I_1 和尺寸为 $W\times H\times C_2$ 的特征 I_2，即特征 I_2 相比特征 I_1 拥有两倍的空间分辨率。在融合策略中，首先利用双线性插值法对特征 I_1 进行步幅 $stride=2$ 的空间上采样，使之与特征 I_2 的空间分辨率相同。接着，将上采样后的特征输入扩张率 $rate=2$ 的 3×3 空洞卷积进行操作，细化上采样后的特征，新特征的空间分辨率为 $W\times H\times C_1$。空洞卷积可以从一系列原始邻接像素中合并特征信息，相比转置卷积上采样，使用双线性插值法紧跟空洞卷积操作只需要很小的卷积核就能获得同样大小的感受野，从而占用更少的计算资源。将所得的特征与 I_2 按通道维度进行串联拼接，得到 $W\times H\times(C_1+C_2)$ 的特征。为了进一步细化，最后用 $rate=16$ 的 3×3 空洞卷积操作串联后的特征，得到最终分辨率为 $W\times H\times C_3$ 的融合特征。本文模型中的 $C_1=N$，即目标类别的个数；$C_2=512$，即第 5 个池化模块的输出；$C_3=4096$。将融合后的特征继续送入当前分支的最后两层卷积层，得到该层分支最终提取到的深度特征。该特征融合策略很好地解决了其他方法中不同层级之间的特征简单求和而无视差异性导致的目标局部识别错误问题，能够使分割结果包含相对明确的目标边界，在图像语义分割领域具有重要的价值。

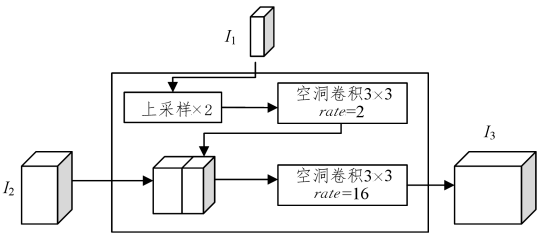

图 4 特征融合策略

Fig. 4 Strategy for feature fusion

3.3 损失函数设计

本节对经典 FCN 算法的损失函数进行了改进，采用一种逐层标签监督策略提升了网络模型中每一层分支的学习能力，利用与空间金字塔中各层图像等分辨率的理想分割图标签来监督每一层分支提取的输出特征。每一层输出特征的分辨率分别为 1/64，1/32，1/16，1/8，并经过步幅 $stride=8$ 的转置卷积上采样恢复到与对应分支原始图层相同的分辨率大小 1/8，1/4，1/2，1。

在给定的 $T=4$ 个分支中，每层提取的特征的通道数即训练集中的类别数为 N。第 t 个分支末尾上采样后的特征 F^t 的空间分辨率为 $W_t \times H_t$，其对应特定坐标位置 (w,h,n) 的值为 $F^t_{w,h,n}$。对每层分支加入带权重的 softmax 交叉熵损失，对应权重为 λ_t。将 F^t 输入到 softmax 函数中，计算图像中每个像素归属不同类别的概率 $\hat{P}^t_{w,h,n}$，softmax 函数的具体公式为：

$$\hat{P}^t_{w,h,n}=\frac{e^{F^t_{w,h,n}}}{\sum_{n=1}^{N} e^{F^t_{w,h,n}}} \tag{8}$$

将 $\hat{P}^t_{w,h,n}$ 映射到真实标签 $P^t_{w,h,n}$ 上，最终用于训练的损失函数如式(9)所示：

$$L=-\sum_{t=1}^{T}\lambda_t \frac{1}{W_t \times H_t}\sum_{w=1}^{W_t}\sum_{h=1}^{H_t}\sum_{n=1}^{N}(P^t_{w,h,n}\log \hat{P}^t_{w,h,n}+ (1-P^t_{w,h,n})\log(1-\hat{P}^t_{w,h,n})) \tag{9}$$

逐层标签监督策略使得梯度优化更加平滑，模型也更容易训练。监督下的每一层分支各自拥有强大的学习能力，能够学到各个层级丰富的语义特征，通过融合使得最终得到的分割图精度不依赖于任意单独的分支。

4 实验结果与分析

本节在 ImageNet 预训练过的 VGG16 基础上微调网络模型的权重，将前端的网络与后端的条件随机场分开训练，通过随机梯度下降算法对网络模型的所有参数进行训练。在设置条件随机场参数时，固定前端网络输出的一元项不变。在 PASCAL VOC 2012 和 PASCAL CONTEXT 两个数据集上评估本文提出的模型，实验硬件平台是 Core i7 处理器，3.6GHz 主频，48GB 内存，GPU 为 NVIDIA GTX 1080，代码运行在 TensorFlow 深度学习框架上。

4.1 数据集与评价指标

用于语义分割的数据集主要有 PASCAL VOC，Cityscapes，Microsoft COCO，CamVid 等。本文主要使用 PASCAL VOC 2012 和 PASCAL CONTEXT 两种数据集。

4.1.1 PASCAL VOC 2012

PASCAL VOC 2012 是目前语义分割领域中最常用的数据集[21]，其分割基准涉及 20 个前景类别，包括飞机、自行车、船、公共汽车、汽车、摩托车、火车、瓶子、椅子、餐桌、盆栽、沙发、电视机、鸟、猫、牛、狗、马、羊、人，以及 1 个背景类别。最初该数据集包含 1464 张用于训练的图像，1449 张用于验证的图像和 1456 张用于测试的图像，随后 Hariharan 等[22]为该数据集提供了额外的注释增强，将训练集的数量扩充至 10582 张。本节实验中使用的是 PASAL VOC 2012 扩充数据集，由于其数据量较大，应用范围比原始 PASCAL VOC 2012 数据集更加广泛。

4.1.2 PASCAL CONTEXT

PASCAL CONTEXT 数据集[23]是比 PASCAL VOC 2012 数据集更加具有挑战性的自然数据集，提供了对整幅图像的标注，同时标注了不同的类别和场景，包括 4998 张用于训练的图像和 5105 张用于验证的图像。该数据集总共包含 457 个类别，大多数类别的出现次数极少，因此通常只对其中 59 个出现较为频繁的类别进行评估。与 PASCAL VOC 数据集不同，PASCAL CONTEXT 数据集的分割任务既包含对目标类别如飞机、自行车、鸟、船、瓶子等的分割，还包括对背景类别如天花板、地板、草地、地面的分割。因为其对目标考虑得更加全面，划分得更加细致，多数标签具有相似的上下文，包含更多易混淆的类别，对模型的分类和分割能力提出了更大的挑战。

4.1.3 评价指标

已经有许多评估标准被提出用于评估语义分割技术的精度，这些指标通常是像素精度和交并比(IOU)的变体。其中平均交并比(mIOU)是语义分割评价指标常用的标准，其计算的是两个集合的交集与其并集的重合比例：

$$mIOU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{10}$$

其中，k 是前景对象的个数，p_{ij} 是指原本属于第 i 类却被分类到第 j 类的像素的数量。

4.2 实验参数设置

在超参数设置方面，为了生成图像金字塔中的第 $i+1$ 层，用式(11)所示的高斯核 k_{Gaussian} 对 G_i 进行预处理并删除每个偶数行和列，生成的图像是其前驱的 1/4。对于边界点而言，把已有的点拷贝到另一面的对应位置便可以模拟出完整的矩阵。

$$k_{\text{Gaussian}}=\frac{1}{256}\begin{bmatrix}1&4&6&4&1\\4&16&24&16&4\\6&24&36&24&6\\4&16&24&16&4\\1&4&6&4&1\end{bmatrix} \tag{11}$$

在网络模型的训练过程中采用图 5 所示的学习率策略，学习率遵循迭代协议：

$$lr=lr_0\times(1-\frac{iter}{max_{\text{iter}}})^{power} \tag{12}$$

相比以固定的步长减小学习率，迭代策略更加高效。实验设置初始学习率 $lr_0=0.01$，$power=0.9$。$iter$ 为当前训练迭代的次数，网络的性能随着迭代次数的增加可以逐渐提升，设置最大迭代次数 $max_{\text{iter}}=600\,000$，训练中批处理图像的张数为 20。

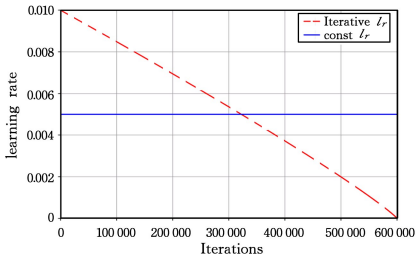

图 5 学习率迭代策略

Fig. 5 Strategy for learning rate iteration

为了防止过拟合，损失函数中加入了 L2 正则项做惩罚约束，正则项的权值衰减设为 0.000 1，并设置动量 $\upsilon=0.9$。在数据增强方面，实验会预先对整个数据集进行预处理，包括随机翻转、0.5 到 1.5 倍的随机缩放、−10°到 10°的随机旋转、对训练集数据进行逐样本均值削减，以及在每个样本上减去整个训练集的统计平均值等操作。将空洞卷积、初始特征提取、特征融合以及辅助监督的加权损失函数等技巧堆叠在一起，在 PASCAL VOC 2012 以及 PASCAL CONTEXT 数据集上进行训练。图 6 给出了两个数据集上网络训练的收敛情况。可以看出，目标的优化过程并非一帆风顺，加权损失函数在训练过程中有所震荡，经过次数较多的迭代之后才会逐渐显现出整体收敛趋势。最终经过 600 000 次迭代逐渐收敛至某个较为优化的区域。

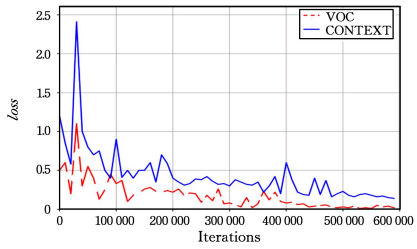

图 6 损失函数的收敛

Fig. 6 Convergence of loss function

完成前端网络的训练后，将得分图送入全连接条件随机场做后端优化。式(4)中的一元势函数 $\psi_u(x_i)$ 是由前端训练好的卷积网络分类器为每个像素 i 计算的概率得分 $P(x_i)$ 得到的，分类器计算每个像素 i 分配的标签为 x_i 的概率，从而产生一个分布。本节实验只使用由前端网络分类器为像素 i 计算得到的标签分配概率 $P(x_i)$，得到一元势函数 $\psi_u(x_i)=-\log P(x_i)$ 与由颜色向量 $\boldsymbol{I}_i$ 和 $\boldsymbol{I}_j$ 及空间位置 p_i 和 p_j 所定义的对比度敏感的双核势函数。颜色向量由 RGB 三维向量组成，位置向量由水平和垂直两个方向组成。定义在颜色向量 $\boldsymbol{I}_i$ 和 $\boldsymbol{I}_j$ 以及空间位置 p_i 和 p_j 上的两个核函数 k_c 和 k_p 分别为：

$$k_c=\exp\left(-\frac{|p_i-p_j|^2}{2\theta_\alpha^2}-\frac{|\boldsymbol{I}_i-\boldsymbol{I}_j|^2}{2\theta_\beta^2}\right) \tag{13}$$

$$k_p=\exp\left(-\frac{|p_i-p_j|^2}{2\theta_\gamma^2}\right) \tag{14}$$

最终得到双核势函数：

$$k(f_i,f_j)=w^{(1)}k_c+w^{(2)}k_p \tag{15}$$

其中，$w^{(1)}$ 为 k_c 核函数项的权重，$w^{(2)}$ 为 k_p 核函数项的权重。k_c 项同时定义在像素的空间位置和颜色向量上，其因基于相似颜色的邻近像素可能属于同一类别的假设被称为外观核函数，其中像素的邻近与相似程度由超参数 θ_α 和 θ_β 控制。k_p 项只定义在像素的空间位置上，用于执行强制平滑以去除孤立的小区域，因此被称为平滑核函数，超参数 θ_γ 的大小决定 k_p 的平滑程度。

利用高维滤波算法[20]对式(13)－式(15)中的参数进行交叉验证。使用默认的超参数 $w^{(2)}=5$，$\theta_\gamma=3$，然后从验证集中选出 100 张图像的子集进行交叉验证以搜索最佳的 $w^{(1)}$，θ_α，θ_β 值。在交叉验证中搜索的超参数取值区间设为 $w^{(1)}\in[5,10]$，$\theta_\alpha\in[50,100]$，$\theta_\beta\in[3,10]$，$w^{(1)}$ 和 θ_β 每次的取值间隔为 1，而 θ_α 每次的取值间隔为 10。平均场迭代次数一般固定为 10，实际上迭代 5 到 8 次模型就已经基本收敛。

图 7 给出了训练过程中整个模型的性能提升过程。由图可知，在 PASCAL VOC 2012 以及 PASCAL CONTEXT 测试集上的 *mIOU* 随目标函数的优化而逐渐变高，模型的输出精度增加。PASCAL VOC 2012 的 *mIOU* 变化相对平缓，而 PASCAL CONTEXT 数据集更加复杂，目标类别划分细致且易混淆，以致 *mIOU* 性能变化曲线不断震荡，直至接近迭代次数尽头才逐渐平缓，没有再出现较大的变化幅度。

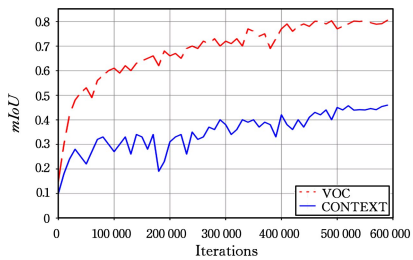

图 7 模型性能的提升

Fig. 7 Improvements of models' performance

4.3 性能分析

本节在 PASCAL VOC 2012 数据集和 PASCAL CONTXET 数据集上对几种方法进行实验比较分析。首先，在 PASCAL VOC 2012 测试集上将本文算法与现有的 FCN，DeepLab，DPN，BoxSup，Piecewise 以及 LRR 等算法进行了对比，各方法在测试集上的逐类别准确率如表 1 所列。为了分析不同对象的分割，表 1 列出了 PASCAL VOC 2012 数据集中所有对象的分割效果。可以看出，最好的方法并不是对所有对象的分割效果都是最佳的，一些特殊或者被局部遮挡导致不连通的对象的分割难度较大。如台式计算机的主机和显示器分离，分割算法在视觉上很难将两者同时标注为同属一个对象的事物；再如自行车车轮为圆环形状，所以在视觉上其中间包含背景或其他对象，算法有时会错误地将其分割为整个圆的效果。

在 mIoU 上，本文方法对于有一半左右的类别分割准确率高于 DeepLab，且部分类别的分割准确率远远高于 DeepLab，最终总的准确率略高于 DeepLab。与前沿的 LRR 方法相比，本文方法在大部分类别上具有较高的准确率，在自行车、船、瓶子、椅子、盆栽、沙发、电视等类别上比 LRR 高出

3%，有的甚至高出15%到20%，这些类别都是分割难度较大且易混淆的类别，本文方法由粗到细融合了多个层级的特征，因此在处理有较多细节的自行车、椅子、盆栽等类别时具有特征提取上的优势，可以分割出较为精细的目标。对于奶牛、羊、狗等具有相似外观的类别目标，也能够分割出复杂语义类别的精确像素。

表1 PASCAL VOC 2012测试集上逐类别的准确率

Table 1 Accuracies by category on PASCAL VOC 2012 test set

（单位：%）

Category	FCN	DPN	BoxSup	Context	Piecewise	LRR	DeepLab	Ours
Aero	76.8	87.7	89.8	**94.1**	**94.1**	92.4	92.6	85.1
Bike	34.2	59.4	38.0	40.4	40.7	45.1	60.4	**63.8**
Bird	68.9	78.4	89.2	83.6	84.1	**94.6**	91.6	85.9
Boat	49.4	64.9	68.9	67.3	67.8	65.2	63.4	**74.2**
Bottle	60.3	70.3	68.0	75.6	75.9	75.8	76.3	**81.0**
Bus	75.3	89.3	89.6	93.4	93.4	**95.1**	95.0	92.7
Car	74.7	83.5	83.0	84.4	84.3	**89.1**	88.4	86.6
Cat	77.6	86.1	87.7	88.7	88.4	92.3	**92.6**	91.5
Chair	21.4	31.7	34.4	41.6	42.5	39.0	32.7	**60.7**
Cow	62.5	79.9	83.6	86.4	86.4	85.7	**88.5**	88.3
Table	46.8	62.6	67.1	63.3	64.7	**70.4**	67.6	68.2
Dog	71.8	81.9	81.5	85.5	85.4	88.6	89.6	**89.9**
Horse	63.9	80.0	83.7	89.3	89.0	89.4	**92.1**	85.8
Mbike	76.5	83.5	85.2	85.6	85.8	**88.6**	87.0	85.0
Person	73.9	82.3	83.5	86.0	86.0	86.6	**87.4**	84.0
Plant	45.2	60.5	58.6	67.4	67.5	65.8	63.3	**73.4**
Sheep	72.4	83.2	84.9	90.1	**90.2**	86.2	88.3	86.7
Sofa	37.4	53.4	55.8	62.6	63.8	57.4	60.0	**67.1**
Train	70.9	77.9	81.2	80.9	80.9	85.7	**86.8**	**86.8**
Tv	55.1	65.0	70.7	72.5	73.0	77.3	74.5	**80.3**
mIoU	62.2	74.1	75.2	77.8	78.0	79.3	79.7	**80.5**

可以看出，从FCN模型到LRR和DeepLab模型，大多数算法在准确率方面都在不断提升，其中DeepLab和DPN使用了条件随机场做后端处理操作，但受到基本卷积网络特征表达能力的限制，虽然能够较好地识别目标类别，但在对象边界的细节问题上往往缺乏一致性。而本文方法从粗粒度图像语义分割网络中获取了各个层级丰富的语义特征和细节特征，最终取得了更好的语义分割效果，在PASCALVOC 2012上取得了80.5%的准确率，比DeepLab高出0.8%，比LRR高出1.2%，说明提出的特征融合策略的确改进了卷积网络在语义分割问题上的有效性。部分方法也采用了类似空洞卷积的结构，但未在此基础上平衡对象分类的准确率与目标边界的精度之间的矛盾，本文方法在这一方面拥有明显的优势。

本文方法与其他几种对比方法的图像语义分割效果如图8所示。

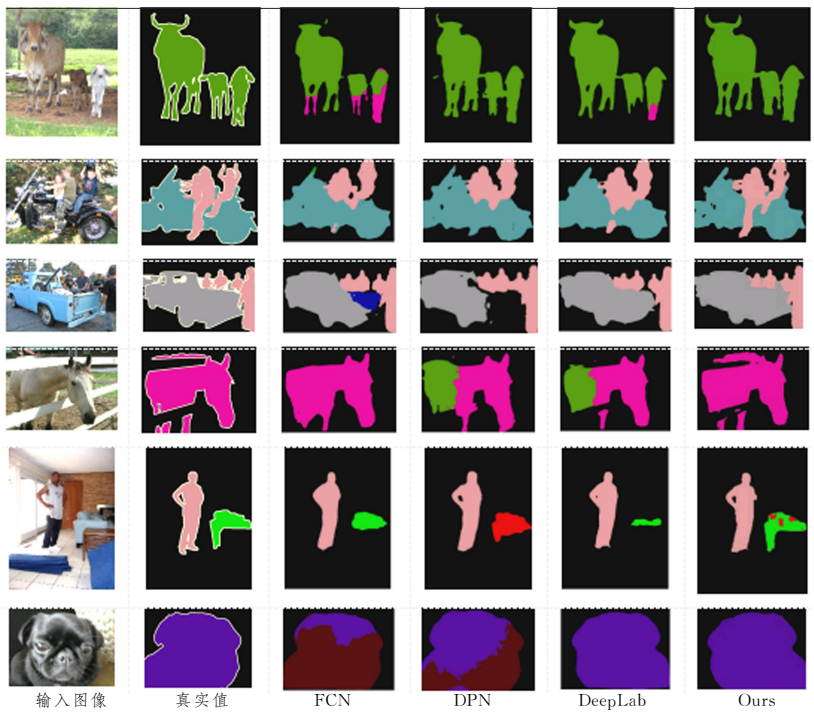

图8 部分可视化比较

Fig. 8 Partial visualization comparison

从图 8 可以看出，DPN 和 DeepLab 方法对于栏杆边上的马的处理有明显的分割错误，栏杆没有被完全识别，而马的后半身被误识别为沙发，而本文方法相对而言效果更好，体现出了强大的图像语义细节识别能力。对于人和沙发的图像，沙发部分被遮挡，沙发的视觉外观和椅子又非常相似，二者属于易混淆类别。对此，FCN 只简单识别出了沙发却不能很好地分割像素，DPN 更是错误地将沙发识别为椅子，DeepLab 在此处的效果甚至低于 FCN，而本文方法除去细微的误识别之外，基本能够准确分割，体现出层级特征融合细化对象分割的有效性。但本文方法也存在不足，在处理大目标、多目标时存在分割不一致的问题。

除 PASCAL VOC 2012 数据集外，本文还在 PASCAL CONTEXT 数据集常用的 59 个类别上进行了实验，并将测试结果与 FCN，Piecewise，DeepLab 等进行了比较，如表 2 所列。

表 2 PASCAL CONTEXT 测试集上的准确率比较

Table 2 Accuracy comparison on PASCAL CONTEXT test set

（单位：%）

Method	*mIoU*
FCN	39.10
CAMN[24]	41.20
Piecewise	43.30
VeryDeep[25]	44.50
DeepLab	45.70
CRFasRNN[26]	39.28
Ours	**45.93**

在众多比较方法中，算法采用的架构起到了比较重要的作用，使用更深层主干网络的方法，如 VeryDeep[25] 和 DeepLab 等都取得了较好的结果。2017 年 CVPR 中的 CAMN[24] 方法在 FCN 的基础上加入了细化后处理和融合上下文的 RNN 网络，准确率有了较大提升。本文方法的准确率比 VeryDeep 高 1.4%，比 DeepLab 高 0.2%，在道路、地面、草地、树等语义近似而细节纹理不同的类别上表现出了更好的分类性能，对存在较多难分类以及易混淆样例的数据集具有很好的鲁棒性。与 PASCAL VOC 2012 相比，PASCAL CONTEXT 数据集包含更多相似的上下文，若融合更多的全局上下文特征，则性能上还有提升空间。实验结果证明，本文提出的特征融合架构与辅助监督的加权损失函数大大加强了网络的判别能力，在对复杂场景进行语义分割时，起到了显著作用。

最后，实验在同一台机器上对几种分割算法的时间性能进行了对比，所用图像取自 PASCAL VOC 2012 测试集。图 9 给出了几种算法处理一帧图像所需的推理时间。

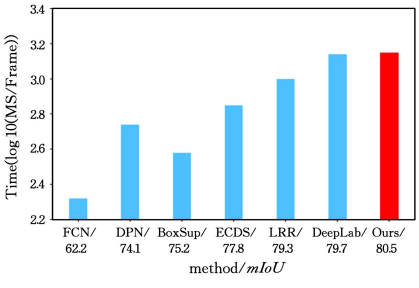

图 9 推理单帧图像所需的时间

Fig. 9 Time required to infer a single frame image

可以看出，本文方法与 DeepLab 时间性能接近，两者都存在因结合了条件随机场而牺牲了时间性能的问题。与其他方法相比，两者解析单帧图像都需要更多的时间，来换取准确率上的明显优势。

结束语 图像语义分割是计算机视觉和机器学习领域的研究热点之一，正为越来越多的视觉应用提供精确且高效的分割机制。为了解决当前流行的全卷积模型应用于图像语义分割时存在的分割结果对目标边界定位细节不敏感的问题，本文从深度特征提取入手，在分析了相关研究的基础上，对所设计的网络模型架构中的多尺度上下文特征提取、逐层特征融合策略和改进损失函数进行了详细阐述。同时，讨论了网络架构后端的全连接条件随机场联合图像像素的空间位置信息和颜色向量信息以用于优化模型捕捉空间细节能力的建模。最后，分析了条件随机场后端处理模块带来的时间开销问题。下一步研究将考虑如何将条件随机场融入卷积网络中参与训练，实现整体的端到端识别系统，从而提高系统的时间性能。

参 考 文 献

[1] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE Press，2015：3431-3440.

[2] YU F，KOLTUM V. Multi-Scale Context Aggregation by DilatedConvolutions[C]// Proceedings of International Conference on Learning Representations. Puerto Rico：IEEE Press，2016：397-410.

[3] WANG P，CHEN P，YUAN Y，et al. Understanding Convolution for Semantic Segmentation[C]// Proceedings of IEEE Winter Conference on Applications of Computer Vision. Santa Rosa：IEEE Press，2017：1451-1460.

[4] LIU Z，LI X，LUO P，et al. Semantic Image Segmentation via Deep Parsing Network[C]// Proceedings of IEEE International Conference on Computer Vision. Santiago Chile：IEEE Press，2015：1377-1385.

[5] NGUYEN K，FOOKES C，SRIDHARAN S. Deep Context Modeling for Semantic Segmentation[C]// Proceedings of IEEE Winter Conference on Applications of Computer Vision. Santa Rosa，California，United States：IEEE Press，2017：56-63.

[6] GHIASI G，FOWLKES C C. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation[C]// Proceedings of European Conference on Computer Vision. Cham：Springer Press，2016：519-534.

[7] BERTASIUS G，TORRESANI L，YU S X，et al. Convolutional Random Walk Networks for Semantic Image Segmentation [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，Hawaii：IEEE Press，2017：6137-6145.

[8] DAI J，HE K，SUN J. BoxSup：Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation [C]// Proceedings of IEEE International Conference on Computer Vision. Santiago，Chile：IEEE Press，2015：1635-1643.

[9] WANG G, LUO P, LIN L, et al. Learning Object Interactions and Descriptions for Semantic Image Segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE Press, 2017: 5235-5243.

[10] MAURO D D, FURNARI A, PATANE G, et al. Scene Adaptation for Semantic Segmentation using Adversarial Learning [C]// Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand: IEEE Press, 2018: 1-6.

[11] ZHANG Y H, QIU Z F, YAO T, et al. Fully Convolutional Adaptation Networks for Semantic Segmentation[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 6810-6818.

[12] TSAI Y H, HUNG W C, Schulter S, et al. Learning to Adapt Structured Output Space for Semantic Segmentation[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 7472-7481.

[13] LENG J X, LIU Y, ZHANG T L, et al. Context-Aware U-Net for Biomedical Image Segmentation[C]// Proceedings of IEEE International Conference on Bioinformatics and Biomedicine. Madrid, Spain: IEEE Press, 2018: 2535-2538.

[14] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(12): 2481-2495.

[15] BULO S R, NEUHOLD G, KONTSCHIEDER P. Loss Max-Pooling for Semantic Image Segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE Press, 2017: 7082-7091.

[16] LIN G, SHEN C, HENGEL A V, et al. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, United States: IEEE Press, 2016: 3194-3203.

[17] LIN G, SHEN C, HENGEL A V, et al. Exploring Context with Deep Structured Models for Semantic Segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(6): 1352-1366.

[18] 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 & Machine Intelligence, 2016, 40(4): 834-848.

[19] PHILIPPK, KOLTUN V. Parameter Learning and Convergent Inference for Dense Random Fields[C]// Proceedings of International Conference on International Conference on Machine Learning. Atlanta, GA, USA: ACM Press, 2013: 513-521.

[20] ADAMS A, BAEK J, DAVIS M A. Fast High-Dimensional Filtering Using the Permutohedral Lattice[J]. Computer Graphics Forum, 2010, 29(2): 753-762.

[21] EVERINGHAM M, ESLAMI S M A, Van G L, et al. The PASCAL Visual Object Classes Challenge: A Retrospective [J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[22] HARIHARAN B, BOURDEV L, ARBELAEZ P, MALIK J, et al. Semantic Contours from Inverse Detectors[C]// Proceedings of IEEE International Conference on Computer Vision. Barcelona: IEEE Press, 2011: 991-998.

[23] MOTTAGHI R, CHEN X, LIU X, et al. The Role of Context for Object Detection and Semantic Segmentation in the Wild [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: ACM Press, 2014: 891-898.

[24] ABDULNABI A H, SHUAI B, WINKLER S, et al. Episodic CAMN: Contextual Attention-Based Memory Networks with Iterative Feedback for Scene Labeling[C]// Proceedings ofIEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE Press, 2018: 6278-6287.

[25] WU Z, SHEN C, ANTONV D H. Bridging Category-Level and Instance-Level Semantic Image Segmentation[J]. International Journal of Computer Vision, 2016, 111(1): 140-155.

[26] ZHENG S, JAYASUMANA S, VINEET V, et al. Conditional Random Fields as Recurrent Neural Networks[C]// Proceedings of IEEE International Conference on Computer Vision. Santiago, Chile: IEEE Press, 2015: 1529-1537.

ZHOU Peng-cheng, born in 1992, postgraduate. His main research interests include digital image processing, computer vision and pattern recognition.

GONG Sheng-rong, born in 1966, Ph.D, professor, Ph.D supervisor, is the vice chairman of Suzhou CCF Association. His main research interests include machine learning and computer vision.