

深度融合网结合条件随机场的遥感图像语义分割

肖春姣^{1,2}, 李宇¹, 张洪群¹, 陈俊¹

1. 中国科学院遥感与数字地球研究所, 北京 100094;

2. 中国科学院大学, 北京 100049

摘要: 为了充分利用遥感图像中丰富的细节信息和上下文信息, 提高图像语义分割精度, 提出一种深度融合网结合条件随机场模型的遥感图像语义分割方法。方法在全卷积神经网络框架中增加反卷积融合结构, 搭建深度融合DFN (Deep Fusion Networks) 网络, 通过深层网络自动获取多尺度特征, 避免人工设计和选择特征, 提高模型的泛化能力; 同时借助反卷积融合结构, 利用多尺度信息, 将浅层细节信息和深层语义信息相融合, 提高模型的处理精度。由全连接条件随机场引入空间上下文信息, 更好地定位边界, 得到最终的语义分割结果。在遥感图像数据集上的实验结果显示: (1) 随着不同尺度细节信息的融入, 结果的边缘轮廓越精确、接近标签图像; (2) 增加了空间上下文信息后, 语义分割结果边缘更细化、准确, 精度更高。实验表明, 该方法可以有效提高遥感图像语义分割的精度, 改善结果的过平滑现象。

关键词: 遥感图像语义分割, 全卷积网络, 条件随机场, 融合结构, 反卷积

引用格式: 肖春姣, 李宇, 张洪群, 陈俊. 2020. 深度融合网结合条件随机场的遥感图像语义分割. 遥感学报, 24(3): 254-264

Xiao C J, Li Y, Zhang H Q and Chen J. 2020. Semantic segmentation of remote sensing image based on deep fusion networks and conditional random field. *Journal of Remote Sensing (Chinese)*, 24(3): 254-264 [DOI: 10.11834/jrs.20208298]

1 引言

图像语义分割ISS (Image Semantic Segmentation) 是指将图像分割成几组具有特定语义含义的像素区域, 并识别出每个区域的类别, 最终得到一幅具有像素级语义标注的图像 (魏云超和赵耀, 2016)。遥感图像语义分割即图像语义分割在遥感图像上的应用。随着遥感技术的发展, 遥感图像的空间分辨率不断提高, 呈现的结构、纹理等细节信息更加清晰, 地物边界和空间布局等上下文信息日益丰富, 为图像语义分割提供了良好的数据条件。同时, 遥感图像信息量巨大、数据复杂、特征信息丰富等特点给图像语义分割带来困难和挑战。

传统的图像语义分割方法主要包括3部分: 图像分割、特征提取和表达、语义分割模型训练。Shotton等 (2008) 通过随机森林分类器提取语义纹元森林特征进行语义分割。Yang等 (2010) 利

用Logistic回归分类器提取颜色、纹元特征, 通过CRF (Conditional Random Field) 模型训练实现语义分割。Arbeláez等 (2012) 综合提取全局外观特征、语义轮廓特征等进行SVM (Support Vector Machine) 模型训练, 应用到VOC-2010自然图像集的语义分割中。Volpi和Ferrari (2015) 提出在CRF模型训练中采用基于环的类别相互作用势函数, 实现城市高分遥感影像语义分割。Montoya-Zegarra等 (2015) 通过增加城市遥感影像中建筑物等类别的形状特征, 提高利用CRF模型进行语义分割的精度。但是传统图像语义分割方法不仅特征的提取和表达需要依靠先验知识进行人工选择和设计, 而且在建立相应语义分割模型的过程中由于人工设计的特征和高层语义特征间“语义鸿沟” (Smeulders等, 2000) 的存在, 最终建立的语义分割模型泛化能力较差。

近年来, 卷积神经网络方法CNN (Convolutional

收稿日期: 2018-07-20; 预印本: 2018-11-07

基金项目: 国家自然科学基金(编号: 61501460); 广东省现代视听信息工程技术研究中心开放基金

第一作者简介: 肖春姣, 1994年生, 女, 硕士研究生, 研究方向为遥感图像处理、深度学习。E-mail: xiaocj@radi.ac.cn

通信作者简介: 李宇, 1986年生, 女, 助理研究员, 研究方向为遥感图像处理。E-mail: liyu@radi.ac.cn

Neural Networks) 因其可以避免传统算法中人工显式的特征提取与表达的需要以及从海量样本图像集中学习到更深层次、更具本质性的特征, 越来越受到研究者的关注 (常亮等, 2016; 肖进胜等, 2017; 张康等, 2018)。Long 等 (2015) 提出了基于全卷积网络 FCN (Fully Convolutional Networks) 的图像语义分割模型 FCN8s, 将原来 CNN 结构中的全连接层用卷积层替换, 通过反卷积操作上采样至原始图像大小, 并结合了中间池化层信息, 实现像素到像素的图像语义分割。Badrinarayanan 等 (2017) 提出基于编码—解码结构的 SegNet 深度卷积网络模型, 该网络将编码过程中最大池化操作的像素索引传输到解码器中, 保留了部分细节信息, 提高了语义分割精度。Sun 和 Wang (2018) 提出将 FCN 模型和数字地表模型结合, 通过高程信息的增加增强了高分辨率遥感影像语义分割效果。但是上述基于 FCN 模型、SegNet 模型的方法由于仅利用了图像部分细节信息以及缺乏对空间、边缘等上下文信息的约束, 语义分割结果较粗糙, 存在过平滑现象。Chen 等 (2018) 将全连接条件随机场引入到 FCN 中, 提出 DeepLab 模型, 增加了上下文信息, 进一步提升了语义分割的性能。Zheng 等 (2015) 又将 CRF 的学习推理过程以 RNN (Recurrent Neural Networks) 的形式嵌入 FCN 模型中, 提出 CRFasRNN 模型, 但该方法的位置信息提取度不如 DeepLab 模型。Wei 等 (2017) 将 CRFasRNN 模型中全连接条件随机场的一阶势由全卷积网络输出结果换成 SegNet 模型输出结果, 在 Vaihingen 城市机载航空影像的语义分割实验中得到了较好的结果。Zhang 等 (2018) 将 DeepLab 模型得到的初步语义分割结果映射到 3D 点云数据, 有效地提高了建筑物等类别的语义分割精度。以上基于 DeepLab 模型、CRFasRNN 模型的方法由于模型训练中的池化操作损失了图像细节信息, 最终的语义分割精度较低。

为了充分利用遥感图像的细节信息和空间上下文信息, 提高图像语义分割精度, 本文提出一种深度融合网 DFN (Deep Fusion Networks) 结合条件随机场模型的遥感图像语义分割方法。方法在全卷积神经网络框架中加入反卷积融合结构搭建 DFN 网络, 不仅利用全卷积网络智能提取深度语义特征的优势, 提升了模型泛化能力; 而且通过反卷积融合结构引入模型训练中的多尺度特征,

将浅层细节信息和深层语义信息结合起来做模型预测, 有效地提高模型的处理精度; 之后借助全连接条件随机场二阶势函数进一步引入上下文信息, 充分考虑像素与像素之间的关系, 在细化语义分割结果边缘的同时提高语义分割精度。

2 模型方法

本文提出的遥感图像语义分割算法流程如图 1 所示, 主要分为 2 个阶段: 模型训练和测试阶段。

模型训练阶段, 首先采用全卷积网络结合反卷积融合结构搭建 DFN 网络模型, 然后利用制作好的图像数据集和标签数据集, 采用随机梯度下降进行模型参数学习, 完成 DFN 模型的训练。测试阶段, 即图像语义分割, 对待处理的遥感图像, 首先利用训练阶段得到的 DFN 模型完成初步语义分割; 然后通过全连接条件随机场模型进行语义分割结果优化, 将 DFN 模型输出作为一阶势, 借助能量函数的构造将上下文信息融入 CRF 模型构建过程, 更好地定位边界, 最后采用平均场算法推断出各像素标签得到最终的语义分割结果。

2.1 全卷积神经网络

Krizhevsky 等 (2012) 使用深度卷积神经网络模型赢得大规模图像识别比赛的冠军。自此, 卷积神经网络技术开始被国内外研究者们关注并应用到相关领域中 (吕永标等, 2017; 唐智川等, 2017; 张洪群等, 2017)。卷积神经网络的优势在于多层结构自动学习不同层次特征, 网络中浅层特征图具有更丰富的细节信息, 经过多次卷积和池化操作提取出的特征图具有更深层次的语义信息。

本文中的全卷积神经网络主要由 4 个部分组成: 输入层, 卷积层, 池化层, 输出层。

输入层: 读取数据为 $h \times g \times d$ 的 d 维数组, 其中 h 和 g 为数据尺寸大小, d 为特征或通道维数。

卷积层: 卷积层又叫做特征提取层, 是使用具有特定权重的滤波器对输入数据进行卷积的过程, 不同卷积核提取的特征各不相同, 卷积结果经过激活函数输出, 本文采取的激活函数为 ReLU (Rectified Linear Units), 定义为

$$f(x) = \max(0, x) \quad (1)$$

池化层 (pool): 也叫下采样层。为避免训练过多参数发生过拟合问题, 使用池化层在保留有用信息的基础上减少数据处理量, 加快网络训练

的速度。常用的池化操作有最大池化, 平均池化。卷积层和池化层可定义为 (Long等, 2015)

$$y_{ij} = f_{ks} \left(\left(x_{si + \delta i, sj + \delta j} \right)_{0 \leq \delta i, \delta j < k} \right) \quad (2)$$

式中, x_{ij} 为位置 (i, j) 处的数据向量, y_{ij} 为经过

层操作后 (i, j) 处的输出向量, k 为卷积核尺寸, s 为步长或下采样因子, f_{ks} 为层的类型 (代表卷积层的矩阵乘或者平均池化操作及 ReLU 函数输出层)。

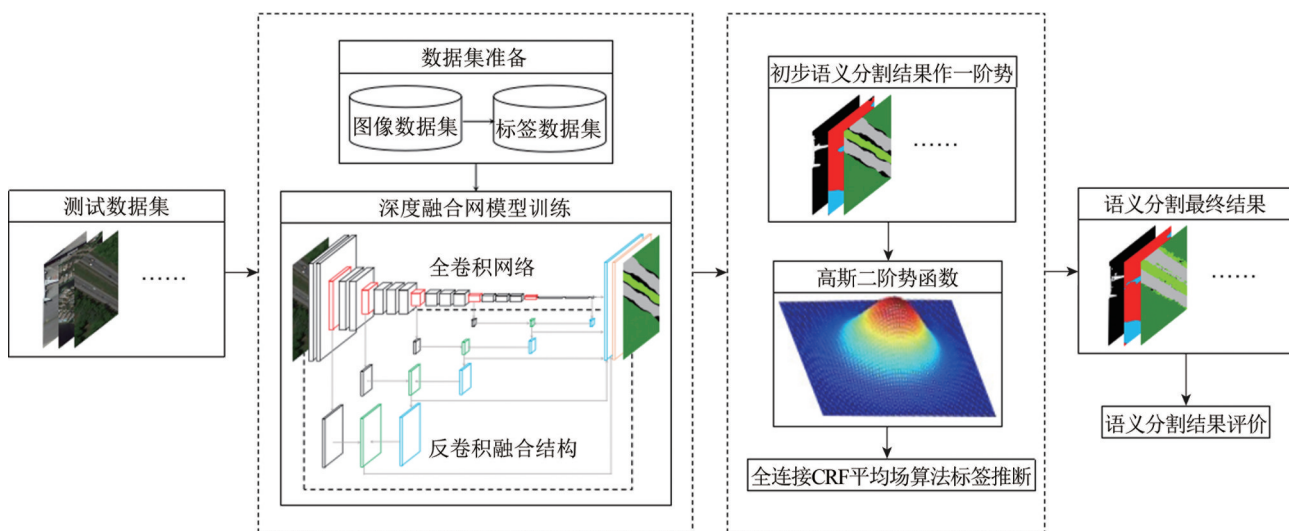


图1 遥感图像语义分割算法流程图

Fig.1 Flow chart of remote sensing image semantic segmentation

输出层: 通过 Softmax 分类器输出每个像素点属于各语义类别的概率向量。

2.2 反卷积融合结构

全卷积网络模型训练中的池化操作不仅缩小了图片尺寸, 同时损失了图像中丰富的细节信息, 使得语义分割精度较低。因此, 本研究采用反卷积操作来恢复原始图像大小, 实现像素到像素的图像语义分割。同时, 为了获得更高精度的语义分割结果, 全卷积网络加入了融合结构, 通过结合不同尺度的池化层结果和最终的卷积层结果引入网络模型中的浅层细节信息。

2.2.1 反卷积操作

反卷积操作即在神经网络模型的正向和反向传播中做与卷积操作相反的运算, 以 Caffe 框架中的反卷积操作为例:

首先分析卷积操作的前向和反向传播过程, 前向传播过程为

$$O = A \times B \quad (3)$$

式中, A 为卷积核矩阵, B 为图像特征矩阵, O 为输出矩阵。反向传播时, 根据矩阵微分公式

$$\frac{\partial Mx + b}{\partial x} = M^T \quad (4)$$

式中, M 表示任意矩阵, M^T 为任意矩阵的转置矩阵, b 为任意常数。推出

$$\frac{\partial Loss}{\partial B} = \frac{\partial Loss}{\partial O} \cdot \frac{\partial O}{\partial B} = A^T \frac{\partial Loss}{\partial O} \quad (5)$$

式中, $Loss$ 为损失函数, A^T 为 A 的转置矩阵。所以反卷积操作就是正向传播时左乘 A^T , 反向时左乘 $(A^T)^T$, 即 A 的运算。

2.2.2 融合结构

为了充分利用图像的细节信息, 采用结合浅层细节信息和深层语义信息的融合结构。该融合结构主要由卷积层、反卷积层以及结合层组成, 其中结合层是将来自不同层的结果求和后输出。全卷积网络模型最终的卷积层结果仅包含了最后一层池化层的信息, 直接将这一结果进行反卷积操作得到的模型称为 DFN-5。根据所结合池化层深度的不同, 记为模型 DFN-4、DFN-3、DFN-2、DFN-1, 分别表示结合池化层深度到 pool4、pool3、pool2、pool1。

以 DFN-4 为例, 融合结构即将最后卷积层的结果反卷积到 pool4 层输出结果尺寸, 与 pool4 层卷积后结果通过结合层输出, 再将结合层结果反卷积到输入图像尺寸利用 Softmax 分类器得到最终的语义分割结果。模型 DFN-3 即将 DFN-4 中结合层

结果反卷积到 pool3 层输出结果尺寸，与 pool3 层卷积后结果通过结合层输出，再将结合层结果反卷积到输入图像尺寸进行 Softmax 分类，模型 DFN-2、DFN-1 以此类推。

深度融合网络模型如图 2：图 2 中池化层从左到右依次为 pool1、pool2、pool3、pool4、pool5。各层旁边标注为卷积（或池化、反卷积）核大小以及核个数。所有卷积层步长为 1；池化层核大小为 2，步长为 2；反卷积层步长为反卷积核尺寸一半。

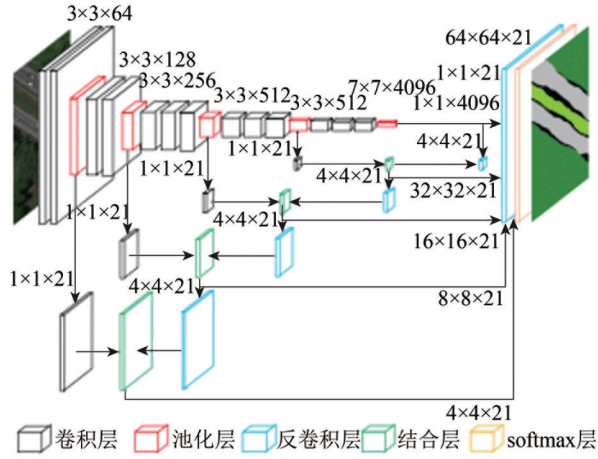


图 2 深度融合网络模型图

Fig.2 Model map of deep fusion networks

2.3 全连接条件随机场

条件随机场模型最早是由 Lafferty 等（2001）提出的一种判别式概率无向图学习模型，后来 Kumar 和 Hebert（2003）将 CRF 模型扩展到 2 维图像处理领域，进一步推动了 CRF 模型在图像分割、图像分类、目标识别（李祖传 等，2011；Zhao 等，2015；刘威 等，2017）等领域的快速发展。CRF 模型同时利用原始图像和标签图像的上下文信息，直接对类后验概率进行建模。在 DFN 模型之后使用 CRF 模型，不仅可以利用 DFN 模型提取像元本质特征和充分使用图像细节信息的优势，而且通过 CRF 模型的二阶势函数考虑空间上下文信息，提高图像语义分割精度。

对于观测随机场 $\mathbf{X}$: $\{X_1, X_2, \dots, X_n\}$ 和 $\mathbf{I}$: $\{I_1, I_2, \dots, I_n\}$ ，当给定标记随机场 $\mathbf{I}$ ， $\mathbf{X}$ 的条件分布满足马尔可夫性质

$$p(X_i | \mathbf{I}, X_j, j \neq i, j \in S) = p(X_i | \mathbf{I}, X_j, j \in N_i) \quad (6)$$

式中， S 为像素集合， N_i 为像素 i 的邻域位置集合，那么 $(\mathbf{X}, \mathbf{I})$ 就称为一个条件随机场。对于遥感

图像语义分割而言，假定 $\mathbf{I}$ 表示给定的遥感图像， $\mathbf{X}$ 表示相应的标签图像，总共有 n 个像素，其标签类别有 K 个， I_i 为像素 i 的光谱信息， X_i 为像素 i 的类别标签，其取值范围为 $L = \{l_1, l_2, \dots, l_K\}$ 。

根据 Hammersley-Clifford 定理（Li 等，2009），标记场 $\mathbf{X}$ 的后验概率服从 Gibbs 分布

$$P(\mathbf{X} | \mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp \left\{ - \sum_{c \in \mathbf{C}} \phi_c(\mathbf{X}_c | \mathbf{I}) \right\} \quad (7)$$

$$Z(\mathbf{I}) = \sum_{\mathbf{X}} \exp \left\{ - \sum_{c \in \mathbf{C}} \phi_c(\mathbf{X}_c | \mathbf{I}) \right\} \quad (8)$$

式中， $Z(\mathbf{I})$ 为归一化因子， ϕ_c 是定义在基团 c 上带有待估参数的势函数，根据基团所包含变量个数不同，势函数通常分为一阶势、二阶势以及高阶势。标记值 $x \in L^K$ 对应的 Gibbs 能量表示如下

$$E(\mathbf{x} | \mathbf{I}) = \sum_{c \in \mathbf{C}} \phi_c(\mathbf{x}_c | \mathbf{I}) \quad (9)$$

标记推断的目标就是求出使后验概率 $P(\mathbf{x} | \mathbf{I})$ 最大，Gibbs 能量 $E(\mathbf{x} | \mathbf{I})$ 最小的标记 $\mathbf{x}^*$ 。本文采用的 Gibbs 能量函数 $E(\mathbf{x} | \mathbf{I})$ （Krähenbühl 和 Koltun，2012）表示如下（以下公式中均默认给 $\mathbf{I}$ ）

$$(E(\mathbf{x}) =) \sum_i \psi_1(x_i) + \sum_{i \in S, j \in N_i} \psi_2(x_i, x_j) \quad (10)$$

$$\psi_1(x_i) = -\ln P(x_i) \quad (11)$$

$$\psi_2(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^K \omega^{(m)} k^{(m)}(f_i, f_j) \quad (12)$$

$$k^{(m)}(f_i, f_j) = \exp \left(-\frac{1}{2} (f_i - f_j)^T \Lambda^{(m)} (f_i - f_j) \right) \quad (13)$$

式中， $\psi_1(x_i)$ 为一阶势函数，描述单像素点观测信息与其相应标记之间的关系。本文中 $P(x_i)$ 为上一步 DFN 模型计算出的像素 i 位置被分配为各标签的概率。 $\psi_2(x_i, x_j)$ 为二阶势函数，描述像素之间的关系，鼓励相似的像素分配相同的标签，相差较大的像素分配不同的标签。本文采用高斯核函数的线性组合作为二阶势函数，构成全连接条件随机场模型。式（12）中如果 $x_i \neq x_j$ ，那么 $\mu(x_i, x_j) = 1$ ，否则为 0。 $\omega^{(m)}$ 为各高斯核函数相应的权重， $k^{(m)}(f_i, f_j)$ 为高斯核函数， f_i, f_j 表示像素 i 位置和像素 j 位置的特征向量， $\Lambda^{(m)}$ 为与核 $k^{(m)}$ 对应的对称的正定精度矩阵。本文采用的包含两个高斯核函数的线性组合 $k(f_i, f_j)$ 如下

$$k(f_i, f_j) = \omega^{(1)} \exp \left(-\frac{\|g_i - g_j\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2} \right) + \omega^{(2)} \exp \left(-\frac{\|g_i - g_j\|^2}{2\sigma_\gamma^2} \right) \quad (14)$$

式中, 像素 i 的光谱信息 I_i 和位置信息 g_i 构成了特征向量 f_i 。参数 σ_α 、 σ_β 和 σ_γ 控制高斯核的尺度。式 (14) 的第一项鼓励特征相似的像素具有相同标签, 第二项主要是在推断标记时考虑空间相似度问题。

由于全连接条件随机场边连接数量较大导致直接求解后验概率比较困难, 本文采用近似推断的平均场算法 (Krähenbühl 和 Koltun, 2012) 来实现, 通过最小化 KL-距离 $D(Q||P)$, 以方便实现计算的近似分布 $Q(X)$ 代替精确分布 $P(X)$

$$Q(X) = \prod_i Q_i(X_i) \quad (15)$$

子分布 $Q_i(X_i)$ 通过迭代更新的方法计算

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp(-\psi_i(x_i) - \sum_{l' \in L} \mu(l, l') \sum_{m=1}^K \omega^{(m)} \sum_{j \neq i} k^{(m)}(f_i, f_j) Q_i(l')) \quad (16)$$

具体算法如下:

(1) 初始化 Q_i

$$\frac{1}{Z_i} \exp(-\psi_i(x_i)) \rightarrow Q_i(x_i), \forall i \in S$$

(2) $Q(X)$ 和 $P(X)$ 间的 K-L 散度不收敛时迭代更新 Q_i

1) 消息传递

$$\sum_{j \neq i} k^{(m)}(f_i, f_j) Q_j(l) \rightarrow \tilde{Q}_i^{(m)}(l), \forall m \in [1, K]$$

2) 滤波结果求加权和

$$\sum_m \omega^{(m)} \tilde{Q}_i^{(m)}(l) \rightarrow Q'_i(l)$$

3) 兼容性转换

$$\sum_{l \in L} \mu(x_i, l) Q'_i(l) \rightarrow \hat{Q}_i(x_i)$$

4) 一元项加入

$$-\psi_i(x_i) - \hat{Q}_i(x_i) \rightarrow \bar{Q}_i(x_i)$$

5) 类别概率归一化

$$\frac{1}{Z_i} \exp(\bar{Q}_i(x_i)) \rightarrow Q_i(x_i)$$

3 实验与分析

实验主要基于 Linux 系统的 Caffe 框架下完成, 使用 GPU 加速, 服务器处理器为 Intel Xeon (R) CPU E5-2620 0 @ 2.00GHz, 显卡为 NVIDIA GeForce GTX TITAN X, 采用 Python 2.7.13 与 Matlab R2016b 混合编程。

3.1 实验数据

选取 20 种遥感地物语义类别进行实验, 分别为: 飞机、菱形棒球场、灌木丛、圆形农田、密集住宅区、沙漠、森林、高速公路、高尔夫球场、湖泊、草地、中等密集住宅区、山、停车场、矩形农田、河流、海冰、稀疏住宅区、梯田、湿地。模型训练及测试数据来自公开的遥感数据集 NWPU (Cheng 等, 2017), 每类 700 张图像, 包含 R、G、B 共 3 个波段数据, 大小为 256×256 像素。NWPU 数据集在视角、对象姿势、空间分辨率、光照、阴影等方面都有丰富的变化, 具有较高的类内多样性和类间相似性, 能够很好地检验语义分割方法的可靠性。同时, 为验证本文方法模型的泛化能力, 本文还采用来自 UC Merced Land-Use 数据集 (Yang 等, 2010) 的遥感图像进行了语义分割实验, 该数据集来源为美国 USGS (United States Geological Survey) 国家城市地图中的航空遥感影像, 分辨率为 0.3 m。

3.2 评价准则

从定性和定量两个方面进行语义分割结果评价。定性评价主要从直观感觉如分割边缘是否明确、轮廓是否清晰等方面进行评估; 定量评价主要采用像素平均精度 PA (Pixel Accuracy)、类别平均精度 CA (Class Accuracy) 以及平均像素交叠率 MIOU (Mean Intersection Over Union) 3 个指标, 具体计算方法如下

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_i} \quad (17)$$

$$CA = \frac{1}{n_c} \sum_i \frac{n_{ii}}{t_i} \quad (18)$$

$$MIOU = \frac{1}{n_c} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \quad (19)$$

式中, n_c 为数据集语义类别数, n_{ij} 为类别 i 被预测为类别 j 的像素总数, $t_i = \sum_j n_{ij}$ 。

3.3 实验结果与分析

3.3.1 细节信息影响分析

为了分析细节信息对遥感图像语义分割结果的影响, 对具有结合不同深度池化层的融合结构的网络模型进行对比实验, 包括模型 DFN-5、

DFN-4、DFN-3、DFN-2 和 DFN-1。实验随机选取数据集集中的 12000 张遥感图像进行训练，其余作为测试图像。各模型迭代 50000 次的训练时间及一幅图像的平均测试时间如表 1 所示。其中深度融合网络的训练参数设置为：学习率固定为 1×10^{-10} ，动量设置为 0.99，权重衰减参数为 0.0005，采用随机梯度下降 SGD (Stochastic Gradient Descent) 优化算法进行网络训练，反卷积层权重初始化采用双线性插值法。

表 1 各模型训练及测试时间

Table 1 Training and testing time of each model					
模型	DFN-5	DFN-4	DFN-3	DFN-2	DFN-1
训练时间/min	341	4828	4904	5346	5749
测试时间/s	0.58	0.64	0.69	0.70	0.72

结果如图 3 所示。其中 DFN-5 的语义分割结果非常粗糙（在图 3 (c) 第 1 行图中基本看不出飞机的轮廓，第 2 行图中草地与高速公路没有分割

开，第 3 行图中右上角湖泊未识别出且其余两个湖泊缺失部分较多）。DFN-4 模型中加入 pool4 层的细节信息，其语义分割结果较 DFN-5 在以上 3 处不足都有明显的改善（图 3 (d) 第 1 行图中飞机轮廓渐显，第 2 行图中草地与公路也分割开，第 3 行图中 3 处湖泊及 DFN-5 结果缺失部分基本上都识别出来），但图 3 (d) 中仍不能明显看出物体为飞机，且公路间的草地部分未识别出。DFN-3 为在 DFN-4 基础上又融合 pool3 层细节信息进行语义分割，在加入不同尺度的细节信息后，图 3 (e) 中已经可以较清楚地看出飞机轮廓，公路间的草地也被识别出来。DFN-2 和 DFN-1 各自又在前一个模型的基础上融合 pool2 层和 pool1 层信息，从结果图中可以看出 DFN-2 和 DFN-1 的语义分割结果边缘更加精细（如图 3 (f)、图 3 (g) 中，飞机尾部更细化、最右侧公路与森林逐渐分割开、左下角湖泊识别部分增加）。

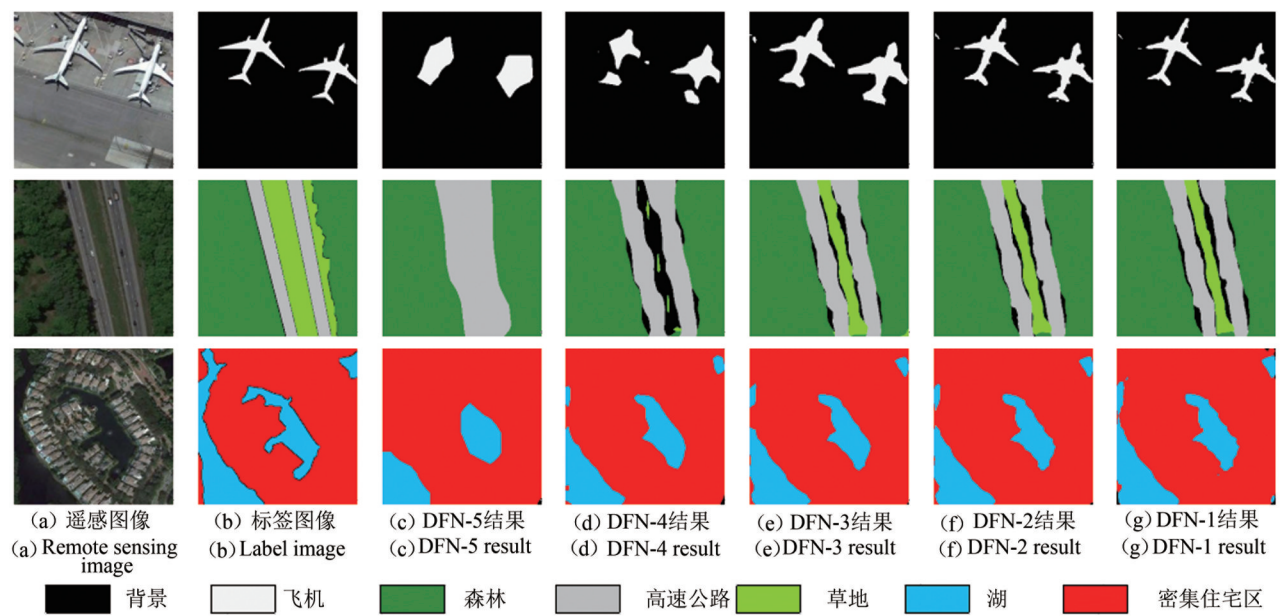


图 3 采用不同层级融合结构的语义分割结果对比图背景
Fig.3 Comparison diagram of semantic segmentation results using different levels of fusion structure

从图 3 中可以看出，DFN-5 模型的语义分割结果最差，这是由于 DFN-5 为 pool5 层输出直接上采样的语义分割结果，没有加入任何细节信息；从 DFN-4 到 DFN-1，随着融合层深度的增加，由于在模型训练的过程中加入了不同尺度的细节信息，语义分割结果边缘更加细化、精确，逐渐接近标签图像。

各模型语义分割精度如表 2 所示。由表 2 可知，从 DFN-5 到 DFN-3，PA、CA、MIOU 值都显著增加，与定性评价结果一致。从 DFN-3 到 DFN-1，MIOU 值逐渐提高，PA 和 CA 值略微降低。为了进一步分析，图 4 展示了不同模型各类地物的 PA 值，从图 4 中可以看出，对于不同地物类别，并非融合层深度越深，模型的 PA 值越高，同时可以看出，

细节信息的融合深度对于具有丰富细节信息的地物类别的影响也会更大。以飞机类别为例,图5展示了不同模型(从DFN-3到DFN-1)的语义分割结果和标签结果叠加示意图。从图中可以看出,从DFN-3到DFN-1,由于加入更多不同尺度的细节信息,大幅减少了结果图中被错分为类别飞机的像素数量,但细化后的边缘使得类别飞机被分为其他类别的像素数量微量增加。PA值计算公式中分母值一定,分子仅考虑类别*i*被正确语义分割为类别*i*的像素数量,所以从DFN-3到DFN-1,PA值略微下降。当被错分的像素(类别*i*被错分的像素或被错分为类别*i*的像素)数量较少时,对所有测试的标签图像中像素总数量多的类别的PA值影响较小(如图4中类别密集住宅区、森林等),对像素总数量少的类别的PA值影响较大(如图4中类别飞机、高速公路、河等)。CA值为所有类别PA值的均值,所以语义分割结果更细化的DFN-1、DFN-2模型中像素数量多的类别的PA值增加的总程度小于像素数量小的类别的PA值减少的总程度,CA值下降。

表2 DFN模型语义分割结果定量评价

Table 2 DFN models' quantitative evaluation of semantic segmentation results

模型	PA	CA	MIOU
DFN-5	0.908352	0.854187	0.766236
DFN-4	0.910000	0.872659	0.796386
DFN-3	0.918074	0.896049	0.806497
DFN-2	0.914048	0.887421	0.810139
DFN-1	0.912465	0.883090	0.811262

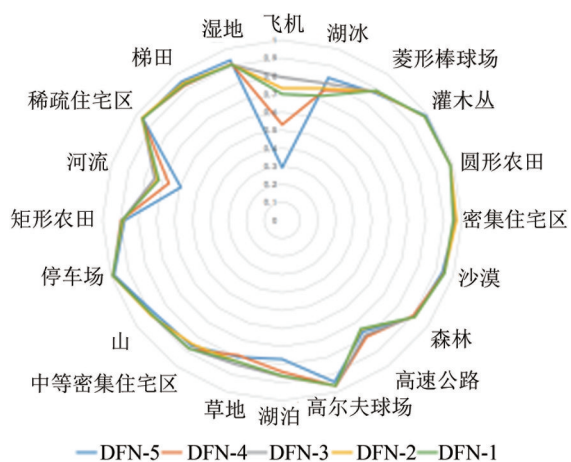


图4 各模型结果所有类别PA对比图

Fig.4 PA contrast diagram of different model's results on all categories

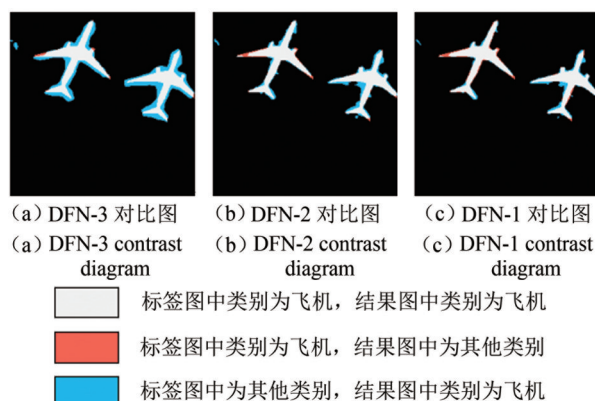


图5 语义分割结果与标签图像叠加图

Fig.5 Superposition graph of semantic segmentation result and label image

MIOU 由于同时考虑了类别*i*被错分的像素与被错分为类别*i*的像素,所以相比于PA和CA,MIOU更符合本文对语义分割结果的评价需求,因此后续对比试验中采用MIOU指标进行实验结果的对比评价。

3.3.2 上下文信息优化分析

以MIOU指数为指标,深度融合网络采用DFN-1模型为基础进行语义分割实验。为分析上下文信息对遥感图像语义分割的优化效果,对深度融合网络模型DFN-1及其结合全连接条件随机场引入上下文信息的模型DFN-1-CRF进行对比实验。首先对全连接条件随机场模型的6个参数: $\omega^{(1)}$ 、 σ_α 、 σ_β 、 $\omega^{(2)}$ 、 σ_γ 、MaxIteration(依次为二元项权重、二元项的高斯核尺度参数 σ_α 和 σ_β 、一元项权重、一元项高斯核尺度参数、模型最大迭代次数)进行了6组对比实验选取最优参数。实验先选取 $\omega^{(1)}$ 参数作变量,其他参数依次分别初始化为49、5、3、3、10;选择出最优的 $\omega^{(1)}$ 参数后,固定 $\omega^{(1)}$ 参数,选取 σ_α 参数作变量,剩余参数初始化同上,以此类推,得出最优参数,如图6所示(各柱状图中,横坐标为参数取值,纵坐标为语义分割结果评定的MIOU值)。最终选取参数 $\omega^{(1)}$ 、 σ_α 、 σ_β 、 $\omega^{(2)}$ 、 σ_γ 、MaxIteration分别为5、61、3、2、2、12的全连接CRF模型进行语义分割结果优化。全连接CRF模型处理1张测试图像平均用时0.69 s。

最终的语义分割结果如图7所示,可以明显看出,加入空间上下文信息后,语义分割结果边缘更细化,分割对象轮廓更清晰。对比图7(c)的DFN-1结果图,图7(d)DFN-1-CRF结果图的第1行中飞机已经与标签图像基本一致,同时还正确

标识出了左上角的飞机机翼部分；图7（d）第2行图中高速公路边界更明确，最右边草地区域也都被识别出来；图7（d）第3行图中左下角湖泊与密集住宅区分割开来，接近标签图像，其余两处

湖泊的轮廓也相较DFN-1结果更加细化。DFN-1与DFN-1-CRF定量评价如表3所示。从表3中可以看出，DFN-1-CRF较DFN-1模型，MIOU提高了0.90%左右，与定性评价结果一致。

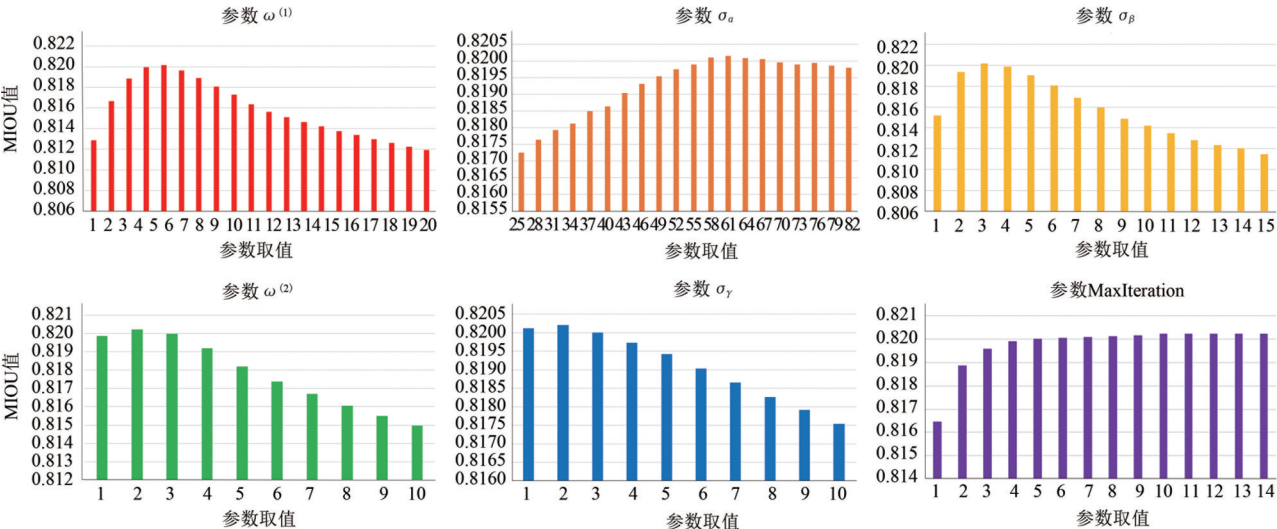


图6 全连接CRF模型参数选取对比实验结果图
Fig.6 Comparison of experimental results for parameter selection of fully connected CRF model

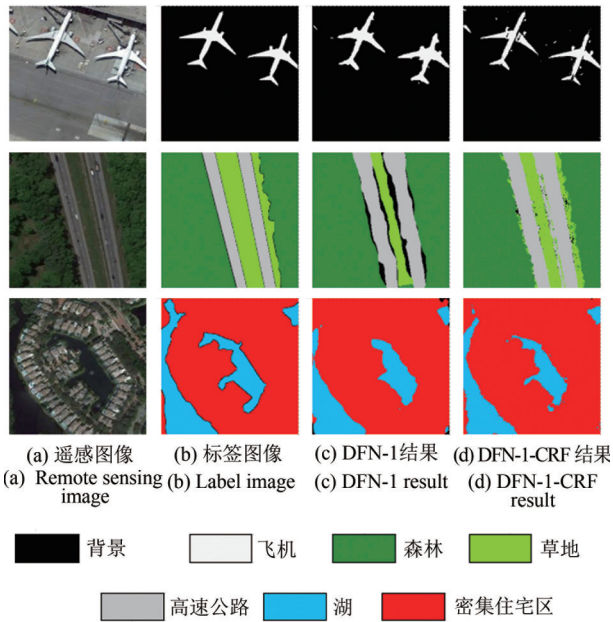


图7 DFN-1与DFN-1-CRF实验结果对比图
Fig.7 Comparison diagram of experimental results between DFN-1 and DFN-1-CRF

表3 语义分割结果定量评价

Table 3 Quantitative evaluation of semantic segmentation results

模型	训练时间/min	MIOU
DFN-1	5749	0.811262
DFN-1-CRF	5861	0.820219

3.3.3 与FCN8s、DeepLab模型对比实验

将本文方法与目前常用的基于卷积神经网络的FCN8s（Long等，2015）、DeepLab（Chen等，2018）模型方法进行对比实验，训练集和测试集保持一致。结果如表4所示，可以看出，DeepLab模型的训练时间最短，但精度最低；本文方法相比FCN8s模型训练时间增加了0.2倍，但结果精度更高。通过融合结构和全连接条件随机场加入了细节信息和空间上下文信息的DFN-1-CRF模型相比DeepLab模型、FCN8s模型MIOU值分别提高了3.68%和1.50%左右，在遥感图像数据集上取得了更高的精度。

3.3.4 泛化能力验证实验

实验中模型训练采用的遥感图像数据来自NWPU数据集，为了验证方法的泛化能力，采用UC Merced Land-Use数据集集中的航空遥感影像进行了语义分割实验，实验结果如图8所示。从实验结果图8中可以看出本文方法能够很好地将飞机、高速公路、河流与周围环境分割开来，且边缘精确、轮廓清晰；图8中的草地由于类内差异太大，大部分像素能够被正确语义分割为草地，但仍存在少量的误差。整体来看，本文方法在对与模型训练

数据集不同数据源的遥感图像进行语义分割时也 能够较好地实现，具有较强的泛化能力。

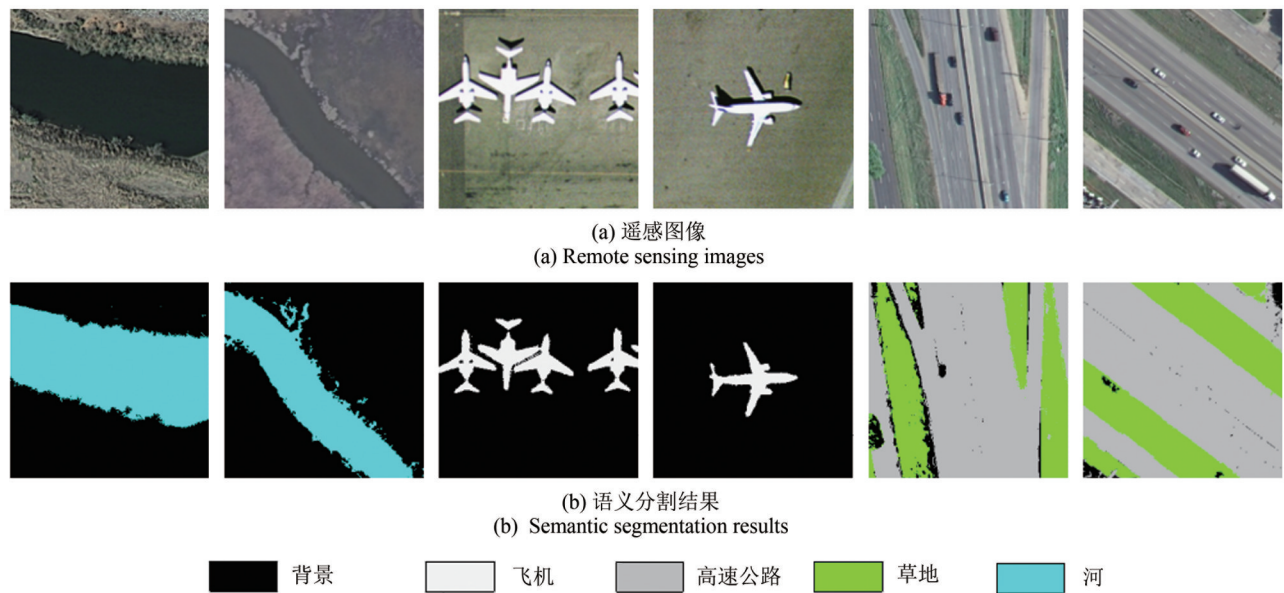


图8 UC Merced Land-Use数据集实验结果
Fig.8 Experimental results of UC Merced Land-Use data set

表4 本文方法与其他方法结果对比

Table 4 The comparison between our method and other methods

模型	训练时间/min	MIOU
DeepLab	1692	0.783401
FCN8s	4917	0.805267
DFN-1-CRF	5861	0.820219

4 结 论

针对目前常用的基于卷积神经网络的图像语义分割方法，处理过程中由于图像细节信息和上下文信息利用不充分，造成图像语义分割精度不高的问题提出了一种深度融合网结合条件随机场模型的遥感图像语义分割方法。方法首先采用全卷积神经网络模型经过多层卷积和池化得到图像深层语义特征，避免人工显式提取特征；然后将全卷积网络输出上采样到输入图像大小，再由融合结构结合不同尺度的池化层结果引入更丰富的细节信息输出；最后将深度融合网结果作为全连接条件随机场一阶势，通过高斯二阶势函数引入空间上下文信息更好地定位边界。在遥感图像数据集上实验表明：（1）随着融合结构结合池化层的深度的增加，所引入的细节信息更丰富，结果的边缘轮廓越接近标签图像，有效改善了语义分

割结果过平滑现象；（2）全连接条件随机场模型进一步综合了遥感图像的全局信息与局部信息，增加了空间上下文信息，从而细化了语义分割结果边缘，提高了最终语义分割的效果和精度。如何将上下文信息融入模型构建过程，优化网络模型，提升模型的处理速度和精度，是本文进一步的研究方向。

参考文献(References)

Arbeláez P, Hariharan B, Gu C H, Gupta S, Bourdev L and Malik J. 2012. Semantic segmentation using regions and parts//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE: 3378-3385 [DOI: 10.1109/CVPR.2012.6248077]

Badrinarayanan V, Kendall A and Cipolla R. 2017. SegNet: a deep convolutional encoder-decoder architecture for scene segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495 [DOI: 10.1109/TPAMI.2016.2644615]

Chang L, Deng X M, Zhou M Q, Wu Z K, Yuan Y, Yang S and Wang H A. 2016. Convolutional neural networks in image understanding. Acta Automatica Sinica, 42(9): 1300-1312 (常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 王宏安. 2016. 图像理解中的卷积神经网络. 自动化学报, 42(9): 1300-1312) [DOI: 10.16383/j.aas.2016.c150800]

Chen L C, Papandreou G, Kokkinos L, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE

- Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]
- Cheng G, Han J W and Lu X Q. 2017. Remote sensing image scene classification: benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865-1883 [DOI: 10.1109/JPROC.2017.2675998]
- Krähenbühl P and Koltun V. 2012. Efficient inference in fully connected CRFs with Gaussian edge potentials. *Advances in Neural Information Processing Systems* 24: 109-117
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//*Proceedings of the 25th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada Curran Associates Inc: 1097-1105
- Kumar S and Hebert M. 2003. Discriminative random fields: a discriminative framework for contextual interaction in classification//*Proceedings of the 9th IEEE International Conference on Computer Vision*. Nice, France: IEEE: 1150-1157 [DOI: 10.1109/ICCV.2003.1238478]
- Lafferty J D, McCallum A and Pereira F C N. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data//*Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc: 282-289
- Li S Z. 2009. Markov Random Field Modeling in Image Analysis. 3rd ed. London: Springer
- Li Z C, Ma J W, Zhang R and Li L W. 2011. Classifying hyperspectral data using support vector machine conditional random field. *Geomatics and Information Science of Wuhan University*, 36(3): 306-310 (李祖传, 马建文, 张睿, 李利伟. 2011. 利用SVM-CRF进行高光谱遥感数据分类. *武汉大学学报(信息科学版)*, 36(3): 306-310) [DOI: 10.13203/j.whugis.2011.03.009]
- Liu W, Zhou T, Yuan H and Zhao H. 2017. 3D spatial layout understanding from image based on multiple CRFs. *Acta Electronica Sinica*, 45(2): 328-336 (刘威, 周婷, 袁淮, 赵宏. 2017. 基于多条件随机场模型的图像3D空间布局理解. *电子学报*, 45(2): 328-336) [DOI: 10.3969/j.issn.0372-2112.2017.02.010]
- 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, MA, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Lü Y B, Zhao J W and Cao F L. 2017. Image denoising algorithm based on composite convolutional neural network. *Pattern Recognition and Artificial Intelligence*, 30(2): 97-105 (吕永标, 赵建伟, 曹飞龙. 2017. 基于复合卷积神经网络的图像去噪算法. *模式识别与人工智能*, 30(2): 97-105) [DOI: 10.16451/j.cnki.issn1003-6059.201702001]
- Montoya-Zegarra J A, Wegner J D, Ladický L and Schindler K. 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4: 127-133 [DOI: 10.5194/isprsannals-II-3-W4-127-2015]
- Shotton J, Johnson M and Cipolla R. 2008. Semantic texon forests for image categorization and segmentation//*Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587503]
- Smeulders A W M, Worring M, Santini S, Gupta A and Jain R. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1349-1380 [DOI: 10.1109/34.895972]
- Sun W W and Wang R S. 2018. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geoscience and Remote Sensing Letters*, 15(3): 474-478 [DOI: 10.1109/LGRS.2018.2795531]
- Tang Z C, Zhang K J, Li C, Sun S Q, Huang Q and Zhang S Y. 2017. Motor imagery classification based on deep convolutional neural network and its application in exoskeleton controlled by EGG. *Chinese Journal of Computers*, 40(6): 1367-1378 (唐智川, 张克俊, 李超, 孙守迁, 黄琦, 张三元. 2017. 基于深度卷积神经网络的运动想象分类及其在脑控外骨骼中的应用. *计算机学报*, 40(6): 1367-1378) [DOI: 10.11897/SP.J.1016.2017.01367]
- Volpi M and Ferrari V. 2015. Semantic segmentation of urban scenes by learning local class interactions//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Boston, MA, USA: IEEE: 1-9 [DOI: 10.1109/CVPRW.2015.7301377]
- Wei X, Guo Y J, Gao X, Yan M L and Sun X. 2017. A new semantic segmentation model for remote sensing images//*Proceedings of 2017 IEEE International Geoscience and Remote Sensing Symposium*. Fort Worth, TX, USA: IEEE: 1776-1779 [DOI: 10.1109/IGARSS.2017.8127319]
- Wei Y C and Zhao Y. 2016. A review on image semantic segmentation based on DCNN. *Journal of Beijing Jiaotong University*, 40(4): 82-91 (魏云超, 赵耀. 2016. 基于DCNN的图像语义分割综述. *北京交通大学学报*, 40(4): 82-91) [DOI: 10.11860/j.issn.1673-0291.2016.04.013]
- Xiao J S, Liu E Y, Zhu L and Lei J F. 2017. Improved image super-resolution algorithm based on convolutional neural network. *Acta Optica Sinica*, 37(3): 0318011 (肖进胜, 刘恩雨, 朱力, 雷俊峰. 2017. 改进的基于卷积神经网络的图像超分辨率算法. *光学学报*, 37(3): 0318011) [DOI: 10.3788/AOS201737.0318011]
- Yang W, Triggs B, Dai D X and Xia G S. 2010. Scene Segmentation with low-dimensional semantic representations and conditional random fields. *EURASIP Journal on Advances in Signal Processing*, 2010: 196036 [DOI: 10.1155/2010/196036]
- Yang Y and Newsam S. 2010. Bag-of-visual-words and spatial extensions for land-use classification//*Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. San Jose, California: ACM: 270-279 [DOI: 10.1145/1869790.1869829]
- Zhang H Q, Liu X Y, Yang S and Li Y. 2017. Retrieval of remote sensing images based on semisupervised deep learning. *Journal of Remote Sensing*, 21(3): 406-414 (张洪群, 刘雪莹, 杨森, 李宇. 2017. 深度学习的半监督遥感图像检索. *遥感学报*, 21(3): 406-414) [DOI: 10.11834/jrs.20176105]

- Zhang K, Hei B Q, Zhou Z and Li S Y. 2018. CNN with coefficient of variation-based dimensionality reduction for hyperspectral remote sensing images classification. *Journal of Remote Sensing*, 22(1): 87-96 (张康, 黑保琴, 周壮, 李盛阳. 2018. 变异系数降维的 CNN 高光谱遥感图像分类. *遥感学报*, 22(1): 87-96) [DOI: 10.11834/jrs.20187075]
- Zhang R, Li G Y, Li M L and Wang L. 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 143: 85-96 [DOI: 10.1016/j.isprsjprs.2018.04.022]
- Zhao J, Zhong Y F and Zhang L P. 2015. Detail-preserving smoothing classifier based on conditional random fields for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5): 2440-2452 [DOI: 10.1109/TGRS.2014.2360100]
- Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z Z, Du D L, Huang C and Torr P H S. 2015. Conditional random fields as recurrent neural networks//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago, Chile: IEEE: 1529-1537 [DOI: 10.1109/ICCV.2015.179]

Semantic segmentation of remote sensing image based on deep fusion networks and conditional random field

XIAO Chunjiao^{1,2}, LI Yu¹, ZHANG Hongqun¹, CHEN Jun¹

1. *Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China;*

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

Abstract: Image semantic segmentation refers to segmenting an image into several groups of pixel regions with different specific semantic meanings and identifying the categories of each region. In recent years, the common semantic segmentation methods that are based on Convolutional Neural Networks (CNN) have realised the pixel-to-pixel image semantic segmentation. They can avoid the problems of artificial design and selection of features in traditional image semantic segmentation methods. As a result of the pooling operation and lack of context information, the detailed information of images is neglected, the precision of the final image semantic segmentation result is low and the segmentation edge is inaccurate. Therefore, this study proposes a semantic segmentation method for remote sensing image on the basis of Deep Fusion Networks (DFN) combined with a conditional random field model.

The method initially builds a DFN model in a Fully Convolutional Network (FCN) framework with a deconvolutional fusion structure. On the one hand, the multiscale features can be extracted through the deep networks, which can avoid the artificial design and selection of features to improve the generalisation ability of the model. On the other hand, the multiscale information is used in the model with the help of the deconvolutional fusion structure. The processing accuracy of the model is also improved by fusing the shallow detail information and deep semantic information. Fundamentally, the fully connected conditional random field is introduced to supplement the spatial context information towards precisely locating the boundary and obtaining final semantic segmentation results.

From this study, we can draw the following conclusions:

(1) With the increase in the depth of the fusion layer, detailed information becomes abundant, the semantic segmentation results become refined and the edge contour becomes close to the label image;

(2) The fully connected conditional random field model synthesises the global and local information of the remote sensing image and further improves the efficiency and accuracy of the final semantic segmentation results.

Key words: remote sensing image semantic segmentation, fully convolutional networks, conditional random field, fusion structure, deconvolution

Supported by National Natural Science Foundation of China (No. 61501460)