融合余弦退火与空洞卷积的遥感影像语义分割

唐振超[1]，韦蔚[2]，罗蔚然[3]，胡洁[2]，张东映[1]

1. 华中科技大学 土木与水利工程学院, 武汉 430074;
2. 黄河勘测规划设计研究院有限公司, 郑州 450003;
3. 郑州大学 水利科学与工程学院, 郑州 450001

摘 要：为了捕捉遥感影像中丰富的上下文信息与多尺度的地物信息，改进集成模型的策略，提高语义分割精度，提出一种融合周期递增余弦退火与多尺度空洞卷积的高分辨率遥感影像语义分割方法。方法引入多尺度并行的空洞卷积，有利于捕捉更大范围的上下文信息，在不增加参数的情况下，提高网络对多尺度对象的辨识能力；使用全连接条件随机场引入空间和边缘的上下文信息，提高网络对遥感影像的细节分割能力；引入周期递增的余弦退火策略调整学习率，获得合适数量的局部最优解，集成局部最优解进一步提升网络在像素上的分类能力。在Gaofen Image Dataset数据集上的实验结果表明，多尺度并行空洞卷积可以充分捕捉遥感影像上的多尺度地物信息，能有效辨识复杂对象；空间和边缘上下文信息的引入使语义分割对象的边界辨识更精准；周期递增余弦退火策略能明显减少集成模型的推理时间，模型的总体精度与Kappa系数均优于目前主流的语义分割模型。

关键词：高分辨率遥感影像，语义分割，周期递增余弦退火，多尺度并行空洞卷积，目标提取，上下文学习，条件随机场，多尺度学习

中图分类号：TP751.1/P2

引用格式：**唐振超，韦蔚，罗蔚然，胡洁，张东映.2023.融合余弦退火与空洞卷积的遥感影像语义分割.遥感学报，27(11)：2579-2592**

Tang Z C, Wei W, Luo W R, Hu J and Zhang D Y. 2023. Remote sensing image semantic segmentation method combining cosine annealing with atrous convolution. National Remote Sensing Bulletin, 27(11):2579-2592[DOI:10.11834/jrs.20211038]

1 引 言

高分辨率遥感影像语义分割作为数据到信息对象化提取的过渡环节与关键步骤，是高分辨率遥感影像解译的典型任务。传统的高分辨率遥感影像解译通常采用人工目视解译方式，费时费力且精度低。近年来，随着人工智能技术的发展，采用深度学习方法实现高分辨率遥感影像解译已成为主流的研究方向（Zhou等，2021）。最近的工作表明，深度卷积神经网络结合条件随机场的方法已在高分辨率遥感影像语义分割任务上取得广泛应用（Li等，2020）。

自从全卷积神经网络FCN（Long等，2015）首次被用于图像的语义分割后，各种网络不断被提出和改进，segnet（Badrinarayanan等，2017）通过保留池化索引提高分割效果，unet（Ronneberger等，2015）基于U型结构使网络融合不同尺度的信息。Sun和Wang（2018）提出全卷积神经网络结合数字高程模型DEM，通过引入高程信息提高遥感影像的语义分割效果。但是标准卷积的感受野较小，缺乏上下文信息。因此，从deeplabv1（Chen等，2015）开始，使用了空洞卷积（Yu和Koltun，2016），空洞卷积在不增加参数的情况下保持分辨率并扩大感受野，有利于捕捉更大范围的上下文信息。Wang等（2020）设计了空洞卷积组块，在结冰湖面误提取，阴影漏提取，以及提取结果完整性等方面，具有较好的效果。但对于上述堆叠空洞卷积组块的模型，容易出现网格效应（Anthimopoulos等，2019），遥感影像的地物对象会呈现出异常的网格区域。Wang等（2018）提

收稿日期：2021-02-09；**预印本：**2021-06-02

第一作者简介：唐振超，研究方向为高分辨率卫星遥感影像智能解译。E-mail：u201715748@hust.edu.cn

通信作者简介：张东映，研究方向为卫星遥感智能解译与定量分析。E-mail：zhangdongying@hust.edu.cn

出标准化结构HDC，按照锯齿状的规律设置膨胀率并堆叠空洞卷积可以缓解网格效应；与HDC的串行结构相反，空洞空间金字塔（Chen等，2018）提出并行结构，该方法使用不同膨胀率的空洞卷积对特征执行并行的卷积计算。

标准卷积与空洞卷积缺乏空间与边缘上下文信息的约束（Teichmann和Cipolla，2019）。全连接条件随机场CRF（Krähenbühl和Koltun，2011）是一种判别式概率无向图学习模型，可充分考虑影像全局结构信息。Zhao等（2020）使用CRF结合Pauli相干分解重建假彩色图，对FCN的输出进行全局像素类别转移获得分割结果，在高分三号C频段PolSAR影像上取得了较好的精度。

深度学习模型训练通常采取学习率递减的优化方式，该策略导致模型收敛于局部最优。余弦退火方法（Loshchilov和Hutter，2017），通过学习率急剧上升帮助模型跳出局部最优解，该策略使学习率递减到一定值再急剧上升，如此往复。snapshot ensembling（Huang等，2017）提出在使用余弦退火策略训练时，保留各个局部最优解，推理时集成局部最优解可以使集成模型的分类精度明显超越单一模型。但经典余弦退火策略使用相同的周期调整学习率，生成过多局部最优模型，导致集成模型所花费的推理时间大幅增加。因此本文引入周期递增余弦退火策略，能有效减少集成模型的推理时间。

为了充分利用遥感影像中丰富的上下文信息，改进集成模型的学习策略，提高语义分割精度，本文提出一种融合周期递增余弦退火与多尺度空洞卷积的高分辨率遥感影像语义分割方法。本文方法采用并行的多尺度空洞卷积充分捕捉遥感影像上的多尺度地物信息，使模型能有效辨识不同尺度的复杂对象；基于全连接条件随机场引入空间和边缘上下文信息，细化语义分割结果；使用周期递增余弦退火方法作为学习策略，以减少集成模型的推理时间，并提高遥感影像的语义分割精度。

2 方 法

2.1 多尺度空洞卷积网络

对于普通的标准卷积，假设有离散的函数F：$\mathbb{Z}^2 \to \mathbb{R}$，有$\boldsymbol{\Omega}_r = [-r, r]^2 \cap \mathbb{Z}^2$，令$k$为一个离散的卷积核：$\boldsymbol{\Omega}_r \to \mathbb{R}$，则以$p$为中心展开的卷积可以描述为

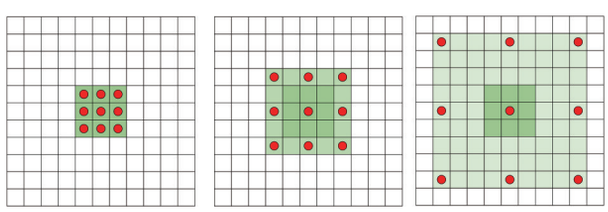
$$(F*k)(p) = \sum_{s+t=p} F(s)k(t) \tag{1}$$

对标准卷积进行扩充，令l表示空洞卷积的膨胀率，则空洞卷积可以描述为

$$(F*_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{2}$$

可见，标准卷积是空洞卷积的特殊形式，当空洞卷积膨胀率为1时，空洞卷积等价于标准卷积。

如图1所示，图1(a)，(b)，(c)分别对应空洞卷积膨胀率为1，2，4的情况，可以看出，当空洞卷积膨胀率逐渐增加，感受野随之增大。



（a）膨胀率为1 （b）膨胀率为2 （c）膨胀率为3
（a）Dilation is one （a）Dilation is two （c）Dilation is three

图1 空洞卷积采样示意图
Fig. 1 Sampling diagram of atrous convolution

空洞卷积可以通过设置膨胀率在特征上稀疏采样，在密集计算任务中，有利于控制感受野，增加上下文信息。空洞卷积膨胀率的设置不影响原始网络参数的结构，有利于模型的迁移学习，因此，可以方便地设置膨胀率并基于原始网络的参数进行微调。

在深层网络提取特征的过程中，拟合残差比拟合恒等映射更加容易，在resnet（He等，2016）中，跳接是实现该结构的方式，将卷积网络跳接并封装成为残差块。多个残差块堆叠可以加深网络并确保模型学习到高层信息。本文基于resnet101作为基本框架，使用到resnet101的第1层至第4层，使用较深的层是为了捕捉到较高层的语义信息，更高层的信息有助于提高分类的准确率。在resnet101中，低层网络使用标准卷积，高层网络的卷积设置膨胀率为2，即利用空洞卷积获取对象的周边信息。网络低层使用标准卷积是为了完整提取对象的特征，如果在低层直接使用空洞卷积，网络会过度关注对象周边的低层信息，削弱网络对真实对象的理解能力；另外，基于特征进行空洞卷积，

有助于网络理解对象周边信息的高层语义。

在深层网络中，连续堆叠相同膨胀率的空洞卷积容易引起网格效应，由于空洞卷积模板在特征上执行的是一种膨胀计算方式，所以卷积过程中会丢失部分特征的信息，信息损失对于空间密集的分割任务来说是不利的（Dumoulin 和 Visin，2016）。另外，当空洞卷积模板尺寸较小，但膨胀率较大时，对于较大目标的对象，空洞卷积依然能够感知到，但对于小目标对象，容易在计算中被忽略。为了改善空洞卷积带来的问题，可以采用对输入特征进行多尺度并行卷积的方法，并行结构可以有效处理多尺度对象，多尺度并行卷积的结构类似于 pspnet 的空间金字塔池化（Zhao 等，2017）和 deeplabv2 的空洞空间金字塔池化。

如图 2 所示，空洞空间金字塔可以对给定输入特征以不同膨胀率的空洞卷积进行采样，在不同尺度上捕捉特征的上下文信息。遥感影像的语义分割对象尺度大小一般很极端，平原上可能草地的尺度远远大于建筑物的尺度，如果使用结构化的 HDC 串行计算会使过分小的特征在网络加深的过程中受到影响，而且堆叠结构化的空洞卷积，在计算上也会存在冗余。因此，为了更好地保留不同尺度的特征，本文使用空洞空间金字塔的并行卷积结构对特征进行计算，基于不同的膨胀率并行地在特征上采样多尺度信息。

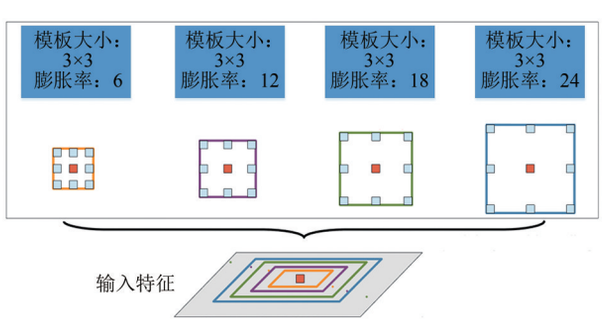


图 2　空洞空间金字塔池化示意图

Fig. 2　The pooling procedure of atrous space pyramid

2.2　网络结构

空洞卷积实际上是在标准卷积的基础上通过模板膨胀对特征进行采样，所以从标准卷积改进到空洞卷积不会改变原始卷积网络的参数。对于语义分割任务，特征提取会降低分辨率（Zuo 等，2020），为了恢复分辨率，需要对特征上采样解码，在 FCN 中，借助跳级结构可以将低层特征用于上采样，因为低层特征具有一定分辨率，包含位置信息。本文以 resnet101 为特征提取主干网络，从较高层网络开始使用空洞卷积，并用空洞空间金字塔捕获不同尺度的特征，在金字塔分支中保留标准卷积操作以关注对象本身的特征，相比 deeplabv3，本文丢弃全局池化以降低过度下采样的影响，并在网络输出端增加全连接条件随机场 CRF 进行后处理。CRF 符合吉布斯分布，使用能量函数为

$$E(x)=\sum_i\theta_i(x_i)+\sum_{ij}\theta_{ij}(x_i,y_j) \tag{3}$$

一元势能函数描述观测序列对标记变量的影响：

$$\theta_i(x_i)=-\log P(x_i) \tag{4}$$

对于像素点 i，$P(x_i)$ 是网络对该像素的分类的概率，二元势能函数描述变量之间的相关性，即像素之间的相关性：

$$\theta_{ij}(x_i,y_j)=u(x_i,y_j)\sum_{m=1}^{K}w_m k^m(f_i,f_j) \tag{5}$$

当 $x_i\neq y_j$ 时，$u(x_i,\ y_j)=1$，否则值为零，可以看出，不同像素之间是全连接的，而 $k^m(f_i,\ f_j)$ 是 f_i 与 f_j 之间的高斯核，f_i 是像素 i 对应的特征向量即颜色信息，w_m 是高斯核的权重。通过最小化能量函数，可以使图像的像素分类更加准确。综合以上描述，可以得到一个详细的网络结构，本文网络结构如图 3 所示。

语义分割是像素级的分类，所以可以用交叉熵计算损失。令 N 为图像中像素的数量，k 为类别的数量，对于某个确定的像素 i，y^i 表示其类别，用 $[z_1^i,\ \cdots,\ z_k^i]$ 表示预测各类别的得分，由于遥感影像数据分类对象分布规律不均衡，为了强迫网络学习到各类对象的分布，需要在交叉熵的每类对象上附加权重 w^i，损失函数计算如下：

$$\text{loss}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}w_j^i y_j^i\log\left(\frac{\exp(z_j^i)}{\sum_{l=1}^{k}\exp(z_l^i)}\right) \tag{6}$$

2.3　余弦退火方法调整学习率

在一般情况下，优化的目标函数是多峰的，存在多个局部最优解，在传统学习策略下，学习率逐步减小会使模型陷入局部最优解，为了跳出局部最优解，可以急剧增大学习率，这被称为热重启随机梯度下降法，重启指的是恢复学习率。

其中较简单的一种热重启方式为余弦退火（Hinton等，2015），余弦退火方法的原理描述为

$$\eta_t = \eta^i_{\min} + \frac{1}{2}\left(\eta^i_{\max} - \eta^i_{\min}\right)\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right) \quad (7)$$

式中，i表示热重启的次数，$\eta^i_{\max}$和$\eta^i_{\min}$限制了第i次热重启的学习率变化范围，可以使$\eta^i_{\max}$和$\eta^i_{\min}$随着热重启次数的上升逐步减小，也可以为了计算简便，保持两者的值不变。T_{cur}表示当前学习经历的次数，T_i表示第i次热重启到第$i+1$次热重启的学习次数，即余弦退火的周期。

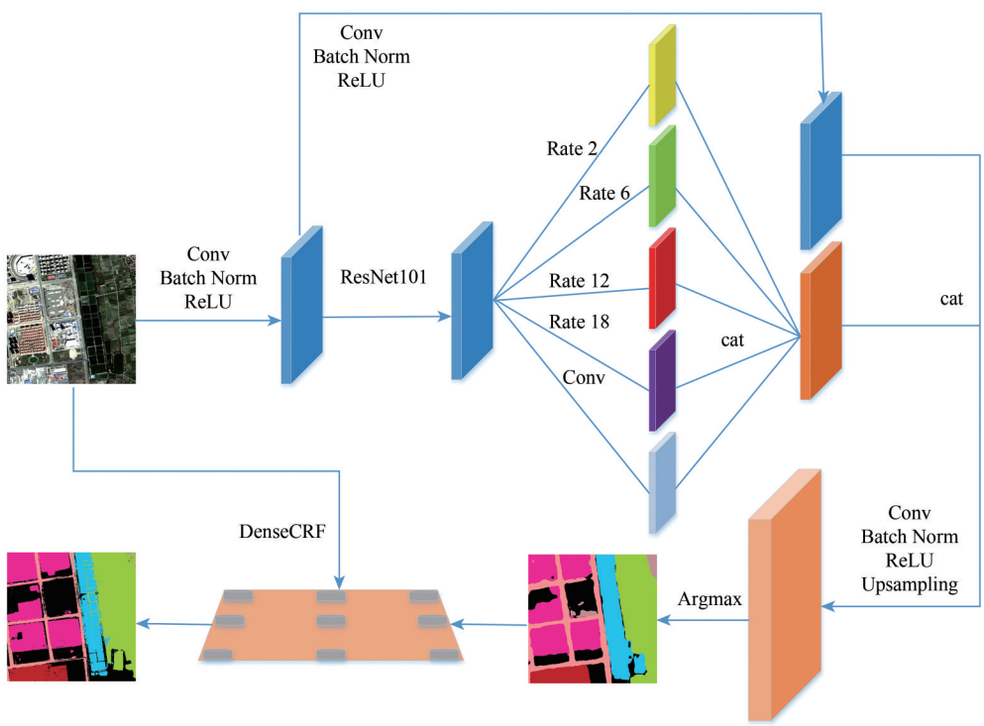


图3 网络结构示意图

Fig. 3 The proposed neural network architecture

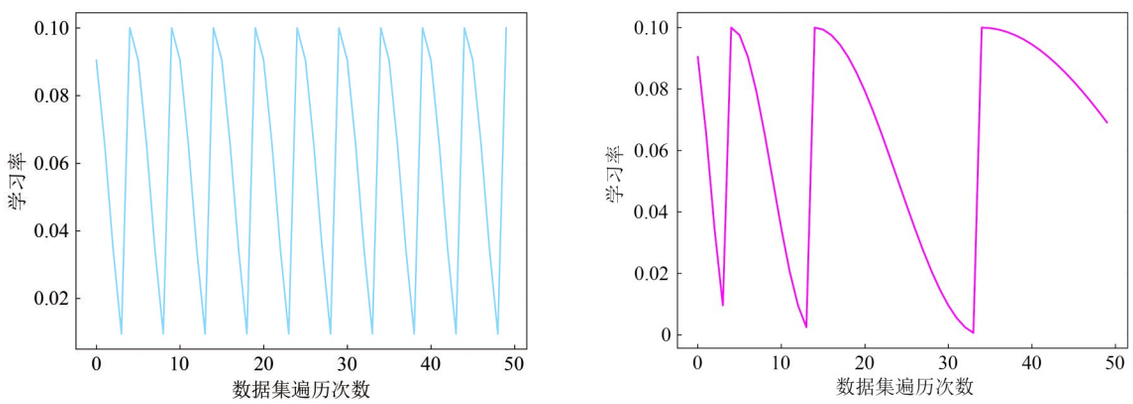
如图4所示，初始学习率从0.1开始，余弦退火方法使学习率逐渐下降又快速上升到初始值。相同周期的余弦退火方法会使网络学习缺少稳定性，因此，本文首次提出周期递增变化的余弦退火方法，采用该方法调整学习率则能够使学习过程相对平缓，图4中周期递增余弦退火的周期是以2为公比的等比数列。等周期的余弦退火使学习率变化频率过快，模型反复跳出局部最优，导致不能找到一个表现较为稳定的局部最优模型，这一现象会影响结果集成的准确程度。很明显周期递增的余弦退火方法相比等周期的余弦退火方法，可以在学习中后期获得训练更平稳的局部最优模型，从而提升结果集成的准确程度。

另外，模型集成必然会增加网络推理的时间，在相同的迭代次数下，周期递增余弦退火策略获得的局部最优模型数量远少于等周期余弦退火策略的模型数量，更少的局部最优模型可以大幅度缩短集成推理的时间（Polino等，2018）。综合看来，周期递增的余弦退火策略可以使模型集成在超越单一模型表现的同时避免过长的推理时间，训练中保存的局部最优模型相比等周期余弦退火的局部最优模型效果会更好更稳定。

训练时，在每次学习率热重启前需要保留局部最优解，语义分割实际上是像素级别的分类任务，所以集成模型可以基于保留下来的局部最优模型，按照得分投票的方式选择最终像素分类结果。

综上所述，本文提出的方法具体分为以下步骤：（1）基于resnet101初始化网络，截取layer1至layer4，layer4的空洞卷积膨胀率为2，layer1至layer3的空洞卷积膨胀率均为1，相当于标准卷积；（2）对resnet101输出的特征做空洞空间金字塔卷积，用不同的膨胀率并行卷积，空间金字塔卷积不进行全局池化，将全局池化分支改用标准卷积代替，从而更深入获取语义信息，提高分类准确率；（3）使用跳

级结构将resnet101中layer1生成的低层特征与线性插值后的空间金字塔卷积结果进行融合，低层特征可以为高层特征带来部分位置信息，对网络输出的粗糙分割结果基于条件随机场进行后处理；（4）使用交叉熵计算损失，由于遥感影像的对象分布不均衡，所以在交叉熵计算时会给每一类对象附加权重，网络的训练采用周期递增的余弦退火方法调整学习率，并保留每个局部最优模型，推理时再集成局部最优模型的结果；（5）高分辨率遥感影像不能一次性完成分割，所以需要先切片再逐一语义分割，拼接各个切片时通过简单的填充孔洞和去除小连通域，修复不合理的预测结果。



（a）周期余弦退火方法的学习率

（a）Learning rate of cosine annealing with equal period

（b）周期递增余弦退火方法的学习率

（b）Learning rate of cosine annealing with increasing period

图4　不同周期下余弦退火方法的学习率

Fig. 4　Learning rates of cosine annealing method with different periods

3　实验设置

3.1　数据集与数据预处理

本文基于GID（Gaofen Image Dataset）（Tong等，2020）评估语义分割方法。GID建立于Gaofen-2卫星遥感影像，GID语义分割对象覆盖范围大，分布广泛且空间分辨率高。大规模分类集涉及5类对象，精细分类集则将分类对象细化至15类。本文在15类精细分类集上选取了包含不同地物信息的10幅高分辨率遥感影像及其对应的标注影像作为训练样本。在GID中，15类对象以外的其他对象所占比例不能忽略，所以要将其视为一类对象考虑，因此，实际分类的数量应该是16类。本文语义分割的类别有：水田，水浇地，旱耕地，园地，乔木林地，灌木林地，天然草地，人工草地，工业用地，城市住宅，村镇住宅，交通运输，河流，湖泊，坑塘以及其他类别。

高分辨率遥感影像的尺寸往往较大，GID精细分类的单幅图像分辨率为（7200，6800）。为了适应计算机视觉模型的实际处理情况，需要对原始高分辨率遥感影像切片处理，在实验中，切片大小的不同没有对模型性能产生显著性影响，考虑到目前主流卷积网络处理的图像分辨率一般是（512，512），因此本文将每幅遥感影像切片至512分辨率，切片步长设置为256以确保切片数据的连续性。为了与常规的深度卷积神经网络相兼容，需要从切片后的遥感影像中提取RGB三通道。遥感影像的地物信息复杂，目标对象形状变化各异，卷积神经网络擅长局部特征的模式匹配，即需要一定的数据增强让网络学习到地物的形变，提高模型的鲁棒性。本文只进行常规的数据增强：随机水平翻转，随机竖直翻转，颜色抖动。在数据增强时，标注图像也要跟随RGB图像做同样的处理。

对于深度神经网络来说，数值较小的张量对反向传播的计算较为友好，且在标准的分布上进行学习会更加容易（Ioffe和Szegedy，2015）。因此，可以根据数据集中不同通道的均值与标准差对输入图像进行标准化。假设数据集一共有m张RGB图像，而这些RGB图像可分成3个通道的张量$[y_1, y_2, y_3]$。

再根据各个通道的均值μ和标准差σ进行标准化得到张量$[z_1, z_2, z_3]$。

$$z_c = \frac{y_c - \mu_c}{\sqrt{\sigma_c^2}}, c \in \{1,2,3\} \tag{8}$$

3.2 语义分割实验设置

本文模型的训练采用周期递增余弦退火方法调整学习率，保留每个局部最优模型，在验证集上通过集成局部最优模型投票决定像素类别。模型训练的优化方法采用Adadelta（Zeiler，2012），初始学习率设置为1×10^{-1}，余弦退火的周期设置以2为公比的等比数列，其余参数采用Adadelta默认值。Adadelta可以在训练初中期取得较快速的效果，当进入训练后期，则会反复在局部最小值附近抖动，此时学习率急剧上升，模型保存局部最优解后，再跳出局部最优解，开始一段新的优化过程。模型的特征提取主干网络是resnet101，在ImageNet（Deng等，2009）上预训练过的resnet101虽然不能直接检测到遥感影像的具体地物信息，但可以有效感知边，角，颜色等低层信息，使网络获得一个良好的初始解；对网络的其他层参数采用服从标准正态分布的随机初始化，空洞卷积的膨胀率分别设置（1，2，6，12，18）。本文模型在遍历整个数据集256次后能够收敛，如果设置批处理大小为8，则训练一共迭代次数为5×10^4。

关于模型的对比实验，本文在并行空洞卷积层调整结构，分别验证使用并行标准卷积，连续堆叠相同膨胀率的空洞卷积，按照HDC结构堆叠空洞卷积和本文网络的语义分割表现。4种网络均使用周期递增的余弦退火方法进行训练。为了有效对比不同卷积结构的影响，4种网络都不使用CRF进行后处理。

关于模型学习策略的对比，以本文网络为基础，设置3种不同的训练模式：使用标准随机梯度下降训练，使用等周期余弦退火方法训练，使用周期递增余弦退火方法训练。比较3种训练方式下，模型的推理时间增长趋势，以及模型在验证数据上的语义分割表现。

对于CRF的影响，本文在所提出的方法上，分别设置是否使用CRF两种情况，在验证数据上对比使用CRF与否得到的语义分割表现。另外，引入近年来常用的语义分割模型：FCN-8s（Long等，2015），segnet（Badrinarayanan等，2017），unet（Ronneberger等，2015），deeplabv3（Chen等，2017）。将主流卷积网络模型与本文方法进行比较。主流模型的训练均采用Adam（Kingma和Ba，2015），训练参数使用Adam方法的默认值。FCN-8s，segnet，unet的网络参数按照文献（Garcia-Garcia等，2017）提出的标准进行设置，deeplabv3按照文献（Kamann和Rother，2020）中使用的参数进行设置。

3.3 模型评价指标

本文使用像素分类的总体精度，具体某一类的分类精度，以及Kappa系数评价实验的语义分割效果。记P_{ab}为将属于a类的像素预测为属于b类的数量，令$t_a=\sum_b P_{ab}$表示属于a类的所有像素数量，$t_b=\sum_a P_{ab}$表示被预测为b类的所有像素数量。则总体精度OA表示为正确分类的像素与图像中所有像素的百分比：

$$\mathrm{OA}=\frac{\sum_a P_{aa}}{\sum_a t_a} \tag{9}$$

对于b类对象的分类精度UA表示所有被分类为b的像素中，被正确分类的像素比例：

$$\mathrm{UA}=\frac{P_{bb}}{t_b} \tag{10}$$

Kappa系数是一个用于衡量预测与真实标签的吻合程度的统计量：

$$\mathrm{Kappa}=\frac{OA-P_c}{1-P_c} \tag{11}$$

式中，

$$P_c=\frac{\sum_k\left(\sum_b P_{kb}\cdot\sum_a P_{ak}\right)}{\sum_a t_a\sum_a t_a} \tag{12}$$

式中，有$k\in[1,K]$，K是分类对象的数量。

为了便于可视化观察各个类别的分类结果，可以使用混淆矩阵清晰反映，混淆矩阵的每一行之和是实际为该类别的样本数量，每一列之和是预测为该类别的样本数量。

4 结果与分析

4.1 不同卷积的实验结果对比

卷积层结构的调整对语义分割结果造成不同意义的影响，在本文方法的并行空洞卷积层调整卷积的结构，不同结构下的语义分割结果对比如图5所示。图5(a)—(f)分别为原图，真实标注，并行标准卷积分割结果，连续堆叠等膨胀率空洞卷积分割结果，按HDC标准堆叠空洞卷积分割结果，并行多尺度空洞卷积分割结果。从图5(c)可以看出，虽然采用并行的结构，但标准的卷积不能较好地学习到图像的像素语义信息，比如错误地将水浇地的像素分类到其他类别，部分住宅被

错分为交通运输。图5(d) 反映了使用连续堆叠相同膨胀率空洞卷积的分割结果，当使用连续堆叠的空洞卷积时，相比标准卷积，分割结果有所改善。由于空洞卷积可以注意到更多上下文信息，因此对比标准卷积，堆叠的空洞卷积可以更广泛地感知到水浇地周围的信息，从而利于水浇地的识别。图5(d) 也可以看出，分割结果是粗糙的，由于连续堆叠的空洞卷积膨胀率相同，在前向计算不断扩张采样区域的同时，导致了网格效应，造成在遥感影像的分割结果中，出现广泛分布的异常区域。使用标准化结构的设计，按照HDC的标准堆叠空洞卷积，改善了网格效应，基于HDC标准的分割结果如图5(e)所示。根据HDC标准，膨胀率呈锯齿状分布的空洞卷积可以在前向计算中弥补信息丢失的风险，从而降低网格效应的影响，结合空洞卷积广泛感知上下文信息的优点，使分割结果得到提升。本文方法采用并行的多尺度空洞卷积，分割结果如图5(f) 所示，相比基于HDC标准的堆叠空洞卷积，其分割结果与真实标注更吻合。并行且多尺度的设计结构一方面可以让模型获得感知多尺度地物信息的能力；另一方面将各个尺度的信息进行融合，在一定程度上弥补了前向计算中的信息丢失，从而降低网格效应的影响。不同膨胀率的空洞卷积让模型在面对同一对象时，可以不同程度地感受到周围信息，加强模型对目标对象的识别能力。并行的结构相比HDC标准下的串行设计具有更高效的计算优势。

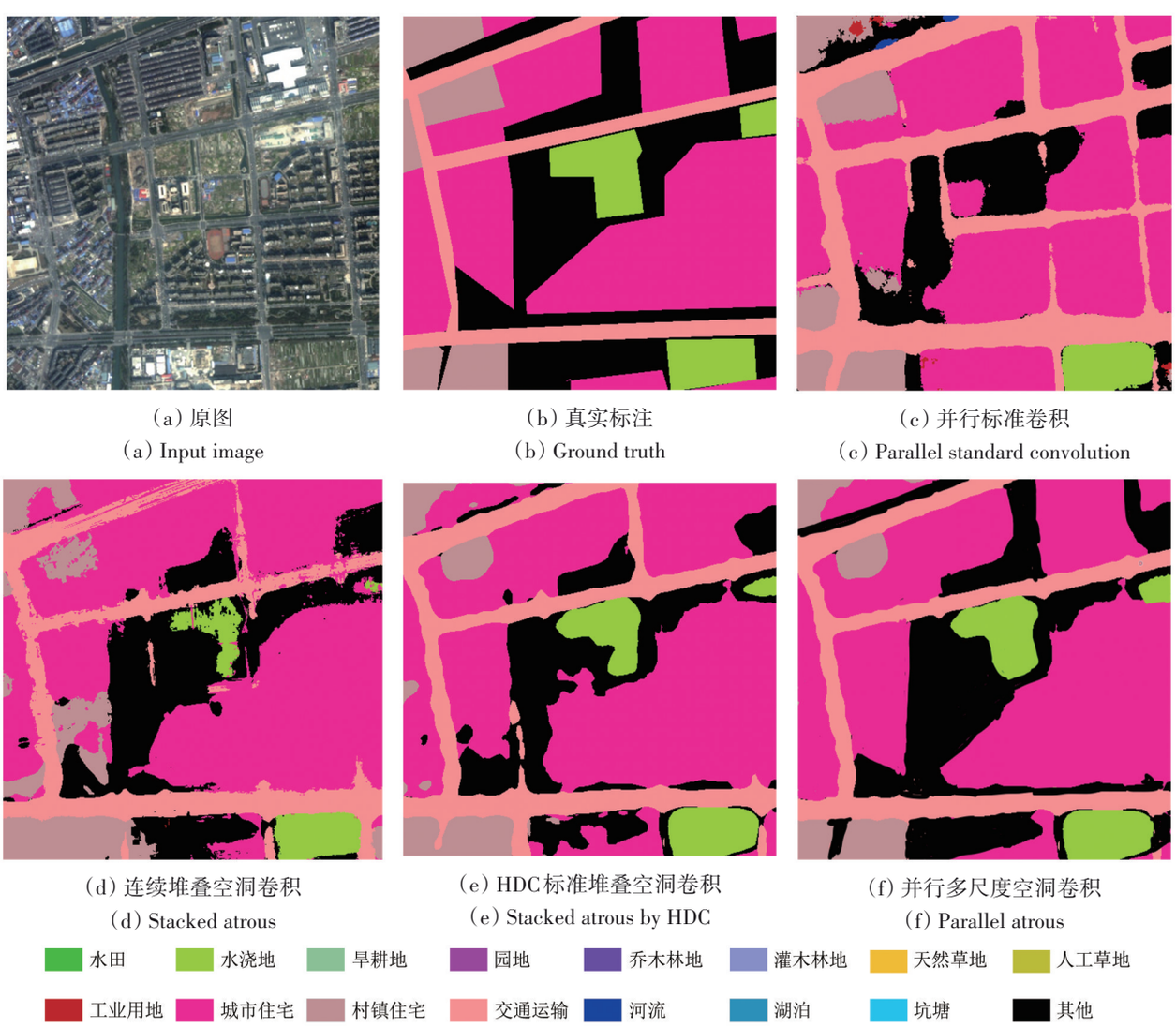


图5 不同卷积的语义分割结果

Fig. 5 Semantic segmentation results of different convolutions

表1为并行标准卷积，连续堆叠相同膨胀率的空洞卷积，按照HDC结构堆叠空洞卷积和本文网络在验证集上的语义分割结果。本文采用的并行多尺度空洞卷积在整体精度与Kappa系数上均优于采用其他卷积结构的模型。

表1 不同卷积的分割结果

Table 1 Segmentation results of different convolutions

/%

卷积结构	整体精度	Kappa系数
并行标准卷积	65.1	60.2
连续堆叠等膨胀率空洞卷积	71.4	64.4
HDC标准堆叠空洞卷积	80.2	75.1
并行多尺度空洞卷积	**84.3**	**79.6**

注：加粗表示最优评价指标。

4.2 学习策略的效率对比分析

模型的集成过程导致推理花费的时间上升，处理的数据量越大，时间花费越显著，使用周期递增余弦退火策略可以避免推理造成过多的时间花费，本文在模型学习时，设置退火周期为一个以2为公比的等比数列，在经过设置的迭代次数后一共得到6个局部最优模型，相比于等周期余弦退火在训练结束后一共得到的17个局部最优模型，推理速度可以获得明显的改善。

表2反映了采用标准随机梯度下降，等周期余弦退火方法和周期递增余弦退火方法训练后，模型在验证数据上的整体精度和Kappa系数。3种策略分别记作策略（1，2，3）。从表2看出，集成模型的效果优于单一模型，且合适数量的局部最优模型也可以接近大量局部最优模型的计算结果。

表2 学习策略对比

Table 2 Comparison of learning strategies

/%

评价指标	学习策略		
	策略1	策略2	策略3
整体精度	81.5	**84.9**	**84.3**
Kappa系数	76.3	**79.1**	**79.6**

注：加粗表示最优评价指标。

图6反映了伴随数据量逐步上升后推理时间变化的趋势，每批数据包含8张切片图像，时间花费以毫秒为单位。从图6可以看出，标准随机梯度下降得到的模型在时间变化程度上最慢，因为在推理时，标准随机梯度下降法训练的模型不需要进行集成。当使用余弦退火训练模型时，模型集成使推理时间快速上升，如果使用周期递增的余弦退火则可以缓解时间花费过高的情况。因此，在使用周期递增余弦退火策略后，一方面通过集成确保结果的准确程度优于标准随机梯度下降法训练的模型，另一方面该策略生成的子模型数量较少，从而确保推理的时间花费不会过高。

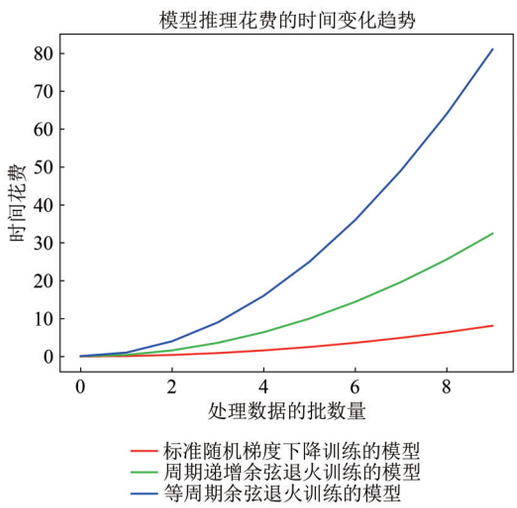


图6 数据量逐步上升的推理时间变化趋势

Fig. 6 The variation trend of inference time by the increasing of data volume

4.3 使用CRF处理与否的对比分析

图7(a)—(d)分别为原图，真实标注，本文方法在不使用CRF情况下的分割结果，以及本文方法使用CRF后处理的分割结果。从图7(c)对比真实标注可以看出，模型能够得到较为精细的分割结果，且保持了一定的分类精度，不论是交通运输这类细致目标对象，还是坑塘，水浇地这类大范围目标对象，由于多尺度的空洞卷积，模型均能够得到合理的分割结果。

本文方法在空洞卷积金字塔层上，取消了deeplabv3中的全局池化，并使用CRF引入空间上下文信息，这可以获得更精细的位置信息。实验过程表明，CRF的迭代次数为5次就可以得到较好结果，图7(d)就是利用模型输出的粗糙分割结果与原图融合并经过条件随机场5次迭代获得的最终语义分割结果。观察原图与真实标注，可以发现，在CRF精细修复后，获得了一个更良好的效果。在验证数据上的分割结果显示，结合CRF后处理，本文模型的整体精度与Kappa系数分别从84.3%和79.6%，提升到86.6%和81.8%。

比较图7(d)与真实标注，使用CRF后，在坑塘等位置存在差异，因为原图的坑塘间本身存在细小的道路，且两者颜色差异较大，这会对CRF计算的分布产生影响，从而造成预测结果与真实标注在坑塘、交通运输与水浇地等位置上的差异。

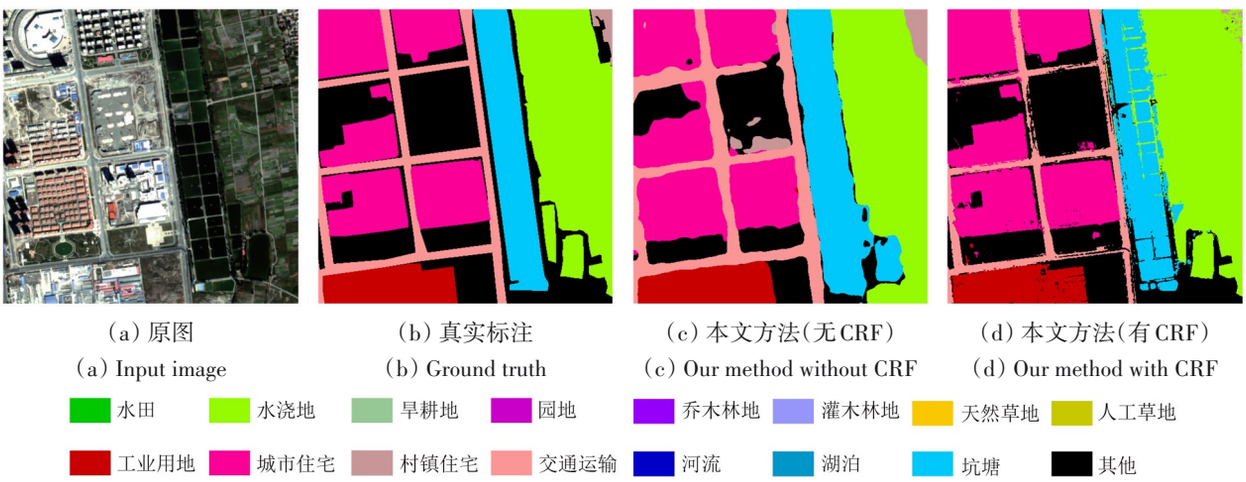


图7 使用CRF与否的语义分割结果对比

Fig. 7 Comparison of semantic segmentation results using CRF or not

4.4 本文方法与主流卷积网络模型的对比分析

本文方法在验证集上优于近年来常用的语义分割方法，分割结果见表3。表3中比较了分别采用FCN-8s，segnet，unet，deeplabv3以及本文方法的总体精度与Kappa系数；另外，对于数据集，根据各类对象的标注像素数量，从16类对象中选择出前5个难分类对象，并在表3中列出其分类精度。前5个难分类对象分别为：人工草地，灌木林地，园地，坑塘，旱耕地。图8以具体场景展示了这5种方法的分割效果，不同类别的对象对应着不同的颜色。其中，图8的(a)—(f)分别代表真实标注图像，FCN-8s分割结果，segnet分割结果，unet分割结果，deeplabv3分割结果，以及本文方法分割结果。

从实验结果中可以看出，FCN-8s与segnet的分割整体精度与Kappa系数均较低，图8(b)与(c)可以直观反映出分割边界比较模糊而且不规整，其次是被错误分类的像素较多，难分类对象的较低分类精度影响了整体精度与Kappa系数。

FCN将VGGnet（Simonyan和Zisserman，2015）修改为全卷积网络，通过对分割任务进行微调，将其学习的特征转移到全卷积网络中，低分辨率语义特征图的上采样使用双线性插值结合滤波器施加卷积操作完成，虽然存在跳级结构，但实际效果表明，FCN在对象的边缘上不能精细分割，网络输出的结果较粗糙。FCN利用标准卷积神经网络作为视觉模型，用标准卷积提取特征，尽管全卷积的架构在语义分割上具有灵活性，但依然有所局限，由于标准卷积固有的平移不变性使网络不能合理考虑上下文信息，这是造成FCN对细节不够敏感的原因之一。

表3 Gaofen Image Dataset数据集语义分割结果

Table 3 Semantic segmentation results using Gaofen Image Dataset

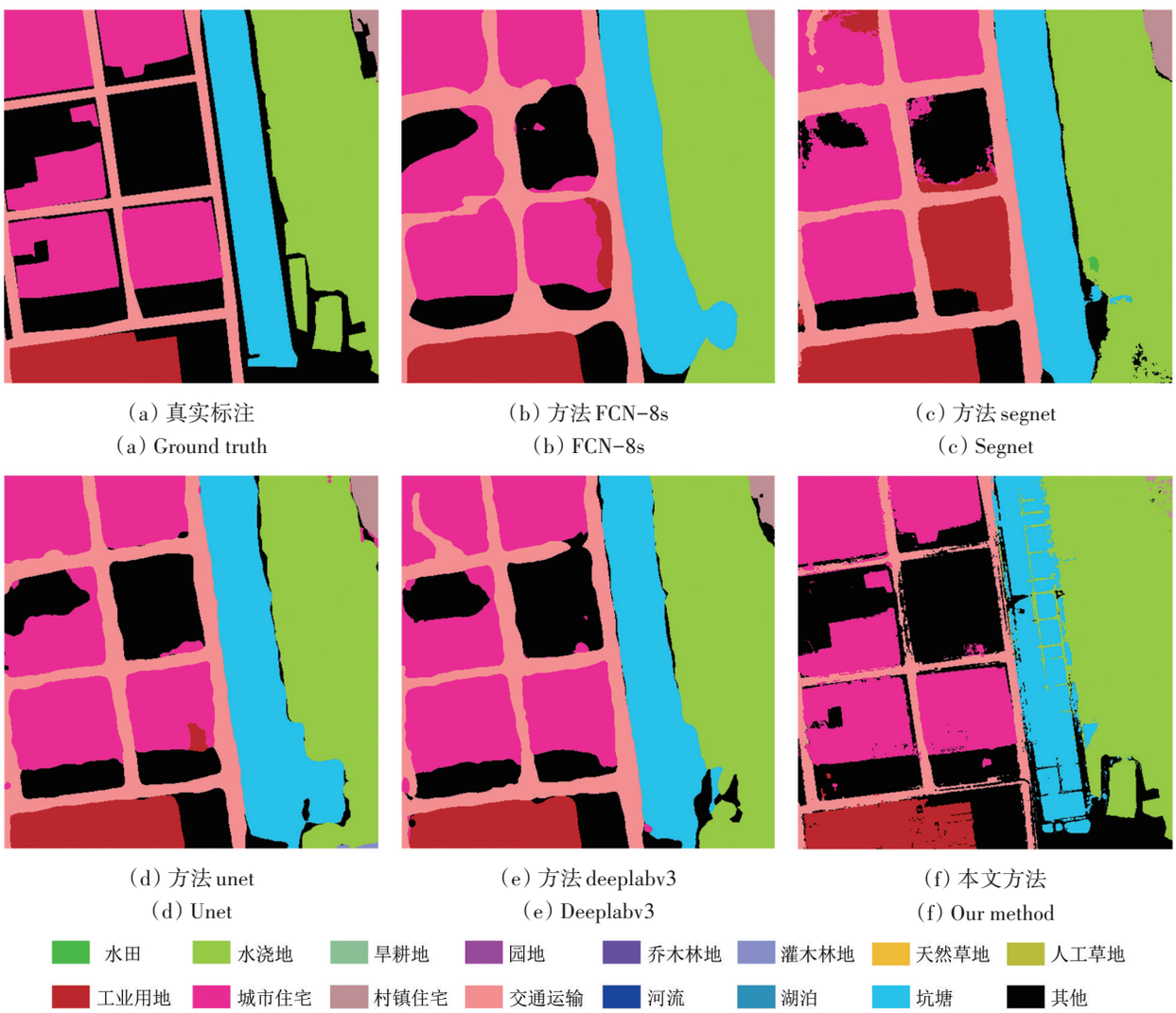
/%

方法	难分类对象的分类精度					整体精度	Kappa系数
	人工草地	灌木林地	园地	坑塘	旱耕地		
FCN-8s	54.1	47.3	81.7	**82.4**	77.4	77.2	70.6
segnet	56.4	65.9	71.5	68.6	71.4	80.1	73.2
unet	77.4	42.5	**82.4**	81.6	92.1	82.4	77.4
deeplabv3	70.1	76.3	74.8	71.2	**93.4**	81.2	76.1
本文方法	**81.1**	**79.4**	77.2	79.8	82.3	**86.6**	**81.8**

注：粗体表示最优评价指标。

对于segnet，在恢复分辨率的解码过程中，使用了在特征提取时的池化位置信息，解码过程中的反池化操作缓解了上采样的学习负担，在分割中保留了高频信息的完整性。经过上采样得到的特征是稀疏的，可以使用卷积再次生成密集的特征，卷积需要学习的是如何修复下采样过程中的

信息损失。利用segnet虽然能很好地保留高频信息，但在低分辨率的特征反池化过程中，必然伴随着邻近信息的丢失，在一定程度上，会影响到对象的分类精度。比如在图8(c)中，有大面积的城市住宅被错分类为工业用地。



(a) 真实标注 (a) Ground truth
(b) 方法FCN-8s (b) FCN-8s
(c) 方法segnet (c) Segnet
(d) 方法unet (d) Unet
(e) 方法deeplabv3 (e) Deeplabv3
(f) 本文方法 (f) Our method

图8 具体场景下的语义分割结果

Fig. 8 Semantic segmentation results under specific scenarios

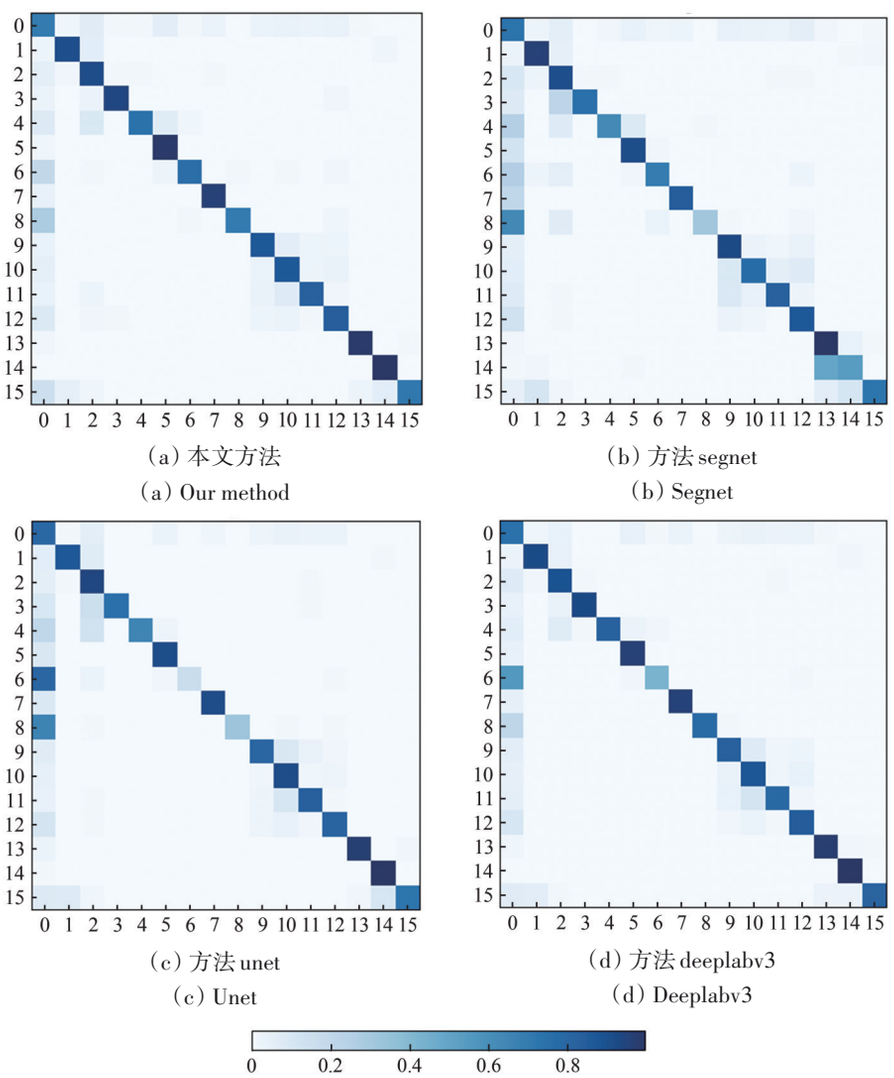
在使用unet架构后，分割的精度与kappa系数得到提升。unet简单地将编码器特征拼接到每个阶段的解码器输出特征上，层与层对应相连接，形成了一个U形结构。网络通过跳接的方式，在每个阶段允许解码器保留编码器下采样过程中丢失的特征。对比segnet，同样是对称的编码与解码，但特征的跳接相比反池化可以让网络具备更完整的上采样能力，在图8(d)中可以看出，unet的分割结果比FCN-8s更加精细，而segnet中被大面积错分类的对象在unet中也得到了改善。

通过deeplabv3实现分割后，整体精度与Kappa系数和unet的结果相接近，虽然表现没有unet良好，但已经超过了FCN-8s与segnet，其中一个很大因素是deeplabv3使用了空洞卷积。空洞卷积帮助网络捕捉到对象的上下文信息，deeplabv3中使用了并行的多尺度空洞卷积，多尺度的空洞卷积让网络的特征变得更加丰富，来自不同尺度的上下文信息帮助网络更准确地分类对象。但deeplabv3没有融合低分辨率的特征，如果仅依靠线性插值实现上采样，网络不容易感知到细节的位置与边缘信息。

本文的模型融合了并行的多尺度空洞卷积，

帮助网络获取大范围的上下文信息，在上采样过程中，利用低层的高分辨率特征进行跳接，弥补了下采样特征提取时丢失的细节位置与边缘信息。网络学习策略基于周期递增余弦退火方法得到多个局部最优解，在推理时集成所有局部最优解的结果，按照投票方式选取最终的像素分类结果，更进一步提高网络在像素分类上的准确率。本文方法在本次实验的数据集上，表现均超过了常用语义分割模型，整体精度与Kappa系数分别为86.6%和81.8%。

混淆矩阵可以更直观地反映各类别的分类结果，本文方法与segnet，unet，deeplabv3在验证集上的混淆矩阵如图9所示，由于遥感影像中不同类别的像素数量分布极不平衡，所以本文对混淆矩阵的结果进行了归一化以便于对比。从图9看出，本文模型在对角线上的分布比segnet，unet，deeplabv3方法的分布更集中，这也反应了模型在验证集上的语义分割结果更加吻合于真实标注。



（a）本文方法
（a）Our method

（b）方法segnet
（b）Segnet

（c）方法unet
（c）Unet

（d）方法deeplabv3
（d）Deeplabv3

图9 本文方法与其他方法的混淆矩阵

Fig. 9 Confusion matrix of the proposed method and other methods

对于前5个难分类对象，本文的语义分割方法与FCN-8s，segnet，unet，deeplabv3相比较，虽然在某些对象上分类精度不够高，但每类对象的分类精度在分布上都更加平均，没有出现过度偏差，即模型的分类结果不会偏向于某些对象，而忽视剩余对象。从表3的分类精度看出，模型的5个难分类对象的分类精度均保持在75.0%以上，这个表现来源于损失函数的改进，本文模型在训练时，根据各类对象的像素数量为交叉熵的每一项赋予权重，迫使网络平衡地捕捉每一类对象的分布。

4.5 完整高分辨率遥感影像的语义分割

在一般计算机的硬件条件下，高分辨率遥感影像计算量过大，不能一次性完成分割，所以必须先切片再逐一语义分割。在拼接各个切片的分割结果时，本文通过简单的填充孔洞和去除小连通域修复不合理的预测结果，对图像先膨胀后腐蚀，连接邻近的物体和断开的轮廓线。对于一幅完整的高分辨率遥感影像，利用本文方法与主流卷积网络模型完成的分割结果如图10所示，各个方法的整体精度与Kappa系数见表4。图10分别展示了遥感影像的RGB通道图像，遥感影像的真实标注，使用本文方法的分割结果，以及使用segnet，unet，deeplabv3分割的结果。

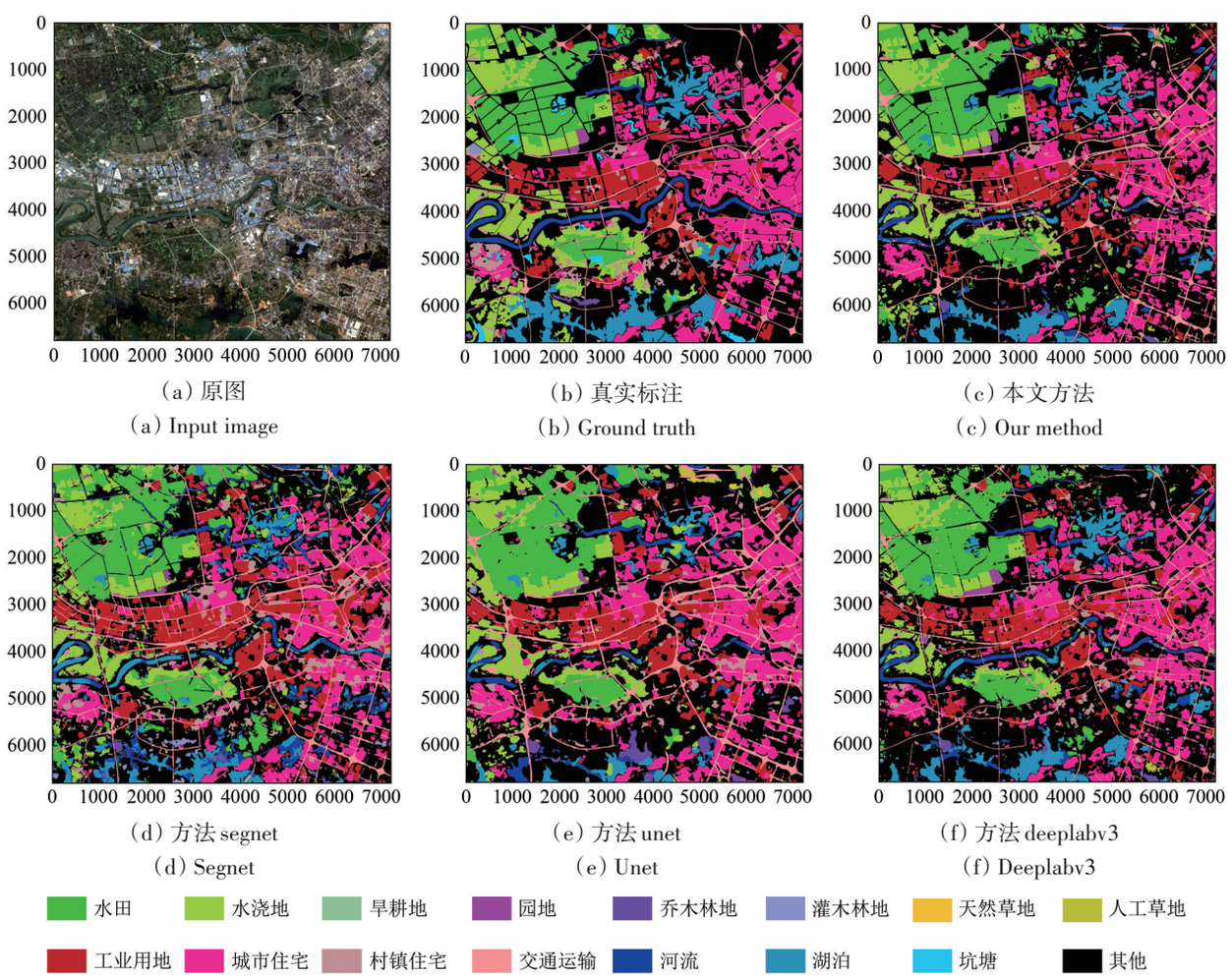
表4 语义分割结果对比

Table 4 Comparison of semantic segmentation results

/%

评价指标	方法			
	本文方法	segnet	unet	deeplabv3
整体精度	**83.6**	79.2	81.1	82.4
Kappa系数	**79.8**	73.5	76.1	77.4

注：粗体表示最优评价指标。



（a）原图 （a）Input image

（b）真实标注 （b）Ground truth

（c）本文方法 （c）Our method

（d）方法segnet （d）Segnet

（e）方法unet （e）Unet

（f）方法deeplabv3 （f）Deeplabv3

图10 高分辨率遥感影像语义分割结果

Fig. 10 Semantic segmentation results using high-resolution remote sensing image

5 结 论

本文研究了多尺度空洞卷积网络架构，并融合周期递增余弦退火方法训练模型，实现了高分辨率遥感影像的语义分割。卷积神经网络已经在图像分割上获得了长足发展，但遥感影像中的复杂对象导致分割能力受到限制。相比之下，本文通过并行的多尺度空洞卷积有效捕捉了复杂地物对象的上下文信息，在不增加参数的情况下扩大感受野，同时保留空间分辨率。全连接条件随机场的引入弥补了细节的位置与边缘信息，将分割结果进一步细化。本文采用周期递增的余弦退火

方法调整学习率，并将局部最优解进行集成，在实验中验证了模型的有效性。与主流语义分割模型FCN-8s、segnet、unet和deeplabv3相比较，本文方法在Gaofen Image Dataset上取得了更好的语义分割效果。然而，本文方法依然存在改进空间，在不简化模型的情况下，集成模型在时间上的花费总是大于单一模型，考虑用知识蒸馏的方法得到近似局部最优模型的简单模型，使集成模型的推理速度接近单一模型的推理速度。

参考文献（References）

Anthimopoulos M, Christodoulidis S, Ebner L, Geiser T, Christe A and Mougiakakou S. 2019. Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE Journal of Biomedical and Health Informatics, 23(2): 714-722 [DOI: 10.1109/JBHI.2018.2818620]

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

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2015. Semantic image segmentation with deep convolutional nets and fully connected CRFs//3rd International Conference on Learning Representations. San Diego: ICLR

Chen L C, Papandreou G, Kokkinos I, 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]

Chen L C, Papandreou G, Schroff F and Adam H. 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706. 05587 [DOI: 10.48550/arXiv.1706.05587]

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Dumoulin V and Visin F. 2016. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603. 07285 [DOI: 10.48550/arXiv.1603.07285]

Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V and Garcia-Rodriguez J. 2017. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv: 1704. 06857 [DOI: 10.48550/arXiv.1704.06857]

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

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503. 02531 [DOI: 10.48550/arXiv.1503.02531]

Huang G, Li Y X, Pleiss G, Liu Z, Hopcroft J E and Weinberger K Q. 2017. Snapshot ensembles: train 1, get m for free//5th International Conference on Learning Representations. Toulon: ICLR

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: JMLR.org: 448-456

Kamann C and Rother C. 2020. Benchmarking the robustness of semantic segmentation models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 8825-8835 [DOI: 10.1109/CVPR42600.2020.00885]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//3rd International Conference on Learning Representations. San Diego: ICLR

Krähenbühl P and Koltun V. 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada: Curran Associates Inc.: 109-117

Li Y, Xiao C J, Zhang H Q, Li X J and Chen J. 2020. Remote sensing image semantic segmentation using deep fusion convolutional networks and conditional random field. Remote Sensing for Natural Resources, 32(3): 15-22 [DOI: 10.6046/gtzyyg.2020.03.03]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Loshchilov I and Hutter F. 2017. SGDR: stochastic gradient descent with warm restarts//5th International Conference on Learning Representations. Toulon: ICLR

Polino A, Pascanu R and Alistarh D. 2018. Model compression via distillation and quantization//6th International Conference on Learning Representations. Vancouver: ICLR

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//3rd International Conference on Learning Representations. San Diego: ICLR

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]

Teichmann M and Cipolla R. 2019. Convolutional CRFs for semantic segmentation//30th British Machine Vision Conference 2019. Cardiff: BMVC: 142

Tong X Y, Xia G S, Lu Q K, Shen H F, Li S Y, You S C and Zhang L P. 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment, 237: 111322 [DOI: 10.1016/j.rse.2019.111322]

Wang P Q, Chen P F, Yuan Y, Liu D, Huang Z H, Hou X D and Cottrell G. 2018. Understanding convolution for semantic segmentation//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE: 1451-1460 [DOI: 10.1109/WACV.2018.00163]

Wang Z W, Wang Z P, You S C, Lei F, Cao L and Yang K J. 2020. Landsat image glacier extraction based on context semantic segmentation network. Acta Geodaetica et Cartographica Sinica, 49(12): 1575-1582 [DOI: 10.11947/j.AGCS.2020.20190313]

Yu F and Koltun V. 2016. Multi-scale context aggregation by dilated convolutions//4th International Conference on Learning Representations. San Juan: ICLR

Zeiler M D. 2012. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 [DOI: 10.48550/arXiv.1212.5701]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhao Q H, Xie K L, Wang G H and Li Y. 2020. Land cover classification of polarimetric SAR with fully convolution network and conditional random field. Acta Geodaetica et Cartographica Sinica, 49(1): 65-78 [DOI: 10.11947/j.AGCS.2020.20190038]

Zhou P C, Cheng G, Yao X W and Han J W. 2021. Machine learning paradigms in high-resolution remote sensing image interpretation. National Remote Sensing Bulletin, 25(1): 182-197 [DOI: 10.11834/jrs.20210164]

Zuo Z C, Zhang W and Zhang D Y. 2020. A remote sensing image semantic segmentation method by combining deformable convolution with conditional random fields. Journal of Geodesy and Geoinformation Science, 3(3): 39-49 [DOI: 10.11947/j.JGGS.2020.0304]

Remote sensing image semantic segmentation method combining cosine annealing with atrous convolution

TANG Zhenchao[1], WEI Wei[2], LUO Weiran[3], HU Jie[2], ZHANG Dongying[1]

1. School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China;
2. Yellow River Survey, Planning, Design and Research Institute Co., Ltd, Zhengzhou 450003, China;
3. School of Water Conservancy and Environment, Zhengzhou University, Zhengzhou 450001, China

Abstract: This study aims to capture the rich context information and multiscale feature information in remote sensing images, improve the integrated model strategy, and enhance the accuracy of semantic segmentation. Thus, this study proposes a high-resolution remote sensing image semantic segmentation method using cosine annealing with increasing period and multiscale atrous convolution.

The multiscale parallel atrous convolution helps the network capture context information in a larger range and improves the ability of the network to recognize multiscale objects without increasing parameters. The method in this study uses the atrous convolution while discarding the pooling operation to maintain the spatial resolution. Meanwhile, the method adopts the fully connected conditional random field to add spatial and edge context information for making up for part of the position information missed by the atrous convolution. As a result, the outline of extraction objects by semantic segmentation fits the ground truth better. Moreover, the cosine annealing strategy with increasing period is introduced to adjust the learning rate and obtain a suitable number of local optimal solutions. We integrate the local optimal solutions in the method to further improve the pixel classification ability of the network.

The overall accuracy and kappa coefficient of the proposed model, which are 86.6% and 81.8%, respectively, are better than those of the current advanced semantic segmentation models.

The experimental results performed on the Gaofen image dataset show that the fusion of image context information and multiscale feature information can effectively identify objects with complex structures. Moreover, the model coupled with the period-increasing cosine annealing strategy could obtain better semantic segmentation accuracy than and less inference time than that coupled with the equal-period cosine annealing strategy.

Key words: high-resolution remote sensing image, semantic segmentation, cosine annealing with increasing period, multi-scale parallel atrous convolution, target extraction, in-context learning, conditional random field, multi-scale learning