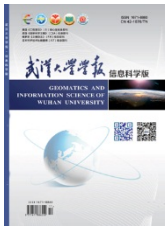


武汉大学学报(信息科学版)
*Geomatics and Information Science of Wuhan University*
ISSN 1671-8860,CN 42-1676/TN

# 《武汉大学学报(信息科学版)》网络首发论文

题目：结合 Atrous 卷积的 FuseNet 变体网络高分遥感影像语义分割
作者：杨军，于茜子
DOI：10.13203/j.whugis20200305
收稿日期：2020-06-22
网络首发日期：2020-10-10
引用格式：杨军,于茜子.结合 Atrous 卷积的 FuseNet 变体网络高分遥感影像语义分割.武汉大学学报(信息科学版). https://doi.org/10.13203/j.whugis20200305

网络首发时间：2020-10-10 15:32:23
网络首发地址：https://kns.cnki.net/kcms/detail/42.1676.TN.20201010.1355.001.html

DOI: 10.13203/j.whugis20200305

# 结合 Atrous 卷积的 FuseNet 变体网络高分遥感影像语义分割

杨军[1]　于茜子[2,3,4]

1 兰州交通电子与信息工程学院，甘肃 兰州 730070

2 兰州交通大学测绘与地理信息学院，甘肃 兰州 730070

3 地理国情监测技术应用国家地方联合工程研究中心，甘肃 兰州 730070

4 甘肃省地理国情监测工程实验室，甘肃 兰州 730070

**摘 要**：针对多模态、多尺度的高分辨率遥感影像的分割问题，提出结合 Atrous 卷积的 FuseNet 变体网络架构对常见的土地覆盖对象类别进行语义分割。首先采用 FuseNet 变体网络将 DSM 图像中包含的高程信息与 RGB 图像的颜色信息融合，其次在编码器和解码器中分别使用空洞卷积来增大卷积核感受野，最后通过对遥感影像逐像素分类，输出遥感影像语义分割结果。实验结果表明，本算法在公开的 ISPRS Potsdam 数据集和 ISPRS Vaihingen 数据集上的 mF1 得分分别达到了 91.6%和 90.4%，优于已有的主流算法，验证了本文算法的有效性。

**关键词**：高分辨率遥感影像；深度卷积神经网络；空洞卷积；语义分割；FuseNet

**中图法分类号**：P237　　**文献标识码**：A

遥感影像语义分割是遥感影像信息获取的关键环节和研究热点，近年来，相关研究成果已经广泛应用于检测土地利用变化、城市扩张、交通监测和灾害预警评估等方面，对政府部门的政策制定以及地理信息数据库的更新具有重要的参考意义[1-2]。高分辨率遥感影像能够表现丰富的地物信息，从而有利于提取地物的复杂特征以识别复杂的人造目标。

传统的遥感图像语义分割主要是通过提取图像的低级特征进行分割，分割结果缺乏语义标注。Shotton 等[3]通过随机森林分类器提取语义特征进行语义分割。Yang 等[4]利用 Logistic 回归分类器提取颜色、纹理特征，通过 CRF(Conditional Random Field)模型训练实现语义分割。然而，传统遥感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，人工设计的特征和高层语义特征间存在差距，因此建立的语义分割模型泛化能力较差。

随着深度学习理论的发展与普及，深度神经网络模型已经被不同行业广泛使用[5]。研究者在遥感影像分析处理中应用深度学习的方法，取得了较为理想的效果[6-7]。全卷积网络[8] (Fully Convolutional Networks, FCN)和 SegNet[9]网络在高分辨率遥感影像语义分割中展现出了较为优异的性能及分割效果，然而 FCN 对像素进行分类时没有考虑到像素之间的关系，忽略了基于像素分类的空间规划步骤，缺乏空间一致性。SegNet 的基本网络结构为编码器-解码器结构。编码器对图像进行高维特征提取和下采样，解码器对提取的特征图进行上采样操作，因此编-解码器结构可以以 1:1 的分辨率进行像素预测，但上采样的过程中易丢失细节信息，使得小目标地物的分割效果较差。文献[10]分别提取 RGB 信息和 DSM 信息，并将它们融合集成到

**收稿日期**：2020-06-22
**项目资助**：国家自然科学基金(61862039)；兰州交通大学优秀平台支持项目(201806)。
**第一作者**：杨军，博士，教授，博士生导师，主要从事计算机图形学、数字图像处理和地理信息系统等方面的研究。yangj@mail.lzjtu.cn

SegNet 结构中再进行语义分割，获得高分辨率的多模态预测 RGB-D 数据用于异构数据源的联合学习。然而该融合策略无法平衡高程信息和颜色信息，导致图像分割不准确。因此，本文针对高分遥感影像中多模态数据融合效果不佳、边缘分割效果不理想、类边界模糊和易产生误分割现象的问题，受编-解码器和文献[11]中 FuseNet 网络结构的启发，对 FuseNet 网络结构进行改进，提出一种结合 Atrous 卷积的 FuseNet 变体网络(Improved FuseNet with Atrous Convolution-CNN, IFA-CNN)模型。主要创新和贡献有：1）在编码器部分，提出虚拟融合单元提高了遥感影像语义分割效果。2）针对遥感影像提取特征部分，引入 Atrous 卷积调整感受野捕获遥感影像多尺度信息，提高目标分割效果。3）在解码器部分，对称链接编码器中的融合特征以提高网络鲁棒性。

# 1 高分遥感影像语义分割模型

## 1.1 多模态遥感数据融合

文献[11]中 FuseNet 采用了编-解码器结构，将二维图像数据融合。网络体系架构如图 1 所示，pooling 为池化操作，conv 为卷积操作，unpooling 为上采样操作。图中使用了两个编码器对 RGB 和 DSM 进行联合编码，将编码后的特征图输入到解码器中进行上采样，然后先由 classifier 进行分类，再通过 softmax 得到最终分割结果。同时，FuseNet 选择深度信息作为辅助特征进行多模态数据融合，如图 2(a)所示。但 FuseNet 在进行多模态数据融合时，DSM 分支与 RGB 分支存在不对称性，使得 DSM 分支仅提取深度特征， RGB 分支需要提取 DSM 与 RGB 数据的融合。此外，这种不对称的融合方案，导致在解码过程中，只使用主分支编码时的索引进行上采样，在一定程度上会影响遥感影像的分割效果。

为了更好地提取 RGB-D 图像的特征，解决主数据源及辅助数据源数据分配不均的问题。本文提出一种“虚拟”分支融合单元(Virtual Branch Fusion Unit, V-Fusion Unit)，即，对主数据源和辅助数据源进行一次卷积运算，从而产生一种“虚拟”模态。将该“虚拟”模态作为融合数据源之一，与 DSM 分支提取的特征和 RGB 分支提取的特征再进行融合，如图 2(b)所示，通过这种方法调整 FuseNet 结构，使其减轻对主数据源和辅助数据源的选择，以解决数据处理不均衡的问题。另外，为解决解码过程中，只使用主分支编码时产生的索引进行上采样，本文将“虚拟”分支融合单元中最大池化操作产生的索引应用于解码阶段的上采样，从而提高语义分割效果。

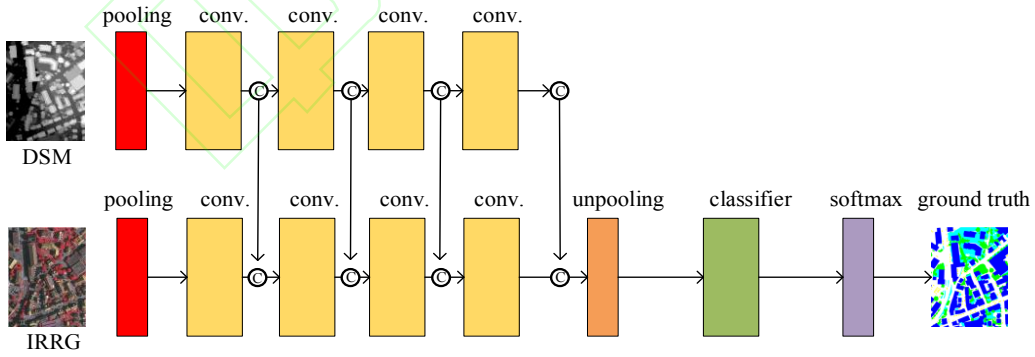
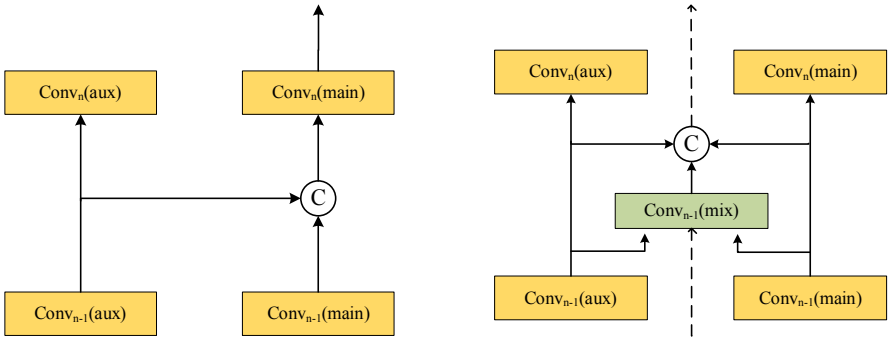



图 1 用于遥感数据融合的 FuseNet 架构[9]

Fig.1 FuseNet architecture for fusion of remote sensing data [9]




(a) Fusion 单元　　(b) V-Fusion 单元

图 2 多模态数据融合策略

Fig.2 Multimodal data fusion strategy

### 1.2 多尺度Atrous卷积

Atrous[12]卷积是在不减小图像大小的同时获得比较大的感受野，所以其主要优势在于允许灵活地调整感受野的大小来捕获多尺度信息，提高多目标分类和分割任务的性能[13]。二维 Atrous 算子定义为：

$$g_{i,j}(x_\ell)=\sum_{c=0}^{C_\ell}\theta_{k,r}^{i,j}*x_\ell^c \qquad (1)$$

式中，$g_{i,j}$ 是对输入特征图的卷积操作 $\mathbb{R}^{H_\ell\times W_\ell\times C_\ell}\to\mathbb{R}^{H_{\ell+1}\times W_{\ell+1}}$，$*$表示卷积算子，$x_\ell\in\mathbb{R}^{H_\ell\times W_\ell\times C_\ell}$ 为在第$i$行和第$j$列中属于通道 $c\in\{0,1,...,C_\ell\}$ 的特征图，$\theta_{k,r}$ 为卷积核大小为 $k$ 和扩张率为$r\in\mathbb{Z}^+$的 Atrous 卷积。在 Atrous 卷积中，卷积核大小 $k$ 增加为 $k$+($k$-1)($r$-1)，当 $r$ = 1 时，Atrous 卷积相当于标准卷积。标准卷积的卷积层感受野与之前所有层卷积核的大小和步长有关，感受野呈线性增长，而 Atrous 卷积感受野为[$2^{r+1}$-1]×[$2^{r+1}$-1]，因此 Atrous 卷积的级联可以实现感受野呈指数增长，使得每个卷积输出都包含较多的信息。

## 2 结合Atrous卷积的FuseNet变体网络

本文使用编-解码器作为基本网络结构，如图3所示。编-解码器是一种输出近似于输入的网络结构。因此，在影像分割阶段，原始图像分辨率与分割图像分辨率保持一致。解码器能够使用反池操作对特征图进行上采样，因此可使输出图像分辨率逼近输入图像分辨率。编码器部分采用VGG- 16架构，其包含5个卷积模块，每个卷积模块分别包含2个或者3个卷积核为3×3的卷积层，然后利用池化核为2×2的最大池化层对每个卷积模块提取的特征进行特征降维。每个卷积层中均使用修正线性单元(Rectified Linear Unit, ReLU)作为激活函数，并利用批归一化(Batch Normalization, BN)使数据服从正态分布。

解码器则是执行上采样和分类的过程。上采样是将编码后的特征图恢复到原始空间分辨率，在解码过程中池化层被反池化层替换，反池化是根据最大池化过程中的索引从较小的特征图映射到一个零填充的上采样特征图。如图4所示，给定一个特征图，定义其大小为4×4，步长为2，通过最大池化操作得到特征图以及特征图中各值在原特征图中的索引。反池化操作是根据索引和特征图进行补 0，这种反池操作将抽象特征转换为几何特征。在反池操作后，卷积块增加稀疏特征图的密度。重复此过程，直到特征图与输入分辨率一致。相比于其他网络结构，降采样操作会丢失细节信息，虽然底层特征具有丰富细节，但判别能力较弱，使得网络对小目标地物分割性能较差。编-解码器结构中通过将上采样操作与跳跃连接相结合，利用反池化操作把浅层信息和高层信息融合，一定程度上缓解了细节丢失问题，使得该基本结构对于分割小目标地物效果也较好。

在编-解码器的特征图处理过程中，如果空间分辨率一致，则可以直接通过跳跃连接进行特征融合；如果空间分辨率不一致，

则将输入特征图通过1×1的卷积核投影成与输出特征图相同的维度。为了保持空间分辨率不变，本文提出的网络保留了初始2×2的最大池化，但需将所有卷积的步长减小为1。为了将特征图恢复到原始分辨率，反池化操作后进行标准卷积操作。最后计算损失函数*loss*，并在像素块上取均值。

$$loss=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}y_j^i\log\left(\frac{\exp(z_j^i)}{\sum_{l=1}^{k}\exp(z_l^i)}\right) \quad (2)$$

其中，$N$ 为输入图像的像素个数，$k$ 为类的个数，对于特定像素 $i$，$y_j^i$ 表示像素 $i$ 属于第 $j$ 类标签，$z_j^i$ 表示为预测矢量。本文在不进行任何空间正则化的情况下，将平均逐像素分类损失降到最低。此外，算法不使用任何影像后处理过程，提高了计算速度。

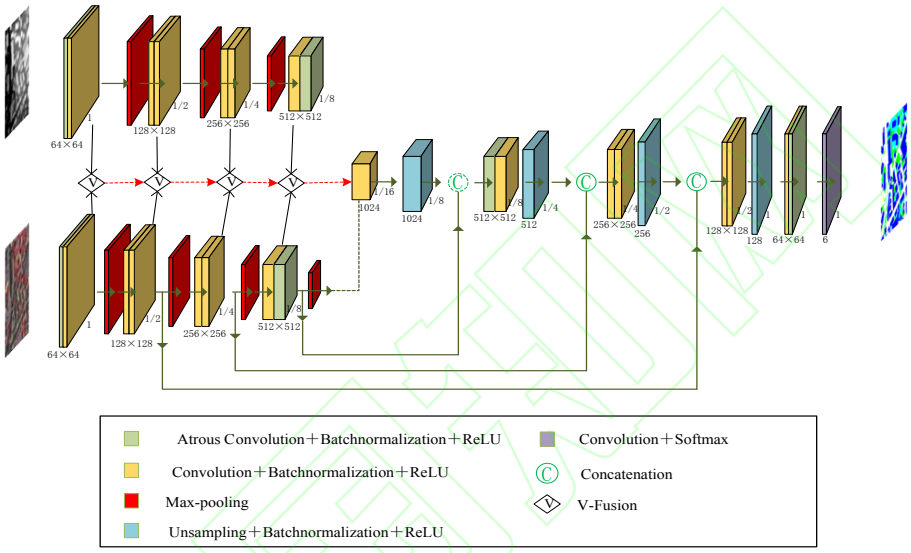



图 3 结合 Atrous 卷积的 FuseNet 变体网络结构

Fig.3 The structure of improved network based on FuseNet and Atrous convolution

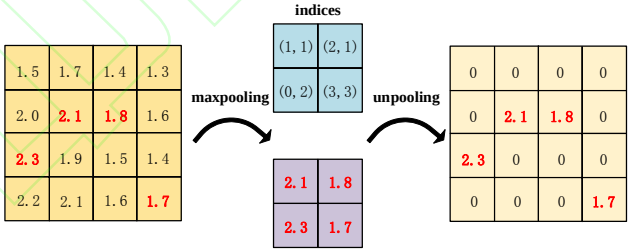



图 4 最大池化和反池化操作对 4×4 特征图的影响

Fig.4 The influence of max pooling and unpooling operation on 4×4 characteristic pattern

## 3 实验结果与分析

### 3.1 数据集

本文在ISPRS[14]航空影像Vaihingen数据集和Potsdam数据集上验证所提出算法的可行性。分别对建筑物、不透水域表面(如道路)、低矮植被、树木、汽车和杂乱背景等六个类别的地物进行语义分割。但在实验中不包含杂乱背景，因为杂乱背景的像素面积仅占总影像像素的0.88%。

Vaihingen数据集是由33张航拍影像组成的，采集于Vaihingen市1.38平方公里的区域内。每幅影像的平均大小为2494×2064像素，空间分辨率为9厘米，含三个波段：近

红外(IR)、红(R)、绿(G)波段。影像中提供物体表面高度的数字表面模型(Digital Surface Models, DSM)作为补充数据。选择29幅影像进行训练，4幅影像用来测试。Potsdam数据集由38幅高分辨率航拍影像组成，其中24幅影像包含真实标签，覆盖面积3.42平方公里，每幅航拍影像由4个波段组成，分别为近红外(IR)、红(R)、绿(G)、蓝(B)，本文使用近红外(IR)、红(R)、绿(G)波段。影像的大小为6000×6000像素，以6个类别的像素级标签作为标注，空间分辨率为5厘米，同样具有DSM补充数据。试验中选择20幅影像进行训练，4幅影像进行测试。

### 3.2 多尺度数据增强及标准化

数据增强的目的是生成新的样本实例。当训练样本较少时，数据增强对于提高网络的泛化能力起到关键性的作用。在Potsdam数据集中，通过对高分遥感影像随机裁剪，得到5000个大小为256×256的图像块，并通过旋转、缩放等操作，扩充数据集的规模用于IFA-CNN网络的训练，从而增强网络的泛化能力。本文使用的高分遥感影像的所有波段(IRRG)都被标准化在[0, 1]区间内。

神经网络的参数和激活函数通常初始化为[0, 1]之间的随机数，需要采用标准化方法避免梯度爆炸、梯度弥散情况的出现。z-score标准化方法[15]将输入图像的像素值逼近于正态分布，有利于提高网络收敛速度。标准化公式为：

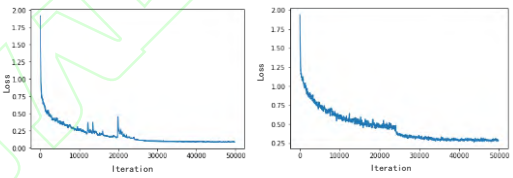
$$X_{out}=\frac{\left(X/Max(X)\right)-\lambda}{\sigma} \tag{3}$$

式中，$X$为输入值，$Max(X)$为输入最大值，$\lambda$、$\sigma$分别为$X$ / $Max(X)$的均值和标准差。

### 3.3 网络训练

由于本文使用的数据集是高分遥感影像数据集，无法在深层网络中直接处理，因此使用滑动窗口的方法来提取 256×256 的小块。滑动窗口的步长也定义了两个连续小块之间重叠区域的大小。在训练时，较小的步长可以提取更多的训练样本，起到数据扩充的作用，Potsdam 数据集和 Vaihingen 数据集的步长分别设定为 64px(pixel, px)和32px。在测试时，较小的步长允许对重叠区域进行平均预测，以提高整体精度，文中分别使用 32px 步长和 16px 步长滑动窗口对 Potsdam 数据集和 Vaihingen 数据集中的测试图像提取 256×256 的小块。

本文设置初始学习率为0.01，每隔5个 *epoch* 将学习率除以 10 直至 0.00001；动量参数为 0.9，权重衰减为 0.0005，批归一化大小为 10。对于编码器-解码器结构，采用迁移学习的方法利用 ImageNet 上训练好的 VGG-16 的权值作为本文初始化编码器的权值，并随机初始化解码器的权值，有效缩短了模型的训练时间。将初始化后权值的学习率设定为新权值学习率的一半，并在每个数据集上对结果进行交叉验证。本文的深度学习网络的损失值曲线如图5所示，图(a)为 Vaihingen 数据集在网络训练过程中的损失值曲线，在25k次迭代后基本处于收敛状态，但当损失值第一次收敛趋近 0.25 时，损失曲线突然上升，其原因为在后期训练中学习率相对过大。图(b)为 Potsdam 数据集在网络训练过程中的损失值曲线。




(a)Vaihingen损失值曲线 (b)Potsdam损失值曲线

图5 损失值曲线

Fig.5 Loss curve

### 3.4 结果与分析

#### 3.4.1 评价标准

为了评估深度学习网络的性能，文中使用以下公式计算 F1 得分：

$$P=TP/\left(TP+FP\right) \tag{4}$$

$$R=TP/\left(TP+FN\right) \tag{5}$$

$$F1=2PR/\left(P+R\right) \tag{6}$$

其中，*TP* 为真正例(True Positive, TP)，表示预测值为 1，真实值为 1；*FP* 为假正例(False Positive, FP)，表示预测值为 1，真实值为 0；*FN* 为假反例(False Negative, FN)，表示预测值为 0，真实值为 1；*P* 为预测正确的正例数占预测为正例总量的比率，即查准率；*R*

为预测正确的正例数占真正的正例数的比率，即查全率。实验中，通过计算 F1 的平均值 mF1 评估网络的分割准确率，mF1 的值越大，表示网络性能越好且分割准确率越高。

此外，文中还利用总体精度(Overall Accuracy，OA)评估算法的分割准确率。总体精度 *OA* 的计算公式如下：

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

其中，*TN* 为真反例(True Negative, TN)，表示预测值为 0，真实值为 0。

3.4.2 实验结果与分析

本文算法(IFA-CNN)得到的部分实验数据结果与Ground Truth之间对比如图6所示。可以看出，IFA-CNN在整体上得到了比较理想的分割结果，尤其是对较大目标地物的分类效果很好，但在图像中也存在一些分割错误的区域。对比Vaihingen分割图像与Ground Truth可以看出，分割错误的区域较少，分割效果较好；但在Potsdam的分割图像中出现了小块区域分割效果不佳的情况，其主要原因为Potsdam数据集地物分布较复杂，且模糊区域较多，而Vaihingen数据集地物分布较均匀，分割难度较低。

表1为本文算法在ISPRS Vaihingen和Potsdam测试集上的分割准确率计算结果。实验结果为五类地物的分割准确率mF1和OA，可以看出，本算法取得了不错的分割结果。

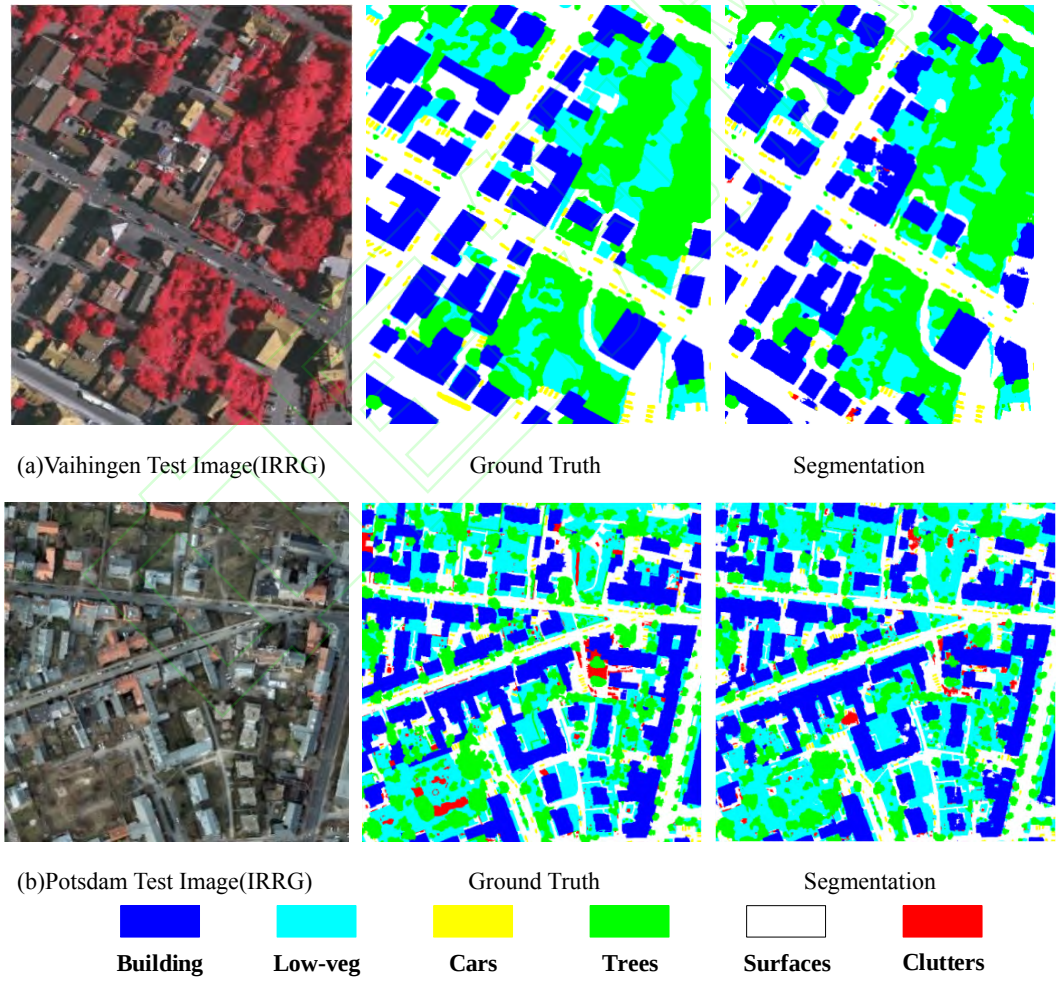



图 6 ISPRS Vaihingen 和 Potsdam 数据集的分割结果

Fig.6 Segmentation results of ISPRS Vaihingen and Potsdam dataset

表1 ISPRS Vaihingen和Potsdam数据集的分割准确率

Tab.1 Segmentation accuracy results on ISPRS Vaihingen and Potsdam dataset

| Dataset | Building (F1-score) | Tree (F1-score) | Low-veg (F1-score) | Surface (F1-score) | Car (F1-score) | OA | mF1 |
|---|---|---|---|---|---|---|---|
| Vaihingen | 0.955 | 0.921 | 0.836 | 0.937 | 0.871 | 0.915 | 0.904 |
| Potsdam | 0.956 | 0.864 | 0.906 | 0.917 | 0.939 | 0.909 | 0.916 |

本文网络与文献[9]中的FuseNet网络在Vaihingen 数据集和 Potsdam 数据集上分别进行实验对比，实验结果如表 2 和表 3 所示。在实验中，除融合单元部分不同外，其他网络结构部分一致。可以看出，本文多模态数据融合策略中使用的虚拟融合单元(V-Fusion)对各类别地物的分割准确率均高于FuseNet网络的Fusion单元，进一步证明了 V-Fusion 单元解决了数据分配不均的问题，它将DSM分支提取的特征与RGB分支提取的特征在此单元进行融合，更好地提取 RGB-D 图像的特征，因此添加V-Fusion单元的FuseNet网络分割准确率更高。

表 2 V-Fusion 单元与 Fusion 单元在 Vaihingen 数据集上的分割准确率比较

Tab.2 Comparison of segmentation accuracy between V-Fusion unit and Fusion unit on Vaihingen dataset

| FuseNet 网络 | Buildings (F1-score) | Tree (F1-score) | Low-veg (F1-score) | Surface (F1-score) | Car (F1-score) | OA (F1-score) | mF1 (F1-score) |
|---|---|---|---|---|---|---|---|
| Fusion 单元 | 0.939 | 0.846 | 0.833 | 0.911 | 0.853 | 0.898 | 0.876 |
| V-Fusion 单元 | 0.955 | 0.921 | 0.836 | 0.937 | 0.871 | 0.915 | 0.904 |

表 3 V-Fusion 单元与 Fusion 单元在 Potsdam 数据集上的分割准确率比较

Tab.3 Comparison of segmentation accuracy between V-Fusion unit and Fusion unit on Potsdam dataset

| FuseNet 网络 | Buildings (F1-score) | Tree (F1-score) | Low-veg (F1-score) | Surface (F1-score) | Car (F1-score) | OA (F1-score) | mF1 (F1-score) |
|---|---|---|---|---|---|---|---|
| Fusion 单元 | 0.942 | 0.863 | 0.828 | 0.909 | 0.930 | 0.882 | 0.894 |
| V-Fusion 单元 | 0.956 | 0.864 | 0.906 | 0.917 | 0.939 | 0.909 | 0.916 |

为探索使用编码器-解码器(encoder-decoder)结构对小目标地物分割准确率的影响，本文网络与文献[16]、[17]、[18]的网络在Vaihingen数据集上进行实验对比，与文献[8]、[19]、[20]的网络在Potsdam数据集上进行实验对比，实验结果如表4和表5所示。可以看出，本文IFA-CNN采用的编-解码器结构对于汽车及低矮植被这两类小目标地物的分割准确率均高于非编-解码器结构网络。由于小目标地物的细节信息较少，相比于其他网络结构，编-解码器结构在编码过程中能够较好地提取高分遥感影像的语义特征，并在解码过程中通过反卷积将特征有效恢复为语义分割预测图，还原小目标地物的语义特征，减少细节信息的丢失。

表4 在Vaihingen 数据集上IFA-CNN与非编-解码器结构网络对小目标地物分割准确率比较

Tab.4 Comparison of the segmentation accuracy for small objects between IFA-CNN and non-encoder-decoder network on Vaihingen dataset

| Classes | UOA[16] | ADL_3[17] | DST_2[18] | IFA-CNN |
|---|---|---|---|---|
| Car(F1-score) | 0.820 | 0.633 | 0.726 | 0.871 |
| Low-veg(F1-score) | 0.804 | 0.823 | 0.834 | 0.836 |

表5 在Potsdam 数据集上IFA-CNN与非编-解码器结构网络对小目标地物分割准确率比较

Tab.5 Comparison of the segmentation accuracy for small objects between IFA-CNN and non-encoder-decoder network on Potsdam dataset

| Classes | FCN[8] | SCNN[19] | RGB+I-ensembel[20] | IFA-CNN |
|---|---|---|---|---|
| Car(F1-score) | 0.893 | 0.912 | 0.892 | 0.939 |
| Low-veg(F1-score) | 0.800 | 0.837 | 0.822 | 0.906 |

为验证本文方法的有效性，IFA-CNN与数据集中使用DSM数据的网络进行实验对比，结果如表6、表7所示。表6为在Vaihingen数据集上IFA-CNN与文献[17，21，22]算法分割准确率对比，IFA-CNN无论从平均F1得分还是总体分割精度OA都取得了比较理想的结果。特别是对Tree类别的F1-score比文献[21，22]提高了0.22%，对Car类别的F1-score比文献[22]提高了0.47%。表7为在Potsdam数据集上IFA-CNN与文献[20，23-25]算法的分割准确率对比，IFA-CNN除Car的分割准确率略低于文献[25]，Surface的分割准确率与文献[23]持平外，其余类别地物的分割准确率均高于其他算法。另外，IFA-CNN的总体分割精度OA和平均F1得分均高于其他算法，证明了IFA-CNN的有效性。

由表6、表7可知，IFA-CNN无论在Vaihingen数据集还是Potsdam数据集上对各类别地物的分割效果都有着较好的表现。相比于其他算法，本文IFA-CNN网络多模态数据融合方案的优点在于多个模式之间的互补性得到了更有效利用，联合特征明显增强，更适用于将较弱的辅助数据(如DSM数据)集成到主学习网络中，并且虚拟融合单元很好地解决了特征融合效果不佳的问题。此外，由于IFA-CNN网络使用了多模态数据融合方案，同时Atrous卷积通过扩大感受野的大小来捕获多尺度信息，提高了多目标分割任务的性能，所以IFA-CNN网络更好地提高了各类别地物的分割准确率。

表6 本文方法与其他方法在Vaihingen上的分割准确率对比

Tab.6 Comparison of the accuracy of the proposed method with other methods on Vaihingen dataset

| Models | Building (F1-score) | Tree (F1-score) | Low-veg (F1-score) | Surface (F1-score) | Car (F1-score) | OA | mF1 |
|---|---|---|---|---|---|---|---|
| ADL_3[17] | 0.932 | 0.882 | 0.823 | 0.895 | 0.633 | 0.880 | 0.833 |
| ONE_7[21] | 0.945 | 0.899 | **0.844** | 0.910 | 0.778 | 0.898 | 0.875 |
| GSN[22] | 0.951 | 0.899 | 0.837 | 0.922 | 0.824 | 0.903 | 0.887 |
| IFA-CNN | **0.955** | **0.921** | 0.836 | **0.937** | 0.871 | **0.915** | **0.904** |
| | (+0.004) | (+0.022) | (-0.008) | (+0.015) | (+0.047) | (+0.012) | (+0.017) |

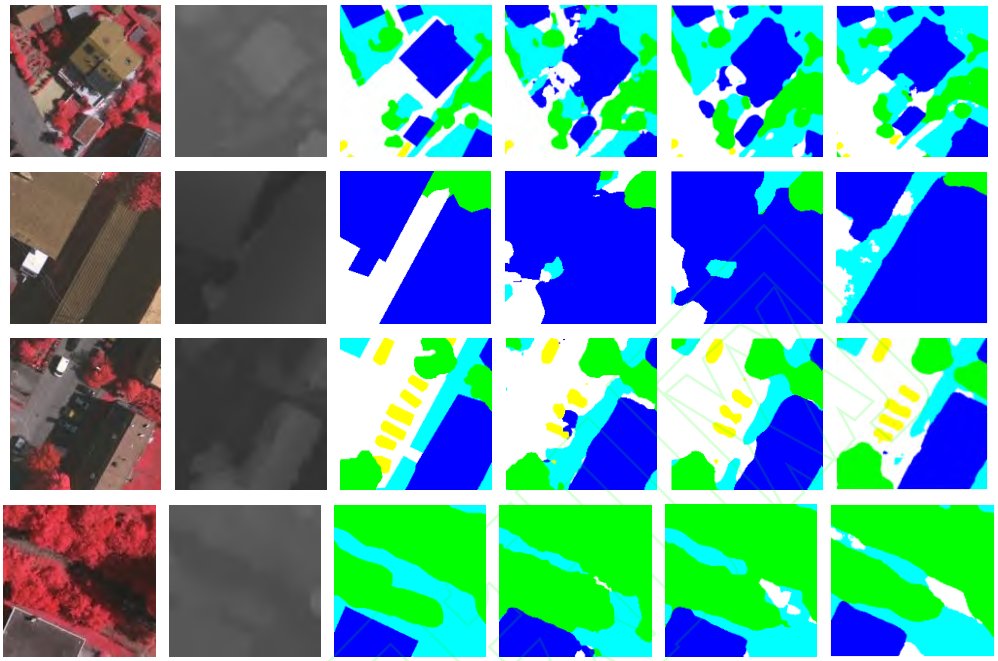
表7 本文方法与其他方法在Potsdam上的分割准确率对比

Tab.7 Comparison of the accuracy of the proposed method with other methods on Potsdam dataset

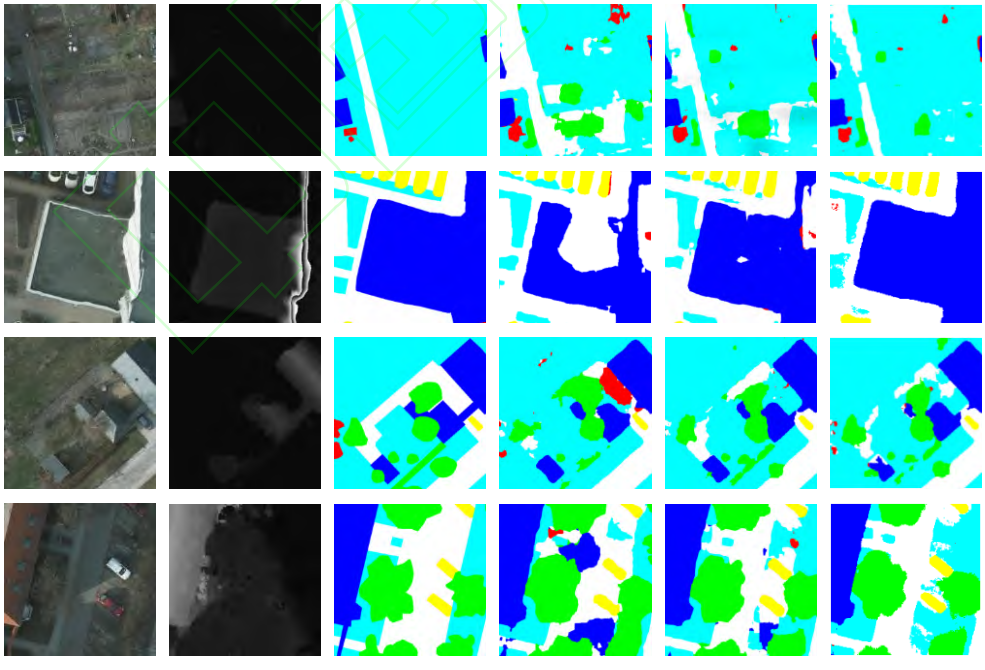
| Models | Building (F1-score) | Tree (F1-score) | Low-veg (F1-score) | Surface (F1-score) | Car (F1-score) | OA | mF1 |
|---|---|---|---|---|---|---|---|
| RiFCN[23] | 0.930 | 0.819 | 0.837 | **0.917** | 0.937 | 0.883 | 0.861 |
| RGB+I-ensemble[20] | 0.936 | 0.845 | 0.822 | 0.870 | 0.892 | 0.900 | 0.873 |
| Hallucination[24] | 0.938 | 0.848 | 0.821 | 0.873 | 0.882 | 0.901 | 0.872 |
| S-RA-FCN[25] | 0.947 | 0.835 | 0.868 | 0.913 | **0.945** | 0.886 | 0.880 |
| IFA-CNN | **0.956** | **0.864** | **0.906** | **0.917** | 0.939 | **0.909** | **0.916** |
| | (+0.009) | (+0.016) | (+0.038) | (---) | (-0.006) | (+0.008) | (+0.036) |

为使实验更具科学性，本文IFA-CNN与 网络结构为编-解码器且使用DSM数据的文

献进行影像分割细节对比，实验结果如图7所示。图中第一列为输入遥感影像的局部细节，第二列为局部DSM细节，第三列为局部真实标签。如图7(a)所示，IFA-CNN与ONE_7[21]和GSN[22]分割细节实例相比，IFA-CNN有效改善了分割影像中的边缘毛刺、细化类的边界，使得目标边缘更加接近场景的真实边缘。在图7(b)中，IFA-CNN与RiFCN[23]和S-RA-FCN[25]相比，对建筑、树木等较大目标地物分割更加准确，有效地减少误分割现象，阴影覆盖区域部分分割效果也较为理想。




(a) Vaihingen数据集分割实例的细节对比




(b) Potsdam数据集分割实例的细节对比

图7 ISPRS Vaihingen和Potsdam数据集分割实例细节对比

Fig.7 Comparison of detailed segmentation results on ISPRS Vaihingen and Potsdam dataset

## 4 结 语

本文提出了一种结合 Atrous 卷积的 FuseNet 变体深度学习网络架构，实现了高分辨率遥感影像语义分割。FuseNet 变体的多模态数据融合可以使网络学习到更强的特征并有效地利用异构数据的互补性，将高分遥感影像的 DSM 信息与 RGB 信息融合。在编-解码器架构中使用了跳跃连接，将高级特征与低级特征结合，使网络的整体分割精度提高。感受野在编-解码器部分均使用了 Atrous 卷积，采用大滤波器的转置卷积进行上采样，获得了更大的接受域。在公开数据集 ISPRS Vaihingen 和 Potsdam 上进行了实验，并与相关文献方法进行对比，实验结果表明本文所提出的深度学习网络取得了较好的分割准确率。然而，本文方法仍存在改进的空间。一方面，对于高分遥感影像的边缘还存在分割不准确的情况，另一方面，尝试在保证分割准确率的情况下减少网络层数，提高网络运算效率。

## 参 考 文 献（References）


[1] Kampffmeyer M, Salberg A B, Jenssen R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks [C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, NV, USA: IEEE, 2016: 1-9.

[2] ZUO Zongcheng, ZHANG Wen, ZHANG Dongying. A Remote Sensing Image Semantic Segmentation Method by Combining Deformable Convolution with Conditional Random Fields[J]. *Acta Geodaetica et Cartographica Sinica,* 2019, 48(6): 718-726. (左宗成, 张文, 张东映. 融合可变形卷积与条件随机场的遥感影像语义分割方法[J]. 测绘学报, 2019, 48(6): 718-726.)

[3] Wang H, Wang Y, Zhang Q, et al. Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images[J]. *Remote Sensing,* 2017, 9(5): 446.

[4] MOU Lichao, HUA Yuansheng, ZHU Xiaoxiang. A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Aerial Scenes[C]. Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 12416-12425.

[5] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 1–9.

[6] Hoffman J, Gupta S, Darrell T. Learning with Side Information Through Modality Hallucination[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 826-834.

[7] Hazirbas C, Ma L, Domokos C, et al. FuseNet: Incorporating Depth into Semantic Segmentation Via Fusion-Based CNN Architecture[C]. Asian Conference on Computer Vision. Taipei, Springer, 2016: 213-228.

[8] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 2014, 39(4): 640-651.

[9] Badrinarayanan V, Kendall A, Segnet R C. A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. *arXiv preprint arXiv*: 1511.00561, 2015.

[10] Sherrah J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery[J]. *arXiv preprint arXiv:*1606.02585, 2016.

[11] Nogueira K, Penatti O A B, Dos Santos J A. Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification[J]. *Pattern Recognition,* 2017, 61: 539-556.

[12] ZHANG Kang, HEI Baoqin, ZHOU Zhuang, et al. CNN with Coefficient of Variation-Based Dimensionality Reduction for Hyperspectral Remote Sensing Images Classification[J]. *Journal of Remote Sensing,* 2018, 22(1): 91-100. (张康, 黑保琴, 周壮, 等. 变异系数降维的 CNN 高光谱遥感图像分类[J]. 遥感学报, 2018, 22(1): 91-100.)

[13] Everingham M, Eslami S M A, Van gool L, et al. The Pascal Visual Object Classes Challenge: A Retrospective[J]. *International Journal of Computer Vision,* 2015, 111(1): 98-136.

[14] Gerke M, Rottensteiner F, Wegner J D, et al. ISPRS Semantic Labeling Contest[J]. *Remote Sensing,* 2020, 12(3): 417-446.

[15] Ngiam J, Khosla A, Kim, et al. Multimodal Deep Learning[C]. Proceedings of the 28th International Conference on Machine Learning. Washington, USA: IMLS, 2011: 689–696.

[16] Chen L C, Papandreou G, Kokkinos I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully

Connected Crfs[J]. *arXiv preprint arXiv*: 1412.7062, 2014.

[17] Luo W, Li Y, Urtasun R, et al. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[C]. Proceedings of the 30th Conference on Advances in Neural Information Processing Systems. Barcelona, Spain: JMLR, 2016: 4898-4906.

[18] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. *arXiv preprint arXiv*: 1511.07122, 2015.

[19] Liu Y, Piramanayagam S, Monteiro S T, et al. Dense Semantic Labeling of Very-High-Resolution Aerial Imagery and Lidar with Fully-Convolutional Neural Networks and Higher-Order CRFs[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA: IEEE, 2017: 76-85.

[20] ZHAO Jun, GUO Feixiao, LI Qi. Fisher-Score Algorithm of WTLS Estimation for PEIV Model[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(2): 214-220. (赵俊, 郭飞霄, 李琦. PEIV 模型 WTLS 估计的 Fisher-Score 算法[J]. 武汉大学学报信息科学版, 2019, 44(2): 214-220.)

[21] Chen G, Zhang X, Wang Q, et al. Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, 11(5): 1633-1644.

[22] Wei Y, Xiao H, Shi H, et al. Revisiting Dilated Convolution: A Simple Approach for Weakly-And Semi-Supervised Semantic Segmentation[C]. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7268-7277.

[23] Lin G, Shen C, Van Den Hengel A, et al. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 3194-3203.

[24] Paisitkriangkrai S, Sherrah J, Janney P, et al. Effective Semantic Pixel Labelling with Convolutional Networks and Conditional Random Fields[C]. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, MA, USA: IEEE, 2015: 36-43.

[25] Audebert N, LE Saux B, Lefèvre S. Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-Scale Deep Networks[C]. Proceedings of 13th Asian Conference on Computer Vision. Taipei, Taiwan: Springer, 2016: 180-196.

# Semantic segmentation of high-resolution remote sensing images based on improved FuseNet combined with the Atrous convolution

*YANG Jun*[1] *YU Xizi*[2,3,4]

1 School of Electronic and Information Engineering, Lanzhou 730070, China

2 School of Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China

3 National-Local Joint Engineering Research Center of Technologies and Applications for National Geographic State Monitoring, Lanzhou 730070, China

4 Gansu Provincial Engineering Laboratory for National Geographic State Monitoring, Lanzhou 730070, China

**Abstract:** Objectives: Semantic segmentation of remote sensing images is a key task in the acquisition of remote sensing image information. In recent years, related research results have been widely applied in detecting land-use changes, urban expansion, traffic monitoring, and disaster warning assessment. Traditional segmentation methods of remote sensing image mainly rely on the manually designed features by experts. Although good results can be achieved for some specific tasks, it has poor generalization ability. The features need to be redesigned when task conditions change resulting in low efficiency and limited applications. With the development and popularization of deep learning theory, deep neural networks are widely used in image analysis and interpretation. The high-resolution remote sensing images have the characteristics of a large amount of information, complex data, and rich feature information, and most of the current semantic segmentation neural networks of the natural image are not completely designed for the characteristics of high-resolution remote sensing images, so it cannot

effectively extract the detailed features of the ground objects in remote sensing images, and the segmentation accuracy needs to be improved. Methods: Aiming at the characteristics of high-resolution remote sensing images, we propose the process of improved FuseNet with the Atrous convolution-CNN (IFA-CNN). Firstly, It uses the improved FuseNet to fuse the elevation information of DSM images with the color information of RGB images. At the same time, there propose a multimodal data fusion scheme to solve the problem of poor fusion of the RGB branch and DSM branch. Second, multiscale features are captured through flexibly adjusting the receptive field by the Atrous convolution. Then, through deconvolution and upsampling, a decoder that increases the feature maps is formed. Finally, the SoftMax classifier is used to procure the semantic segmentation results. Results: Compared with relevant algorithms, IFA-CNN effectively improves the edge burr and thinning boundaries in segmented images, and is more accurate for segmentation of larger objects such as buildings and trees, it also reduces the miss segmentation condition with effect, the segmentation of the shadow covered areas is close to being perfect. The mF1 score achieved when our model is applied to the open ISPRS Potsdam and ISPRS Vaihingen datasets are 91.6% and 90.4% respectively, exceeding by a considerable margin of relevant algorithms. Conclusions: (1) The virtual fusion unit (V-Fusion) used for segmentation by the multimodal data fusion strategy is more accurate than the one used by the FuseNet network, meaning that, by adjusting the architecture of FuseNet and making a prudent choice of the main and auxiliary data sources, the problem of unbalanced data processing can be solved through this unit. (2) The encoder-decoder structure is arranged in such a way that the effective improvement of the segmentation accuracy of small target features is guaranteed, in particular, because during the encoding process better semantic features of high-resolution images can be extracted, the deconvolution carried out during the decoding process ensures that the predicted feature maps of the semantic segmentation as well as the semantic features of small target objects are restored. So, the loss of detailed information can be decreased. (3) While the multimodal data fusion is being carried out by IFA-CNN, the Atrous convolution expands the receptive field accordingly to extract the multiscale information, therefore the performance of the multitarget segmentation is improved.

**Keyword:** high-resolution remote sensing image; deep convolutional neural network; Atrous convolution; semantic segmentation; FuseNet

**First Author:** Yang Jun, PhD, professor, specializes in computer graphics, image processing, and geographic information system. E-mail: yangj@mail.lzjtu.cn

**Foundation Support:** The National Natural Science Foundation of China(61862039); LZJTU EP(201806).