Computer Science and Application 计算机科学与应用, 2022, 12(12), 2657-2666
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212269

Hans汉斯

基于改进HRNet的遥感图像地物覆盖语义分割研究

张琦智，王正勇*，何小海，陈洪刚

四川大学，四川 成都

收稿日期：2022年11月1日；录用日期：2022年11月30日；发布日期：2022年12月5日

摘　要

随着科技的发展，利用深度学习的方法帮助遥感图像地物覆盖语义分割工作取得了很大的进展。然而，针对于遥感语义分割数据集中存在的像素分布不平衡问题，我们提出了基于注意力的HRNet (Attention-based HRNet, AbHRNet)结构。首先，针对于各个类别目标之间的像素数量不平衡问题，本文在特征提取网络中引入了卷积注意力模块，使得网络对于我们感兴趣的目标特征尤其是数量较少的目标特征赋予了更多的关注，并减小了由复杂的背景信息带来的干扰；其次，针对于目标和目标、目标和背景之间像素数量不平衡的问题，在基准网络交叉熵损失的基础上又引入了二元交叉熵损失和Dice Loss，以实现对背景样本的有效监督，并解决由于像素数量不平衡带来的模型难以优化的问题。在LoveDA数据集上的实验结果表明，我们提出的AbHRNet的平均交并比达到了51.14%，相较于基准HRNet模型提升了1.97%，尤其是帮助分割效果很差的荒地类别的精度提升了一倍。

关键词

遥感图像，地物覆盖，语义分割，基于注意力的HRNet，卷积注意力模块

Research on Semantic Segmentation Method for Remote Sensing Image Land Cover Based on Improved HRNet

Qizhi Zhang, Zhengyong Wang*, Xiaohai He, Honggang Chen

Sichuan University, Chengdu Sichuan

Received: Nov. 1st, 2022; accepted: Nov. 30th, 2022; published: Dec. 5th, 2022

*通讯作者。

文章引用: 张琦智, 王正勇, 何小海, 陈洪刚. 基于改进 HRNet 的遥感图像地物覆盖语义分割研究[J]. 计算机科学与应用, 2022, 12(12): 2657-2666. DOI: 10.12677/csa.2022.1212269

Abstract

With the development of science and technology, great progress has been made in the remote sensing image land cover semantic segmentation task by using deep learning methods. However, in view of the unbalanced pixel distribution of remote sensing semantic segmentation dataset, we propose the Attention-based HRNet (AbHRNet) structure. First, in view of the imbalance problem of the number of pixels between various category targets, Convolutional Block Attention Module (CBAM) is introduced into the feature extraction network, so that the network gives more attention to the target features that we are interested in, especially the target features with a small number, and reduces the interference caused by complex background information; second, in view of the problem of the unbalanced pixels number between target and target, target and background, binary cross entropy loss and Dice Loss are introduced on the basis of cross entropy loss of baseline network to achieve effective supervision of background samples and to solve the problem is difficult to be optimized due to the unbalanced pixels number. The experimental results on the LoveDA dataset show that the mean intersection over union (mIoU) of our proposed AbHRNet reaches 51.14%, which has a 1.97% improvement compared to the baseline HRNet model, especially helping the barren category with poor segmentation effect to double the accuracy.

Keywords

Remote Sensing Image, Land Cover, Semantic Segmentation, Attention-Based HRNet, Convolutional Block Attention Module

1. 引言

随着科学技术的不断发展，通过卫星得到的遥感图像的空间分辨率不断提高，而利用高分辨率的遥感图像可以帮助地物覆盖制图工作取得更好的效果，从而清楚地了解土地及其生态环境以做出更准确的城乡规划决策。遥感图像的地物覆盖语义分割作为一个对于生产生活具有重要意义的研究课题，近年来在算法层面开展了许多相关的研究工作，像 Long 等人首先提出了全卷积神经网络(Fully Convolutional Networks, FCN) [1]，通过端到端的训练实现了像素到像素的语义分割；剑桥大学团队提出了 SegNet 网络[2]，采用编码和上采样加反卷积的解码结构，使网络不仅保留了图像的边缘信息还减小了计算量；Noh 等人提出了 DeconvNet 网络[3]，采用了编码和反卷积加反池化的解码结构，通过集成深度反卷积网络和提案式预测，缓解了 FCN 网络的局限性；Ronneberger 等人设计了一个对称的 U 型网络——UNet [4]，同样采用编码－解码结构，并在特征间进行有效的信息融合，使得网络能够在较少的训练图像下产生精确的分割结果；Zhou 等人提出了 UNet++网络[5]，采用深度监督的编码－解码结构，通过一系列嵌套的密集跳跃连接减小了低层特征和高层特征之间的语义间隙，实现了比 UNet 更精准的语义分割结果。可以看到，目前一些主流的语义分割算法采用的都是编码－解码结构，其特点是在编码器中将高分辨率到低分辨率的卷积串联起来，逐步减小特征图的大小；而在解码器中，利用上采样、反卷积等操作，将特征恢复到高分辨率。而最近由 Wang 等人提出的 HRNet [6]则是从一个高分辨率卷积流开始，逐步添加低分辨率的卷积流，并将多个分辨率卷积流并联起来，通过在整个训练过程中保持高分辨率表示，避免从低分

辨率复原到高分辨率时带来的空间误差。

在早期的遥感图像语义分割研究中主要依赖于中、低分辨率的数据集，像 MCD12Q1 [7]、LandCoverNet [8]、GlobeLand30 [9]等，在这些数据集上的工作主要侧重于宏观上的制图分析。随着遥感技术的进步，能够获取到的高分辨率的数据集越来越多，像 ISPRS Potsdam [10]、ISPRS Vaihingen [10]、ZurichSummer [11]和 Zeebruges [12]等，利用这些数据集可以实现更为精准的地物覆盖语义分割。然而，这些数据集的数量较少，且覆盖范围有限。而最近由武汉大学构建的 LoveDA 数据集[13]不仅包含了 5987 张高分辨率图像，并且还覆盖了乡村和城市两种场景，在乡村场景中有很多像森林、河流等自然结构，而在城市场景中以建筑、道路等人造结构为主，在这种具有丰富场景的数据集中开展相关工作具有重要的实际意义。与此同时，丰富的场景还带来了像素分布不均衡的问题。对此，我们提出了基于注意力的 HRNet (Attention-based HRNet, AbHRNet)结构：首先，我们在主干网络中引入了卷积注意力模块[14] (Convolutional Block Attention Module, CBAM)帮助模型关注数量较少的类别的特征，并减小由复杂的背景信息带来的干扰；与此同时，我们还在基准网络交叉熵损失的基础上引入了二元交叉熵损失和 Dice Loss [15]以实现对背景样本的有效监督，并解决目标和背景、目标和目标之间由于面积差距过大带来的难以优化的问题。我们的工作都是基于 CodaLab Competitions 中的 LoveDA Semantic Segmentation Challenge 开展的，在相关的网站提交了我们的结果后，最终的测试得分是由官方平台计算得出的。

2. 网络模型结构

2.1. AbHRNet 的总体结构

如图 1 所示为我们提出的 AbHRNet 的总体结构，包括主干网络和语义分割输出表示网络。

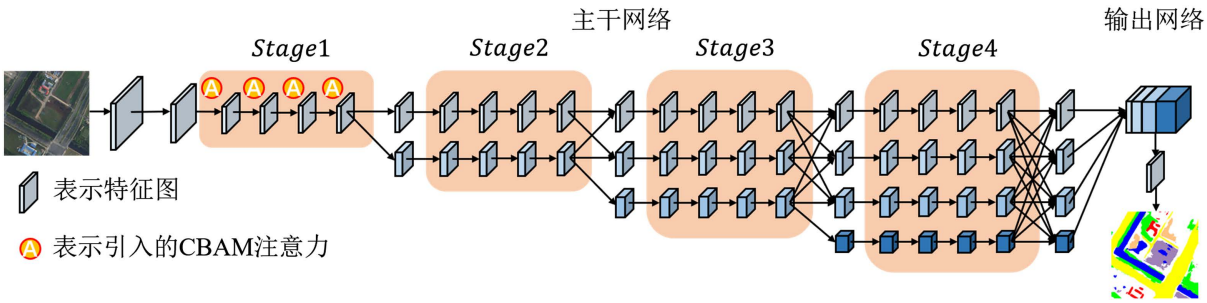

Figure 1. Overall structure of AbHRNet
图 1. AbHRNet 的总体结构

在主干网络中，首先将图像通过 2 个步长为 2 的 3 × 3 卷积，得到分辨率为原图 1/4 的特征，然后将特征送入由多个分辨率卷积流形成的特征提取模块。该模块从一个高分辨率卷积流开始作为第一阶段，然后逐步添加由高分辨率到低分辨率的卷积流形成新的阶段，最终将 I 个分辨率的卷积流并联起来。$F^{s,r}$ 用于表示相关卷积流的特征，其中 s 表示第 s 个阶段，r ($r \in \{1,2,\cdots,I\}$)既表示第 r 个卷积流又表示分辨率的索引，第 1 个卷积流的分辨率索引为 1，则第 r 个卷积流的分辨率为第 1 个卷积流分辨率的 $1/2^{r-1}$，具体的特征表示请见图 2。

而在主干网络多个分辨率卷积流之间还利用了特征融合操作来反复交换信息。其中，低分辨率特征具有较强的语义信息表征能力，而高分辨率特征则具备较强的空间信息表征能力，因此在融合之后，增强了高分辨率特征的语义表示，同时还在高分辨率特征中保持了精准的空间信息。具体实现过程如图 3 所示，我们以融合 3 个分辨率的特征表示为例。输入包括 3 个特征：$F^{3,1}, F^{3,2}, F^{3,3}$，输出特征 $F^{4,2}$ 为三个输入特征转换后的特征之和，计算公式如下式所示：

$$F^{4,2} = f_{12}\left(F^{3,1}\right) + f_{22}\left(F^{3,2}\right) + f_{32}\left(F^{3,3}\right) \tag{1}$$

其中，$f_{xr}(\bullet)$ 代表独立于输入分辨率索引 x 和输出分辨率索引 r 的转换函数。当 $x = r$ 时，$f_{xr}(F) = F$；当 $x < r$ 时，$f_{xr}(\bullet)$ 代表 $(r - x)$ 个步长为 2 的 3 × 3 卷积；当 $x > r$ 时，$f_{xr}(\bullet)$ 则代表 $(x - r)$ 倍的双线性插值上采样操作，并通过 1 × 1 的卷积以保持通道数的一致。

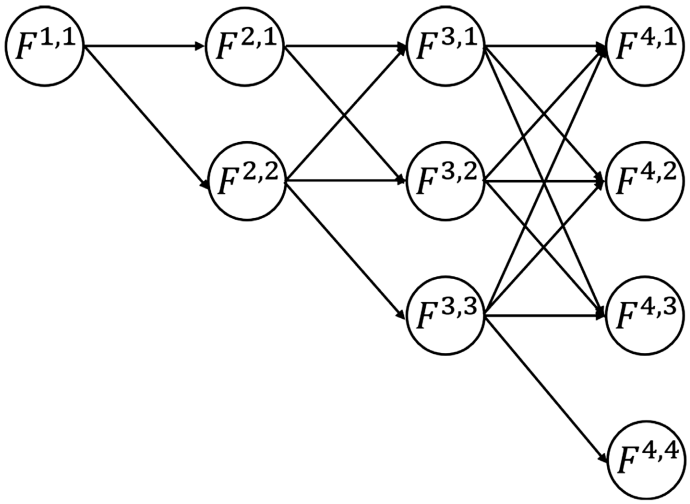

Figure 2. Backbone network structure sketch
图 2. 主干网络结构简图

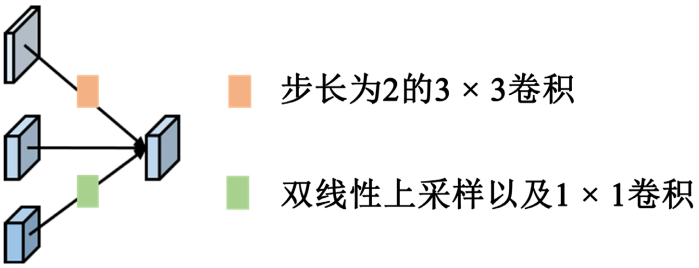

Figure 3. Multi-resolution feature fusion
图 3. 多分辨率的特征融合

而在通过主干网络后将得到的多个分辨率卷积流的输出特征送入语义分割输出表示网络。首先是通过双线性插值上采样操作将多个分辨率的特征缩放到同一尺寸下，保持通道数不变，并沿着通道维度将输出特征拼接起来，最终将拼接后的特征通过 1 × 1 的卷积把通道数转变为要分割的类别数，再通过双线性插值上采样放大到与原图同样大小后作为最终的语义分割输出结果表示。

2.2. 基于卷积注意力的增强特征提取模块

由于在遥感图像中，不仅包括我们感兴趣的地物类别，还包括复杂的背景样本，而为了让模型能够更加关注到我们感兴趣的目标特征信息，并减小由复杂的背景信息带来的干扰，我们在 HRNet 中引入了卷积注意力模块(Convolutional Block Attention Module, CBAM)。在基准 HRNet 网络中，Stage1 的基本组件为 BottleNeck 模块，Stage2、Stage3 和 Stage4 的基本组件为 BasicBlock 模块，我们在 Stage1 的 BottleNeck 模块中引入了 CBAM 注意力，构建了 CBAM-BottleNeck 模块，并把它作为 AbHRNet 中 Stage1 的基本组件，具体的实现过程如图 4 所示。首先在通过 1 × 1 卷积、3 × 3 卷积和 1 × 1 卷积后，将得到的特征 F 依

次通过通道注意力(Channel Attention, CA)模块以及空间注意力(Spatial Attention, SA)模块。在通道注意力模块中，对特征 F 分别进行全局最大池化(Global Maximum Pooling, GMP)操作和全局平均池化(Global Average pooling, GAP)操作，得到聚合了遥感图像空间信息的最大池化特征和平均池化特征。而在将它们通过具有隐藏层的多层感知机(Multi-Layer Perceptron, MLP)共享网络后，利用逐元素求和来融合输出特征向量，从而得到通道注意力特征图 M_C：

$$M_C(F)=\sigma\left(\mathrm{MLP}\left(\mathrm{GMP}(F)\right)+\mathrm{MLP}\left(\mathrm{GAP}(F)\right)\right) \tag{2}$$

其中，σ 表示 Sigmoid 函数。而后将特征 F 与通道注意力特征图 M_C 进行逐元素相乘，得到特征 F'：

$$F'=M_C(F)\otimes F \tag{3}$$

其中，$\otimes$ 表示逐元素相乘操作。

在空间注意力模块中，则是对特征 F' 沿着通道维度分别进行 GMP 操作和 GAP 操作，而后将得到的最大池化特征和平均池化特征进行拼接以得到有效突出信息区域的特征，再通过标准卷积层后输出空间注意力特征图 M_S：

$$M_S(F')=\sigma\left(f^{7\times7}\left(\left[\mathrm{GMP}(F'),\mathrm{GAP}(F')\right]\right)\right) \tag{4}$$

其中，$f^{7\times7}$ 表示滤波器大小为 7×7 的卷积操作，$[\bullet]$ 表示特征的拼接操作。而后将特征 F' 与通道注意力特征图 M_S 进行逐元素相乘，得到特征 F''：

$$F''=M_C(F')\otimes F' \tag{5}$$

将得到的特征 F'' 与输入特征 F 进行逐元素求和，并通过 Relu 激活函数后得到输出特征。

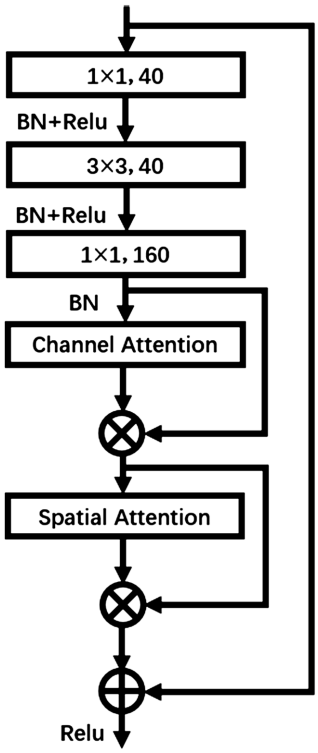

Figure 4. CBAM-BottleNeck module
图 4. CBAM-BottleNeck 模块

2.3. 损失函数

在本文的任务中，除了目标类别以外的地物均归为背景类别，而背景样本数量巨大且形态复杂，使得在该任务中出现了很多目标样本被误分类为背景样本的情况。而为了实现对背景样本的有效监督，我们在基准网络只有交叉熵损失(Cross Entropy Loss, CE Loss)的基础上引入了二元交叉熵损失(Binary Cross Entropy Loss, BCE Loss)和 Dice Loss。其中，BCE Loss 用于对背景样本的有效监督；而 Dice Loss [15]起先是用来计算两个样本之间相似性的度量函数，而现在则是在各种语义分割任务中得到了广泛的使用。通过将 Dice Loss 与交叉熵损失结合在一起，具有以下的优势：1) Dice Loss 和交叉熵损失分别从图像的整体层面和像素层面对网络进行优化；2) 当图像中背景样本和目标样本之间的面积差距过大时，交叉熵损失无法对小面积的目标样本进行优化，而 Dice Loss 则可以帮助各种大小的样本进行优化；3) Dice Loss 可以帮助模型更加关注前景区域的学习。最终，通过将损失之间的优势进行互补，实现有效地性能提升，总损失计算如下：

$$L_{total} = L_{ce} + \omega L_{bce} + \theta L_{dice} \tag{6}$$

其中，ω 和 θ 分别表示 BCE Loss 和 Dice Loss 的权重系数。

3. 实验及结果分析

3.1. 数据集

我们选择 LoveDA 数据集进行实验验证。LoveDA 数据集是武汉大学测绘与遥感信息工程国家重点实验室于 2016 年 7 月在南京、常州和武汉等地通过谷歌地球平台收集的 0.3 米空间分辨率的遥感图像数据集，总面积为 536.15 km^2。该数据集总共包含 5987 张高分辨率遥感图像，以及 166768 个标注对象，图像的尺寸为 1024 × 1024，并将其划分为 7 个类别：背景(RGB: 255, 255, 255)、建筑(RGB: 255, 0, 0)、道路(RGB: 255, 255, 0)、水体(RGB: 0, 0, 255)、荒地(RGB: 159, 129, 183)、森林(RGB: 0, 255, 0)、农田(RGB: 255, 0, 255)。图 5 展示了在 LoveDA 数据集中各个类别的像素数量，可以看到背景占据的像素数量最多，而荒地占据的像素数量最少，总体呈现不均衡的分布。LoveDA 数据集中是从上述 3 个城市内 9 个农村场景和 9 个城市场景中收集的数据，训练集选择了 4 个农村场景和 4 个城市场景共 8 个场景的数据，剩余的数据作为验证集和测试集。

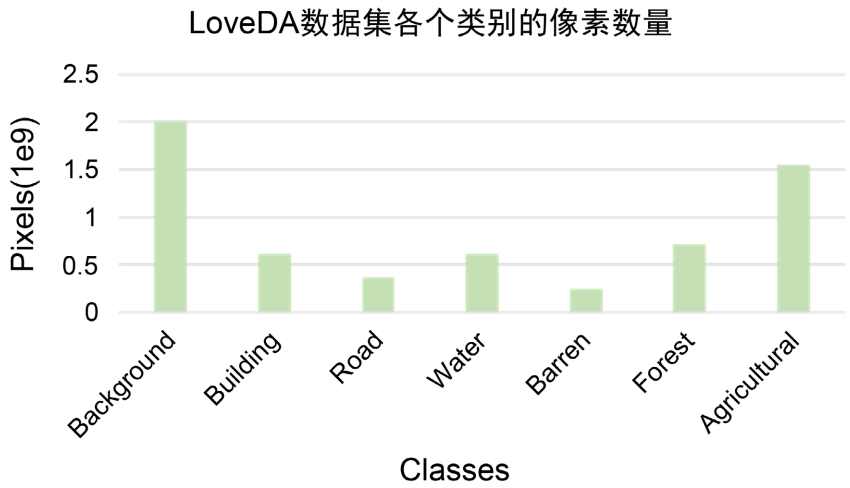

Figure 5. Number pixels of each category in the LoveDA dataset
图 5. 在 LoveDA 数据集中各个类别的像素数量

3.2. 实验设置和评估指标

本实验基于 Pytorch 深度学习框架，GPU 使用了 NVIDIA GeForce GTX 2080Ti 显卡。在训练期间，初始学习率设置为 0.1，并使用动量为 0.9、权重衰减为 10^{-4} 的随机梯度下降(SGD)优化器。训练的迭代次数为 30k，batchsize 设置为 8。对于数据的预处理，采用了随机裁剪、水平/垂直翻转、角度旋转等数据增强方式。实验使用的预训练模型是基于 ImageNet 数据集。我们采用 mIoU (即所有类别 IoU 的平均值)作为评价指标。

3.3. 实验结果及分析

表 1 展示了我们的方法与部分主流的语义分割方法的比较。具体而言，我们选择的对比网络包括：FCN8S [1], DeepLabV3+ [16], PAN [17], UNet [4], UNet++ [5], Semantic-FPN [18], PSPNet [19], LinkNet [20], FarSeg [21], FactSeg [22]。通过比较可以看到在我们提出的 AbHRNet 网络中，mIoU 达到了最高的 51.14%，并且在背景、建筑、道路、荒地、农田等 5 个类别中都达到了最高的分割精度，尤其是帮助精度较低的荒地类别实现了较大的性能提升，这证明了我们提出的改进对于遥感图像地物覆盖语义分割任务是有帮助的。

Table 1. Comparison with mainstream semantic segmentation algorithms on the LoveDA dataset
表 1. 在 LoveDA 数据集上与主流语义分割算法的比较

Method	Backbone	IoU accuracy							mIoU (%)
		Background	Building	Road	Water	Barren	Forest	Agriculture	
FCN8S	VGG16	42.60	49.51	48.05	73.09	11.84	43.49	58.30	46.69
DeepLabV3+	ResNet50	42.97	50.88	52.02	74.36	10.40	44.21	58.53	47.62
PAN	ResNet50	43.04	51.34	50.93	74.77	10.03	42.19	57.65	47.13
UNet	ResNet50	43.06	52.74	52.78	73.08	10.33	43.05	59.87	47.84
UNet++	ResNet50	42.85	52.58	52.82	74.51	11.42	44.42	58.80	48.20
Semantic-FPN	ResNet50	42.93	51.53	53.43	74.67	11.21	44.62	58.68	48.15
PSPNet	ResNet50	44.40	52.13	53.52	76.50	9.73	44.07	57.85	48.31
LinkNet	ResNet50	43.61	52.07	52.53	76.85	12.16	**45.05**	57.25	48.50
FarSeg	ResNet50	43.09	51.48	53.85	76.61	9.78	43.33	58.90	48.15
FactSeg	ResNet50	42.60	53.63	52.79	**76.94**	16.20	42.92	57.50	48.94
AbHRNet	W40	**45.33**	**57.16**	**56.86**	76.64	**17.81**	43.61	**60.60**	**51.14**

为了对我们提出改进的有效性进行验证，我们开展了相关的消融实验。正如表 2 所展示的，在基准网络 HRNet 上，mIoU 为 49.17%，而在主干网络中引入了 CBAM 注意力后，mIoU 达到了 50.79%，提高了 1.62%，尤其是对于分割效果很差的荒地(Barren)类别，分割精度几乎上升了 1 倍，这说明引入的 CBAM 注意力帮助我们的模型关注到了数量较少的类别，并在训练过程中赋予其更多的权重，实现了有效的性能上升。而在基准网络交叉熵损失的基础上，引入了二元交叉熵损失和 Dice Loss 后，mIoU 达到了 51.14%，进一步地提升了 0.42%，说明该损失可以有效地监督背景样本，并帮助极大或极小的目标样本实现更为有效的分割。

Table 2. Ablation experiments on the LoveDA dataset
表 2. 在 LoveDA 数据集上的消融实验

Method	CBAM	CE+ (BCE+Dice)	Background	Building	Road	Water	Barren	Forest	Agriculture	MIoU (%)
Baseline			43.46	**57.65**	53.12	73.82	8.67	**44.39**	**63.07**	49.17
Ours	√		44.26	55.96	**57.55**	**77.32**	15.19	44.21	61.06	50.79
	√	√	**45.33**	57.16	56.86	76.64	**17.81**	43.61	60.60	**51.14**

与此同时，为了直观地证明我们提出方法的有效性，我们还在图 6 中展示了在 LoveDA 数据集上的语义分割可视化结果，图 6(a)为原始的遥感图像，图 6(b)为基准 HRNet 网络的分割结果，图 6(c)为我们提出的 AbHRNet 网络的分割结果。其中第一行展示了在农村场景中的语义分割可视化结果。可以看到，基准 HRNet 网络将大部分的荒地目标样本错误地分类为背景样本，而在我们改进的 HRNet 网络中，则将大部分的荒地目标样本实现了正确的分割。而第二行则展示了在城市场景中的语义分割可视化结果。在图中的右下角可以看到，在我们提出的方法中实现了对森林类别和建筑物类别更精细的分割。最终通过可视化的结果，同样证明了我们提出的方法在遥感图像语义分割任务中实现了性能的提升。

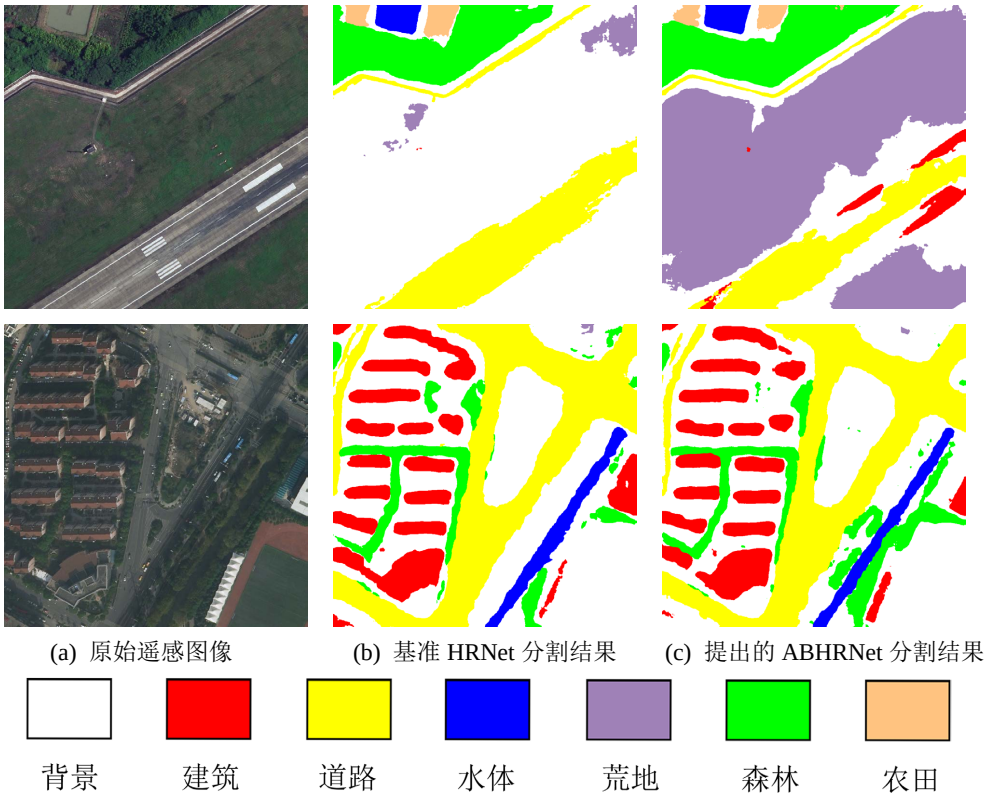

Figure 6. Number pixels of each category in the LoveDA dataset
图 6. 在 LoveDA 数据集中各个类别的像素数量

4. 结论

本文提出了基于注意力的 HRNet (AbHRNet)网络用于遥感图像地物覆盖语义分割任务。首先，通过在特征提取网络中引入卷积注意力(CBAM)模块，帮助模型更加关注我们感兴趣的目标特征信息，并减小

由于复杂的背景信息带来的干扰；同时，我们还在基准网络交叉熵损失的基础上引入了二元交叉熵损失和 Dice Loss，以实现对背景样本的有效监督，并解决了目标和背景、目标和目标之间由于面积差距过大带来的难以优化的问题。最终正如在 LoveDA 数据集上的实验结果所展示的，我们的方法达到了最优的性能，尤其是在分割效果不好的荒地类别中，实现了精度的翻倍。

基金项目

国家自然科学基金 62271336。

参考文献

[1] Shelhamer, E., Long, J. and Darrell, T. (2017) Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 640-651. https://doi.org/10.1109/TPAMI.2016.2572683

[2] 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**, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[3] Noh, H., Hong, S. and Han, B. (2015) Learning Deconvolution Network for Semantic Segmentation. *IEEE International Conference on Computer Vision*, Santiago, 11-18 December 2015, 1520-1528. https://doi.org/10.1109/ICCV.2015.178

[4] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. 18*th International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*), Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[5] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J. (2018) UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: 4*th International Workshop*, *DLMIA* 2018, *and* 8*th International Workshop*, *ML-CDS* 2018, Granada, 11045, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[6] Wang, J., Sun, K., Cheng, T., *et al*. (2021) Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 3349-3364. https://doi.org/10.1109/TPAMI.2020.2983686

[7] Sulla-Menashe, D. and Friedl, M.A. (2018) User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product. USGS, Reston, 1, 18.

[8] Alemohammad, H. and Booth, K. (2020) LandCoverNet: A Global Benchmark Land Cover Classification Training Dataset. https://arxiv.org/abs/2012.03111

[9] Jun, C., Ban, Y. and Li, S. (2014) Open Access to Earth Land-Cover Map. *Nature*, **514**, 434-434. https://doi.org/10.1038/514434c

[10] Mou, L.C., Hua, Y.S. and Zhu, X.X. (2019) A Relation-Augmented Fully Convolutional Network for Semantic Segmentation in Aerial Scenes. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Proceedings*, Long Beach, 16-20 June 2019, 12408-12417.

[11] Volpi, M. and Ferrari, V. (2015) Semantic Segmentation of Urban Scenes by Learning Local Class Interactions. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 Jun 2015, 1-9. https://doi.org/10.1109/CVPRW.2015.7301377

[12] Marcos, D., Volpi, M., Kellenberger, B. and Tuia, D. (2018) Land Cover Mapping at Very High Resolution with Rotation Equivariant CNNs: Towards Small Yet Accurate Models. *ISPRS Journal of Photogrammetry and Remote Sensing*, **145**, 96-107. https://doi.org/10.1016/j.isprsjprs.2018.01.021

[13] Wang, J., Zheng, Z., Ma, A., Lu, X. and Zhong, Y. (2021) LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. https://arxiv.org/abs/2110.08733

[14] Woo, S.H., Park, J., Lee, J.Y. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. 15*th European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[15] Milletari, F., Navab, N. and Ahmadi, S.A. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 4*th IEEE International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3DV.2016.79

[16] Chen, L.C.E., Zhu, Y.K., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-Decoder with Atrous Separable

Convolution for Semantic Image Segmentation. 15*th European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 833-851. https://doi.org/10.1007/978-3-030-01234-2_49

[17] Li, H., Xiong, P., An, J. and Wang, L. (2018) Pyramid Attention Network for Semantic Segmentation. https://arxiv.org/abs/1805.10180

[18] Kirillov, A., Girshick, R., Kaiming, H. and Dollar, P. (2019) Panoptic Feature Pyramid Networks. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Proceedings*, Long Beach, 15-20 June 2019, 6392-6401. https://doi.org/10.1109/CVPR.2019.00656

[19] Zhao, H.S., Shi, J.P., Qi, X.J., Wang, X.G. and Jia, J.Y. (2017) Pyramid Scene Parsing Network. 30*th IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[20] Chaurasia, A. and Culurciello, E. (2017) LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. *IEEE Visual Communications and Image Processing* (*VCIP*), St. Petersburg, 10-13 December 2017, 1-4. https://doi.org/10.1109/VCIP.2017.8305148

[21] Zheng, Z., Zhong, Y.F., Wang, J.J. and Ma, A.L. (2020) Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 4095-4104. https://doi.org/10.1109/CVPR42600.2020.00415

[22] Ma, A., Wang, J., Zhong, Y. and Zheng, Z. (2021) Factseg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Transactions on Geoscience Remote Sensing*, **60**, 1-16. https://doi.org/10.1109/TGRS.2021.3097148