计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(8):51−59 [doi: 10.15888/j.cnki.csa.009593]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于语义引导与多尺度增强的遥感影像分割网络①

孙梓翔[1,2], 钱旭威[3], 杨　平[1,2], 杭仁龙[1,2]

[1](南京信息工程大学 计算机学院, 南京 210044)
[2](江苏省大气环境与装备技术协同创新中心, 南京 210044)
[3](东南大学 计算机科学与工程学院, 南京 211189)
通信作者: 杭仁龙, E-mail: renlong_hang@163.com

摘　要: 遥感影像语义分割在环境监测、土地覆盖分类和城市规划等领域发挥着至关重要的作用. 卷积神经网络及其改进模型是遥感影像语义分割的主流方法, 但此类方法更加关注局部上下文特征的学习, 无法有效建模不同物体之间的全局分布关系, 进而制约了模型的分割性能. 为了解决该问题, 本文在卷积神经网络的基础上, 构建了全局语义关系学习模块, 充分学习不同物体之间的共生关系, 有效地增强了模型的表征能力. 此外, 考虑到同一场景中, 待分割物体的尺度存在差异性, 构建了多尺度关系学习模块, 以融合不同尺度的全局语义关系. 为了评估模型的性能, 本文在 Vaihingen 和 Potsdam 两个常用的遥感影像数据集上进行了充分的实验. 实验结果表明, 本文方法能够获得比已有的基于卷积神经网络的模型更高的分割性能.

关键词: 遥感影像; 语义分割; 全局语义关系; 多尺度融合

引用格式: 孙梓翔,钱旭威,杨平,杭仁龙.基于语义引导与多尺度增强的遥感影像分割网络.计算机系统应用,2024,33(8):51–59. http://www.c-s-a.org.cn/1003-3254/9593.html

Remote Sensing Image Segmentation Network Based on Semantic Guide and Multi-scale Enhancement

SUN Zi-Xiang[1,2], QIAN Xu-Wei[3], YANG Ping[1,2], HANG Ren-Long[1,2]

[1](School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing 210044, China)
[3](School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: Semantic segmentation of remote sensing images plays a crucial role in environmental detection, land cover classification, and urban planning. Convolutional neural networks and their improved models are the mainstream methods for semantic segmentation of remote sensing images. However, these methods focus more on learning local contextual features and cannot effectively model the global distribution relationship between different objects, thereby restricting the segmentation performance of the model. To address this issue, this study constructs a global semantic relationship learning module based on convolutional neural networks, which fully learns the symbiotic relationships between different objects and effectively enhances the model's representation ability. In addition, a multi-scale relationship learning module is constructed to integrate global semantic relationships of different scales, given the scale differences of the objects to be segmented in the same scene. To evaluate the performance of the model, sufficient experiments are conducted on two commonly used remote sensing image datasets, Vaihingen and Potsdam. The experimental results show that the proposed method can achieve higher segmentation performance than existing models based on convolutional neural networks.

Key words: remote sensing image; semantic segmentation; global semantic relationship; multi-scale fusion

① 基金项目: 国家自然科学基金 (U21B2049, 61906096)
收稿时间: 2024-02-07; 修改时间: 2024-03-28; 采用时间: 2024-04-03; csa 在线出版时间: 2024-06-28
CNKI 网络首发时间: 2024-07-01

遥感影像是指通过遥感技术获取的反映地表各类地物电磁波辐射和反射信息的图像, 通过对遥感影像进行处理, 可以获取丰富的地物信息. 近年来, 随着遥感技术的迅速发展, 遥感影像在环境监测[1]、土地覆盖分类[2]、城市规划[3]等多个领域得到了广泛应用. 随着遥感数据量的激增, 传统的识别方法已经无法满足现有需求, 如何准确高效地识别影像中地物成为遥感影像领域的重要研究课题.

相较于传统的分割方法, 如聚类分割[4]、边缘分割[5]和阈值分割[6], 卷积神经网络 (convolutional neural network, CNN) 在遥感影像语义分割领域[7]具有显著的优势. CNN 能够自动学习遥感影像中不同层次的特征, 从而提高分割结果的准确性和鲁棒性. FCN[8]首次采用全卷积网络实现了对任意尺寸的图像分割. U-Net[9]采用编码器-解码器结构来提取图像特征, 并通过跳跃连接实现深度与浅层信息的融合, 以保留更多的细节信息. UNet++[10]在 U-Net 的基础上重新设计跳跃连接路径, UNet3+[11]引入全尺度跳跃连接, 并设计分类指导模块以减少图像的过度分割, RA-UNet[12]通过结合残差和注意力机制, 加强模型对图像特定区域的分割能力. DeepLab[13]将全连接条件随机场引入到 FCN 中, 增加了上下文信息的提取能力. 为提高对遥感影像小尺度目标的识别能力, 张寅等人通过特征增强模块提升目标特征提取能力, 并利用注意力机制与空间注意力模块组成的级联注意力机制, 实现精准地捕获小目标特征[14]. 尽管卷积神经网络在遥感影像语义分割领域取得了显著的进展, 但现有方法往往聚焦于局部上下文特征的学习, 无法有效建模不同物体之间的全局分布关系, 进而制约了模型的分割性能.

为了解决这一问题, 本文提出了语义引导与多尺度增强的分割网络 (semantic guided and multi-scale enhanced segmentation network, SGMESN). 该方法在卷积神经网络的基础上, 构建了全局语义关系学习模块, 充分学习不同物体之间的共生关系, 有效地增强了模型的表征能力. 在此基础之上, 考虑到同一场景中, 遥感影像分割物体的尺度存在差异性, 设计了多尺度学习模块, 以融合不同尺度的全局语义关系. 在 Potsdam 和 Vaihingen 两个常用的遥感影像分割数据集上进行的实验结果表明, 本方法与其他分割模型相比, 能够实现更加精准的分割效果, 证明了全局语义关系在遥感影像分割中的重要性和有效性.

1 SGMESN 网络整体结构

本文提出一种语义引导与多尺度增强的遥感影像分割方法, 如图 1 所示. 网络整体结构由主干网络、全局语义关系学习模块 (GRLM) 和多尺度关系学习模块 (MRLM) 组成.

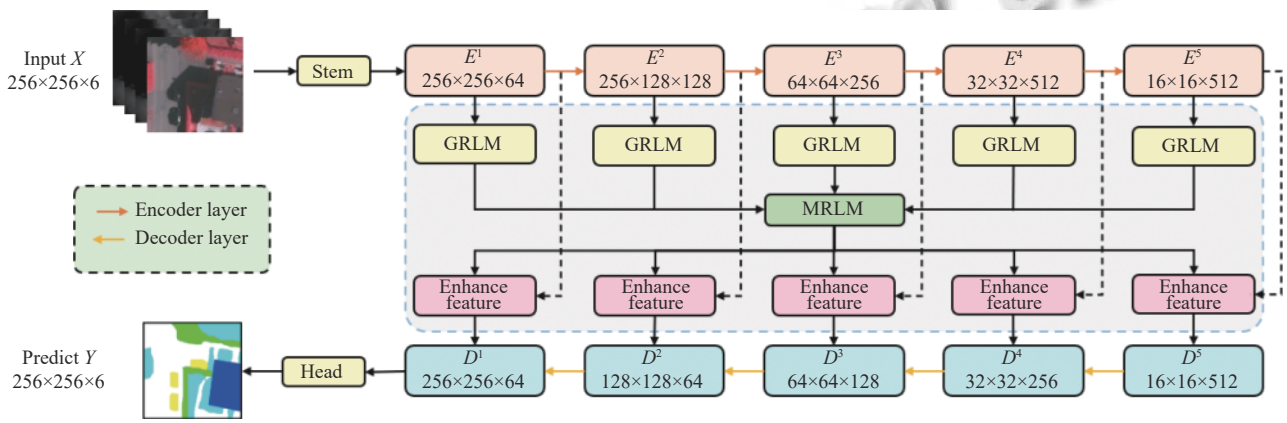

图 1 语义引导与多尺度增强的遥感影像分割网络

2 SGMESN 网络详细介绍

2.1 主干网络

为了充分利用遥感图像中丰富的细节信息和上下文信息, 本文采用 U-Net 作为主干网络进行特征提取, 利用其强大的特征提取能力和对细节信息的关注, 以确保在处理复杂特征的同时能够保留关键细节信息的完整性.

对于输入网络的遥感影像数据 $X \in R^{(H\times W\times C)}$, 首先通过由两个 3×3 的卷积层组成的 Stem 模块, 对图像进行初步的特征提取并增加通道维度, 得到一个大小为

$H\times W\times C$的特征图E^1, 紧接着, 经过由 4 个连续的编码层组成的编码器, 生成 4 个特征图分别表示为: $E^2\in\mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2D}$, $E^3\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4D}$, $E^4\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8D}$, $E^5\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 8D}$. 其中编码层由一个最大池化层和两个串联的3×3卷积层组成. 在解码器中包含了 4 个解码层, 每个解码层由 1 个通过双线性插值的上采样层和 2 个3×3卷积层组成, 并通过跳跃连接将编码层的特征与解码层的特征进行融合, 通过以上步骤获取到的 5 个特征分别表示: $D^5\in\mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 8D}$, $D^4\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 4D}$, $D^3\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 2D}$, $D^2\in\mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times D}$, $D^1\in\mathbb{R}^{H\times W\times D}$. 最后, 将$D^1$输入到由1×1卷积层组成的分割头 (head) 中, 产生图像分割结果$Y\in\mathbb{R}^{H\times W\times N}$, N表示输入的遥感影像数据 X 中包含的目标类别数.

2.2 全局语义关系学习模块

卷积神经网络方法通常专注于提取局部上下文特征的学习, 未能充分利用图像中的不同物体之间的全局分布关系, 进而制约了模型的分割性能. 如图 2 所示在不同区域中, 汽车与周围物体存在全分布关系, 图 2(a) 中汽车周围出现建筑与不透水表面, 图 2(b) 中汽车周围出现低植被、树木和不透水表面, 汽车出现时必然出现不透水表面.

(a) (b)

图 2 不同物体之间存在的全局分布关系

针对此问题本文设计了全局语义关系学习模块, 如图 3 所示. 首先通过 Reshape 操作将输入特征进行转化, 表示为:

$$E^i\in\mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^iD}\rightarrow x^i\in\mathbb{R}^{\frac{HW}{4^i}\times 2^iD} \tag{1}$$

其中, i 代表不同的编码层. 其次利用全连接层 (*Linear*) 将通道维度转换为类别数量 (L), 即:

$$m^i\in\mathbb{R}^{\frac{HW}{4^i}\times L} \tag{2}$$

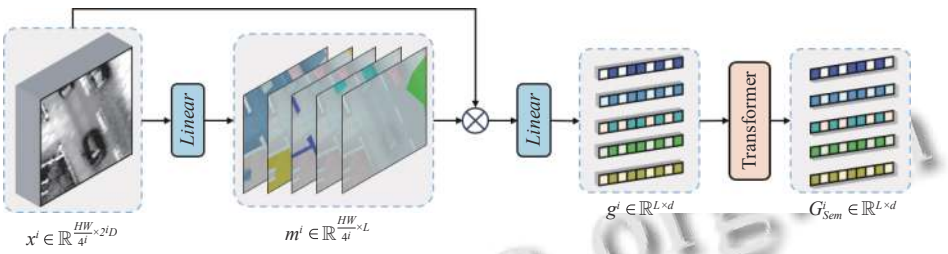

图 3 全局语义关系学习模块

然后经过 *Softmax* 函数对m^i的通道维度进行归一化处理后, 将其与输入特征x^i相乘, 形成了融合类别信息和通道信息的特征$t^i\in\mathbb{R}^{L\times 2^iD}$. 为实现第 2.3 节中不同层之间全局语义特征的融合, 要保持不同编码层中的g^i尺寸一致, 通过使用一个额外的线性层, 将$t^i\in\mathbb{R}^{L\times 2^iD}$ 转换为所需的输出$g^i\in\mathbb{R}^{l\times d}$, 表示为:

$$g^i=Linear[Softmax(Linear(x^i)^{\mathrm{T}}\cdot x^i)] \tag{3}$$

在获取g^i之后, 将其输入到 Transformer[15], 由多头注意力机制和前馈神经网络 (feed-forward neural network, *FFN*) 组成, 用于捕获全局语义特征中类别间的共生关系. 首先q^i和k^i表示g^i通过线性投影得到的 Query 和 Key, 然后q^i和k^i执行向量点积, 并应用 *Softmax* 函数得到类与类之间的权重. 这些权重与相应的g^i相乘, 进一步计算得到输出向量, 并通过前馈神经网络, 得到最终的输出结果G_{Sem}表示为:

$$G_{\mathrm{Sem}}=FFN\left(Attention\left(\frac{q^i\cdot k^{i\mathrm{T}}}{\sqrt{d}}\right)\cdot g^i\right) \tag{4}$$

2.3 多尺度关系学习模块

考虑到同一场景下, 待分割物体的尺度存在差异性. 为了更好地融合不同尺度的全局语义关系, 本文设计了一种多尺度关系学习模块. 该模块将不同编码层中全局语义关系学习模块所获取的$G_{\mathrm{Sem}}\in\mathbb{R}^{L\times d}$都转化为一维向量并融合得到特征$G_{\mathrm{Sca}}\in\mathbb{R}^{5\times Ld}$, 其次利用 Transformer 学习不同尺度下的特征之间的联系, 获取

不同尺度下的特征对于分割目标的重要程度, 并赋予不同的权重. 经过 Transformer 处理后得到G'_{Sca}, 将其分解为与G_{Sem}尺寸一致的 5 个T_i. 然后将T_i与输入特征$x^i \in \mathbb{R}^{wh\times c}$进行融合, 利用$T_i$中的类别间的共生关系和多尺度关系来增强特征$x^i$的表示, 并通过残差连接来弥补细节信息, 融合过程如图 4 所示, 表示为:

$$x^i_{\text{out}} = x^i + Softmax\left(\Phi\left(x^i\right)\cdot\left(\Phi\left(T_i\right)\right)^{\text{T}}\right)\cdot T_i \tag{5}$$

其中, Φ表示全连接层, 用于将输入特征x^i和T_i进行线性变换. *Softmax*函数用于获取输入特征x^i与T_i之间的相似度矩阵, 并与T_i进行加权求和, 以获得每个像素点与所有类别的关系. 通过融合过程, 可以获得优化后的增强特征 (enhance feature) $x^i_{\text{out}} \in \mathbb{R}^{wh\times c}$. 与原始特征相比, 优化后的特征图中包含了更加准确和丰富的全局语义信息和多尺度信息. 最后, 将增强特征x^i_{out}通过 Reshape 操作转化为$\mathbb{R}^{w\times h\times c}$, 以适应后续处理的需要. 通过多尺度关系学习模块, 模型能够更好地融合图像中不同尺度的全局语义关系, 进而实现更准确的分割结果.

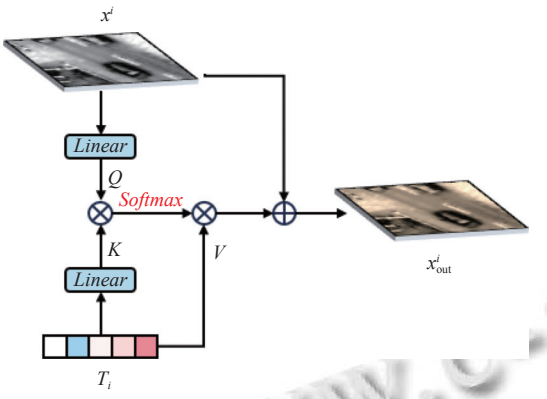

图 4 融合过程

3 实验与分析

3.1 实验数据

为了全面评估本文提出方法的有效性, 在两个公开的数据集进行了系统性地测试. 这两个数据集分别是 Potsdam 和 Vaihingen 数据集.

Potsdam 数据集由 38 幅高质量遥感图像组成, 覆盖了德国勃兰登堡首都上空的区域. 每幅图像由 3 波段组成分别为红外波段、红色波段和绿色波段, 以 TIFF 格式存储, 具有 9 ms 的空间分辨率, 且图像尺寸为 6000×6000 像素. 此外, 该数据集还提供了一个单波段的数字表面模型 (DSM), 为理解地形的三维信息提供了重要依据. 这些图像包括 6 种不同的地物类别: 低植被、不透水表面、建筑、树木、汽车和背景. 本文使用的合成图像包含 6 个通道: 近红外、红色、绿色、归一化植被指数 (*NDVI*)、DSM 和 nDSM. 其中归一化植被指数 (*NDVI*) 的引入是为了提供有关植被的信息, 这对于植被分割尤为重要, *NDVI* 计算如下:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{6}$$

其中, *NIR* 表示近红外中的反射率, *Red* 表示红色波段中的反射率. *NDVI* 的值在−1 到 1 的范围内, 其值越高, 代表植被的密度越大. nDSM 通过滤除 DSM 中的地表信息, 提取出能够真实反映地物高度的信息, 有助于区分不同类别的地物. 该数据集被划分为 24 幅图像的训练集和 14 幅图像的测试集. 在进行训练和验证之前, 对数据集中的图像使用一个步幅为 200 的 256×256 滑动窗口进行裁剪, 最终分别得到 19 342 个训练图像和 7405 个验证图像.

Vaihingen 数据集由 33 幅大小不一的遥感图像组成, 平均尺寸约为 2494×2064 像素. 该数据集的地物类别数与 Potsdam 数据相同, 同样采用 6 通道的合成图像作为模型的输入. 训练集包含 16 幅图像, 剩余 17 幅图像则作为测试集. 在进行训练和验证之前, 对数据集中的图像使用一个步幅为 64 的 256×256 滑动窗口进行裁剪, 最终分别得到 15 268 个训练图像和 1 208 个验证图像. 此外, 为提高模型的泛化能力, 本文还采用了常见的数据增强方法, 如裁剪、缩放、翻转以及对比度增强等, 使数据量扩充至原本的 3 倍.

3.2 评价指标

为了全面评估本文方法的性能, 采用了以下关键的评估指标, 以提供对模型性能的定量评价结果: 总体准确度 (overall accuracy, *OA*)、*F*1 分数 (*F*1-score)、平均 *F*1 分数 (mean *F*1) 以及平均交并比 (mean intersection over union, *mIoU*). 具体 *F*1 和 *mIoU* 指标计算方法如下:

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

$$F1 = 2\cdot\frac{Precision\cdot Recall}{Precision + Recall} \tag{8}$$

$$mIoU = \frac{TP}{FP + FN + TP} \tag{9}$$

其中, *Precision* 代表标记为正样本中被正确标记的比例, *Recall* 代表预测为正样本中被正确预测的比例, *TP* 是正样本被正确识别的数量, *TN* 是负样本被正确识别的数量, *FP* 是误报的负样本数量, *FN* 是漏报的正样本数量.

3.3 实验配置

本文实验在一台配备 64 位体系结构操作系统和 NVIDIA GeForce GTX 3090 GPU 的计算机上进行, 使用 PyTorch 框架搭建模型. 训练过程中, 采用交叉熵损失作为损失函数, 并使用随机梯度下降算法 (SGD) 进行优化, SGD 的动量设置为 0.9, 初始学习率设为 0.01. 此外, 为了更好地调整学习率, 本文采用了自适应学习率调整策略, 其计算公式如下:

$$lr = base_{lr} \cdot \left(1 - \frac{epoch}{epochs}\right)^{power} \tag{10}$$

其中, $base_{lr}$ 设置为 0.01, 功率 (*power*) 设置为 1.0, *epoch* 表示当前迭代, *epochs* 表示迭代的总次数.

3.4 超参数分析

3.4.1 全局语义特征中类别数量的分析

为深入了解全局语义学习模块中类别数量对网络分割性能的具体影响, 在 Potsdam 和 Vaihingen 数据集上进行了一系列实验. 在{2, 6, 12, 18, 24, 30}范围内进行类别数 (*L*) 设定的测试, 同时保持其他的超参数不变, 结果如表 1 所示.

表 1 全局语义特征中类别数量的分析 (%)

类别数量	Potsdam *mIoU*	Vaihingen *mIoU*
2	75.7	78.4
6	76.9	79.2
12	**77.8**	**80.7**
18	77.2	80.4
24	76.7	79.7
30	76.4	79.9

当类别数 (*L*) 从 2 增加到 6 时, 在 Potsdam 和 Vaihingen 数据集上的 *mIoU* 值分别提高了 1.2% 和 0.8%. 进一步增加至 12 个类别时, 可以观察到网络性能达到最高值, 分别提升 0.9% 和 1.5%. 然而, 当类别数继续增加时, 模型的分割性能并未进一步提升, 反而有所下降. 这一结果表明, 一个适中的全局语义特征列别数能够代表特征图的高级语义信息. 因此, 本文在两个数据集上全局语义特征的类别数量均设置为 12.

3.4.2 Transformer 中 head 数量的分析

为了评估全局语义关系学习模块 Transformer 中的 head 数量对于网络分割性能的影响, 本文从{1, 2, 4, 8, 16, 32, 64}中, 依次选择不同的 head 数量, 并固定其他超参数进行实验, 结果如表 2 所示. 随着 head 数量从 1 逐渐增加至 8, 模型在两个数据集上的 *mIoU* 值有明显提升, 分别提升 0.6% 和 0.3%. 特别地, 在 Potsdam 数据集上, 当 head 数量为 16 和 32 时, *mIoU* 指标相近. 然而, 当 head 数量超过 8 后, 两个数据集上的 *mIoU* 值都有出现了下降. 基于这一实验结果, 本文在两个数据集上将全局语义关系学习模块中 Transformer 的 head 数量设置为 8.

表 2 Transformer 中 head 数量的分析 (%)

head数量	Potsdam *mIoU*	Vaihingen *mIoU*
1	77.2	80.4
2	76.9	80.8
4	77.7	80.2
8	**77.8**	**80.7**
16	77.4	80.6
32	77.4	80.5
64	76.8	80.2

3.4.3 测试集的裁剪滑动步长分析

由于裁剪步长决定了模型在处理图像时的空间采样密度, 因此本文为了验证测试集裁剪步长对网络分割性能的影响. 在全局语义关系学习模块设置不同的全局语义特征尺寸的情况下, 依次对测试集进行裁剪滑动步长的实验. 依次从{16, 32, 64, 100, 200}范围内选择不同的值, 并固定其他超参数进行实验. 在表 3 和表 4 中, 展示了两个数据集的测试集在不同的裁剪滑动步长下对 *mIoU* 值的影响.

表 3 Potsdam 数据集上测试集的裁剪滑动步长分析

语义特征的大小	裁切步长	Potsdam *mIoU* (%)
512×12	200	77.1
	100	77.4
	64	77.6
	32	77.4
	16	**77.8**
1024×12	200	77.2
	100	77.4
	64	77.4
	32	77.5
	16	77.5

如表 3 所示, 在 Potsdam 数据集上, 当全局语义特征尺寸设置为 512×12 时, 随着测试集滑动步长从 200 降低到 16 时, *mIoU* 值出现持续增加的趋势. 而在全局语义特征尺寸为 1024×12 时, 随着裁剪滑动步长

逐渐降低, 其 *mIoU* 值也逐渐提升至 77.5%, 略低于 77.8%. 因此, 本文中全局语义特征尺寸被设置为 512×12, 同时测试集的裁剪滑动步长设置为 16. 类似地, 在 Vaihingen 数据集上的实验结果如表 4 所示, 当全局语义特征尺寸设置为 1024×12, 测试集上的裁剪滑动步长设置为 64 时, *mIoU* 值相比表现得更优异.

表 4 Vaihingen 数据集上测试集裁剪滑动步长的分析

语义特征的大小	裁切步长	Vaihingen *mIoU* (%)
512×12	200	79.9
	100	80.0
	64	80.1
	32	80.2
	16	80.3
1024×12	200	80.2
	100	80.3
	64	**80.7**
	32	80.4
	16	80.5

3.5 对比实验

为了全方面评估本文方法的有效性, 与多个先进分割模型进行对比实验, 其中基于卷积神经网络的方法有 FCN-8s、UNet、PSPNet[16]、DeepLabv3+[17]、CGFDN[18] 、CTFNet[19] 、HSDN[20]. 另外一方面, 基于 Transformer 改进的方法包括 Swin-Unet[21]、TransUNet[22]、DC-Swin[23]和 UNetFormer[24]. 所有模型均在 Potsdam 和 Vaihingen 数据集上进行训练和测试.

表 5 和表 6 中的数据展示了不同模型在 Potsdam 数据集和 Vaihingen 数据集上的分割性能结果, 其中加粗数值为每个地物类别中最高的分割精度. 此外, 本文还进行了分割结果的可视化对比, 如图 5 和图 6 所示, 不同地物类型使用不同颜色标注. 通过对比不同模型的分割结果, 可以明显地看出本文方法在分割准确度上的优势.

表 5 Potsdam 数据集上不同模型的性能比较 (%)

方法	不透水表面	建筑	低植被	树木	汽车	*OA*	mean *F*1	*mIoU*
FCN-8s	91.6	96.2	84.7	86.3	94.4	89.6	90.6	74.7
U-Net	92.0	96.0	84.8	86.2	95.4	89.6	90.9	75.3
PSPNet	91.9	96.6	85.4	85.8	95.1	89.9	91.0	75.4
DeepLabv3+	92.2	96.5	85.3	87.3	95.6	90.3	91.4	76.8
CGFDN	92.6	97.0	86.8	87.6	94.9	90.8	91.8	76.3
CTFNet	91.9	96.8	86.1	87.2	93.5	89.9	91.4	76.2
HSDN	92.9	97.1	86.9	87.8	95.2	90.9	92.1	76.9
TransUNet	92.0	96.6	85.5	85.6	94.7	90.0	91.1	74.9
Swin-Unet	91.9	96.8	85.6	86.1	93.5	89.9	90.8	74.2
DC-Swin	91.9	96.5	84.7	85.6	93.6	89.6	90.5	73.8
UNetFormer	92.5	96.8	86.1	87.1	95.1	90.5	91.5	76.1
Ours	**93.3**	**97.3**	**87.2**	**88.4**	**96.1**	**91.2**	**92.5**	**77.8**

表 6 Vaihingen 数据集上不同模型的性能比较 (%)

方法	不透水表面	建筑	低植被	树木	汽车	*OA*	mean *F*1	*mIoU*
FCN-8s	90.7	94.3	80.5	88.0	79.7	89.6	86.6	77.0
U-Net	90.7	94.5	80.8	88.2	82.9	89.9	87.4	78.3
PSPNet	91.2	94.8	81.5	88.5	77.2	90.2	86.6	77.0
DeepLabv3+	91.4	95.1	81.2	88.3	83.9	90.3	88.0	79.0
CGFDN	91.8	94.7	83.1	**89.7**	82.9	91.0	88.4	79.4
CTFNet	90.7	94.4	81.7	87.1	82.7	88.6	87.3	77.8
HSDN	92.0	95.1	**83.9**	88.9	83.5	89.9	88.9	79.8
TransUNet	90.4	93.2	82.7	88.5	72.1	89.6	85.4	75.6
Swin-Unet	91.3	94.3	82.7	89.0	80.2	90.4	87.5	78.5
DC-Swin	91.3	94.9	82.6	88.8	74.9	90.5	86.5	77.2
UNetFormer	91.8	95.6	83.0	89.3	81.9	91.1	88.3	79.6
Ours	**92.2**	**95.8**	83.6	89.4	**84.7**	**91.4**	**89.1**	**80.7**

通过对比实验和分析, 得出以下结论: 首先, 本文选择 FCN-8s 作为基准模型, 并通过 *OA*、mean *F*1 和 *mIoU* 等指标进行定量分析. 在 2 个数据集上, 基于卷积神经网络的方法和基于 Transformer 的方法都表现出了良好的分割性能. 然而, 缺乏全局一致性和物体尺度变换多样导致的分割错误. 相比之下, 本文提出的方

法通过全局语义关系学习模块和多尺度关系学习模块, 有效地克服了这些限制. 特别是在 *OA*、mean *F*1 和 *mIoU* 等关键指标上, 本文方法均取得了最佳表现. 在 Potsdam 数据集上, 本文方法的 *OA*、mean *F*1 和 *mIoU* 分别为 91.2%、92.5% 和 77.8%, 相比最佳的卷积神经网络方法 HSDN 分别提升了 0.3%、0.4% 和 0.9%. 在 Vaihingen 数据集上, 同样实现了最佳性能.

从可视化结果中可以看到, 本文方法在处理具有显著尺度变化的场景时表现更为出色. 与其他方法相比, 本文方法能够更准确地分割出建筑、汽车、树木和低植被等地物类型. 特别是在处理小目标和背景类分割时, 本文方法展现出了更高的准确性和鲁棒性. 综上所述, 通过对比实验和分析, 验证了本文方法的有效性和优越性. 本文提出的全局语义和多尺度关系学习模块显著地提升遥感图像分割的性能, 为该领域的研究和应用提供了有价值的参考.

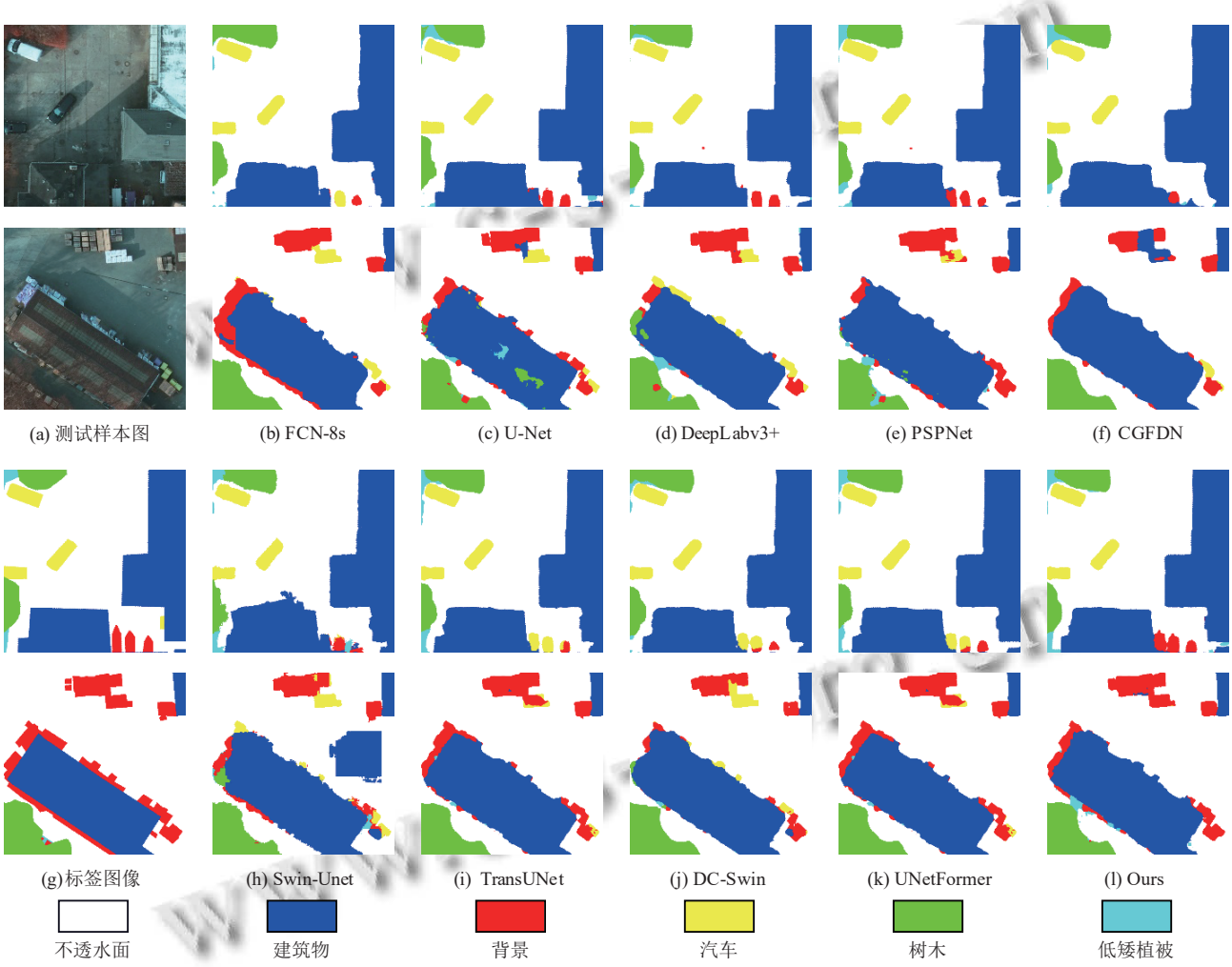

图 5 Potsdam 数据集上不同模型的实验结果对比图

3.6 消融实验

本文提出的方法包括全局语义关系学习模块和多尺度关系学习模块. 前者致力于从特征中提取并学习全局语义信息及其之间的关系, 以帮助模型理解类别的多尺度关系, 并将学习到的类别间共生关系和多尺度关系融入特征中, 从而提高分割精度. 为了评估各模块对整体网络分割精度的影响, 进行模块消融实验. 本实验中将主干网络结构作为 Baseline. GRLM [1, 2, ⋯, *i*] 表示在获取到 E^i 特征后嵌入全局语义关系学习模块.

表 7 和表 8 展示了在 Potsdam 和 Vaihingen 数据集上的消融实验结果.

结果表明与 Baseline 相比, 通过加入全局语义关系学习模块和多尺度关系学习模块, 两个数据集上的 *mIoU* 值分别提升了约 1.0% 和 1.6%.

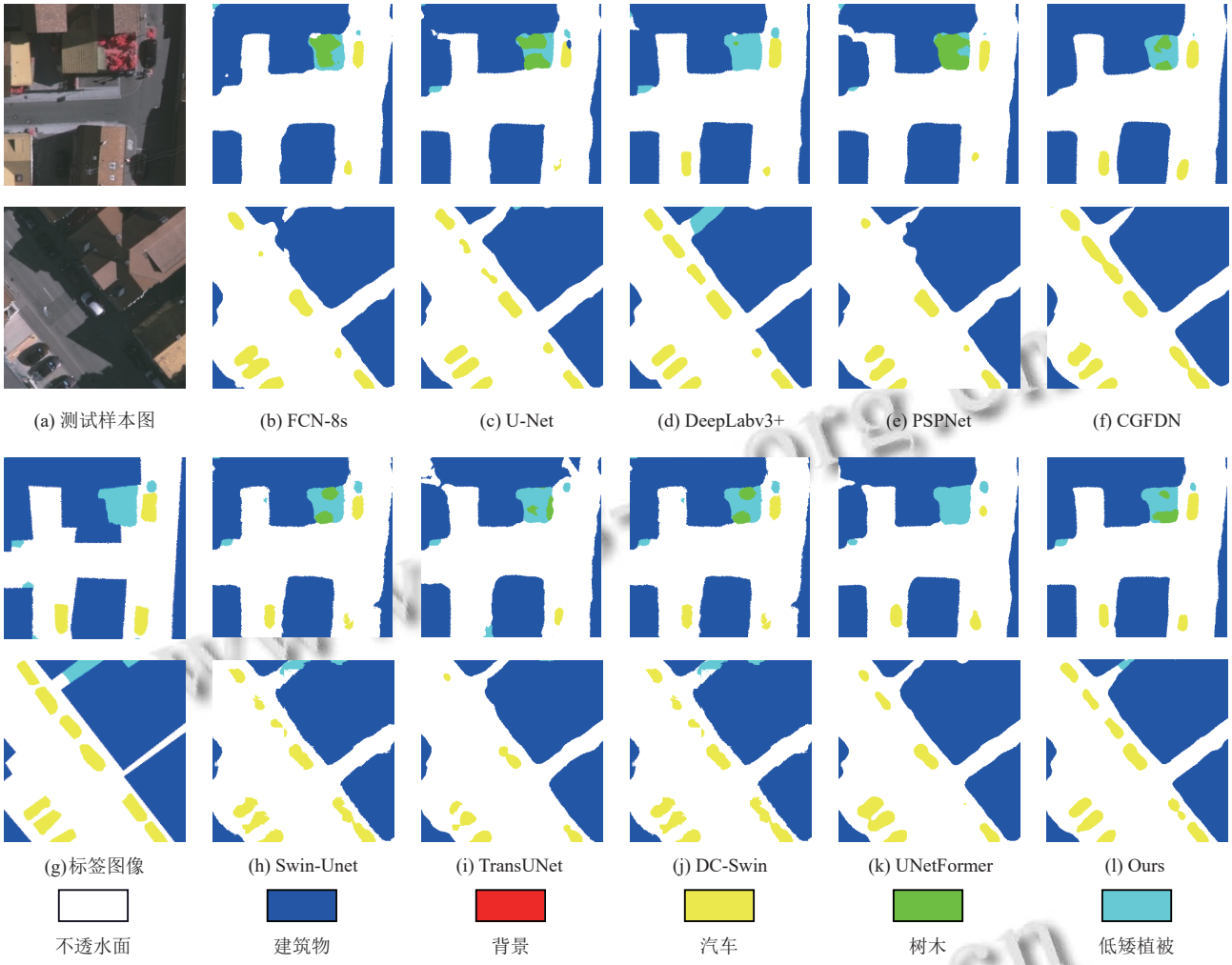

图 6 Vaihingen 数据集上不同模型的实验结果对比图

表 7 Potsdam 数据集上的消融实验

模型	*mIoU* (%)
Baseline	75.2
Baseline + GRLM [1]	75.0
Baseline + GRLM [1,2]	75.5
Baseline + GRLM [1,2,3]	75.3
Baseline + GRLM [1,2,3,4]	74.9
Baseline + GRLM [1,2,3,4,5]	76.2
Baseline + GRLM [1,2,3,4,5] + MRLM	**77.8**

表 8 Vaihingen 数据集上的消融实验

模型	*mIoU* (%)
Baseline	77.8
Baseline + GRLM [1]	78.7
Baseline + GRLM [1,2]	79.1
Baseline + GRLM [1,2,3]	78.8
Baseline + GRLM [1,2,3,4]	78.7
Baseline + GRLM [1,2,3,4,5]	79.4
Baseline + GRLM [1,2,3,4,5] + MRLM	**80.7**

4 结论与展望

通过在遥感图像语义分割任务中, 现有的方法在特征提取中通常面临着缺乏全局一致性和尺度变化多样的挑战, 导致分割结果不准确. 本文提出了一种语义引导与多尺度增强的遥感影像分割方法, 有效应对了这些挑战. 通过主干网络提取多尺度特征, 并嵌入全局语义关系学习模块, 本文方法能够有效捕获类别间的共生关系, 提高了全局一致性; 多尺度关系学习模块进一步增强了模型在处理多尺度变化时的能力, 并将学习到的类别间的共生关系和多尺度关系融入特征中, 显著提升了分割性能. 通过在 Potsdam 和 Vaihingen 两个数据集上的对比实验与消融实验, 验证了本文方法的有效性.

参考文献

1 李云梅, 赵焕, 毕顺, 等. 基于水体光学分类的二类水体水

环境参数遥感监测进展. 遥感学报, 2022, 26(1): 19–31. [doi: 10.11834/jrs.20221212]
2 Hang RL, Yang P, Zhou F, *et al*. Multiscale progressive segmentation network for high-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5412012. [doi: 10.1109/TGRS.2022.3207551]
3 龙丽红, 朱宇霆, 闫敬文, 等. 新型语义分割 D-UNet 的建筑物提取. 遥感学报, 2023, 27(11): 2593–2602.
4 于波, 孟俊敏, 张晰, 等. 结合凝聚层次聚类的极化 SAR 海冰分割. 遥感学报, 2013, 17(4): 887–904.
5 Hu XY, Shen JJ, Shan J, *et al*. Local edge distributions for detection of salient structure textures and objects. IEEE Geoscience and Remote Sensing Letters, 2013, 10(3): 466–470. [doi: 10.1109/LGRS.2012.2210188]
6 吴一全, 吉玚, 沈毅, 等. Tsallis 熵和改进 CV 模型的海面溢油 SAR 图像分割. 遥感学报, 2012, 16(4): 678–690. [doi: 10.11834/jrs.20121192]
7 Zheng Z, Zhong YF, Wang JJ, *et al*. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4096–4105. [doi: 10.1109/CVPR42600.2020.00415]
8 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]
9 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]
10 Zhou ZW, Siddiquee MR, Tajbakhsh N, *et al*. UNet++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis. Granada: Springer, 2018. 3–11. [doi: 10.1007/978-3-030-00889-5_1]
11 Huang HM, Lin LF, Tong RF, *et al*. UNet 3+: A full-scale connected UNet for medical image segmentation. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 1055–1059. [doi: 10.1109/icassp40776.2020.9053405]
12 Jin QG, Meng ZP, Sun CM, *et al*. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. Frontiers in Bioengineering and Biotechnology, 2020, 8: 605132. [doi: 10.3389/fbioe.2020.605132]
13 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]
14 张寅, 朱桂熠, 施天俊, 等. 基于特征融合与注意力的遥感图像小目标检测. 光学学报, 2022, 42(24): 2415001.
15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
16 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2881–2890.
17 Chen GZ, Zhang XD, Wang Q, *et al*. Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(5): 1633–1644. [doi: 10.1109/JSTARS.2018.2810320]
18 Zhou F, Hang RL, Liu QS. Class-guided feature decoupling network for airborne image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(3): 2245–2255. [doi: 10.1109/tgrs.2020.3006872]
19 Wu HL, Huang P, Zhang M, *et al*. CTFNet: CNN-Transformer fusion network for remote-sensing image semantic segmentation. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 5000305. [doi: 10.1109/LGRS.2023.3336061]
20 Zheng CY, Nie J, Wang ZX, *et al*. High-order semantic decoupling network for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5401415. [doi: 10.1109/TGRS.2023.3249230]
21 Cao H, Wang YY, Chen J, *et al*. Swin-Unet: Unet-like pure Transformer for medical image segmentation. Proceedings of the 2022 European Conference on Computer Vision. Tel Aviv: Springer, 2022. 205–218. [doi: 10.1007/978-3-031-25066-8_9]
22 Chen JN, Lu YY, Yu QH, *et al*. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306, 2021.
23 Wang LB, Li R, Duan CX, *et al*. A novel Transformer based semantic segmentation scheme for fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6506105. [doi: 10.1109/lgrs.2022.3143368]
24 Wang LB, Li R, Zhang C, *et al*. UNetFormer: A UNet-like Transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196–214. [doi: 10.1016/j.isprsjprs.2022.06.008]

(校对责编: 张重毅)