

王威,熊艺舟,王新. NHNet:新型层次化遥感图像语义分割网络. 吉林大学学报(地球科学版),2024,54(5):1764-1772. doi:10.13278/j.cnki.jjuese.20230155.

Wang Wei, Xiong Yizhou, Wang Xin. NHNet: A Novel Hierarchical Semantic Segmentation Network for Remote Sensing Images. Journal of Jilin University (Earth Science Edition), 2024, 54 (5): 1764 - 1772. doi: 10.13278/j.cnki.jjuese.20230155.

NHNet——新型层次化遥感图像语义分割网络

王 威,熊艺舟,王 新

长沙理工大学计算机与通信工程学院,长沙 410000

摘要:深度学习分割方法是遥感图像分割领域的热点之一,主流的深度学习方法有卷积神经网络、transformer 神经网络及两者的结合。特征提取是图像分割的重要环节,除了用卷积等方式提取特征,最近的研究聚焦于一些新的特征提取范式,如图卷积、小波变换等。本文利用聚类算法的区域构建属性,将改进的聚类算法用于骨干特征提取模块,同时使用卷积和视觉 transformer 作为辅助模块,以获取更丰富的特征表述;在模块基础上,提出了一种新型层次化遥感图像语义分割网络(NHNet);评估了 NHNet 语义分割的性能,并在 LoveDA 遥感数据集上与其他方法进行比较。结果表明,基于多特征提取的 NHNet 获得了竞争性的性能表现,平均交并比为 49.64%, F_1 分数为 65.7%。同时,消融实验证明辅助模块提高了聚类算法分割的精确性,给 NHNet 分别提升了 1.03%和 2.41%的平均交并比。

关键词:遥感图像;语义分割;聚类算法;卷积神经网络;自注意力

doi:10.13278/j.cnki.jjuese.20230155

中图分类号:TP751.1

文献标志码:A

NHNet: A Novel Hierarchical Semantic Segmentation Network for Remote Sensing Images

Wang Wei, Xiong Yizhou, Wang Xin

School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410000, China

Abstract: Deep learning segmentation method is one of the hot topics in the field of remote sensing image segmentation. The mainstream deep learning methods include convolutional neural networks, transformer neural networks, and a combination of the two. Feature extraction is an important part of image segmentation. In addition to using convolution and other methods to extract features, recent research has focused on some new feature extraction paradigms, such as graph convolution and wavelet

收稿日期:2023-06-26

作者简介:王威(1974—),男,教授,博士生导师,CCF 会员,主要从事计算机视觉、模式识别方面的研究,E-mail: wangwei@csust.edu.cn

通信作者:王新(1976—),女,讲师,主要从事计算机视觉、模式识别方面的研究,E-mail: wangxin@csust.edu.cn

基金项目:湖南省重点研究开发项目(2020SK2134);湖南省自然科学基金项目(2022JJ30625)

Supported by the Key Research and Development Project of Hunan Province (2020SK2134) and the Project of Natural Science Foundation of Hunan Province (2022JJ30625)

transform. In this article, the region construction attribute of clustering algorithms is utilized, and the improved clustering algorithm is used as the backbone feature extraction module while the convolution and visual transformer are used as auxiliary modules to obtain richer feature representations. On the basis of the module, a new hierarchical remote sensing image semantic segmentation network (NHNet) is proposed. The performance of NHNet semantic segmentation is evaluated and compared with other methods on the LoveDA remote sensing dataset. The results show that NHNet based on multi-feature extraction achieved competitive performance, with an average intersection-to-union ratio of 49.64% and a score of 65.7%. At the same time, ablation experiments show that the auxiliary module improves the accuracy of clustering algorithm segmentation, increasing the average intersection-to-union ratio of NHNet by 1.03% and 2.41%, respectively.

Key words: remote sensing images; semantic segmentation; clustering algorithm; convolutional neural network; self attention

0 引言

遥感技术已经广泛应用于农业、林业、水资源管理、城市规划、环境保护、天气预报、国土资源调查等领域。如何利用日益丰富且完善的遥感数据集对地球进行智能观测,是遥感领域的巨大机会和挑战,如利用卫星数据反演海底地形^[1]和进行变化检测^[2]等。其中遥感语义分割^[3-5]利用遥感技术获取的图像,通过计算机视觉技术和深度学习等算法对图像中的目标进行分类,将图像中的每个像素点标记为相应的类别,实现对图像的语义理解和识别。这要求网络必须能够捕获遥感场景的整体视图,同时保留细节和语义信息。语义分割技术可以快速准确地提取出图像中的各种地物信息,如建筑、道路、河流、森林等,从而对地表特征进行精细化分析和研究,也可以辅助农业、林业、水资源管理等领域做出决策,对于遥感数据集的利用有重大意义。

传统的语义分割方法通常采用基于区域^[6]的方法,例如使用区域生长算法或聚类算法将像素分成不同的区域。这些方法在一定程度上可以实现图像中物体的分割,但是很难准确地标记物体边界,对复杂场景的处理效果并不理想。近年来,深度学习方法已经成为语义分割领域中的主流方法。通过使用卷积神经网络(convolutional neural network, CNN)等深度学习模型,可以有效提取图像特征,并进行精确的像素级别分割。其中,全卷积网络^[7](fully convolutional networks, FCN)、编码器-解码器网络(encoder-decoder)^[8-10]、空洞卷积网络^[11](dilated convNet)和分割网络^[12](segmentation network, SegNet)等模型被广泛应用于语义分割任务中。

最近,ViT(vision transformers)^[13]将基于纯自注意力的 transformer 引入了视觉领域,并在各种视觉任务上取得了优异的性能,比如 SegFormer^[14]。同时,为了使 ViT 更加适用于密集视觉任务,SwiT(Swin transformer)^[15]采用了类似于卷积的滑动窗口注意力,并且在网络中加入了层次结构。NAT(neighborhood attention transformer)^[16]重新讨论了滑动窗口注意力,提出了邻域注意力,将像素的自注意力定位到不固定的最近邻,具有平移等方差的额外优势。DiNAT(dilated neighborhood attention transformer)^[17]在邻域注意力的基础上加入了膨胀邻域注意力,扩大了 NAT 的感受野。但是,ViT 需要强大的计算资源,对移动设备不友好,这引发了轻量级语义分割网络的研究。研究通过一系列的参数压缩、技术优化实现了轻量化和高效率的特点^[18-23],如 TopFormer^[24]、SeaFormer^[25]采用卷积神经网络和 ViT 结合的方式发挥他们各自的优势,在计算成本和性能表现之间获得平衡。

为了避免增量改进陷阱,一些研究将目光聚焦于新的特征提取范式,如 ViG(vision graph neural network)^[26]。ViG 为了捕捉不规则和复杂的物体,将图像划分为一些像素块并视为图节点,利用图卷积和 FFN(feed-forward network)模块提取图形特征,首次将图神经网络用于大规模的视觉任务。上下文聚类网络^[27](context clusters, CoCs)将图像视为一组无组织的点,通过聚类算法提取特征,尽管没有使用任何卷积和注意力,也在基准测试上取得不弱于基于卷积和注意力算法的性能。

上述研究表明,随着更多的特征提取方式被引入到视觉领域,传统的卷积网络和自注意力网络不

再是构建神经网络的唯一选择。本文利用聚类算法强大的区域构建能力,借鉴上述研究中卷积和自注意力的有效结合方式,提出一种新型层次化遥感图像语义分割模型(NHNet)。首先利用聚类算法的区域构建属性,将改进聚类块作为网络的基本特征提取模块,搭建由聚类块组成的骨干网络,使神经网络的感受野能够灵活变动,适应遥感图像固有的对象物体不规则的特点;然后提出层次化遥感图像分割架构(NHNet),在浅层使用基于卷积的C模块(C block, CB)增强其空间细节特征,在深层使用基于多头注意力的T模块(T block, TB)提取全局语义信息;最后设计特征注入模块(feature injection module, FIM),以消除三种特征提取方法的语义差距,实现更加平滑的特征融合。并且在LoveDA^[28]数据集上与其他方法进行比较,以期为遥感图像语义分割提供解决分割问题的新视角。

1 网络结构

1.1 总体架构

模型的总体架构如图1所示。为了获得区域构建能力,基于区域对图像进行分割,本文将聚类块作为基本的特征提取模块。图像经过4个聚类块处理之后,为了获得空间细节信息和全局语义信息,分别使用基于卷积和ViT的特征提取模块对原特征进行特征提取,并将提取的特征通过特征注入模块进行融合注入。由于不同层次之间的特征存在差异,所以对浅层和深层特征采用不同的处理方法。浅层网络输出的特征含有丰富的空间细节和纹理信息,使用基于卷积的3个倒置瓶颈模块组成C模块捕

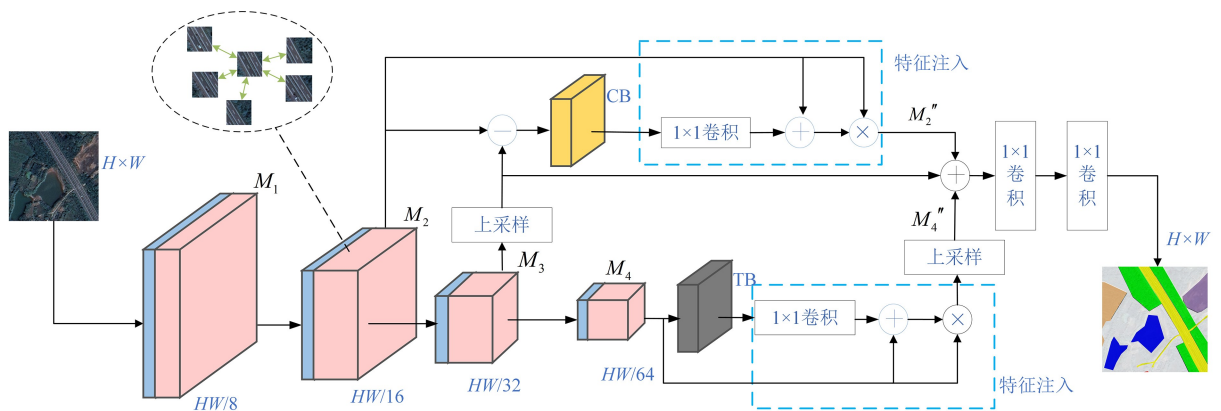
获更精确的分割细节;深层网络输出的特征含有丰富的全局语义信息,使用基于 transformer 的3个ViT模块组成T模块捕获语义特征信息。为了使特征更加平滑地融合,使用特征注入模块,将特征提取模块中的细节和全局特征与相对应的原尺度特征进行融合,以增强表示。

1.2 基本特征提取模块

为了使网络具有基于区域进行分割的能力,网络的基本特征提取模块为聚类块。一个聚类块包含两个子模块,其中:第一个子模块为区域划分模块,利用特征相似矩阵实现特征图的区域划分;第二个子模块为特征聚合与更新模块,在划分好的区域内进行像素点特征的更新计算。整个网络共分为四个阶段,每个阶段都包含一个下采样和 N 个聚类块,下采样模块用来降低特征图的分辨率,增加特征图的维度。如图1所示,四个阶段的通道维度为 $\{32, 64, 196, 320\}$,输出特征图定义为 $\{M_1, M_2, M_3, M_4\}$ 。

1)区域划分。给定一组像素点 $P \in \mathbb{R}^{n \times d}$,其中, n 为像素总数, d 为像素的特征维度。首先进行相似性运算,在空间均匀地提出 c 个中心点,通过平均其周围 k 个相邻点计算中心特征;然后计算像素点和中心点集的像素值,得到余弦相似度 $s \in \mathbb{R}^{c \times n}$,将每个像素点分配到最相似的中心,得到 c 个不重叠区域。

2)特征聚合与更新。划分区域之后,为了进一步计算区域中每个像素点的特征值,需要在每个区域中进行特征聚合。假定一个区域中有 m 个像素点,与中心点的相似度为 $s \in \mathbb{R}^m$,本文将这些像素



H 、 W 为输入图像的高和宽; M_1 、 M_2 、 M_3 、 M_4 、 M_2'' 、 M_4'' 为每个阶段的输出特征图。

图 1 网络结构图

Fig.1 Network structure diagram

点映射到一个值空间 $P_V \in \mathbb{R}^{m \times w}$, 其中 w 是维度数。聚合特征 g 的公式表达如下:

$$g = \frac{V_c}{C} + \frac{1}{C} \sum_{i=0}^m \text{sig}(\alpha s_i + \beta) V_i; \quad (1)$$

$$C = 1 + \sum_{i=0}^m \text{sig}(\alpha s_i + \beta). \quad (2)$$

式中: V_c 为值空间的中心点; V_i 为值空间 P_V 中的第 i 个点; $\text{sig}(\cdot)$ 表示激活函数 sigmoid; α 和 β 为可学习的因子; s_i 为相似度矩阵中的第 i 个元素; 为了控制大小, 聚合特征归一化为 C 。

之后根据像素点与中心点的相似性, 自适应地将 g 分配到区域中的每个像素点。如此, 这些点可以相互通信, 并共享来自区域中所有点的特征。利用 g 对区域内每个点 P_i 的更新方式表示为

$$P_i' = P_i + \text{FC}(\text{sig}(\alpha s_i + \beta))g. \quad (3)$$

式中: P_i' 为进行特征更新之后的第 i 个像素点; $\text{FC}(\cdot)$ 为全连接层。

3) 根据式(3)计算图像的所有像素点, 得到聚类块的计算方式。在网络结构(图1)中, 将一个下采样和重复多次的聚类块作为一个阶段, 不同阶段的聚类块重复次数决定了网络的学习能力和规模, 将每个阶段的聚类块数量设置为 $\{2, 2, 6, 2\}$ 。

聚类块将图像视为一系列无序的点, 每个点都包括原始特征和位置信息, 通过聚类算法对深度特征进行分层分组和特征提取。这种方式为图像和视觉表示提供了一个新的视角, 聚类算法通过特征相似性将所有像素点归纳为几个簇, 在簇内进行特征聚合和特征更新。由于一个簇中像素点的数量是不固定的, 有可能一个簇中的像素点总数为零, 所以一个像素点会在一个不规则不固定的感受野中与其他像素点产生交互, 这种特性更契合图像分割中的感受野要求。

1.3 特征增强模块

特征增强模块如图2所示。

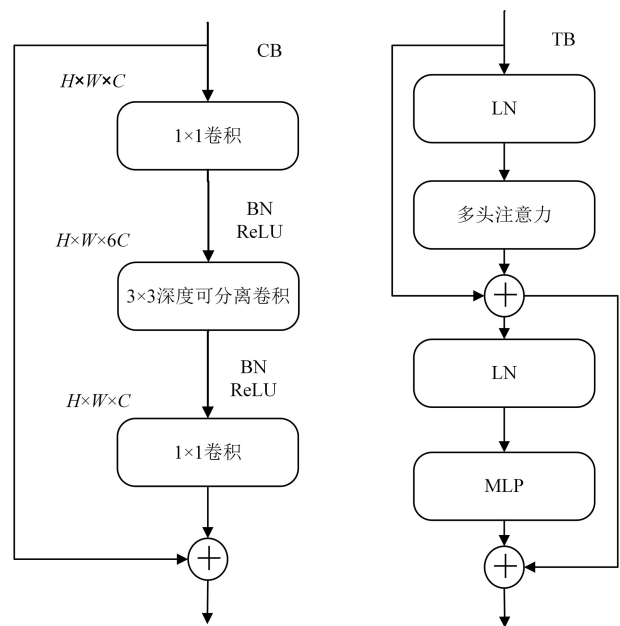
1) T 模块。在骨干网络生成的特征图中, 第4阶段生成的特征图 M_4 具有较大的感受野, 含有丰富的语义特征, 因此本文使用3个T模块对关键语义信息进行过滤提取。T模块由两个残差层组成, 分别为多头自注意力层和线性化层, 其表达如下:

$$M_4 = M_4 + \text{MulAtt}(\text{Norm}(M_4)); \quad (4)$$

$$M_4' = M_4 + \text{Mlp}(\text{Norm}(M_4)). \quad (5)$$

式中: $\text{Norm}(\cdot)$ 表示归一化层; $\text{MulAtt}(\cdot)$ 表示多头注意力层; $\text{Mlp}(\cdot)$ 表示线性化层, 由一个深度可分离的 3×3 卷积和两个 1×1 卷积组成, 进行通道

间的信息交流; M_4' 为最终的输出。



BN, batch normalization; ReLU, rectified linear unit; LN, layer normalization; MLP, multilayer perceptron.

图2 特征增强模块结构图

Fig.2 Structure diagram of feature enhancement module

2) C 模块。 M_2 具有高分辨率特征, 没有 M_3 、 M_4 经过下采样而导致的空间细节信息损失, 对得到更精确的分割结果而言至关重要, 比如更精确的房屋边缘信息、更连续的长距离河流分割信息; 但是对高分辨率特征的过度学习会导致计算成本的上升。因此, 本文采用3个基于深度可分离卷积的C模块完成高分辨率特征的提取。浅层特征图会包含一些噪音和冗余信息, 所以使用 $\text{CR}(\cdot)$ 模块对 M_2 进行预处理, 加强特征多样性, 使C模块学习与 M_3 、 M_4 互补的空间细节信息。整个过程表达如下:

$$\text{CR}(M_2) = \text{Conv}_{1 \times 1}(M_2) - \text{UP}(\text{Conv}_{1 \times 1}(M_3)); \quad (6)$$

$$M_2' = \text{Conv}_{1 \times 1}(\text{DW}_{3 \times 3}(\text{Conv}_{1 \times 1}(\text{CR}(M_2)))). \quad (7)$$

式中: $\text{Conv}_{1 \times 1}(\cdot)$ 为卷积核为1的普通卷积; $\text{UP}(\cdot)$ 为双线性插值上采样; $\text{DW}_{3 \times 3}(\cdot)$ 为卷积核为3的深度可分离卷积; M_2' 为最终的输出。每个卷积都固定一个BN层和ReLU激活层, 没有在公式中写出。 M_2 首先进入预处理模块 $\text{CR}(\cdot)$, 减去 M_3 上采样之后的特征图, 得到骨干网深层损失的特征信息; 然后用 1×1 普通卷积对其进行升维, 用深

度可分离的 3×3 卷积进行特征运算,用 1×1 卷积降低维度。升维幅度设置为 6,C 模块的重复次数设置为 3。

1.4 特征注入模块

由于特征提取方式不同,本文设计了特征注入模块来融合这些特征。 M_4 首先经过 T 模块得到全局语义信息特征;然后分别经过一个 1×1 卷积层、归一化层和激活层,得到消除差异的特征信息 M_4' ;之后将这些特征信息的分辨率和通道数与 M_4 统一,通过一个矩阵加法和矩阵乘法把这些特征信息注入到 M_4 ,最终得到了获得语义增强的特征图 M_4'' 。 M_2 的注入方式与 M_4 基本一致,不同点在于,因为 C 模块中利用了 M_3 ,通道数被上升到了 M_3 的大小,所以 M_2 的通道数首先也要被上升到 M_3 的大小,然后进行特征注入。方法的公式表达如下:

$$M_2' = \text{Conv}_{1 \times 1}(\text{CB}(M_2)); \quad (8)$$

$$M_2'' = M_2' \cdot \text{UP}(M_2) + \text{UP}(M_2); \quad (9)$$

$$M_4' = \text{Conv}_{1 \times 1}(\text{TB}(M_4)); \quad (10)$$

$$M_4'' = M_4' \cdot M_4 + M_4. \quad (11)$$

式中: $\text{CB}(\cdot)$ 为 C 模块; $\text{TB}(\cdot)$ 为 T 模块。

融合的特征具有丰富的空间细节信息和全局语义信息,这是遥感图像分割的性能基础,在此基础上添加分割头。分割头由两个 1×1 卷积层组成,每个卷积层后固定一个 BN 层和 ReLU 激活层,通过平滑的通道维度下降输出结果。

2 实验结果与分析

在本节中,首先在 LoveDA 公共数据集上进行实验,并将结果与其他方法进行比较。然后进行消融研究,以分析特征增强模块和特征注入模块的有效性。

2.1 数据集

为了验证 NHNet 的可行性,本文在数据集 LoveDA 上进行了实验。LoveDA 数据集包含来自三个不同城市的 5 987 张 0.3 m 高分辨率影像和 166 768 个标注语义对象^[20]。与现有的其他遥感数据集相比,LoveDA 同时包含了城市和农村地区的不同影像,这增加了数据集的复杂性和处理的困难程度。城乡景观有不同的分布:人口密度高的城市景观中包含了大量的人工物体,如建筑和道路;相比之下,乡村景观包含了更多的自然元素,如林地和水。这种类间不一致的分布造成了一定问题。

LoveDA 数据集中主要标注了 7 个类别:背景、建筑、道路、水域、荒地、林地和耕地,每个图像经过几何配准和图像预处理,变成 $1\,024 \times 1\,024$ 分辨率的影像,总图像数量为 5 987 张,划分为训练集 2 522 张、验证集 1 669 张、测试集 1 796 张。

2.2 参数设置

实验框架为 MMSegmentation 和 Pytorch。具体实验时,为了保持一致性,所有的实验设置都是统一的:使用 AdamW 优化器训练 4 万次迭代,批量大小为 5,初始学习率为 0.000 1,权重衰减为 0.001,poly 学习策略因子为 0.9。数据集上,采用相同的图像增强策略进行公平比较:将训练图像分辨率调整为 $1\,024 \times 1\,024$,然后进行 50% 概率的随机翻转和随机裁剪;将测试图像分辨率调整为 $1\,024 \times 1\,024$,然后进行概率为 50% 随机翻转。在实验进行时,每 4 000 次迭代评估一次网络的平均交并比和最高精度,到第 4 万次迭代时有十次评估,取最大值。所有实验配备一张内存为 20 GB 的 NVIDIA A10 显卡,软件设备如表 1 所示。

表 1 实验环境

Table 1 Experimental environment

配置	版本
操作系统	Ubuntu 20.04
Python	3.8.16
Pytorch	1.8.0
CUDA	11.1.1
MMSegmentation	1.0.0
MMEngine	0.7.3

2.3 评估指标

实验使用平均交并比、整体像素精度和 F_1 分数作为评估指标。假设一个数据集的总类别为 u ,上述三种指标的公式表达如下:

$$M = \frac{1}{u} \sum_{i=1}^u \frac{p_{ii}}{\sum_{j=1}^u p_{ij} + \sum_{j=1}^u p_{ji} - p_{ii}}; \quad (12)$$

$$F_1 = 2 \frac{PR}{P + R}; \quad (13)$$

$$P = \sum_{i=1}^u \frac{p_{ii}}{p_{ii} + \sum_{j=1}^u p_{ji}}; \quad (14)$$

$$R = \sum_{i=1}^u \frac{p_{ii}}{p_{ii} + \sum_{j=1}^u p_{ij}}; \tag{15}$$

$$O = \frac{\sum_{i=1}^u p_{ii}}{T}. \tag{16}$$

式中: M 为平均交并比; p_{ii} 为预测正确,将 i 类别预测为 i 类别的数目; p_{ij} 为预测错误,将 i 类别预测为 j 类别的数目; p_{ji} 为预测错误,将 j 类别预测为 i 类别的数目; P 为精确度; R 为召回率; O 为整体像素精度; T 为像素点总数。

2.4 实验结果

将 NHNet 在 LoveDA 数据集上与现有的高性能语义分割方法进行了对比实验,结果如表 2 所示。NHNet 在平均交并比和 F_1 分数两个指标上获得了最优(平均交并比达 49.64%, F_1 分数达 65.70%),基于轻量级神经网络 MobileNetV2 的 DeepLabV3+ 在指标整体像素精度上获得了最优(达 67.31%)。表 3 详细记录了每个类别的交并比,与其他分割方法相比,NHNet 对背景、建筑、水域、荒地这四个类别上的识别度最高,但对耕地的识别程度不高(比 DeepLabV3+ 低 5.44%)。

为了更直观地展示实验结果,本文将部分数据集进行了可视化,结果如图 3 所示。从图 3 中可以看出,真实标签中建筑物间的空地较大,NHNet 尽可能地识别出了这些在建筑物中的空地,基于 transformer 的 SegFormer 也有这种识别能力,但是基于卷积的 DeepLabV3+ 识别的建筑物基本上是连续的状态;同时发现 NHNet 对荒地的识别程度不高,从而导致了整体识别度的变低。

2.5 消融研究

本文通过消融实验进一步验证 NHNet 中 T 模

块、C 模块和特征注入模块的影响。所有的结果都在 LoveDA 测试集上进行验证和评估,如表 4 所示。从表 4 中可以看出,基线的平均交并比为 46.20%;当加上注重于浅层空间特征的 C 模块时,平均交并比达到了 47.23%,相比于基线提升了 1.03%;之后在网络中加上注重于高层语义信息的 T 模块,平均交并比达到了 49.64%,相比于 C 模块提升了 2.41%。这证明 C 模块、T 模块对聚类算法具有很好的辅助提升作用。但是这两个模块对于整个网络架构的影响不同,从表 4 中可以看出,T 模块对于网络效果的提升作用比 C 模块高 1.38%。本文认为造成这种现象的原因是基于相似度的聚类算法缺乏对语义信息的有效提取,而基于多头自注意力的 T 模块强化了高层特征图的全局语义信息,加强了网络对相似类别和困难类别的辨别能力。另一方面,聚类算法具有较强的空间细节信息捕捉能力,所以注重于空间信息特征的 C 模块对其的强化弱于 T 模块。

表 2 LoveDA 数据集上的指标评估
Table 2 Evaluation of metrics on LoveDA dataset

方法	编码器	平均交并 比/%	整体像素 精度/%	F_1 /%
BiSeNetV2 ^[29]	BiSeNetV2	40.30	60.60	56.94
STDC ^[30]	STDCNet	45.19	65.24	61.12
SegFormer ^[14]	Mit - B0	47.37	65.60	63.42
DeepLabV3+ ^[10]	MobileNetV2	49.61	67.31	65.69
SETR - MFPD ^[5]	ViT - Base	33.64	57.07	48.91
NHNet	NHNet	49.64	67.25	65.70

注:BiSeNet, bilateral segmentation network; STDC, short-term dense concatenate; SETR, segmentation transformers; MFPD, multiscale feature pyramid decoder。

表 3 LoveDA 数据集上的实验结果
Table 3 Experimental results on LoveDA dataset

方法	编码器	交并比/%						
		背景	建筑	道路	水域	荒地	林地	耕地
BiSeNetV2 ^[29]	BiSeNetV2	46.89	45.73	43.89	42.57	21.83	38.89	42.29
STDC ^[30]	STDCNet	50.13	53.81	49.60	57.61	17.58	41.52	46.05
SegFormer ^[14]	Mit - B0	50.79	60.28	52.16	60.08	25.56	39.86	42.89
DeepLabV3+ ^[10]	MobileNetV2	51.09	59.42	52.59	63.47	32.24	38.85	49.61
SETR - MFPD ^[5]	ViT - Base	46.22	35.46	25.86	51.11	9.74	29.95	37.16
NHNet	NHNet	52.54	61.92	50.07	64.28	32.97	41.50	44.17

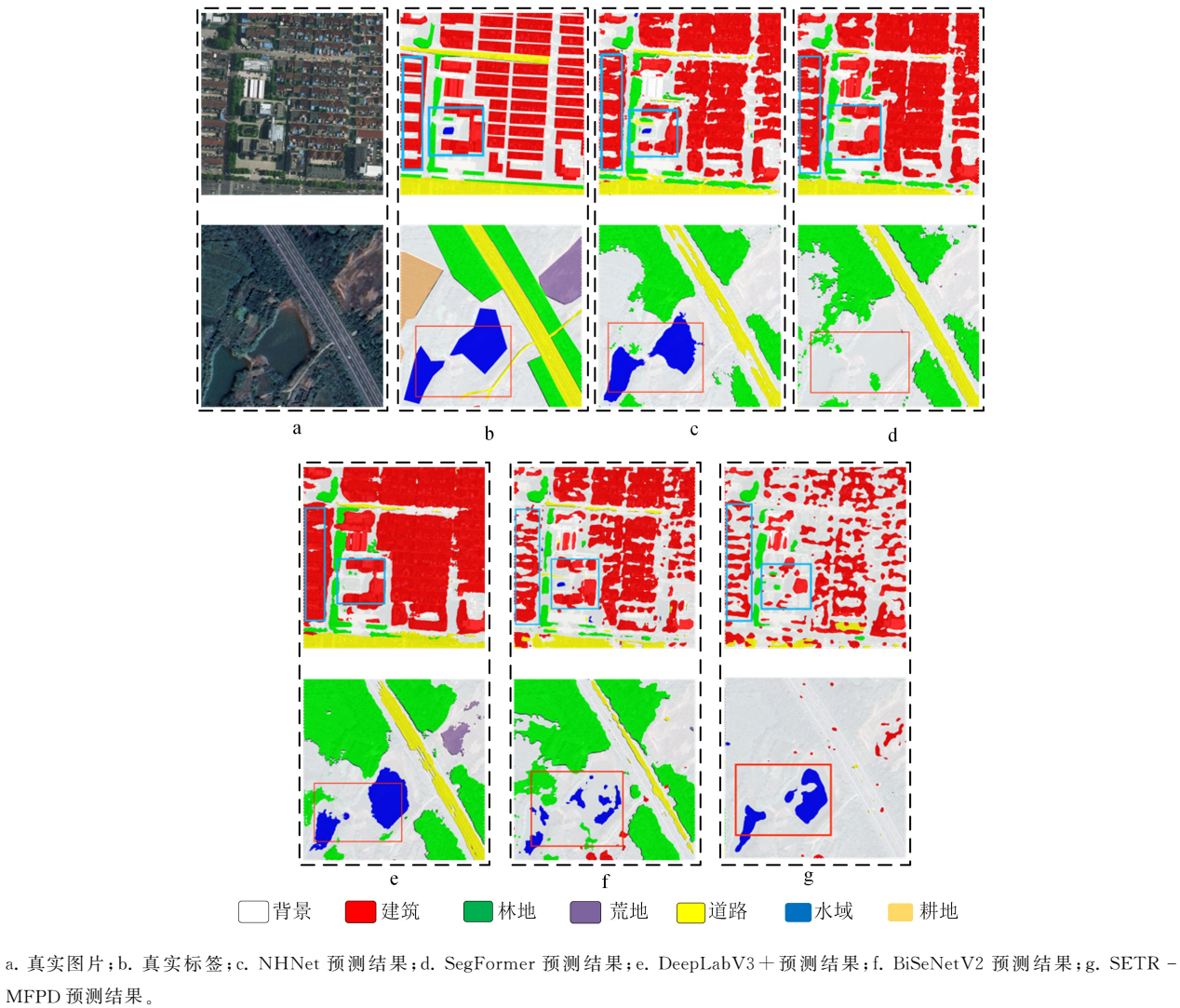


图 3 部分数据集可视化

Fig.3 Partial dataset visualization

表 5 给出了特征注入模块的消融实验结果,表中 CFIM 代表使用了特征注入,但是注入方式与 FIM 不同,CFIM 也放弃了原有的特征,但是将处理之后的特征注入到了 M_3 中,然后进入分割头。从表 5 中可以看出,无注入时平均交并比比使用特征注入模块时低 1.64%,说明特征增强只具有辅助作用,不能完全代替原本骨干网络学习到的特征信息,需要特征注入模块来缓解不同特征提取方法间的差异。使用 CFIM 的情况比使用 FIM 的平均交并比低 1.90%,这与直观上的理解一致,注入不同尺度的特征,尽管通过上采样达到了相同的分辨率,仍不能弥补它们之间存在的差异,所以这也是三个对比实验中效果最不好的。

表 4 T 模块、C 模块的影响			
Table 4 Impact of T block and C block			
基线	CB	TB	平均交并比/%
✓			46.20
✓	✓		47.23
✓	✓	✓	49.64

表 5 特征注入模块的影响			
Table 5 Impact of feature injection module			
无注入	CFIM	FIM	平均交并比/%
✓			48.00
	✓		47.74
		✓	49.64

3 结语

本文提出了一种结合多种特征提取方式的新型遥感图像分割网络,将聚类算法、卷积算法和多头自注意力算法这三个特征提取方式结合使用,利用聚类算法的区域构建属性实现感受野的合理划分,搭建了由聚类块组成的基本特征提取模块;基于层次化的遥感图像分割架构,使用基于卷积的 C 模块促进浅层空间细节信息的精确分割,使用基于自注意力的 T 模块加强网络的语义特征提取能力,利用特征注入模块平衡不同特征间的差异,实现了更加平滑的特征融合。本文在实验部分证明了这种方法的有效性,并且在公开数据集上取得了良好的效果。

但是 NHNet 多种特征提取方式的结合体现在整体网络架构上,还没有更深层次的融合促进。因此未来的研究方向会集中于更深层次的融合促进,比如将特征增强模块加入骨干网络中,在网络运行的每一个阶段进行实时的双向融合,或者在特征增强模块之后设计一个更加彻底的特征混洗,设计关系感知融合模块。

参考文献 (References):

[1] 蒋涛,姜笑,郭金运,等.利用卫星测高数据反演全球海底地形研究进展[J]. 吉林大学学报(地球科学版), 2023,53(6):2029-2044.
Jiang Tao, Jiang Xiao, Guo Jinyun, et al. Review on Research Progress of Recovering Bathymetry from Satellite Altimetry-Derived Data[J]. Journal of Jilin University (Earth Science Edition), 2023, 53(6): 2029-2044.

[2] 李美霖,芮杰,金飞,等.基于改进 YOLOX 的遥感影像目标检测算法[J]. 吉林大学学报(地球科学版), 2023,53(4):1313-1322.
Li Meilin, Rui Jie, Jin Fei, et al. Remote Sensing Image Target Detection Algorithm Based on Improved YOLOX[J]. Journal of Jilin University (Earth Science Edition), 2023, 53(4): 1313-1322.

[3] Guo S, Liu L, Gan Z, et al. Isdnet: Integrating Shallow and Deep Networks for Efficient Ultra-High Resolution Segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: Computer Vision Foundation, 2022; 4361-4370.

[4] Ji D, Zhao F, Lu H, et al. Ultra-High Resolution Segmentation with Ultra-Rich Context: A Novel

Benchmark [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver:Computer Vision Foundation, 2023; 23621-23630.

[5] Wang W, Tang C, Wang X, et al. A ViT-Based Multiscale Feature Fusion Approach for Remote Sensing Image Segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[6] Adams R, Bischof L. Seeded Region Growing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(6): 641-647.

[7] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: Computer Vision Foundation,2015: 3431-3440.

[8] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [C]//Medical Image Computing and Computer-Assisted Intervention: MICCAI 2015: 18th International Conference. Munich: Springer International Publishing, 2015; 234-241.

[9] Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: Computer Vision Foundation,2017: 2881-2890.

[10] Chen L C, Zhu Y, Papandreou G, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[EB/OL]. [2023-03-21]. <https://doi.org/10.48550/arXiv.1802.02611>.

[11] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions [EB/OL]. [2024-04-17]. <https://arxiv.org/pdf/1511.07122.pdf>.

[12] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Aegmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[EB/OL]. [2024-04-17]. <https://arxiv.org/pdf/2010.11929.pdf>.

[14] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[15] Liu Z, Lin Y, Cao Y, et al. Swin Transformer:

- Hierarchical Vision Transformer Using Shifted Windows[EB/OL]. [2023 - 03 - 15]. <https://doi.org/10.48550/arXiv.2103.14030>.
- [16] Hassani A, Walton S, Li J, et al. Neighborhood Attention Transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: Computer Vision Foundation, 2023: 6185 - 6194.
- [17] Hassani A, Shi H. Dilated Neighborhood Attention Transformer[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2209.15001.pdf>.
- [18] Mehta S, Rastegari M. Mobilevit: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2110.02178.pdf>.
- [19] Yang C, Wang Y, Zhang J, et al. Lite Vision Transformer with Enhanced Self-Attention [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: Computer Vision Foundation, 2022: 11998 - 12008.
- [20] Chen Y, Dai X, Chen D, et al. Mobile-Former: Bridging Mobilenet and Transformer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: Computer Vision Foundation, 2022: 5270 - 5279.
- [21] Mehta S, Rastegari M. Separable Self-Attention for Mobile Vision Transformers[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2206.02680.pdf>.
- [22] Wadekar S N, Chaurasia A. Mobilevitv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features [EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2209.15159.pdf>.
- [23] Li Y, Yuan G, Wen Y, et al. Efficient Former: Vision Transformers at MobileNet Speed [J]. Advances in Neural Information Processing Systems, 2022, 35: 12934 - 12949.
- [24] Zhang W, Huang Z, Luo G, et al. TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: Computer Vision Foundation, 2022: 12083 - 12093.
- [25] Wan Q, Huang Z, Lu J, et al. SeaFormer: Squeeze-Enhanced Axial Transformer for Mobile Semantic Segmentation[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2301.13156.pdf>.
- [26] Han K, Wang Y, Guo J, et al. Vision GNN: An Image Is Worth Graph of Nodes[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2206.00272.pdf>.
- [27] Ma X, Zhou Y, Wang H, et al. Image as Set of Points[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2303.01494.pdf>.
- [28] Wang J, Zheng Z, Ma A, et al. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation[EB/OL]. [2024 - 04 - 17]. <https://arxiv.org/pdf/2110.08733.pdf>.
- [29] Yu C, Gao C, Wang J, et al. BiSeNet v2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation[J]. International Journal of Computer Vision, 2021, 129: 3051 - 3068.
- [30] Fan M, Lai S, Huang J, et al. Rethinking BiSeNet for Real-Time Semantic Segmentation [EB/OL]. [2023 - 04 - 06]. <https://doi.org/10.48550/arXiv.2104.13188>.