引文格式：刘琦,杨小华,陈柏行,等. 基于Transformer的成都市地物分类技术[J]. 地理空间信息, 2023, 21(12): 51-53. doi:10. 3969/j.issn. 1672-4623.2023.12.011

基于Transformer的成都市地物分类技术

刘　琦[1]，杨小华[1]，陈柏行[1]，邱姝月[1]，张　攀[2]，杨　茜[2]
（1. 成都市规划编制研究和应用技术中心，四川 成都 610041；2. 航天宏图信息技术股份有限公司，北京 100195）

摘　要：常用的人工判读方式需要耗费大量的人力资源，成本高且周期长。针对这些问题，基于视觉Transformer的深度学习网络模型对成都市重点区域开展典型地物要素分类，智能提取耕地、林地、道路、大棚、水体等8类地物的分布情况，并通过规则化算法对建筑提取结果进行优化。实验结果表明该方法的分类精度达到84.25%，较U-Net、HRNet、DeeplabV3+分类精度更高，且对成都市约6 400 km^{2}重点区域的分类时间约为2 h，具有较高的精度和效率，能够辅助成都市高效率、大范围、高频次的监测重点区域地物。

关键词：深度学习；遥感；语义分割；Transformer；地物分类
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2023）12-0051-03

Terrain ObjectClassification Technology of Chengdu City Based on Transformer

LIU Qi[1], YANG Xiaohua[1],CHENBohang[1], QIU Shuyue[1], ZHANG Pan[2], YANGQian[2]
(1. Chengdu Planning Reasearch and Application Technology Center, Chengdu 610041, China;
2. PIESAT Information Technology Co., Ltd., Beijing 100195, China)

Abstract: The commonly used manual interpretation method requires a lot of human resources, high cost and long time period. To solve these problems, we carried out the classification of typical feature elements in key areas of Chengdu based on the deep learning network model of visual Transformer. Then, we intelligently extracted the distribution of eight feature elements such as farmland, forest, road, greenhouse and water body, and optimized the classification result of building by regularization algorithm. The experimental results show that the classification accuracy of this method reaches 84.25%, which is higher than U-Net, HRNet and DeeplabV3+, and the classification time for key areas in Chengdu of about 6 400 km^2 is about two hours, which has high efficiency and accuracy, and can meet the demand of monitoring terrain objectof key area in Chengdu with high efficiency, large scale scope and high frequency.

Key words: deep learning, remote sensing, semantic segmentation, Transformer, terrain object classification

Transformer是一种基于注意力机制的新型神经网络，最早应用于自然语言处理领域[1-2]。相比于卷积网络，Transformer具有长距离特征捕获能力，基于Transformer的模型在各种视觉任务中具有强大的竞争力，甚至表现更优[3]。因此本文以成都市重点区域为研究区，采用Transformer模型实现遥感影像地物分类目标，并通过实验验证分割效果。

1　研究区与数据集

1.1　研究区域

成都市是中国西部地区重要中心城市，地处四川盆地西部，介于102°59′~104°54′E、30°05′~31°27′N之间，地势由西北向东南倾斜，具有平原、丘陵、高山多地貌特征，自然生态环境多样。成都市属于亚热带季风气候，年平均气温16℃，年降雨量1 000 mm左右。本文以成都市重点区域为研究区，地物类型丰富。

1.2　数据源及预处理

本次实验的数据为2021年成都市重点区域7~8月的JL1GF02A和JL1KF01A影像，其空间分辨率为0.8 m，共17景。对其进行辐射定标、正射校正、影像镶嵌等预处理工作，得到9 300.4 km^{2}的成都市重点区域遥感影像。为保证模型适用于成都市不同类别、不同特征的地物要素准确提取，实验选取了25个具有不同典型特征区域的遥感影像作为样本，将其划分为耕地、大棚、园地、林地、草地、建筑、道路、水体等8种地物类型，样本采用人工标

收稿日期：2022-04-21。
项目来源：四川省测绘地理信息协会开放基金资助项目（CCX202101）。
第一作者简介：刘琦（1975—），高级工程师，研究方向为摄影测量与遥感，E-mail：1316034612@qq.com。

注，经不重叠裁剪，共获取7 718张大小为512×512的样本数据，按照8∶2的比例划分为训练集和测试集。为避免模型过拟合[4]，在训练过程中样本会进行随机缩放、翻转、添加噪声等数据增强[5]操作。

2 研究方法

本文采用基于视觉Transformer的语义分割模型SegFormer[6]进行成都市内建筑、林地、水体、建筑、道路、耕地等典型地物要素的提取。

首先，针对成都市典型地物地类的分布情况，选取遥感影像进行地物要素样本采集；其次，将采集的样本输入到SegFormer网络模型进行训练，按照网络的特征提取结构提取地物的特征，网络的分类结构输出样本中像素的预测结果，通过计算预测值与真实值之间的损失迭代更新网络的参数；最后保存精度最优的模型。在模型的预测阶段通过滑窗的方式读取影像中固定大小的区域，预测区域的结果。预测结果中存在建筑物边缘粗糙、建筑角度不规则的情况，通过建筑轮廓规则化方法对建筑边界进行改善。

2.1 地物分类模型

本文采用的SegFormer网络使用自然语言处理（nature language processing，NLP）中Transformer代替原本的卷积特征提取模块，结合轻量级的All-MLP解码器，形成了一种更加高效、准确、鲁棒的语义分割网络。该网络包含以下2个主要模块，模型结构如图1所示。

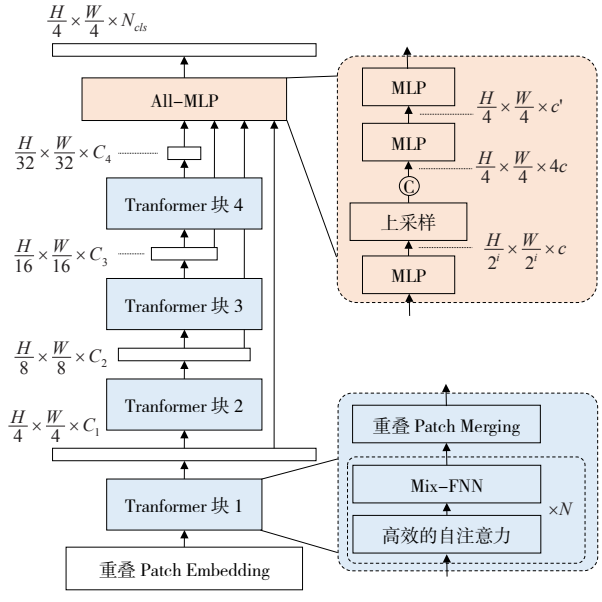


图1 SegFormer网络结构示意图

2.1.1 分级Transformer编码器

该编码器由4个Transformer块级联构成，与CNNs相似，每个Tranformer块可提取一个特征图，从浅层到深层，这些特征图的分辨率依次降低，分别为原始影像大小的$\frac{1}{4}$、$\frac{1}{8}$、$\frac{1}{16}$、$\frac{1}{32}$，特征图通道依次提高，图像特征从粗糙变精细。每个Transformer块包含以下3个部分：

高效的自注意力：在原本的多头注意力的基础上使用序列缩减方法[7]，如式（1）、（2），其中N为序列长度，K为被缩减的序列，R为缩减比，$\mathrm{Linear}(C_{\mathrm{in}}, C_{\mathrm{out}})(\cdot)$为输入通道数为$C_{\mathrm{in}}$，输出通道数为$C_{\mathrm{out}}$的线性操作，$\mathrm{Reshape}(\frac{N}{R}, C\cdot R)(K)$为将$K$变形为$\frac{N}{R}\times(C\cdot R)$。通过变形和线性层将原本为$N\times C$大小的序列缩减到$\frac{N}{R}\times C$，从而使自注意力的计算复杂度从$O(N^2)$降低至$O(\frac{N^2}{R})$，突破来自注意力层的计算瓶颈。

$$\hat{K}=\mathrm{Reshape}(\frac{N}{R}, C\cdot R)(K) \tag{1}$$

$$K=\mathrm{Linear}(C\cdot R, C)(\hat{K}) \tag{2}$$

Mix-FFN：将一个3×3卷积和MLP混入每一个前馈网络（FFN）中，为Transformer提供位置信息，如式（3）所示，其中Conv为卷积，x_{in}为自注意力输出特征，GELU为激活函数：

$$x_{\mathrm{out}}=\mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{MLP}\left(x_{\mathrm{in}}\right)\right)\right)\right)+x_{\mathrm{in}} \tag{3}$$

x_{in}通过MLP将x_{in}映射到更高维度的特征空间，经过3×3卷积和GELU非线性激活，再恢复到原始维度，并于输入特征相加。该结构避免了测试影像分辨率和训练分辨率不同时精度下降的问题。

重叠Patch Merging：给定图像块（patch）$2\times2\times C_i$，可以通过Patch Merging将其转化为$1\times1\times C_{i+1}$的向量，从而获得层次图。同理，我们可以将特征$F_i(\frac{H}{2^i}\times\frac{W}{2^i}\times C_i)$收缩为$F_{i+1}\left(\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_{i+1}\right)$，$i\in\{1,2,3,4\}$。为了保证每个patch周围的局部连续性，这里使用了重叠的Patch Merging，相邻patch的步幅S小于patch大小K，同时进行填充以保证合并过程中特征的空间大小一致。

2.1.2 轻量级All-MLP解码器

All-MLP解码器共包含4个步骤：①通过MLP层统一4个特征图F_i的通道数；②将特征上采样到$\frac{H}{4}\times\frac{W}{4}$大小并拼接在一起；③输入到MLP层，融合拼接后的特征；④输入到另一个MLP层，得到预测分割掩膜。

2.2 参数设置

学习率为训练过程中调整网络权重的参数，学习率的大小与网络的收敛速度相关。本实验中模型的学

习率初始化为0.000 06，训练过程中学习率会不断自我更新。学习率在初始的1 500次迭代中采用线性增加策略，然后学习率采用多项式衰减策略，为线性衰减，直至衰减为0。模型的梯度更新方式为AdamW[8]。为了防止GPU显存溢出，模型的batch_size设置为4，每次迭代读取4张图像进行训练。模型的总迭代次数为40 000，训练过程中模型的精度会逐渐提升，最终保留测试集上预测精度最高的模型。

2.3　建筑规则化

地物分类结果中建筑存在边界不平滑，形状不符合实际情况等问题。为了进一步提高建筑的分割结果，利用建筑轮廓规则化算法[9]对地物分类中建筑的结果进行修正。本文采用的规则化算法，分为粗糙处理和细节处理两部分完成：①粗糙处理主要处理图像中不合理的顶点、边和多边形。②细节处理主要修改建筑的形状。

3　实验结果与分析

本实验利用最优模型对成都市重点区域的建筑、耕地、林地、道路、大棚、水体、园地、草地8类地物进行分类，该模型在划分的测试数据集的指标如表1。8类地物的平均准确率达到84%，平均召回率达到84%，满足总体精度超过80%的需求。同时，该模型相较传统CNNs深度学习方法精度更高，如表2。在145 759×124 736大小的遥感影像上完成全类别推理仅需2 h左右，模型的解译结果经过少量人工修正后能够满足实际应用需求，相比于人工判读的方式具有更高的效率，能够有效减少成都市重点区域的解译周期。

表1　成都市测试数据分类精度

地物类型	Accuracy	F1-score	Precision	Recall
耕地	84.88	85.36	85.84	84.88
园地	79.45	81.62	83.91	79.45
林地	92.35	91.23	90.15	92.35
草地	69.9	71.52	73.22	69.9
建筑	84.51	83.44	82.39	84.51
道路	80.06	80.81	81.57	80.06
水体	93.84	93.68	93.53	93.84
大棚	93.28	87.06	81.63	93.28

表2　不同方法的分类精度

方法	Accuracy	F1-score	Precision	Recall
U-Net	80.45	78.67	80.02	77.56
HRNet	83.66	82.53	82.49	82.6
DeeplabV3+	83.96	81.39	81.20	81.73
本文算法	84.25	83.88	83.63	84.25

经过建筑轮廓规则化算法后，分割结果中不合理的误检区域被剔除，建筑的边缘更加平滑，建筑的角度更加符合实际情况。实验结果说明建筑轮廓规则化算法能够根据建筑的真实轮廓进行修正，不会大量降低建筑的分类精度。因此该算法适用于改善建筑不合理边界，可以有效减少建筑物提取后的人工修正时间。

4　结　论

本文利用基于视觉Transformer的语义分割方法对成都市重点区域进行地物要素分类，该方法通过分级结构的Transformer输出多尺度特征，通过MLP解码器聚合不同尺度特征、融合局部和全局注意力，具有强大的表达能力。本文方法的分类结果经过较少的人工修正能够达到成都市实际应用的精度需求和时间需求，为成都市快速、高效、高频地提取重点区域的地物信息提供技术支持，从而为土地资源的调查提供数据支持。

参考文献

[1] Braşoveanu A M P, Andonie R. Visualizing Transformers for Nlp: a Brief Survey[C]//2020 24th International Conference Information Visualisation (IV). IEEE, 2020:270-279

[2] Jin Y, Han D, Ko H. Trseg: Transformer for Semantic Segmentation[J]. Pattern Recognition Letters, 2021(148):29-35

[3] Han K, Wang Y, Chen H, et al. A Survey on Vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022(102):361-369

[4] Rice L, Wong E, Kolter Z. Overfitting In Adversarially Robust Deep Learning[C]//International Conference on Machine Learning PMLR, 2020:8 093-8 104

[5] Chlap P, Min H, Vandenberg N, et al. A Review of Medical Image Data Augmentation Techniques for Deep Learning Applications[J]. Journal of Medical Imaging and Radiation Oncology, 2021, 65(5):545-563

[6] Xie E, Wang W, Yu Z, et al. Seg Former: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. Advances in Neural Information Processing Systems, 2021(2):45-48

[7] Wang W, Xie E, Li X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:568-578

[8] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization[J]. arXiv Preprints arXiv:1711. 05101, 2017(4):23-29

[9] Wei S, Ji S, Lu M. Toward Automatic Building Footprint Delineation from Aerial Images Using CNN and Regularization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(3):2 178-2 189