结合对象单元和Transformer网络的城市功能区分类

鲁伟鹏[1,2]，贺清康[1]，李佳铃[1]，李诗逸[1]，陶超[1]

1. 中南大学 地球科学与信息物理学院, 长沙 410012;
2. 香港理工大学 土地测量及地理资讯学系, 香港 999077

摘 要：准确识别各类城市功能区并全面掌握其分布情况，对合理规划和科学管理城市至关重要。针对该问题，本文提出一种结合对象单元和Transformer网络的城市功能区分类方法。该方法首先以多尺度分割所获得的过分割对象作为最小分析单元，以避免出现同一分析单元包含多种城市功能区的情况。在此基础上，针对现有方法着重于对分析单元内部特征提取而忽略了分析单元之间的空间关系问题，提出利用Transformer框架和对象地理属性作为位置编码对不同分析单元之间的空间关系进行建模，从而实现兼顾分析单元内部特征和不同分析单元之间空间关系的城市功能区分类。结果表明，使用过分割对象作为最小分析单元能够更加准确地获取城市功能区地边界，从而避免基于规则格网单元所导致的锯齿状边缘及基于路网单元所导致地无法区分路网内不同功能区的问题；与仅考虑分析单元内部特征的传统方法相比，通过对不同分析单元之间的分析单元进行建模可有效提升城市功能区分类精度。

关键词：城市功能区，遥感，深度学习，空间关系建模，Transformer网络

中图分类号：P231/P2

引用格式：鲁伟鹏,贺清康,李佳铃,李诗逸,陶超.2024.结合对象单元和Transformer网络的城市功能区分类.遥感学报,28(8):1927-1939

Lu W P, He Q K, Li J L, Li S Y and Tao C. 2024. Object units and Transformer networks combined with urban functional zone classification method. National Remote Sensing Bulletin, 28(8): 1927-1939[DOI: 10.11834/jrs.20233036]

1 引 言

城市功能区UFZ（Urban Functional Zone）是指在城市发展过程中形成的承担特定社会经济功能的区域，包含工业区、商业区、住宅区等。准确掌握其空间分布情况对城市可持续发展具有重要意义（Chen等，2020；赵伍迪 等，2021）。现有的城市功能区规划图由于受居民生活习惯、经济发展水平等外部因素影响，与实际的城市功能区分布存在不一致的现象，很难直接作为城市功能管理的参考。

传统实地测量与调研的方式存在人工成本高、耗时长的不足，很难进行大范围的城市功能区分布情况调查（Du等，2020）。近年来，一些基于地理大数据的城市功能区分类方法如感兴趣点、街景图片等被提出。谷岩岩等（2018）利用重尾打断法和和密度分析对感兴趣点进行统计建模，实现功能区的分类。Zhao等（2022）则提出了一种自上而下的语义信息检测模型，从街景中提取出4类典型城市功能区但是，这些数据均存在着一些短板，例如感兴趣点数据大多由用户上传，由于上传数量多，很难实现人工检核，其质量无法得到保障（Zhao和Fan，2022），而街景数据则覆盖不全面，城市小区、公园等内部道路、城郊的街景数据获取难度大（Biljecki和Ito，2021）。因此，快速与准确地进行城市功能区分类是当前城市管

收稿日期：2023-02-21；**预印本：**2023-08-14

基金项目：湖南省杰出青年基金（编号：2022JJ10072）；湘江实验室开放基金一般项目（编码：22XJ03007）；国家自然科学基金（编号：42171376,41771458）；湖南省自然科学基金（编号：2021JJ30815）；中南大学高性能计算平台

第一作者简介：鲁伟鹏，研究方向为城市遥感、遥感影像智能解译。E-mail：weipeng.lu@connect.polyu.hk

通信作者简介：陶超，研究方向为遥感影像智能解译和机器学习。E-mail：kingtaochao@126.com

理中的所面临的关键问题之一。

伴随着遥感技术手段的飞速发展，利用高分辨率遥感图像实现城市功能区的分类成为可能（李德仁 等，2014；舒弥和杜世宏，2022； 陶超 等，2021）。当前基于遥感影像提取城市功能区所选取的研究单元可以分为以下3种：以规则格网为最小单元（Lu等，2022）、以封闭路网为最小单元（Zhang等，2018）、以对象为最小单元（Du等，2019）。但是上述的3种分析单元均存在不足之处。以规则格网为分析单元先通过格网影像裁剪为若干的矩形子块，而后对子块进行逐一分类。这使得最终所获取的城市功能区具有明显的锯齿状边界，与城市功能区实际情况吻合度较差；以封闭路网为最小单元则会出现同一封闭道路中可能存在多个功能区的问题；由于城市中功能区的大小不是统一的，且功能区内部的特征复杂多变，因此在进行面向对象分割时很难确一个统一的分割参数将所有的功能区完整且独立地分割出来（Zhou等，2020）。考虑到上述问题，Du等（2021）提出了使用过分割对象作为最小分析单元，通过对过分割对象先进行分类，而后合并相邻同类单元的方法提取出了与实际较为吻合的城市功能区。其中过分割即通过一个较小的分割参数，不直接将独立的功能区分割出来，而是将影像分割为更加细小的具有统一内部视觉特征的语义对象（Troya-Galvis等，2015）。本文研究将参考其方法，选择过分割的地理对象作为城市功能区分类的最小单元。

此外，当前的城市功能区分类方法，特别是以卷积神经网络CNNs（Convolutional Neural Networks）为代表的端到端的深度学习方法，受到如显卡内存等硬件条件的限制，在进行模型训练时往往只能输入较小的分析单元（如256×256像素），同时只关注每个分析单元内部的特征，而忽略了单元之间的关系（龚健雅 等，2022）。例如公园和住宅区的共现概率要远大于公园和工业区的共现概率，因此为了实现准确的功能区分类，有必要进一步分析功能区内部对象之间的空间关系（Tao等,2021;Zhang等,2017）。2020年，Dosovitskiy等（2021）提出ViT（Vision Transformer）模型，并在图像分类上取得了优异的成绩（Vaswani等，2017）。ViT通过自注意力分析解决了CNNs网络权重固化的问题，更重要的是其能够通过自注意力模型和位置编码建立各个分析单元之间的空间关系，为空间关系的建模提供了解决途径。然而，ViT模型相较于CNNs而言，具有更高的计算复杂度（Sabater等，2022；Li等，2022）。因此，如何进行高效的特征表达，使得ViT能够在有限的计算资源下进行空间关系建模是本文考虑的第一个问题。此外，ViT中的位置编码仅适合规则裁剪的棋盘式子块（patch），对于不规则的地理对象而言并无法提供有用的空间位置信息。因此，如何选择合适的位置编码来引导ViT对不规则分布的地理对象进行空间建模是本文所考虑的第二个问题。

综上，本文提出一种结合对象单元和Transformer网络的城市功能区分类方法。其主要包括4个环节：（1）考虑城市功能区的颜色信息、形状信息和纹理信息，采用多尺度分割生成过分割对象。（2）利用CNNs提取对象的内部特征。（3）融合对象内部特征与对象地理属性。（4）利用Transformer对对象特征进行空间关系建模与城市功能区分类。其中，考虑到计算资源有限的问题，本文将利用预训练的CNNs对对象进行特征提取，因此本文所提出的方法并非一个端到端的模型。同时，出于现有位置编码方法无法对不规则分布的对象进行位置描述的考虑，本文同时提出了基于对象地理信息的编码方法，以期为对象的空间关系分析提供几何与位置信息参考。

2 研究区域与方法

2.1 研究区域及数据

如图1所示本文研究区域为北京市六环内及其周边地区，覆盖面积约3300 km^2。本文影像数据来自BING地图，影像尺寸为53248 × 69632像素，分辨率为1 m。基于现有研究中所使用的城市功能区分类体系（Liu等，2021）和GB50137—2011《城市用地分类与规划建设用地标准》（planning.org.cn/law/uploads/2013/1383993139.pdf［2023-02-21］）设计了10类城市功能区，其包括商业、住宅、机构、工业、交通、绿地、待开发、林地、农用地和水体。为了进行模型训练，本文从OpenStreetMap（OSM）上收集了对应区域的多边形数据，并主要依据设施类型（Amenity）、建筑类型（Building）、土地利用类型（Landuse）等10个字段进行重分类，以“商业区”为例，满足以下

任意一条件即标记为商业区：（1）Landuse 包括 commercial 或 retail；（2）Building 为 commercial；（3）Shop 为 wholesale。具体的重分类依据及最终所收集的样本数量（70% 作为训练样本，30% 作为测试样本）详见表1。

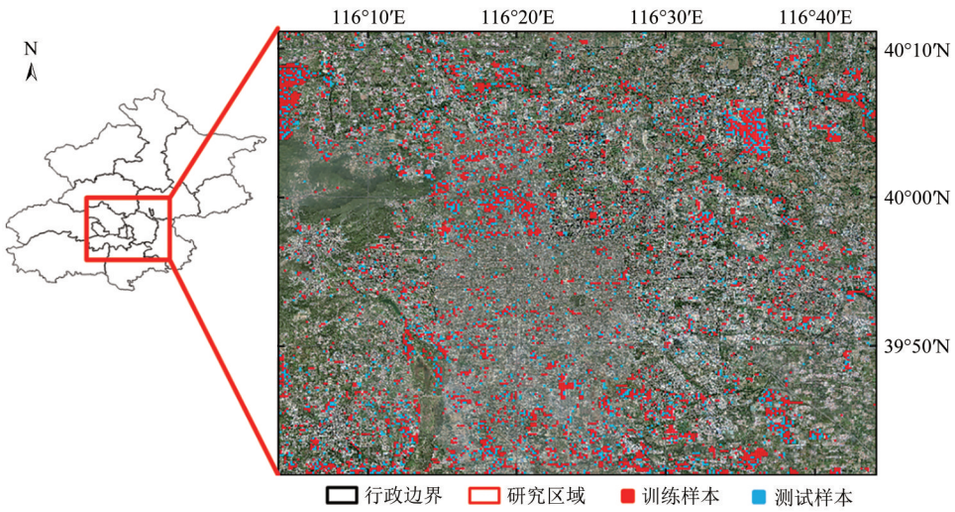


图1 研究区域影像及标签分布

Fig. 1 The imagery and label distribution of the research region

表1 OSM多边形数据重分类规则及标签类别分布

Table 1 Reclassification rule of OSM polygon and the distribution of category labels

功能区类别	判断字段	判断条件	标签数量（对象个数）	
			训练集	测试集
商业区	Landuse	包含 commercial、retail	454	193
	Building	等于 commercial		
	Shop	等于 wholesale		
住宅区	Landuse	等于 residential	1419	579
机构	Amenity	包含 kindergarten、school、college、university、language school	644	317
	Office	等于 educational institution		
	Landuse	包含 education，military		
工业区	Landuse	等于 industrial	299	128
	Water	等于 wastewater		
	Man_made	等于 reservoir covered		
	Building	包含 factory、industrial		
交通	Aeroway	不为空值	836	435
	Name	包含 airport		
	Landuse	等于 depot		
绿地	Name	包含 park	594	292
	Landuse	包含 plant nursery		
	Amenity	包含 park		
待开发	Landuse	包含 construction	1256	581
林地	Landuse	包含 forest、shrubland	284	140
农用地	Landuse	包含 farmland	930	438
水体	Natural	等于 water	288	154

2.2 研究方法

本文技术路线如图2所示。可见其主要包括面向对象的多尺度分割、内部特征编码、对象内部特征与地理信息融合及空间关系建模与分类4个部分。本文采用面向对象的多尺度分割方法，以克服深度学习方法中使用图像子块为最小分割单元而导致的无法准确地勾勒出城市功能区轮廓的问题；同时提出了基于对象地理信息的位置编码方法，使得面向对象的Transformer编码得以实现，达到精细化划分城市功能区的目的。

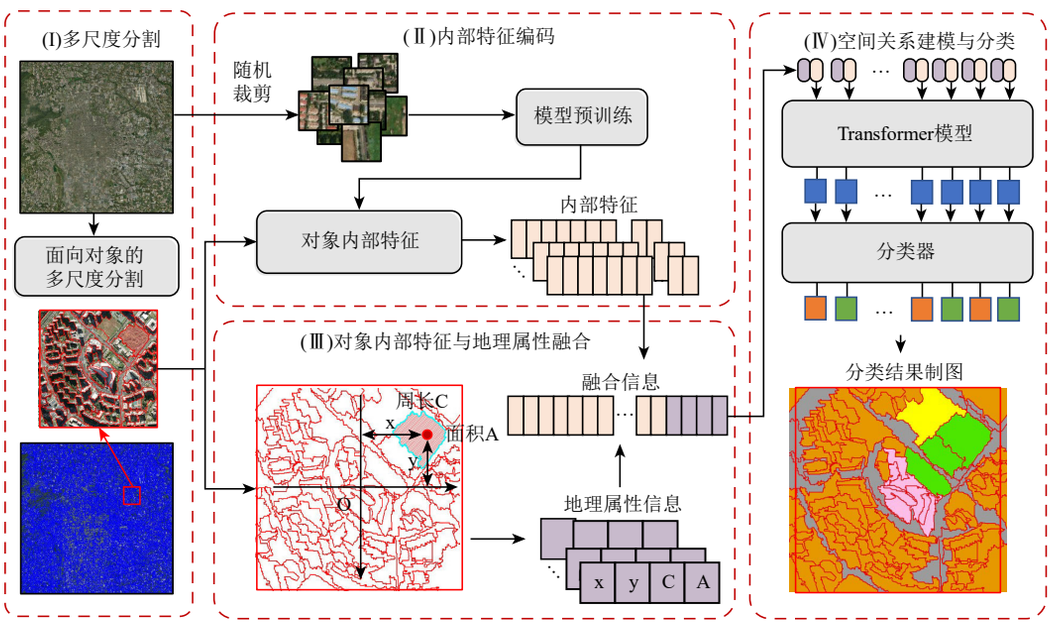
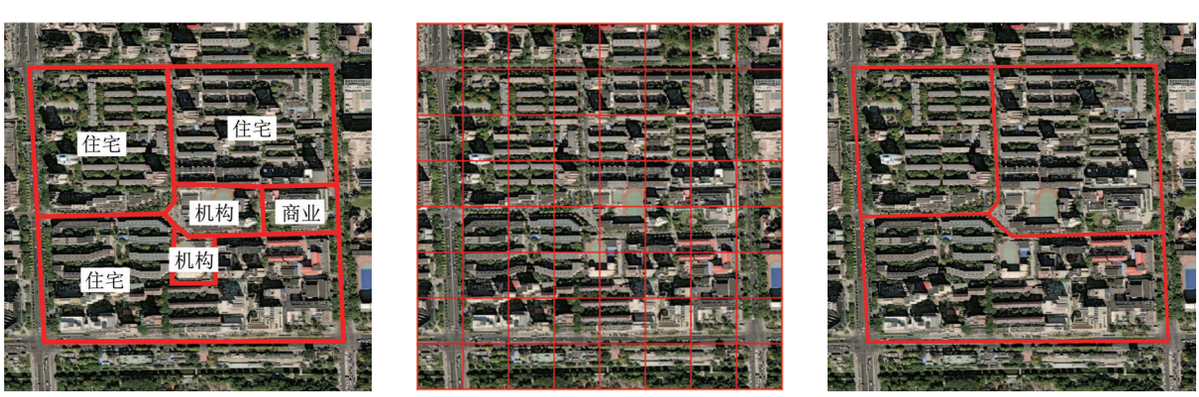


图2 本研究技术路线图

Fig. 2 The flowchart of this study

2.2.1 多尺度分割

图3展示了基于规则格网、封闭路网和不同分割尺度参数下多分辨率分割MRS（Multi Resolution Segmentation）（Baatz和Schäpe，2000）获得的分类单元。考虑到当前广泛使用的规则格网单元、封闭路网单元和对象单元存在与实际的城市功能区边界吻合度较差的情况。参考Du等（2021）的方案，本文选用较小的分割参数对影像进行过分割，并以过分割对象作为最小分析单元。通过对比不同的分割尺度参数，最终选择了尺度参数s = 100来进行对象的分割。其能尽可能地将影像分割成细小的对象，避免出现多个功能区被分割入一个对象中的情况，同时也可以避免因为分割出的对象过多而增加计算负担。



（a）实际的功能区边界

（a）Actual UFZ boundary

（b）规则格网单元

（b）Grid unit

（c）封闭路网范元

（c）Roadblock unit

（d）对象单元，分割尺度参数400
（d）Object units with segmentation scale of 400

（e）对象单元，分割尺度参数200
（e）Object units with segmentation scale of 200

（f）对象单元，分割尺度参数100
（f）Object units with segmentation scale of 100

图3 同一遥感影像基于不同分割单元的比较
Fig. 3 Comparison of the different segementation units of the same image

2.2.2 内部特征编码

CNNs凭借其强大的特征提取能力，已经成为遥感影像特征提取的首选模型之一（Gao等，2022；骆剑承 等，2021；赵伍迪 等，2021）。本文将采用ResNet50（He等，2016）作为特征编码器的主干网络进行城市功能区对象内部特征提取。其特征提取流程主要包括两步：（1）对象内部特征提取器预训练；（2）对象内部特征提取。其中，预训练所使用到的数据随机裁剪自图1中的训练区域，其尺寸为256 × 256像素，其标签则由裁剪区域中标签的众数所决定。

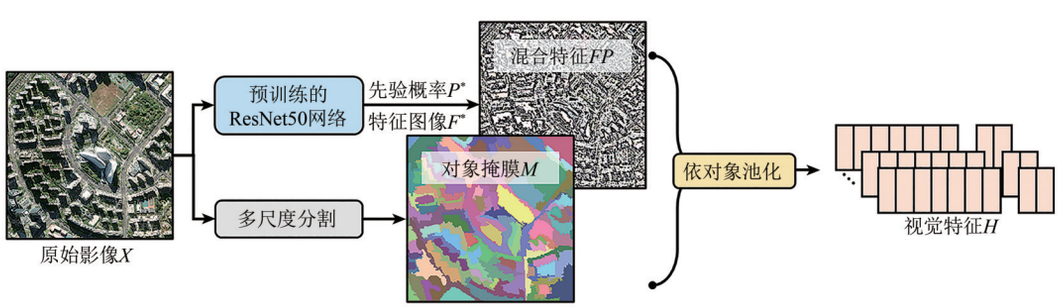


图4 对象内部特征提取流程图
Fig. 4 The flowchart of the object's visual feature extraction

（1）对象内部特征提取器预训练。对于包含K个影像样本$\boldsymbol{x}_i \in \mathbb{R}^{3\times l\times l}$和与之对应的独热编码（one-hot）标签$\boldsymbol{y}_i$的城市功能区分类数据集$\boldsymbol{D}=\left\{\left(\boldsymbol{x}_i,\ \boldsymbol{y}_i\right)\right\}_{i=1}^{K}$，本文以ResNet50为主干网络设计了一个CNN模型。该模型包含：1）ResNet50中的50个卷积层及附属的池化、归一化、激活函数，其能将$\boldsymbol{x}_i$编码为一个尺寸为$\mathbb{R}^{2048\times(l/32)\times(l/32)}$的特征图像；2）一个点对点卷积层（1 × 1卷积层），用于将高维的特征压缩至64维以减小后续运算负担；3）一个全局平均池化层GAP（Global Average Pooling）用以聚合特征；4）一个带有*SoftMax*函数的全连接层用计算$\mathbf{x}_i$的类别概率分布$\boldsymbol{p}_i$。为了方便后文描述，此处将1）和2）记为$f\left(\cdot|\boldsymbol{\theta}_f\right)$，3）记为GAP(·)将4）记为$g\left(\cdot|\boldsymbol{\theta}_g\right)$，$\boldsymbol{\theta}_f$和$\boldsymbol{\theta}_g$分别表示2个模型的参数，模型的正向传播过程可以用如下公式表示：

$$\boldsymbol{p}_i = g\left(\mathrm{GAP}\left(f\left(\boldsymbol{x}_i|\theta_f\right)\right)|\theta_g\right) \tag{1}$$

本文采用交叉熵损失函数对特征提取器进行端到端的优化，其表达式如下：

$$\mathcal{L}_1 = -\sum_{i=1}^{k}\sum_{c=1}^{\mathrm{cls}} y_{i,c}\log\left(p_{i,c}\right) \tag{2}$$

式中，cls表示类别总数，$p_{i,c}$表示样本$\boldsymbol{x}_i$为类别c的预测概率。

（2）对象内部特征提取。本文所构建的对象

特征将由特征图像$\boldsymbol{F}^*$和先验概率$\boldsymbol{P}^*$两部分聚合而成。由于分类卷积神经网络模型仅能获得图像级别的特征，为了获得对象级别的内部特征，本文首先使用上一步中的卷积层及点对点卷积进行特征图像的提取，即$f\left(\cdot|\theta_f\right)$。对于大尺寸遥感影像$\boldsymbol{X}\in\mathbb{R}^{3\times L\times L}$，$f$将其编码为一张尺寸为$\mathbb{R}^{64\times(L/32)\times(L/32)}$的特征图像。而后通过双线性内插将其还原为与输入影像同样大小的特征图像$\boldsymbol{F}^*\in\mathbb{R}^{64\times L\times L}$。同时，本文使用$g\left(\cdot|\boldsymbol{\theta}_g\right)$对$\boldsymbol{F}^*$进行逐像素计算，获得影像$\boldsymbol{X}$每个像素对应城市功能区类别的先验概率$\boldsymbol{P}^*\in\mathbb{R}^{64\times L\times L}$以帮助后续的分类。由此，影像$\boldsymbol{X}$的混合特征$\boldsymbol{FP}$可以表示为

$$\boldsymbol{FP}=\left[\boldsymbol{F}^*;\boldsymbol{P}^*\right]\in\mathbb{R}^{(64+\mathrm{cls})\times L\times L} \tag{3}$$

在本文中，城市功能区被分为10个类别，因此cls = 10 + 1，其中1表示数据集中“未标注”的像素。

基于第一步的多尺度分割，本文获得$\boldsymbol{X}$的对象掩膜$\boldsymbol{M}\in\mathbb{R}^{L\times L}$，和与之对应的对象集$O=\left\{o_i\right\}_{i=1}^{K}$，其中$K$表示$\boldsymbol{X}$经多尺度分割后的对象数目，这一数目与具体的$\boldsymbol{X}$有关，并非一个定值。掩膜$\boldsymbol{M}$像素的取值范围为1—$K$的整数，其表示像素所属对象的编号。对于对象$o_j$，本文按照其对象掩膜对$\boldsymbol{FP}$进行池化，即其内部特征$\boldsymbol{v}_j\in\mathbb{R}^{64+cls}$可以表示为其所有对应像素特征的平均池化值：

$$h_j=\frac{\sum_{M(x,y)=j}FP\left(x,y\right)}{\sum_{M(x,y)}T\left(j,M\left(x,y\right)\right)} \tag{4}$$

$$\mathrm{T}\left(j,M\left(x,y\right)\right)=\begin{cases}1, & M\left(x,y\right)=j\\ 0, & M\left(x,y\right)\neq j\end{cases} \tag{5}$$

式中，T函数用于判断像素$\left(x,\ y\right)$是否属于对象o_j。对象集O的特征即可表示为$\boldsymbol{H}=\left[\boldsymbol{h}_1,\ \boldsymbol{h}_2,\ \cdots,\ \boldsymbol{h}_K\right]\in\mathbb{R}^{K\times(64+cls)}$。

2.2.3 对象内部特征和地理属性融合

由于ViT模型只能针对规则格网进行空间关系建模，而无法对不规则分布的过分割对象进行位置标定，本文提出了使用对象地理信息，即对象相对对象集几何中心的坐标来确定对象之间的相对位置关系。虽然面向对象的分割方法可以将研究区域分割为无重叠的对象，但对于凹多边形而言，其几何中心可能位于多边形外部，因此如果直接使用对象的几何中心坐标来表示对象位置则有可能出现两个对象位置重合的情况。为避免这种情况的出现，本文使用ArcGIS中提供的多边形的标注点（label point）来锚定对象的位置。多边形的标注点往往用于地图学上的标注，是一个多边形的视觉中心，其一定落在多边形内部。同时，由于对象特征经过池化计算聚合为一个向量导致其几何信息丢失，本文同时也考虑使用对象的面积、周长来在一定程度上还原其几何属性。综上，本文将使用对象的标注点坐标$\left(m_i,\ n_i\right)$（IArea.LabelPoint Property （ArcObjects .NET 10 SDK）（arcgis.com）［2023-02-21］）、面积s_i以及周长c_i来替代ViT模型的位置编码以进行城市功能区对象之间的空间关系建模，即对象o_i的地理属性特征可以用$\boldsymbol{g}_i=\left[m_i,\ n_i,\ s_i,\ c_i\right]$表示。其中对象的几何中心坐标采用如下的方法确定：在对象集$\boldsymbol{O}=\left\{o_i\right\}_{i=1}^{K}$中以对象集几何中心为原点，东方向为x轴，南方向为y轴建立平面直角坐标系，取每个对象o_i的标注中心坐标$\left(m_i,\ n_i\right)$表示对象的空间位置。对象集O的地理信息$\boldsymbol{G}$即可表示为$\left[\boldsymbol{g}_1,\ \boldsymbol{g}_2,\ \cdots,\ \boldsymbol{g}_K\right]$。通过将对象集的内部特征与地理信息特征拼接，即可获得带有位置编码的对象集特征$\boldsymbol{F}$：

$$\boldsymbol{F}=\left[\boldsymbol{H};\boldsymbol{G}\right]\in\mathbb{R}^{K\times D} \tag{6}$$

式中，D = 64 + cls + 4。

2.2.4 空间关系建模和分类

在经历上述3个步骤后，影像中每个对象的特征已经可以用一个向量进行表示。但是，到目前为止，对象的特征仍旧只能表示其内在特征和地理属性。正如我们前文所说，不同的对象在不同的功能区中的共现概率和空间分布是不同的。因此，为了进行准确的城市功能区分类，有必要对对象之间的空间关系进行进一步的分析。本文方法将使用多层Transformer模型对对象特征进行进一步的深层次特征提取与空间关系建模，其中Transformer模型结构如图5，每一层Transformer编码器包含两个归一化模块Norm（Normalization）、一个多头注意力模块MHA（Multi-Head Attention）和一个多层感知机模块MLP（Multi-Layer Perceptron）（刘静 等，2013）。其中归一化模块用于控制计算过程中所获得的特征值域，排除量纲的作用。

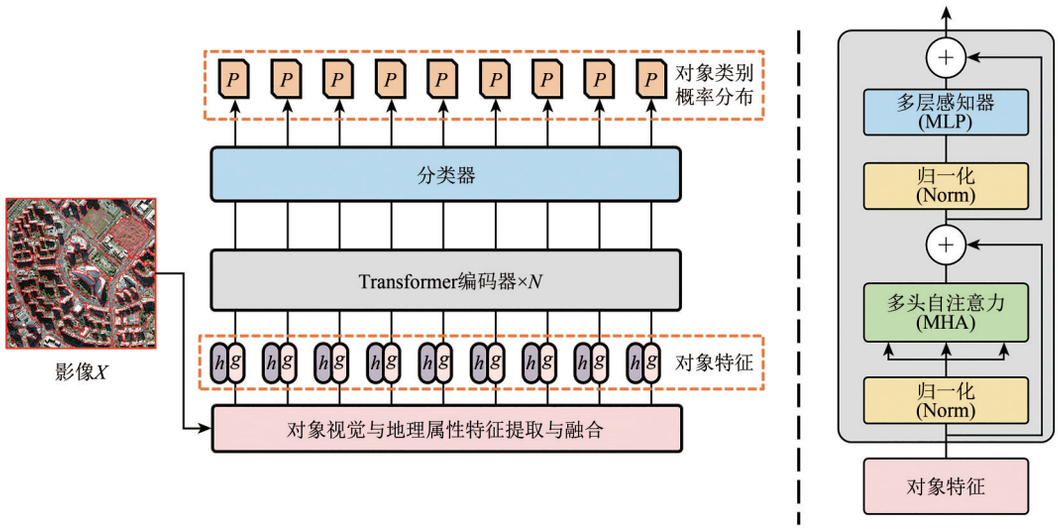


图5 本文所提出的Transformer模型结构图

Fig. 5 The architecture of the Transformer proposed in this paper

此处我们以第一层Transformer为例阐述进行空间关系建模的过程。对于对象集特征$\boldsymbol{F}$，Transformer模型首先对其进行层归一化得到归一化对象集特征$\boldsymbol{F}_{\text{norm}}$。而后通过多个自注意力模块（图6）计算对不同对象之间的关系权重以挖掘不同对象特征之间的空间关系，其计算过程可以用下面2个公式概括：

$$\boldsymbol{F}'_i = \text{SoftMax}\left(\frac{\boldsymbol{QK}^{\text{T}}}{\sqrt{D}}\right)\boldsymbol{V} \tag{7}$$

$$\boldsymbol{F}' = \text{concatenate}(\boldsymbol{F}'_1,\boldsymbol{F}'_2,\cdots,\boldsymbol{F}'_n)\boldsymbol{W}_F \tag{8}$$

式中，$\boldsymbol{K}$，$\boldsymbol{Q}$，$\boldsymbol{V}$分别是$\boldsymbol{F}_{\text{norm}}$的3个线性映射，$\boldsymbol{QK}^{\text{T}}$所计算的是对象两两之间特征的点积，由于此特征同时包含对象内部特征和对象地理属性，其结果可以视为对象两两之间在空间上的相关关系。其归一化结果左乘$\boldsymbol{V}$则是将这个关系嵌入到原始的对象集特征中。concatenate表示矩阵拼接，通过多个子注意力模块的叠加，可以获得更加丰富的空间关系信息。

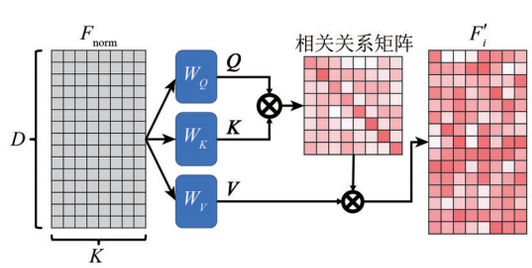


图6 多头注意力模块中的自注意力模块结构图

Fig. 6 The architecture of the self-attention module in MHA

在实际的ViT结构中往往将多个Transformer编码器叠加使用以提升特征的提取效果。此处将多个Transformer编码器的结果记为$\boldsymbol{Z}$，在模型的最后将通过一层线性层对$\boldsymbol{Z}$进行分类以获得最终每个对象的类别概率分布。此处同样使用交叉熵损失函数（式（9））对ViT模型进行优化：

$$\mathcal{L}_1 = -\log\sum_{i=1}^{k}\sum_{c=1}^{\text{cls}} y_{i,c}\log(p_{i,c}) \tag{9}$$

此处的$y_{i,c}$当且仅当对象o_i属于类别c时为1，反之为0；$p_{i,c}$表示对象o_i属于类别c的预测概率。

3 实验设置与结果分析

3.1 实验设置

本文的实验在Ubuntu18.04操作系统下进行，其搭载有NVIDIA Tesla V100S GPU与Intel Xeon Gold 6248R CPU，内存容量为128 GB。在影像分割环节，分割尺度参数s是影响分割效果最主要的参数。本文通过观察局部分割效果，在保证不会将多个功能区分割在一个对象内和设备可支持的情况下，选择一组尽可能小的分割参数进行多尺度分割（$s=100$，$\omega_{\text{shape}}=0.5$，$\omega_{\text{compactness}}=0.5$）(Du等，2021)。

模型实现方面，为了避免过拟合，同时满足GPU显存的限制，本文通过设置较小的Transformer层数和分支头数构建了一个轻量模型，其层数设置为5层，多头注意力分支数目为8，批处理大小（batch size）为8，使用Adam作为优化

器，学习率为1×10^{-5}，不设置额外的梯度衰减策略，训练总共经历200个批次（epoch），总共花费时间约为2.1小时。进行整个研究区域的预测耗时约2.5分钟（以上时间不包括数据预处理、面向对象分割、内部视觉特征编码和融合所消耗的时间）。

精度评价方面，由于不同的分类方法基于的分类单元不同，为了对所有的分类方法进行公平的评价，本文对于所有的方法均以像素为最小评价单元。为了能够同时对分类的准确度和提取的完整度进行评价，本文采用Kappa系数和F1 score两个在遥感影像分类中常用的分类指标来评价整体的分类精度和各个类别的分类精度。二者的值域均为[-1，1]（大多数情况下会大于0），其值越大表示分类的精度越高。

为了在分割效果和分类精度上同时体现出本文方法的优势，本文选择了当下在遥感影像分析中被广泛使用的ResNet50网络和U-Net（Ronneberger等，2015）网络，以及SOCNN（Zhou等，2020）作为对比方法。其中ResNet50是经典的图像分类网络，在场景分类、土地覆盖分类等多种任务中作为主干网络使用；U-Net是一种语义分割网络，能够实现对影像的像素级分类；SOCNN是一个以对象为最小分类单元的城市功能区分类方法，通过在研究区域进行随机裁剪，预测裁剪区域中心点的功能区类型生成带有标签的投票点，而后对每个对象中的投票点进行统计确定对象的功能区类型。

3.2 结果及分析

表2为ResNet、U-Net、SOCNN及本文方法在测试区域上的精度评价结果。可见在同样以像素为单位的评价体系下，本文方法在整体上取得了最好的结果：相较于次优的SOCNN，本文方法的Kappa系数有13.9%的提升；同时，本文方法在绝大多数的类别上的F1 Score均为最优。其中，本文方法在如住宅、交通等社会属性强的类别上具有超过30%的提升。

在功能区分类结果的平均交并比mIoU（mean Intersection over Union）这一精度评价指标方面，本文方法的mIoU为0.488，而ResNet、U-net及SOCNN的IoU分别为0.211、0.255和0.292。相较于其他方法，本文方法有显著提升。这定量地表明，以对象为分析单元在城市功能区分类中对边界提取有显著改善作用。

表2 城市功能区分类结果F1 Score定量评价表

Table 2 Quantitative result of F1 Score on UFZ classification

类别	ResNet	U-net	SOCNN	本文方法
商业	0.084	0.101	0.287	0.346
住宅	0.479	0.683	0.610	0.829
机构	0.029	0.229	0.449	0.537
工业	0.220	0.307	0.327	0.358
交通	0.332	0.559	0.550	0.796
绿地	0.418	0.408	0.526	0.669
待开发	0.445	0.460	0.607	0.680
林地	0.594	0.674	0.552	0.590
农田	0.327	0.286	0.433	0.570
水域	0.504	0.443	0.401	0.466
Kappa	0.305	0.451	0.553	0.630

图7为ResNet50、U-Net、SOCNN及本文方法所提取的北京市二环内城市功能区地图。由图7（a）及其局部放大图可见：对于以子块为最小分类单元的ResNet模型，其总体精度较低的主要原因是以子块为单元的功能区边界与实际的功能区边界存在较大的出入，这一现象可以更直观地从可视化的分类结果图中观察到。由图7（b）及其局部放大图可见：在以像素为最小分类单元的U-Net模型中，虽然子块内部各个功能区之间的边界相较于ResNet而言更加符合实际的功能区边界情况，但是在子块与子块之间存在很明显的拼接现象。由图7（d）及其局部放大图可见：本文采用对象为最小分类单元，其生成过程中没有经过格网裁剪，因此不会出现如ResNet和U-Net类似的锯齿状边缘，更加吻合实际的功能区边界情况，且由于引入空间关系建模的原因，其分类的准确度上有明显的提升。例如图7（d）中的火车铁轨和水体的分类结果无论从边界效果来看，还是从提取的完整度来看均优于U-Net图7（b）。在与同样使用对象为分类单元的SOCNN（图7（c））的对比中可以明显看出，本文方法无论是在最终分类边界的吻合度还是在分类结果的准确率上均

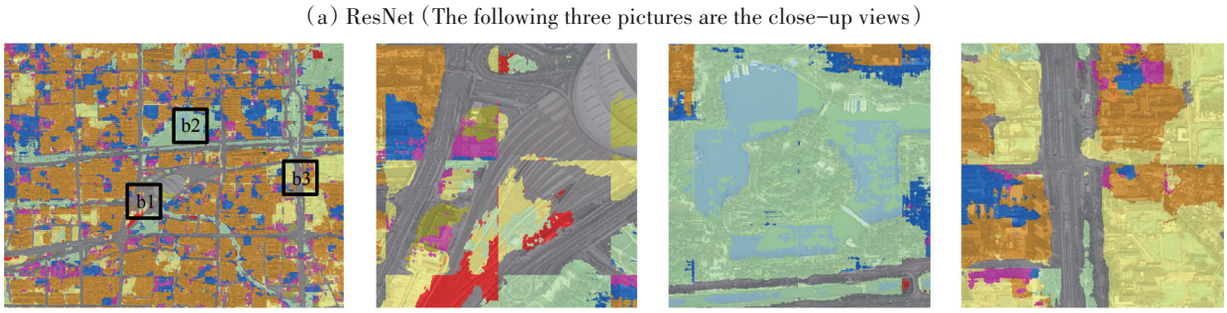
取得了较好的结果，其中的主要原因可能在于SOCNN虽然是通过对象内的投票点来确定对象的功能区类别，但是投票点的类别是由以其为中心的正方形区域决定的，其本质上还是基于子块的分类，并非真正的以对象为最小分析单元。

3.3 讨论

在利用Transformer空间关系建模方面，除本文提出的地理信息编码，与目前的正余弦编码、LPE编码和无位置编码等3种方式进行对比，其城市功能区分类结果的Kappa系数分别为0.590、0.586和0.587。相较于ResNet50和U-Net 2种基于CNNs的分类模型，其他引入Transformer进行对象空间关系建模的方法在总体的分类精度上具有明显的提升。但同时，现有的3种位置编码方式之间的Kappa系数差距并不明显。对比的结果总结如下：（1）Transformer的自注意力机制可以聚合对象之间的信息，以帮助更准确地进行城市功能区的分类；（2）现有的位置编码方式对于分规则分布的地理对象而言，并不能提供有效的空间位置信息，因此其不能实现更为准确的空间关系建模。此外，与同样使用以对象为最小分类单元的SOCNN相比，引入Transformer的分类方法，无论是使用何种位置编码，抑或是不适用任何位置编码，均取得了更好的表现。其主要的原因是Transformer即便在没有地理属性引导的情况下也可以进行对象两两之间的相关关系分析，这在一定程度上可以帮助优化最后的分类结果。

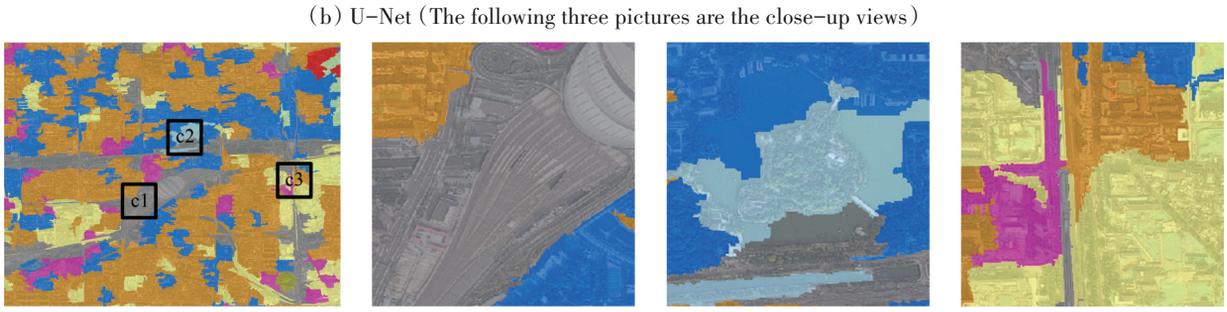
（a）ResNet（后面3个图为局部放大图）

（a）ResNet（The following three pictures are the close-up views）



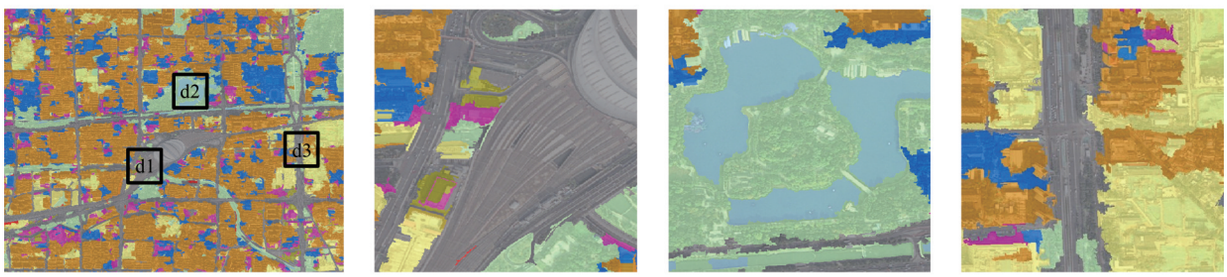
（b）U-Net（后面3个图为局部放大图）

（b）U-Net（The following three pictures are the close-up views）



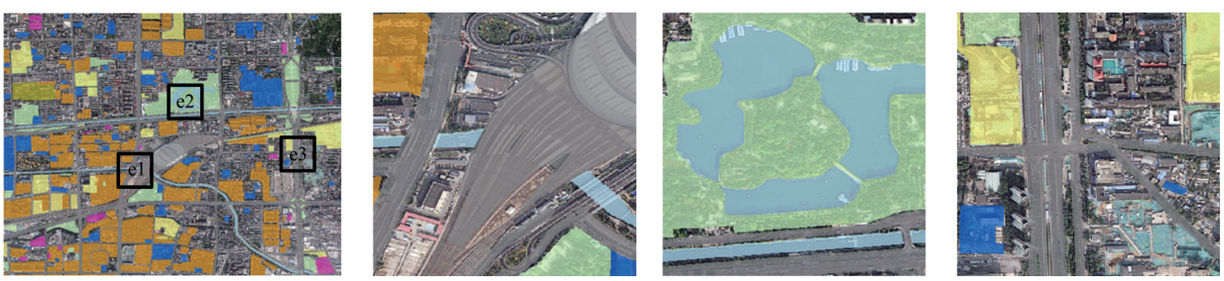
（c）SOCNN（后面3个图为局部放大图）

（c）SOCNN（The following three pictures are the close-up views）



（d）本文方法（后面3个图为局部放大图）

（d）The proposed method（The following three pictures are the close-up views）



（e）标签（后面3个图为局部放大图）

（e）Ground truth（The following three pictures are the close-up views）

图7 城市功能区分类结果图

Fig. 7 Result of UFZ classification

此外，从表2的精度评价结果中可以看出虽然商业区、工业区的F1-score相交于其他对比方法而言有一定的提升，但是相较于其他的住宅区、交通等类别，其分类效果仍旧不够理想。城市功能区的类型不仅与物理特征密切相关，而且与社会语义有着密切的关系（Wu等，2021；Zhong等，2020）。本文所使用的所有特征均是基于遥感影像所提取的，其可以较好地表征功能区地物理特征，但是在社会语义地表达方面则不够理想。对于商业区和工业区这种社会语义明显的功能区，仅仅依靠遥感影像所提供的物理特征很难进行准确地辨别，因此融合多源数据，从物理特征和社会语义两方面进行城市功能区的分类是值得进一步研究的方向（Wu等，2023）。

4 结 论

针对当前城市功能区分类方法在功能区边界提取效果差，及忽略对象间空间关系的问题，本研究首先采用多尺度影像分割方法提取的过分割对象作为最小分析单元，以获取相对准确的城市功能区边界。在此基础上，提出面向对象的Transformer模型，以对象地理信息作为位置编码进行对象间空间关系建模，以弥补现有城市功能区分类方法往往只关注分析单元内部特征的不足，实现了在深度学习框架下的对象空间关系建模和功能区的精细化分类。主要结论如下：（1）过分割对象可以更加准确地描述城市功能区的边界；（2）分析单元之间的空间关系能有效提高城市功能区分类的准确度。

参考文献（References）

Baatz M and Schäpe A. 2000. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. Angewandte Geographische Informationsverarbeitung XII, Heidelberg: Wichmann: 12-23

Biljecki F and Ito K. 2021. Street view imagery in urban analytics and GIS: a review. Landscape and Urban Planning, 215: 104217 [DOI: 10.1016/j.landurbplan.2021.104217]

Chen Z L, Zhou L L, Yu W H, Wu L and Xie Z. 2020. Identification of the urban functional regions considering the potential context of interest points. Acta Geodaetica et Cartographica Sinica, 49(7): 907-920 (陈占龙, 周路林, 禹文豪, 吴亮, 谢忠. 2020. 顾及兴趣点潜在上下文关系的城市功能区识别. 测绘学报, 49(7): 907-920) [DOI: 10.11947/j.AGCS.2020.20190315]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszko-

reit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 [DOI: 10.48550/arXiv.2010.11929]

Du S H, Du S H, Liu B and Zhang X Y. 2021. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. Remote Sensing of Environment, 261: 112480 [DOI: 10.1016/j.rse.2021.112480]

Du S J, Du S H, Liu B and Zhang X Y. 2019. Context-enabled extraction of large-scale urban functional zones from very-high-resolution images: a multiscale segmentation approach. Remote Sensing, 11(16): 1902 [DOI: 10.3390/rs11161902]

Du S J, Du S H, Liu B, Zhang X Y and Zheng Z J. 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. GIScience and Remote Sensing, 57(3): 411-430 [DOI: 10.1080/15481603.2020.1724707]

Gao L R, Han Z, Hong D F, Zhang B and Chanussot J. 2022. CyCU-Net: cycle-consistency unmixing network by learning cascaded autoencoders. IEEE Transactions on Geoscience and Remote Sensing, 60: 5503914 [DOI: 10.1109/TGRS.2021.3064958]

Gong J Y, Zhang M, Hu X Y, Zhang Z, Li Y S and Jiang L C. 2022. The design of deep learning framework and model for intelligent remote sensing. Acta Geodaetica et Cartographica Sinica, 51(4): 475-487 (龚健雅, 张觅, 胡翔云, 张展, 李彦胜, 姜良存. 2022. 智能遥感深度学习框架与模型设计. 测绘学报, 51(4): 475-487) [DOI: 10.11947/j.AGCS.2022.20220027]

Gu Y Y, Jiao L M, Dong T, Wang Y D and Xu G. 2018. Spatial distribution and interaction analysis of urban functional areas based on multi-source data. Geomatics and Information Science of Wuhan University, 43(7): 1113-1121 (谷岩岩, 焦利民, 董婷, 王艳东, 许刚. 2018. 基于多源数据的城市功能区识别及相互作用分析. 武汉大学学报(信息科学版), 43(7): 1113-1121) [DOI: 10.13203/j.whugis20160192]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Li D R, Zhang L P and Xia G S. 2014. Automatic analysis and mining of remote sensing big data. Acta Geodaetica et Cartographica Sinica, 43(12): 1211-1216 (李德仁, 张良培, 夏桂松. 2014. 遥感大数据自动分析与数据挖掘. 测绘学报, 43(12): 1211-1216) [DOI: 10.13485/j.cnki.11-2089.2014.0187]

Li Y Y, Yuan G, Wen Y, Hu J, Evangelidis G, Tulyakov S, Wang Y Z and Ren J. 2022. EfficientFormer: vision transformers at MobileNet speed//. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.: 12934-12949

Liu B H, Deng Y B, Li M, Yang J and Liu T. 2021. Classification schemes and identification methods for urban functional zone: a review of recent papers. Applied Sciences, 11(21): 9968 [DOI: 10.3390/app11219968]

Liu J, Zuo X Q, Wu L M, Huang L and Lu Z Y. 2013. Water extraction of satellite images considering space and feature relationship. Science of Surveying and Mapping, 38(3): 163-165 (刘静, 左小清, 吴俐民, 黄亮, 卢昭羿. 2013. 顾及地物空间关系的卫星影像水系信息提取. 测绘科学, 38(3): 163-165) [DOI: 10.16251/j.cnki.1009-2307.2013.03.050]

Lu W P, Tao C, Li H F, Qi J and Li Y S. 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. Remote Sensing of Environment, 270: 112830 [DOI: 10.1016/j.rse.2021.112830]

Luo J C, Hu X D, Wu T J, Liu W, Xia L G, Yang H P, Sun Y W, Xu N, Zhang X, Shen Z F and Zhou N. 2021. Research on intelligent calculation model and method of precision land use/cover change information driven by high-resolution remote sensing. National Remote Sensing Bulletin, 25(7): 1351-1373 (骆剑承, 胡晓东, 吴田军, 刘巍, 夏列钢, 杨海平, 孙营伟, 徐楠, 张新, 沈占锋, 周楠. 2021. 高分遥感驱动的精准土地利用与土地覆盖变化信息智能计算模型与方法研究. 遥感学报, 25(7): 1351-1373) [DOI: 10.11834/jrs.20219402]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//18th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Munich: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sabater A, Montesano L and Murillo A C. 2022. Event Transformer. A sparse-aware solution for efficient event data processing//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE: 2676-2685 [DOI: 10.1109/CVPRW56347.2022.00301]

Shu M and Du S H. 2022. Forty years' progress and challenges of remote sensing in national land survey. Journal of Geo-Information Science, 24(4): 597-616 (舒弥, 杜世宏. 2022. 国土调查遥感40年进展与挑战. 地球信息科学学报, 24(4): 597-616) [DOI: 10.12082/dqxxkx.2022.210512]

Tao C, Lu W P, Qi J and Wang H. 2021. Spatial information considered network for scene classification. IEEE Geoscience and Remote Sensing Letters, 18(6): 984-988 [DOI: 10.1109/LGRS.2020.2992929]

Tao C, Yin Z W, Zhu Q and Li H F. 2021. Remote sensing image intelligent interpretation: from supervised learning to self-supervised learning. Acta Geodaetica et Cartographica Sinica, 50(8): 1122-1134 (陶超, 阴紫薇, 朱庆, 李海峰. 2021. 遥感影像智能解译: 从监督学习到自监督学习. 测绘学报, 50(8): 1122-1134) [DOI: 10.11947/j.AGCS.2021.20210089]

Troya-Galvis A, Gançarski P, Passat N and Berti-Équille L. 2015. Unsupervised quantification of under- and over-segmentation for object-based remote sensing image analysis. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(5): 1936-1945 [DOI: 10.1109/JSTARS.2015.2424457]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Wu H, Lin A Q, Xing X D, Song D X and Li Y. 2021. Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method. International Journal of Applied Earth Observation and Geoinformation, 103: 102475 [DOI: 10.1016/j.jag.2021.102475]

Wu H, Luo W T, Lin A Q, Hao F H, Olteanu-Raimond A M, Liu L F and Li Y. 2023. SALT: a multifeature ensemble learning framework for mapping urban functional zones from VGI data and VHR images. Computers, Environment and Urban Systems, 100: 101921 [DOI: 10.1016/j.compenvurbsys.2022.101921]

Zhang X Y, Du S H and Wang Q. 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS Journal of Photogrammetry and Remote Sensing, 132: 170-184 [DOI: 10.1016/j.isprsjprs.2017.09.007]

Zhang X Y, Du S H and Wang Q. 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sensing of Environment, 212: 231-248 [DOI: 10.1016/j.rse.2018.05.006]

Zhao K, Liu Y K, Hao S Y, Lu S X, Liu H B and Zhou L J. 2022. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. IEEE Transactions on Geoscience and Remote Sensing, 60: 5602817 [DOI: 10.1109/TGRS.2021.3064316]

Zhao W D, Li S S, Li A, Zhang B and Chen J. 2021. Deep fusion of hyperspectral images and multi-source remote sensing data for classification with convolutional neural network. National Remote Sensing Bulletin, 25(7): 1489-1502 (赵伍迪, 李山山, 李安, 张兵, 陈俊. 2021. 结合深度学习的高光谱与多源遥感数据融合分类. 遥感学报, 25(7): 1489-1502) [DOI: 10.11834/jrs.20219117]

Zhao Z Y and Fan H C. 2022. Towards exploring patterns of editing behavior on OpenStreetMap. Journal of Geodesy and Geoinformation Science, 5(2): 85-97 [DOI: 10.11947/j.JGGS.2022.0209]

Zhong Y F, Su Y, Wu S Q, Zheng Z D, Zhao J, Ma A L, Zhu Q Q, Ye R C, Li X M, Pellikka P and Zhang L P. 2020. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: a case study of Chinese cities. Remote Sensing of Environment, 247: 111838 [DOI: 10.1016/j.rse.2020.111838]

Zhou W, Ming D P, Lv X W, Zhou K Q, Bao H Q and Hong Z L. 2020. SO－CNN based urban functional zone fine division with VHR remote sensing image. Remote Sensing of Environment, 236: 111458 [DOI: 10.1016/j.rse.2019.111458]

Object units and Transformer networks combined with urban functional zone classification method

LU Weipeng[1,2], HE Qingkang[1], LI Jialing[1], LI Shiyi[1], TAO Chao[1]

1. School of Geosciences and Info-Physics, Central South University, Changsha 410012, China;
2. Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong 999077,China

Abstract: Urban Functional Zones (UFZs) refer to specific areas within a city that have distinct functionalities and land uses. These zones are designated based on their primary activities and the roles they play in the urban environment. Accurate extraction of UFZs and a comprehensive understanding of their spatial distribution play an important role in urban planning and management. Traditional Convolutional Neural Networks (CNNs) focus on local features through convolutions, but they often miss the broader spatial relationships. Vision Transformer (ViT), while advanced, still has limitations; its tokenization method and learnable position encoding do not effectively represent geographical entities and their spatial relationships, which is a crucial feature in geospatial analysis.

This study proposes a UFZ classification method combining object units and ViT to address this issue. First, this method utilizes over-segmented objects generated from a multi-scale segmentation approach as analysis units to avoid the presence of multiple kinds of UFZs within a single object. Over-segmentation helps in creating smaller, more homogeneous units, thereby increasing the precision of the classification process. Then, considering that current methods often focus on the inherent analysis of objects while ignoring their spatial relationships, ViT is employed for spatial relationship modeling between objects, with the geographic attributes of objects serving as position embeddings. In this way, both the inherent features of a single analysis unit and the inter-spatial features among objects are considered for UFZ classification. Position embeddings using geographic coordinates allow the model to understand spatial proximity and relationships, which are crucial for accurate classification. We chose Beijing as the study area and downloaded imagery of the area within the Sixth Ring Road from Bing Maps. We also collected labels from OpenStreetMap and reclassified them into 10 typical urban functional zones according to the “Code for classification of urban land use and planning standards of development land (GB 50137-2011)”. This

dataset provided a comprehensive and diverse set of examples that are representative of different urban functionalities.

Experimental results show that, firstly, compared with the results of existing methods, over-segmented objects can improve boundary accuracy. This enhancement avoids the jagged boundaries resulting from grid units and the presence of multiple UFZs within a single unit due to road-block units. The improved boundary accuracy ensures that the functional zones are delineated more precisely, reflecting true urban layouts and reducing classification errors. Secondly, the accuracy of UFZ classification increases by 13.9% compared to the method that employs objects as analysis units while ignoring their spatial relationships. This significant improvement highlights the importance of considering spatial relationships in UFZ classification. Additionally, the traditional position encoding method achieved similar accuracy to the method without position encoding, indicating that traditional position encoding does not effectively provide positional information. The kappa coefficient of the proposed method, which uses geographic coordinates for encoding, shows an average improvement of 0.042 compared to the traditional Transformer position encoding method. This demonstrates that the introduction of geographic coordinates can effectively provide spatial relationship information, leading to better classification results. The kappa coefficient is a measure of classification accuracy adjusted for chance agreement, and an improvement in this metric underscores the robustness of the proposed method.

Key words：urban functional zone, remote sensing, deep learning, spatial relationship modeling, transformer networks

Supported by Natural Science Foundation of Hunan for Distinguished Young Scholars (No. 2022JJ10072); Open Project of Xiangjiang Laboratory (No. 22XJ03007); National Natural Science Foundation of China (No. 42171376, 41771458); Natural Science Foundation of Hunan (No. 2021JJ30815); High-Performance Computing Center of Central South University